UNIVERSITÀ DI PARMA

UNIVERSITA' DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN
Biotecnologie e Bioscienze

CICLO XXXVII

Machine learning analysis of enzyme neofunctionalization following gene duplication in vertebrate evolution

Coordinatore:
Chiar.ma Prof. Elena Maestri

Tutore:
Chiar.mo Prof. Riccardo Percudani

Dottorando: Carlo De Rito

Anni Accademici 2021/2022 - 2023/2024

Index

Index....1
Nomenclature and Abbreviations....3
Abstract....4
Introduction....5
1. Gene Duplication: Mechanisms and Evolutionary Implications....5
1.1 Historical Perspective on Gene Duplication....5
1.2 Mechanisms Driving Gene Duplication....7
1.3 Evolutionary Consequences of Gene Duplication....8
1.4 Natural Selection and the Evolution of Duplicated Genes....10
1.5 Gene Duplication as a Catalyst for Evolutionary Innovation....13
1.6 The Role of Gene Duplication in Vertebrate Evolution....15
2. Homology: Concepts, Tools, and Functional Insights....17
2.1. Core Concepts: Homology, Orthology, and Paralogy....17
2.2 Sequence Alignments and Substitution Matrices....19
2.3 Conserved Residues and Their Functional Roles....20
2.4 Bioinformatics Tools for Homology Detection and Analysis....21
3. Neofunctionalization: Molecular Pathways and Analytical Approaches....23
3.1. Molecular Mechanisms Behind Neofunctionalization in Proteins....23
3.2. Analytical Approaches to Studying Neofunctionalization....28
4. Machine Learning in Molecular Biology: Applications and Innovations....32
4.1. Introduction to Machine Learning in Biological Sciences....32
4.2. Machine Learning for Protein Structure Prediction....34
4.3. Protein Sequence Embeddings and Their Analytical Uses....38
5. Integrating Concepts: From Gene Duplication to Functional Divergence....43
Materials and Methods....45
1. Sequence Collection and Mapping....45
1.1. Homologous Sequence Collection....45
1.2. Chromosomal Mapping and Gene Pair Visualization....45
2. Orthologous Sequence Collection from Ensembl Compara....46
3. Multiple Sequence Alignment and Residue Classification....47
3.1. MSA Generation and Cleaning....47
3.2. Alignment Scoring Metrics: ESPript and Custom Approach....48
3.3. Residue Classification: Robust, Adaptive, Plastic....50
4. Divergence Scoring and Functional Analysis....51
4.1. Residue-Level Divergence Scoring....51
4.2. Global Metrics Summarization....54
4.3. RHEA Truth Set Construction....55
4.4. Functional Divergence Identification Framework....56
4.5. Probability and Permutation Analysis for Validation....57
5. Expression Analysis and Functional Enrichment....58
5.1. Tissue-Specific Expression Analysis....58

5.2. Functional Enrichment: Gene Ontology and KEGG Pathways..... 58
6. Autonomous Functional Divergence Pipeline..... 59
7. Case Studies and Experimental Procedures..... 61
7.1. Case Study: AADAC-AADACL2 Docking and Evolutionary Insights..... 61
7.2. AADACL2 Expression and Solubility in *E. coli*..... 62
7.3. AADACL2 Resuspension from Inclusion Bodies in *E. coli*..... 64
7.4. AADACL2 Activity Assay (4-Nitrophenyl Acetate)..... 65
7.5. AADACL2 Expression in *Pichia pastoris*: Propagation and Transformation..... 65
7.6. AADACL2 Expression and Solubility in *Pichia pastoris*..... 68
7.7. Lysis Protocols in *Pichia pastoris*..... 69
8. Pipeline Overview..... 70
Results..... 72
1. Duplicated Gene Pairs: Numbers and Chromosomal Distribution..... 72
2. Duplication Events Across Taxonomic Classes..... 74
3. Alignment Quality Evaluation..... 76
3.1. Visual Assessment of Alignment Quality Improvement..... 76
3.2. Quantitative Analysis of Alignment Metrics..... 79
4. Divergence Metrics and Functional Insights..... 80
4.1. Residue-Level Divergence Metrics Analysis..... 80
4.2. Validation of Divergence Metrics with RHEA Truth Set..... 83
4.3. Impact of KDE-Derived Thresholds on Metrics Performance..... 84
4.4. Identification of Functionally Divergent Protein Pairs..... 88
4.5. Probability and Permutation Validation Analysis..... 89
5. Gene Expression and Functional Enrichment..... 90
5.1. Tissue-Specific Expression Analysis..... 90
5.2. Functional Enrichment: Gene Ontology and KEGG Pathways..... 92
6. Results from Autonomous Pipeline: Divergence of AOC2 and AOC3..... 94
7. Case Studies and Experimental Results..... 98
7.1 Case Study: AADACL2 Functional Divergence and Evolutionary Insights..... 98
7.2. AADACL2 Induction and Expression in *E. coli*..... 105
7.3. AADACL2 Induction and Activity Assay in *Pichia pastoris*..... 107
Discussion..... 112
1. Duplicated Gene Pairs: Numbers and Chromosomal Distribution..... 112
2. Duplication Events Across Taxonomic Classes..... 113
3. Alignment Quality and Residue-Level Divergence Analysis..... 116
4. Validation of Divergence Metrics with RHEA: Thresholds and Probability Analysis... 120
5. Functional Enrichment and Tissue-Specific Expression Analysis..... 125
6. Autonomous Pipeline: Functional Divergence of AOC2 and AOC3..... 127
7. Case Study: AADACL2 Functional Divergence and Expression..... 128
Bibliography..... 142

Nomenclature and Abbreviations

ISc - In-Group Score
XSc - Cross-Group Score
DSc - Differential Score
TSc - Total Score
dsclen - DSc Length (metric for divergence)
hdelta - Delta Differential Score
hdsc - Hot Differential Score
pdsc - P2Rank Differential Score
CDS - Coding Sequence
FPLC - Fast Protein Liquid Chromatography
MWCO - Molecular Weight Cutoff
AUC - Area Under the Curve (ROC metric)
F1 Score - Harmonic mean of precision and recall
Precision - True Positives / (True Positives + False Positives)
Recall - True Positives / (True Positives + False Negatives)
ROC - Receiver Operating Characteristic Curve
KDE - Kernel Density Estimate
RMSD - Root Mean Square Deviation
BRH - (Best Reciprocal Hit)
PNPA - (4-Nitrophenyl Acetate)

BLASTP - (Basic Local Alignment Search Tool for Proteins)
Clustal Omega (CLUSTALO) - Multiple sequence alignment tool
ESPript - Easy Sequencing in PostScript (alignment visualization)
HPC - High Performance Computing
ProtTrans T5 - Transformer-based Protein Embedding Model
RAxML - Randomized Axelerated Maximum Likelihood (phylogenetic software)
QuickGO - Gene Ontology Database maintained by EBI
NetNGlyc 1.0 - N-glycosylation prediction tool
NetOGlyc 4.0 - O-glycosylation prediction tool
AlphaFold - AI-based protein structure prediction tool
P2Rank - Ligand Binding Site Prediction Tool
PyMOL - Molecular visualization software

UniProtKB - Universal Protein Knowledgebase
Pfam - Protein Families Database
RHEA - Curated biochemical reactions database
EC - Enzyme Commission (classification of enzymes)
HPA - Human Protein Atlas
GO - Gene Ontology
KEGG - Kyoto Encyclopedia of Genes and Genomes
PDB - Protein Data Bank

AADAC - Arylacetamide Deacetylase
AADACL2 - Arylacetamide Deacetylase Like 2
pET-28a(+)-TEV - Expression plasmid with a TEV cleavage site
pPICZ A - Pichia pastoris expression vector with Zeocin resistance
AOX1 - Alcohol Oxidase 1 (methanol-inducible promoter in Pichia pastoris)
c-myc - Epitope tag for detection with anti-myc antibodies
6xHisTag - Polyhistidine tag for protein purification
BL21-CodonPlus - E. coli strain for protein expression
pGro7 - Plasmid for GroES/GroEL co-expression
pG-KJE8 - Plasmid for chaperone co-expression
TUNER - E. coli strain for controlled expression
GS115 - Pichia pastoris strain used for protein expression

Abstract

We investigated the evolutionary and functional divergence of duplicated genes in vertebrates, focusing on enzyme neofunctionalization. Starting with approximately 2,400 enzymatic proteins sharing identical PFAM domain architectures, we identified orthologous gene pairs across major vertebrate clades using a specialized pipeline. High-quality multiple sequence alignments were analyzed to compute per-residue evolutionary metrics, such as In-group and Differential Conservation Scores, utilizing the BLOSUM62 substitution matrix to pinpoint residues contributing to functional divergence. Context-based metrics from ProtTrans embeddings and functional hotspot predictions via BindEmbed21, refined with AlphaFold models and P2Rank pocket predictions, further elucidated potential functional changes. By aggregating these per-residue scores into per-protein metrics, we systematically assessed functional divergence across gene pairs. Thresholds established using a truth set from the Rhea database revealed that 35% of the analyzed gene pairs exhibited strong evidence of neofunctionalization. Enrichment analyses incorporating tissue-specific expression data and functional annotations provided biological context for the observed divergence patterns. A case study on the skin-expressed *AADACL2* gene illustrated our approach. Compared to its paralog *AADAC*, the protein encoded by *AADACL2* possesses additional functional pocket residues, suggesting a unique lipase function potentially involved in ceramide processing for cornified envelope formation. Experimental validation through heterologous expression faced challenges in protein solubility, leading us to consider ancestral sequence reconstruction for enhanced protein stability. Our findings advance the understanding of enzyme neofunctionalization in vertebrates and offer a framework for detecting functional divergence in duplicated genes.

Introduction

1. Gene Duplication: Mechanisms and Evolutionary Implications

1.1 Historical Perspective on Gene Duplication

The diversity of life on Earth reflects the processes that drive evolution. Among these processes, gene duplication stands out as a fundamental mechanism that contributes to genomic complexity and the emergence of novel functions. Gene duplication provides the raw genetic material upon which natural selection can act, allowing organisms to adapt to new environments and challenges. The concept was first introduced by Susumu Ohno in 1970, who posited that gene duplication is a primary source of evolutionary innovation (Ohno, 1970). Since then, subsequent research has expanded our understanding of the mechanisms, outcomes, and implications of gene duplication in evolutionary biology. This chapter describes the historical development of the concept, elucidates the molecular mechanisms by which gene duplication occurs, and examines the evolutionary trajectories of duplicated genes. By exploring these aspects, we provide a basis for understanding gene duplication's role in evolutionary innovation. Specifically, we investigate how gene duplication influences neofunctionalization in vertebrates and enzyme evolution.

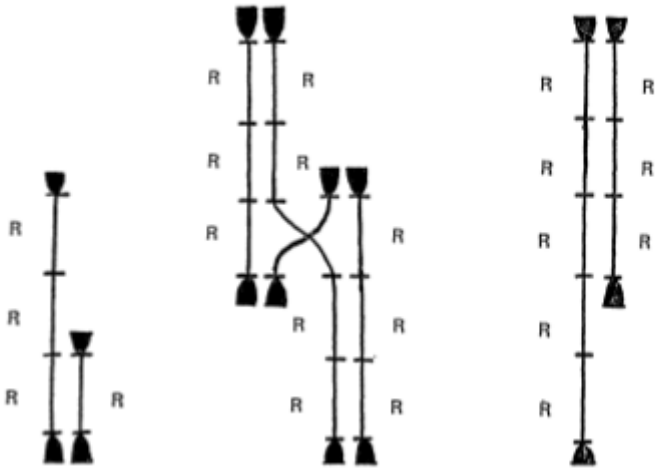


Figure 1. The consequence of unequal crossing-over between duplicated segments, resulting in variations in the number of ribosomal RNA genes. Reproduced from Ohno (1970).

Gene duplication was first recognized as a major driver of evolution thanks to Susumu Ohno's influential work. In his book, Evolution by Gene Duplication, Ohno argued that duplicating genes is crucial for evolution because it provides the raw material needed to

create new functions. Without duplication, the evolution of new genes would be limited, as mutations in essential genes often cause harm. Ohno observed gene families and protein functions, suggesting that a duplicated gene could safely accumulate mutations while the original gene continued its function. This "extra" gene could eventually take on a new role—a process called neofunctionalization (Ohno, 1970).

Ohno's ideas initially faced doubt due to limited evidence and the genetic beliefs of the time. However, as DNA sequencing and genome analysis advanced, strong support emerged for gene duplication as a widespread and important evolutionary force (Lynch & Conery, 2000; Zhang, 2003). Research showed that gene duplication occurs across many species and has expanded gene families involved in various biological functions.

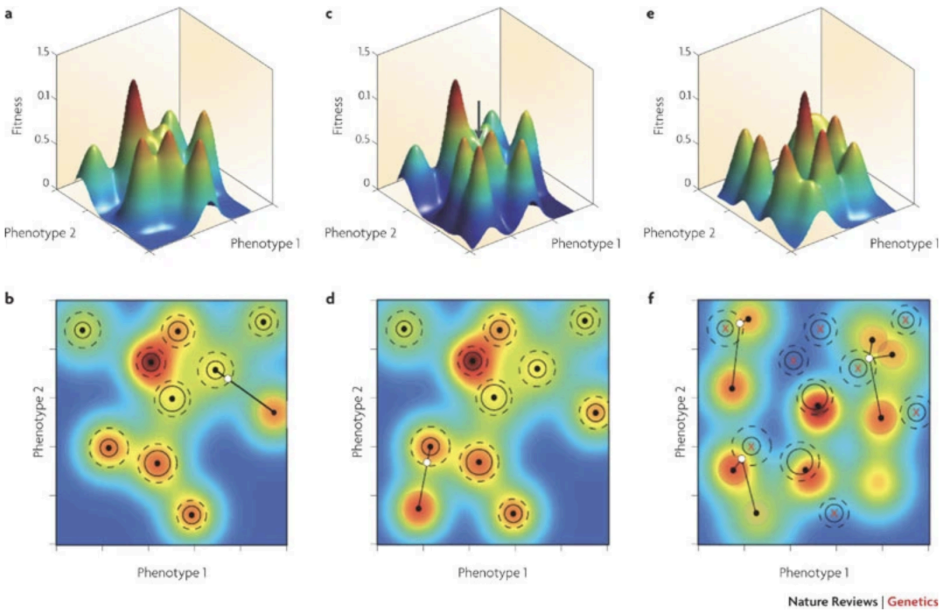


Figure 2. A simplified fitness landscape model illustrating the role of polyploidy in evolutionary innovation and niche occupation. The upper and lower panels show phenotype space, where red regions represent highly adapted organisms and blue regions represent unviable areas. Polyploid organisms may develop phenotypic innovations, allowing them to colonize new niches or survive environmental changes. Reproduced from Van de Peer et al. (2009).

Ohno's work also had a big impact on how we understand genome evolution. His ideas cleared the way for studying whole-genome duplications (WGDs) and how they increase organism complexity. For example, the "2R hypothesis" suggests that two rounds of WGD happened early in vertebrate evolution, contributing to the complexity of vertebrate genomes (**Fig. 2**) (Dehal & Boore, 2005; Van de Peer et al., 2009). Ohno's insights helped us understand the link between genome structure and evolutionary potential.

1.2 Mechanisms Driving Gene Duplication

Gene duplication can occur through several molecular mechanisms, each contributing uniquely to genomic evolution. Understanding these mechanisms is essential for appreciating how gene duplication generates the genetic material necessary for evolutionary innovation.

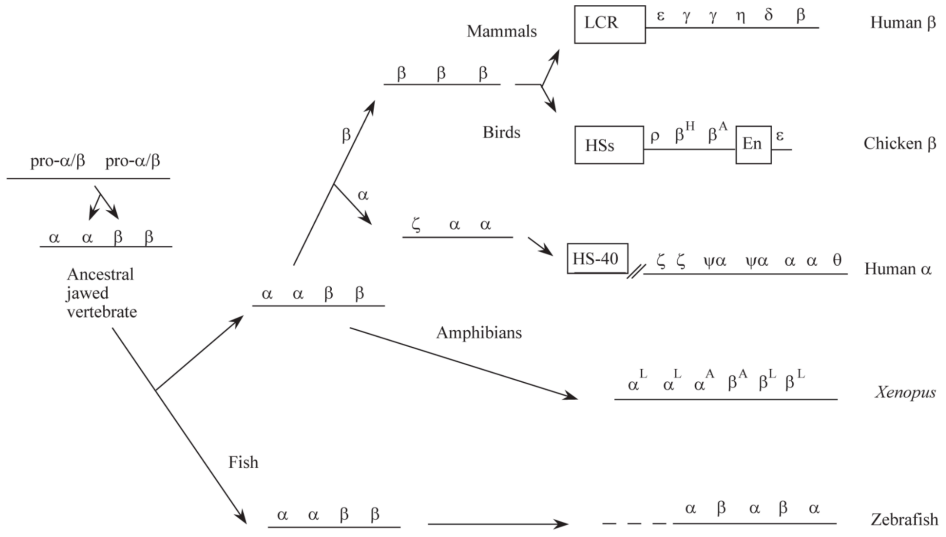


Figure 3. Evolution of globin gene clusters in vertebrates. Each Greek letter represents a globin gene. These gene clusters likely originated from a single ancestral globin gene, although it is possible that the ancestor had multiple globin genes with distinct regulation. Reproduced from Hardison (1998).

One primary mechanism is unequal crossing over during meiosis. When homologous chromosomes misalign, crossing over can result in one chromosome with a duplication of a gene segment and another with a corresponding deletion (**Fig. 1**) (Ohno, 1970; Bailey et al., 2002). This process often leads to tandem gene duplications and contributes significantly to the expansion of gene families. For example, the globin gene family in vertebrates has expanded through such tandem duplications, resulting in genes that encode different globin proteins with distinct oxygen-binding properties (**Fig. 3**) (Hardison, 1998).

Another mechanism is retroposition, also known as retroduplication. This involves the reverse transcription of mRNA transcripts into cDNA, which can then integrate into the genome at a new location (Kaessmann et al., 2009). Retrogenes typically lack introns and may acquire new regulatory elements at their insertion sites, potentially leading to novel expression patterns and functions. An example is the Jingwei gene in *Drosophila*, which

originated through retroposition and has acquired a new function related to alcohol dehydrogenase activity (Long & Langley, 1993).

Segmental duplications involve the duplication of large genomic regions, often encompassing multiple genes. These duplications can arise from replication errors or chromosomal rearrangements and contribute to genomic instability and variation (Bailey et al., 2002). They play a significant role in the evolution of gene families and can lead to dosage imbalances that drive evolutionary change.

Whole-genome duplications (WGDs) are events where the entire genome is duplicated, leading to polyploidy. WGDs have been particularly influential in the evolution of plants and vertebrates (Van de Peer et al., 2009). In vertebrates, evidence suggests that two rounds of WGDs occurred early in their evolutionary history, contributing to the complexity of their genomes (Dehal & Boore, 2005). These duplications provide vast amounts of genetic material for diversification and innovation.

Replication slippage during DNA replication is another mechanism that can lead to small-scale gene duplications. This process occurs when the DNA polymerase slips and re-replicates a portion of the DNA strand, resulting in copy number variations (Streisinger et al., 1966). Although these duplications are typically small, they can have significant impacts on gene function and regulation. These mechanisms show the flexibility of genomes and the continual generation of genetic diversity through gene duplication. Each mechanism contributes differently to the structure and function of genomes, influencing the potential for evolutionary innovation.

1.3 Evolutionary Consequences of Gene Duplication

Following duplication, genes may undergo different evolutionary outcomes that influence genetic diversity and organismal complexity. The three primary outcomes are nonfunctionalization, subfunctionalization, and neofunctionalization, each contributing differently to evolutionary processes. Nonfunctionalization is perhaps the most common fate, where one copy of the duplicated gene accumulates deleterious mutations and becomes a pseudogene (Li et al., 1981; Zhang, 2003). Since the other copy still performs the essential function, mutations in the duplicate are often neutral, leading to its loss of function. Pseudogenes can remain in the genome and sometimes serve as raw material for future

evolutionary changes, such as by acting as templates for gene conversion or recombination. In a subfunctionalization event, the original functions of the ancestral gene are divided between the two duplicated copies. According to the Duplication-Degeneration-Complementation (DDC) model proposed by Force et al. (1999), complementary degenerative mutations in regulatory or coding regions can lead to each gene copy retaining a subset of the original functions. Together, these copies retain the full functional range of the ancestral gene, which allows them both to persist in the genome (**Fig. 4**).

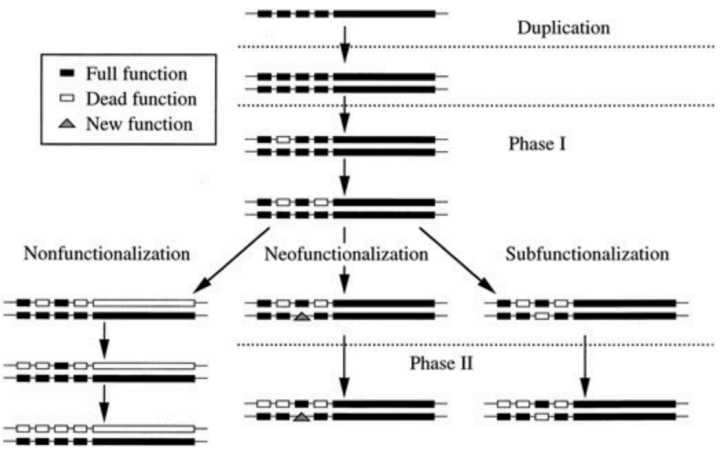


Figure 4. Three possible evolutionary fates of duplicate gene pairs. Small boxes represent regulatory elements, large boxes represent transcribed regions, and triangles indicate the evolution of new functions. Solid boxes denote intact regions, while open boxes represent null mutations. On the left, one gene copy becomes a nonfunctional pseudogene due to accumulated mutations. On the right, both copies are preserved because each retains essential regulatory regions. In the center, one gene copy evolves a new function while the other retains the original function. Reproduced from Force et al. (1999).

Subfunctionalization is particularly relevant in complex organisms where genes often have multiple functions or expression patterns. Neofunctionalization occurs when one of the duplicated genes acquires mutations that confer a new function not present in the ancestral gene (Ohno, 1970; Hughes, 1994). This new function can provide a selective advantage, leading to the retention of the gene in the population. Neofunctionalization is a critical mechanism for the evolution of novel traits and increased organismal complexity.

For example, the antifreeze proteins in Antarctic notothenioid fish evolved through neofunctionalization of a duplicated pancreatic trypsinogen gene, allowing the fish to survive in freezing temperatures (**Fig. 5**) (Chen et al., 1997). Another factor influencing the retention

of duplicated genes is the dosage balance hypothesis. This hypothesis suggests that genes encoding proteins involved in multi-subunit complexes or tightly regulated pathways are more likely to be retained after duplication to maintain stoichiometric balances (Papp et al., 2003; Freeling & Thomas, 2006). Disruption of these balances can be deleterious, so both gene copies are maintained under purifying selection. The evolutionary fate of duplicated genes is determined by a combination of genetic drift, selection pressures, and functional constraints. Understanding these outcomes provides insights into how genomes evolve and how new genetic functions arise.

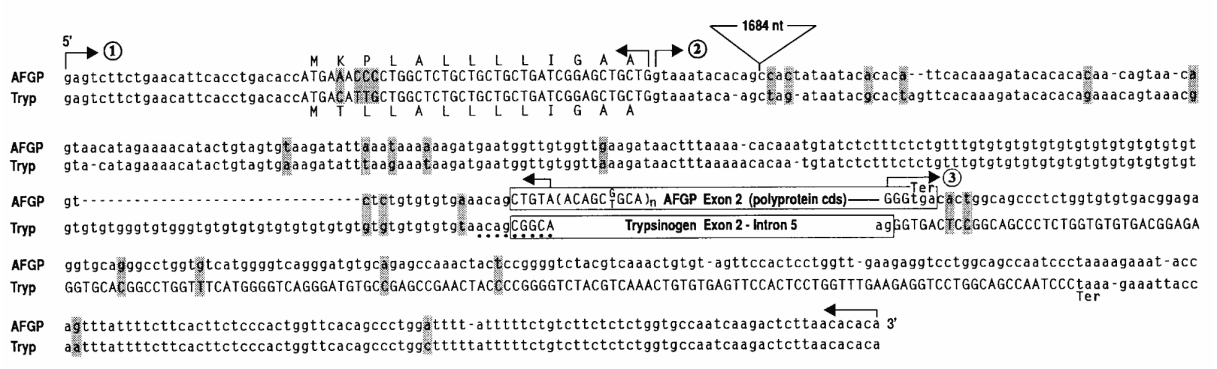


Figure 5. Alignment of genomic sequences from *D. mawsoni* antifreeze glycoprotein (*AFGP*) and trypsinogen (*Tryp*), showing three regions of high sequence identity: (1) 5' UTR and signal peptide coding sequences, (2) intron I, and (3) 3' UTR. Gene-specific sequences are boxed, and nucleotide differences are highlighted. This alignment suggests the amplification of a Thr-Ala-Ala coding element in trypsinogen led to the repetitive sequence in *AFGP*. Reproduced from Chen et al. (1997).

1.4 Natural Selection and the Evolution of Duplicated Genes

Natural selection is a key force in evolution, shaping genetic variation within populations. It plays a crucial role in determining the fate of duplicated genes by influencing whether they are preserved, diverge to acquire new functions, or are eliminated from the genome. Two primary forms of natural selection—positive selection and purifying selection—act on gene duplicates in different ways, impacting their evolutionary trajectories.

Positive selection, also known as diversifying or Darwinian selection, occurs when genetic variants confer a selective advantage, increasing an organism's fitness. In the context of gene duplication, positive selection can drive the neofunctionalization of duplicated genes by

favoring beneficial mutations that enhance or alter gene function (Hughes, 1994; Kondrashov et al., 2002). After a gene duplication event, one copy may acquire mutations that lead to a new function advantageous in a specific environmental context. Positive selection acts on these advantageous mutations, promoting their fixation in the population. This process is critical for the evolution of new gene functions and contributes to adaptation and speciation. A notable example of positive selection acting on duplicated genes is observed in the *RNASE1* gene duplication in leaf-eating monkeys. Zhang et al. demonstrated that a duplicated pancreatic ribonuclease gene underwent positive selection to improve its ability to degrade bacterial RNA in the acidic environment of the small intestine, aiding in nutrient acquisition from bacterial fermentation (**Fig. 6**) (Zhang et al., 2002).

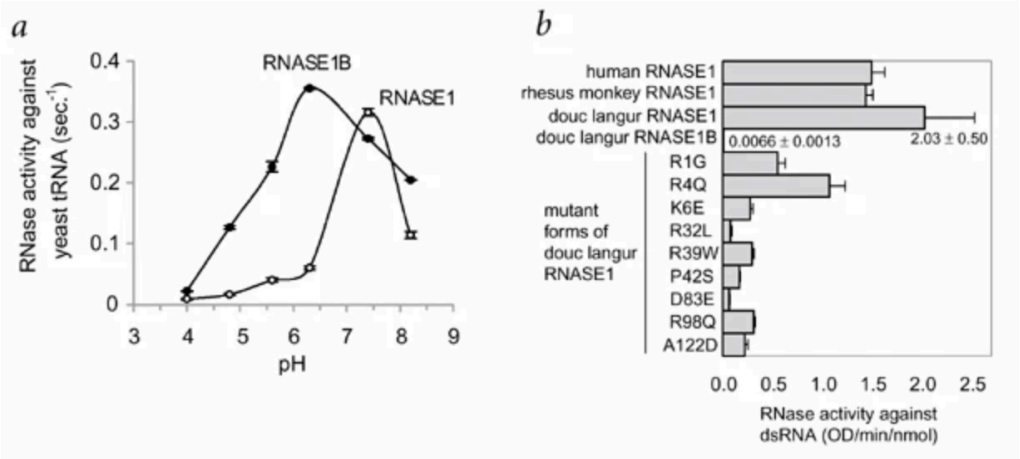


Figure 6. Enzyme activities of recombinant RNASE1B, RNASE1, and mutant forms of RNASE1. (a) RNase activity against yeast tRNA at different pH levels, and (b) RNase activity against double-stranded RNA (dsRNA). Mutant forms are indicated by the formula XyZ, where amino acid X is replaced by Z at position y. Error bars represent 1 standard error of the mean (s.e.m.). Reproduced from Zhang et al. (2002).

Purifying selection, or negative selection, removes deleterious mutations from a population to keep essential genes functional. In the context of gene duplication, purifying selection can act to maintain the ancestral function in one or both copies of the duplicated gene (Lynch & Conery, 2000). When both gene copies are essential, purifying selection prevents the accumulation of harmful mutations, preserving their functions. This is especially important for genes in key biological processes or when extra gene copies provide an advantage. An example is the retention of duplicated histone genes, which are highly conserved due to their essential role in DNA packaging and regulation (Marzluff et al., 2002). Purifying selection maintains the integrity of these genes across multiple copies to ensure proper chromatin structure and gene expression. Purifying selection also plays a role in subfunctionalization by

maintaining different essential functions in each gene copy. Complementary degenerative mutations in regulatory regions can partition the ancestral gene's functions between the duplicates. Purifying selection preserves these functions by eliminating mutations that would further degrade essential aspects of the genes (Force et al., 1999). In some cases, balancing selection maintains genetic diversity within a population by preserving multiple alleles at a locus. For duplicated genes, balancing selection can act when heterozygotes have a selective advantage or when different environmental conditions favor different alleles (Fischer et al., 2014). This can contribute to the maintenance of gene duplicates with complementary or condition-specific functions. The evolutionary fate of duplicated genes is also influenced by genetic drift, particularly in small populations where random fluctuations in allele frequencies can have significant effects (**Fig. 7**) (Ohta, 1987).

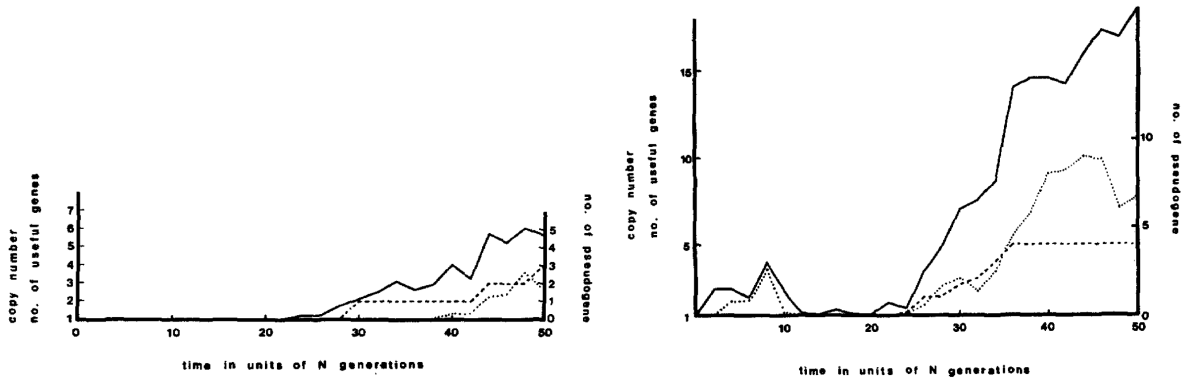


Figure 7. Simulation of gene duplication dynamics under positive selection. (Left panel) The simulation shows an increase in copy number, pseudogene number, and allele diversity over time with a selection coefficient (2Ns = 40) and an unequal crossing-over rate (y = 0.0025). (Right panel) The same simulation with a higher unequal crossing-over rate (y = 0.005) shows a faster increase in these parameters. Adapted from Ohta (1987).

Genetic drift can lead to the fixation or loss of gene duplicates independent of selective pressures. However, the interplay between natural selection and genetic drift determines the overall trajectory of duplicated genes. In large populations, selection tends to have a stronger influence, while in small populations, drift can override selection (Kimura 1968). For example, slightly deleterious mutations in a duplicated gene may become fixed due to drift, leading to nonfunctionalization despite purifying selection.

Natural selection, through positive and purifying selection, plays a role in shaping the evolution of duplicated genes. Positive selection drives the acquisition of new functions, contributing to neofunctionalization and adaptive evolution. Purifying selection maintains essential gene functions and can preserve duplicated genes when they are advantageous or

necessary for survival. Understanding these selective forces enhances our comprehension of how gene duplication contributes to genetic diversity, adaptation, and evolutionary innovation.

1.5 Gene Duplication as a Catalyst for Evolutionary Innovation

Gene duplication has profound implications for evolutionary innovation, serving as a catalyst for the development of new gene functions, complex regulatory networks, and increased organismal complexity.

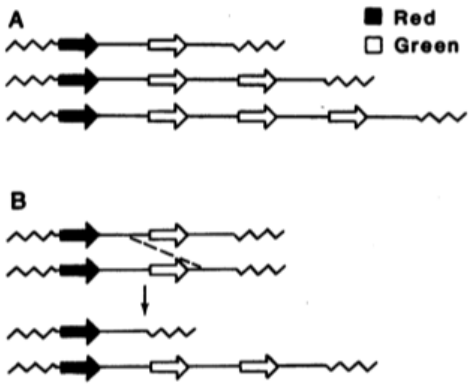


Figure 8. (A) Arrangement of green and red pigment genes in color-normal males. Each arrow represents a gene, with thin straight lines indicating duplicated intergenic spacers and zigzag lines showing single-copy flanking DNA. (B) Unequal crossing-over in the intergenic region can alter the copy number of green pigment genes, producing chromosomes with fewer or more green pigment genes. Reproduced from Nathans et al. (1986).

By providing additional genetic material, gene duplication allows for the exploration of new functional landscapes without compromising existing essential functions. One of the primary implications is the source of genetic novelty. Duplicated genes are free from the constraints of purifying selection that act on single-copy genes, enabling them to accumulate mutations that may lead to new functions or regulatory patterns (Ohno, 1970; Long et al., 2003). This process is essential for the evolution of complex traits and adaptations. For instance, the evolution of color vision in primates involved the duplication and subsequent divergence of opsin genes, allowing for trichromatic vision (**Fig. 8**) (Nathans et al., 1986).

Gene duplication also facilitates the evolution of regulatory networks. Duplication of transcription factors and other regulatory genes can lead to the expansion and diversification of regulatory pathways, impacting development and physiological processes (Teichmann & Babu, 2004). This diversification can result in increased complexity and specialization of gene expression patterns, contributing to morphological and functional diversity. Furthermore, gene duplication contributes to morphological complexity. The expansion of gene families involved in developmental processes, such as the Hox gene clusters, has been linked to the evolution of new body plans and structures (**Fig. 9**) (Carroll, 1995; Holland et al., 1994).

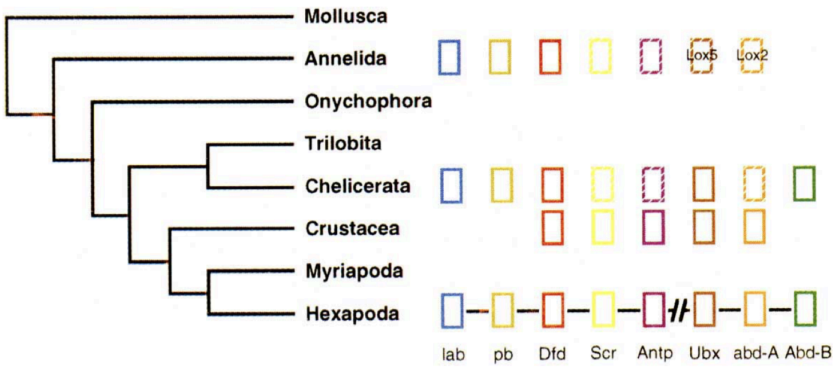


Figure 9. Hox genes and arthropod phylogeny. The comparison shows the array of Hox genes in arthropods and other invertebrates, alongside Drosophila Hox genes. Most Hox genes originated before the annelid/arthropod and insect/crustacean divergence. Solid boxes represent confirmed Hox genes, stippled boxes indicate genes that are too similar to distinguish, and missing boxes denote the absence of data. Reproduced from Carroll (1995).

The duplication and divergence of these genes allow for the modification of developmental pathways, leading to the emergence of novel morphological features. Additionally, gene duplication plays a role in adaptive evolution. Duplicated genes can acquire functions that confer selective advantages in specific environmental contexts. For example, the duplication of the amylase gene (*AMY1*) in humans has been linked to increased starch digestion capacity, which is believed to have provided a dietary advantage as humans adapted to agricultural environments. Populations with high-starch diets tend to have more copies of the *AMY1* gene, illustrating how gene duplication can facilitate adaptation to new diets by increasing enzyme production (Perry et al., 2007). This demonstrates how gene duplication can play a crucial role in adaptive evolution by providing increased gene product and enabling rapid response to environmental pressures. Overall, gene duplication expands the functional repertoire of organisms, providing opportunities for evolutionary innovation and

adaptation. It is a fundamental mechanism underlying the diversification of life and the complexity of biological systems.

1.6 The Role of Gene Duplication in Vertebrate Evolution

In vertebrates, gene duplication has been influential in shaping genomic architecture and facilitating the evolution of complexity. The early vertebrate lineage experienced two rounds of whole-genome duplication (WGDs), known as the 2R hypothesis, which significantly expanded the genetic material available for evolutionary diversification (Ohno, 1970; Dehal & Boore, 2005). These WGDs allowed for the retention and divergence of numerous gene copies, leading to the expansion of key gene families. For example, the Hox gene clusters, which play key roles in body plan development, expanded from a single cluster in invertebrates to four clusters in most vertebrates (Holland et al., 1994). This expansion is associated with the increased morphological complexity observed in vertebrates compared to their invertebrate ancestors. The expansion of other gene families, such as those encoding signaling molecules, transcription factors, and components of the immune system, has also contributed to vertebrate complexity. The diversification of the immunoglobulin superfamily genes, for instance, has been essential for the development of the adaptive immune system unique to vertebrates (**Fig. 10**) (Flajnik & Kasahara, 2010).

Gene duplication enabled new traits to evolve in vertebrates, allowing them to adapt to a wide range of ecological niches. The evolution of color vision in primates, electric organs in electric fish, and antifreeze proteins in Antarctic fish are examples of how gene duplication and subsequent divergence have led to specialized adaptations (Chen et al., 1997; Zakon 2006; Li & Graur, 1991). Moreover, gene duplication has contributed to the redundancy and robustness of vertebrate genomes. The presence of multiple gene copies can buffer against deleterious mutations and provide flexibility in gene regulation and expression (Kafri et al., 2006). This redundancy allows for greater experimentation at the genetic level, facilitating evolutionary innovation. Overall, gene duplication has been a driving force in vertebrate evolution, contributing to the complexity, diversity, and adaptability of this lineage.

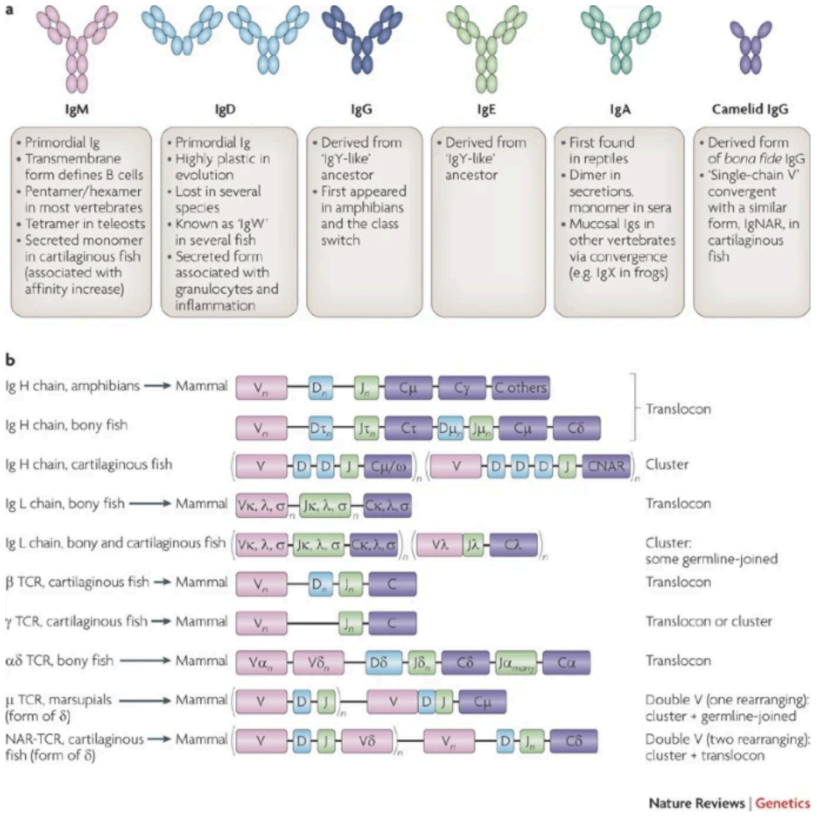


Figure 10. Antigen receptor proteins and genes in jawed vertebrates. (a) Different B cell receptor (*BCR*) isotypes in mammals and their counterparts in other vertebrates. Each oval represents an immunoglobulin superfamily (IgSF) domain. *IgD* is shown in two forms for mouse and human, reflecting its structural plasticity. (b) *BCR* and T cell receptor (*TCR*) genes throughout vertebrate phylogeny, showing recombination signal sequences and the deletion of antigen receptor families upon V-D-J rearrangement. Reproduced from Flajnik & Kasahara (2010).

This chapter has covered the basics of gene duplication and its crucial role in driving evolution. Starting with Susumu Ohno's historical insights, we reviewed the different ways gene duplication happens, such as through unequal crossing over, retroposition, segmental duplications, and whole-genome duplications. The various paths that duplicated genes can take—nonfunctionalization, subfunctionalization, and neofunctionalization—demonstrate how they can contribute to new genetic functions and complexity. Gene duplication is a catalyst for evolutionary change by adding genetic material that natural selection can shape. In vertebrates, gene duplication has played a major role, helping expand important gene families, build complex structures, and lead to new traits. Early whole-genome duplications in vertebrate history have left lasting effects on genome structure and the complexity of these organisms.

2. Homology: Concepts, Tools, and Functional Insights

2.1. Core Concepts: Homology, Orthology, and Paralogy

Gene duplication and homology relationships are essential in evolutionary biology and genomics. Accurately detecting and describing these duplications requires strong computational methods. This chapter covers the concepts of homology, orthology, and paralogy, which are fundamental for categorizing gene relationships. It also examines the computational tools and techniques—like sequence alignments and phylogenetic analysis—used to identify duplicated genes. By combining these methods and focusing on conserved residues and substitution matrices, we could better understand gene evolution and the functional impacts of gene duplication.

Homology refers to the similarity between genes or proteins that arises from a shared ancestor. Homologous genes originate from a common ancestral gene and retain structural and functional similarities across different species (Fitch, 1970). Understanding homology is crucial for inferring evolutionary relationships and predicting functional similarities between genes across diverse taxa (**Fig. 11**).

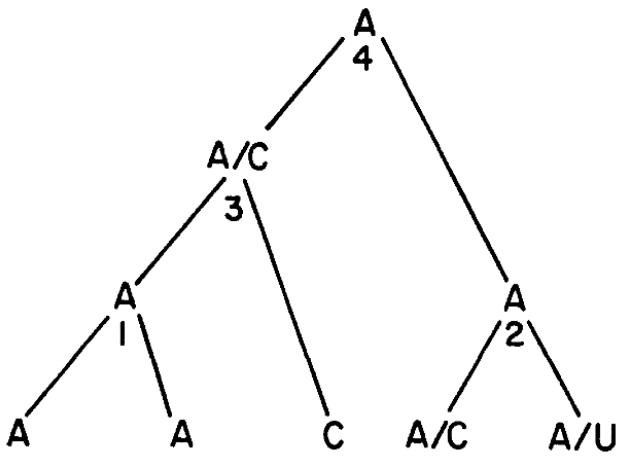


Figure 11. Reconstruction of ancestral gene coding. The tree illustrates the method used to infer ancestral nucleotide sequences at different evolutionary nodes. At each apex, the most likely nucleotide is selected based on the commonality between descendants, with ambiguous positions marked when no common nucleotide exists. This approach minimizes the number of mutations required to explain the observed sequences. Reproduced from Fitch (1970).

Homologous genes can be further categorized into orthologs and paralogs, based on the evolutionary events that have shaped their divergence. Orthologs are homologous genes

found in different species that originated from a single gene in the most recent common ancestor. They are typically separated by a speciation event and often retain conserved functions across species (**Fig. 12**) (Koonin, 2005). For example, the hemoglobin alpha genes in humans and mice are orthologs, maintaining similar roles in oxygen transport. In contrast, paralogs are homologous genes within the same genome that arose from gene duplication events.

Paralogs may diverge in function through processes such as neofunctionalization, where one copy acquires a new function, or subfunctionalization, where each copy retains a subset of the original gene's functions (Kondrashov & Kondrashov, 2006). An illustrative example of paralogy is the human hemoglobin alpha and beta genes, which originated from a duplication event and specialized for optimizing oxygen binding and transport. Differentiating between orthologs and paralogs is fundamental in comparative genomics and functional genomics. Orthology implies functional conservation. This makes orthologs valuable for transferring functional annotations across species. Paralogy provides insights into functional divergence and the evolutionary processes that generate new gene functions (**Fig. 12**) (Koonin, 2005).

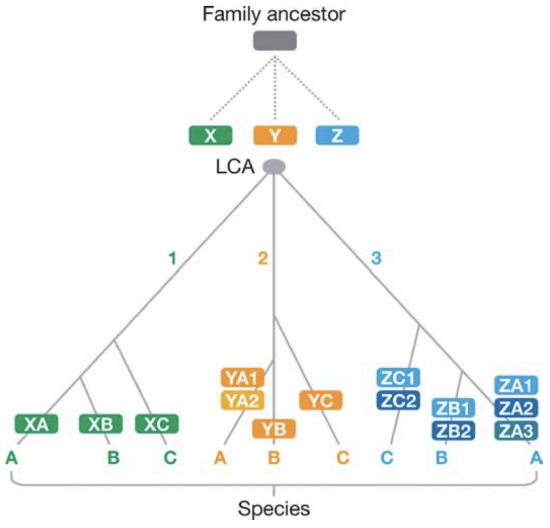


Figure 12. A hypothetical phylogenetic tree showing orthologous and paralogous relationships between three ancestral genes and their descendants across three species. LCA indicates the last common ancestor of the species being compared. Reproduced from Koonin (2005).

However, defining orthologous and paralogous relationships can be complicated by subsequent evolutionary events, such as gene loss, horizontal gene transfer, and lineage-specific expansions (Gabaldón & Koonin, 2013). Additionally, the presence of xenologs—genes acquired through horizontal gene transfer—and ohnologs—genes

duplicated via whole-genome duplication—further adds complexity to gene classification (Wolfe, 2000). Accurate classification of these relationships is essential for understanding evolution and gene functions.

2.2 Sequence Alignments and Substitution Matrices

Sequence alignment is a foundational technique in bioinformatics that involves arranging sequences of DNA, RNA, or proteins to identify regions of similarity (Notredame, 2007). These similarities often indicate functional, structural, or evolutionary relationships between the sequences. Alignments can be pairwise, comparing two sequences at a time, or multiple, involving multiple sequences to find conserved regions across a gene family. Central to the effectiveness of sequence alignments are substitution matrices, such as PAM (Point Accepted Mutation) and BLOSUM (Blocks Substitution Matrix). These matrices score alignments based on the likelihood of amino acid substitutions occurring over evolutionary time, reflecting the evolutionary pressures acting on protein sequences (**Fig. 13**) (Henikoff & Henikoff, 1992).

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
C	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
S	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
T	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
P	-3	-1	-1	7		2	0	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	G
A	0	1	0	-1	4		3	-1	-1	0	0	1	-1	0	-1	0	-1	0	0	0	N
G	-3	0	-2	-2	0	6		2	-1	-1	-1	0	-1	0	0	0	0	2	1	3	D
N	-3	1	0	-2	-2	0	6		1	0	0	2	2	1	-1	0	0	2	2	4	E
D	-3	0	-1	-1	-2	-1	1	6		0	-2	0	1	1	-1	0	0	1	3	3	Q
E	-4	0	-1	-1	-1	-2	0	2	5		2	-1	0	1	0	-1	0	1	2	2	H
Q	-3	0	-1	-1	-1	-2	0	0	2	5		-1	-1	0	-1	1	0	1	3	-4	R
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8		1	-2	-1	1	1	2	3	1	K
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5		-2	-1	-1	0	1	2	4	M
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5		-1	1	0	0	1	3	I
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5		-1	0	-1	1	2	L
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4		0	1	2	4	V
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4		-1	-2	1	F
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		-1	2	Y
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		-1	W
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure 13. BLOSUM 62 substitution matrix (lower panel) and the difference matrix (upper panel) obtained by subtracting the PAM 160 matrix, position by position. Both matrices have identical relative entropies (0.70), with expected values of -0.52 for BLOSUM 62 and -0.57 for PAM 160. Reproduced from Henikoff & Henikoff (1992).

For instance, PAM matrices are derived from closely related sequences and are more effective for detecting alignments between similar sequences, while BLOSUM matrices are

based on conserved blocks of sequences and are better suited for identifying alignments between more divergent sequences (Dayhoff et al., 1978). The choice of substitution matrix significantly impacts alignment results, influencing both sensitivity and specificity in detecting homologous sequences. Accurate alignment is crucial for subsequent analyses, such as phylogenetic tree construction and functional annotation.

2.3 Conserved Residues and Their Functional Roles

Conserved amino acids identified through sequence alignments often signify critical functional or structural roles within proteins. Studies have demonstrated that the most conserved residues are frequently involved in active sites, binding interfaces, or maintaining the structural integrity of proteins (Chothia & Lesk, 1986).

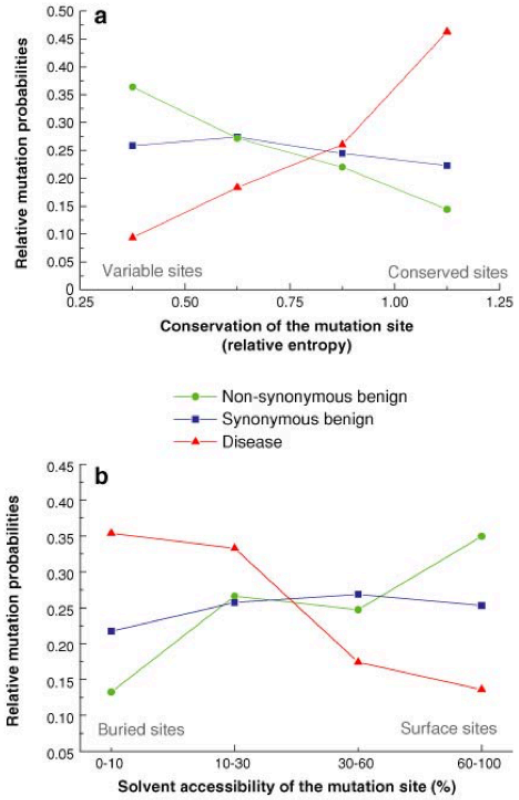


Figure 14. (a) The probability that a mutation will cause disease increases with the degree of site conservation, while benign nonsynonymous SNPs show the opposite trend. Synonymous SNPs, which do not alter amino acids, have a uniform probability across sites regardless of conservation. (b) Disease-causing mutations are more likely to occur in the protein interior, while benign SNPs are more common on the surface. Reproduced from Ng & Henikoff (2006).

This correlation underscores the utility of substitution matrices in highlighting regions under strong evolutionary constraints, where substitutions are less likely to be tolerated due to their impact on protein function (Henikoff & Henikoff, 1992). For example, in enzyme families, conserved residues typically constitute the catalytic core essential for substrate binding and catalysis (Bartlett et al., 2002). Similarly, in structural proteins, conserved residues are essential for maintaining the protein's three-dimensional conformation (Lesk & Chothia, 1980). The functional significance of conserved residues is further supported by studies showing that mutations in these regions often lead to loss of function or disease (Ng & Henikoff, 2006). Incorporating the analysis of conserved residues into sequence alignment workflows enhances the accuracy of homology detection and functional annotation (**Fig. 14**) (Eisen, 1998).

2.4 Bioinformatics Tools for Homology Detection and Analysis

The identification and classification of homologous genes rely heavily on a suite of computational tools and specialized databases. These resources facilitate the detection of orthologs and paralogs, reconstruction of phylogenetic histories, and functional annotation of gene families. Sequence Alignment Tools: Tools like BLAST (Basic Local Alignment Search Tool) are indispensable for identifying homologous sequences through local alignments. Variants such as BLASTP (protein-protein), BLASTN (nucleotide-nucleotide), and BLASTX (translated nucleotide-protein) cater to different types of sequence searches, enabling researchers to detect potential homologs across diverse datasets (Altschul et al., 1990). The effectiveness of BLAST is enhanced by substitution matrices like BLOSUM62, which balance sensitivity and specificity for detecting both closely related and moderately divergent sequences (**Fig. 15**).

Phylogenetic Reconstruction Tools: Tools such as MEGA (Molecular Evolutionary Genetics Analysis), RAxML (Randomized Axelerated Maximum Likelihood), and MrBayes are used for building and interpreting phylogenetic trees. These tools utilize multiple sequence alignments to infer evolutionary relationships and can be used for distinguishing between orthologous and paralogous genes based on tree topology (Felsenstein, 1981; Huelsenbeck & Ronquist, 2001).

RAxML is particularly efficient for constructing large phylogenetic trees using maximum likelihood methods, making it suitable for extensive gene family analyses (Stamatakis, 2014). Orthology Databases: Databases like Ensembl Compara (**Fig. 16**), OrthoDB, OMA

(Orthologous MAtrix), and InParanoid provide catalogs of orthologous and paralogous genes across multiple species.

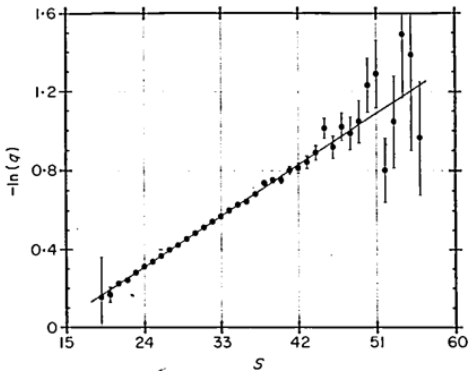


Figure 15. The probability (q) of BLAST missing a random maximal segment pair as a function of its score (S). Reproduced from Altschul et al. (1990).

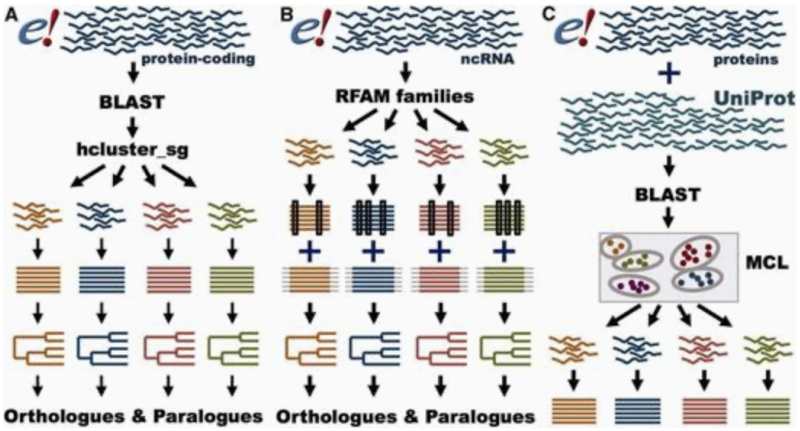


Figure 16. GeneTree and Ensembl Protein Family pipelines. (A) GeneTree pipeline for protein-coding genes: sequences are grouped into gene families using BLAST scores, aligned with MCoffee or MAFFT, and phylogenetic trees are built with TreeBeST. Orthologs and paralogs are inferred from the tree. (B) GeneTree pipeline for ncRNA genes: short ncRNA genes are grouped by RFAM classification and aligned using Infernal and PRANK, followed by phylogenetic tree construction with TreeBeST. (C) Ensembl Protein Family pipeline: sequences from Ensembl and UniProt are grouped by similarity using BLAST scores and MCL, then aligned with MAFFT. Reproduced from Herrero et al. (2016).

Ensembl Compara leverages phylogenetic methods to generate gene trees that depict evolutionary relationships, including duplication and speciation events, while OrthoDB offers hierarchical orthology assignments that facilitate deeper insights into gene conservation across evolutionary lineages (Vilella et al., 2009; Herrero et al., 2016; Kriventseva et al.,

2019). OMA predicts orthology based on sequence similarity and evolutionary distances, enabling the inference of relationships between genes from different species (Altenhoff et al., 2015). InParanoid focuses on identifying recent paralogs and orthologs between specific species pairs, providing valuable information on recent duplication events (Ostlund et al., 2010).

Gene Family Databases: Pfam and InterPro are essential for identifying conserved protein domains and functional units within genes. Pfam catalogs protein families and domains, facilitating the identification of shared functional motifs across different proteins (Finn et al., 2016). InterPro integrates data from multiple sources to provide annotations on protein families, domains, and functional sites, supporting the functional classification and annotation of newly discovered genes (Mitchell et al., 2019). These computational tools and databases play a key role in the workflow of gene duplication analysis, providing the necessary data for identifying homologous relationships, reconstructing evolutionary histories, and annotating gene functions with high accuracy.

In summary, the concepts of homology, orthology, and paralogy form the foundation for our investigation into gene duplication and its evolutionary consequences. By accurately classifying homologous relationships, we can infer functional similarities and divergences across species. Sequence alignments and substitution matrices, such as PAM and BLOSUM, are critical in identifying conserved residues and assessing the evolutionary pressures shaping protein sequences. Furthermore, the computational tools and databases we utilize—ranging from BLAST for homology detection to phylogenetic reconstruction tools like RAxML and orthology databases such as Ensembl Compara—provide an essential framework for analyzing gene duplications and their functional implications.

3. Neofunctionalization: Molecular Pathways and Analytical Approaches

3.1. Molecular Mechanisms Behind Neofunctionalization in Proteins

Neofunctionalization is a fundamental evolutionary process whereby duplicated genes acquire new functions not present in the ancestral gene. Changes at the molecular level can lead to the emergence of new enzymatic activities, structural features, and regulatory mechanisms. Understanding neofunctionalization at the protein level provides critical insights into how genetic diversity and complexity arise, driving adaptation and speciation.

Neofunctionalization in proteins occurs through various molecular mechanisms that alter amino acid sequences, leading to modifications in protein structure and function. Key mechanisms include point mutations, insertions and deletions (indels), domain shuffling, and changes affecting post-translational modifications and protein interactions. To understand these changes, it is important to examine the role of enzymes in biological systems, as they are often directly impacted by neofunctionalization. Enzymes, which serve as biological catalysts, accelerate chemical reactions by lowering the activation energy required. They play key roles in various biochemical processes, including metabolism, DNA replication, and cellular signaling. Each enzyme has a specific region known as the active site, where substrate molecules bind and the chemical reaction occurs.

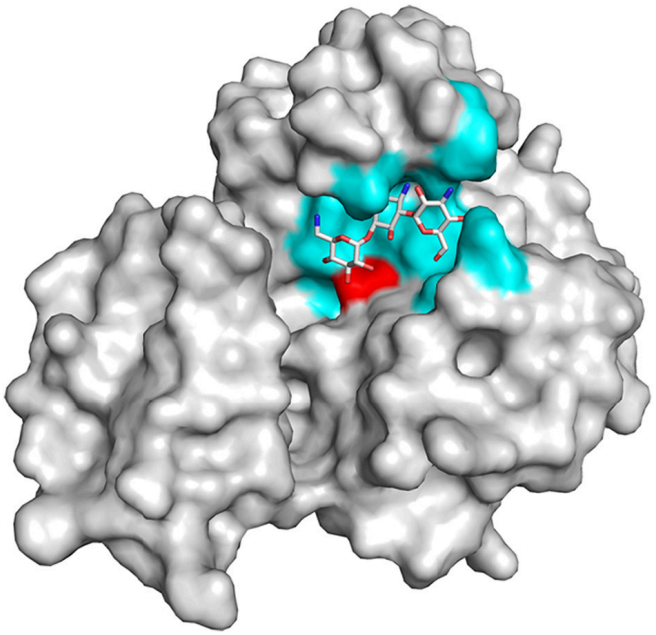


Figure 17. Surface representation of aminoglycoside-3′-phosphotransferase-IIa in complex with kanamycin (PDB ID 1ND4). Substrate-binding residues (D159, E160, R211, D227, E230, E262, F264) are shown in cyan, while the catalytic residue D190, which facilitates proton removal prior to phosphoryl transfer, is shown in red. Kanamycin is represented as sticks. Reproduced from Yu (2022).

The active site is typically composed of key amino acid residues involved in substrate recognition, binding, and catalysis (**Fig. 17**). These residues often include nucleophilic amino acids such as serine, threonine, or cysteine, as well as acidic residues like aspartate or glutamate, which can act as proton donors or acceptors. Histidine residues are frequently involved in proton transfers, while polar residues such as lysine or arginine may stabilize charged intermediates during the reaction. The precise arrangement of these residues creates an environment optimized for catalysis, contributing to the enzyme's specificity and efficiency (Branden & Tooze, 1999; Kraut, 1988).

Point mutations, specifically missense mutations, involve single nucleotide substitutions that result in the replacement of one amino acid with another. These mutations are a primary driver of neofunctionalization in proteins, with their effects varying depending on the nature and location of the amino acid change (Li et al., 1985). Functional outcomes of such mutations include altered enzymatic activity, where changes within the active site of enzymes can modify substrate specificity or catalytic efficiency. Even a single point mutation can lead to the loss or change in catalytic activity, as demonstrated by an example in which a point mutation in thiopurine S-methyltransferase (*TPMT*) resulted in a dramatic reduction in catalytic activity due to altered protein folding (Krynetski et al., 1995). For instance, in ribonuclease enzymes of primates, missense mutations have enhanced enzymatic activity under specific physiological conditions (Zhang et al., 2002).

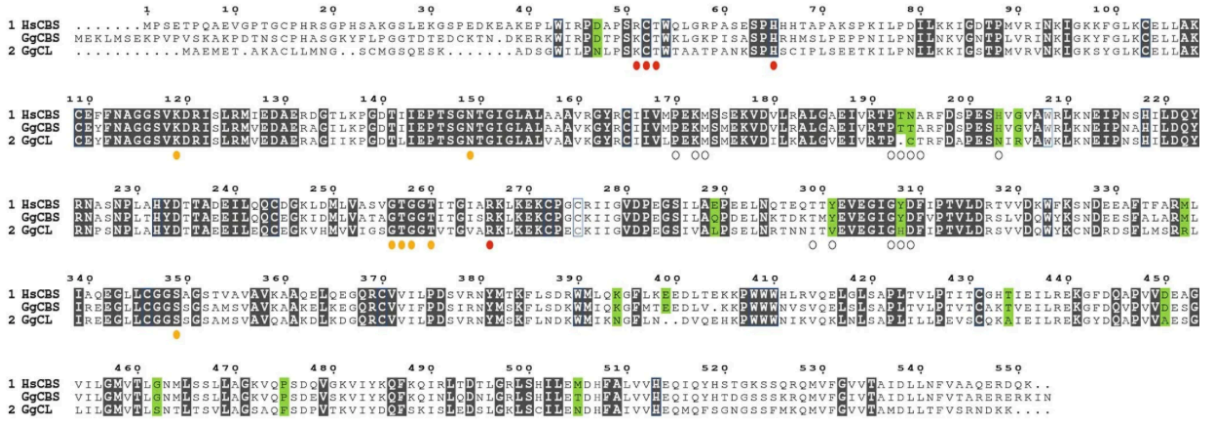


Figure 18. Multiple alignment of *H. sapiens* CBS (HsCBS) with *G. gallus* CBS (GgCBS) and CL proteins (GgCL). Filled circles indicate residues involved in binding heme (red), PLP (yellow), and serine (white) in the CBS structure (PDB code 3PC4). Conserved residues across 8 CL and 22 CBS sequences from vertebrates are shaded in black, while conserved differences between CBS and CL groups are shaded in green. Reproduced from Malatesta et al. (2020).

Additionally, amino acid substitutions can lead to different substrate preferences, enabling proteins to interact with new substrates. For example, mutations in protein tyrosine phosphatases (PTPs) have been shown to alter substrate interactions, leading to different physiological roles. Specifically, substrate-trapping mutants of *PTP1B* were generated by mutating the catalytic aspartate, which allowed identification of new substrates through altered binding properties (Flint et al., 1997). Such mutations illustrate how specific changes can enhance the ability of proteins to interact with a broader array of substrates, thereby contributing to neofunctionalization. Moreover, the cystathionine β-synthase (*CBS*) gene serves as an example of neofunctionalization where, after gene duplication, one copy evolved a distinct function due to key mutations, leading to new metabolic roles (**Fig. 18**)

(Malatesta et al., 2020). Mutations in active sites can significantly alter enzymatic activity and substrate specificity. For example, engineering of *E. coli* transketolase demonstrated that targeted mutations in phylogenetically and structurally defined sites expanded the enzyme's substrate range, illustrating how such mutations can enhance protein adaptability to new substrates (**Fig. 19**) (Yu et al., 2022).

Moreover, mutations distant from the active site can modify allosteric regulation, affecting how proteins respond to regulatory molecules (Gunasekaran et al., 2004). Finally, mutations can also alter kinetic properties such as Km and Vmax, optimizing proteins for various metabolic conditions. Increased protein flexibility due to mutations can lead to modified kinetic properties and the emergence of new enzymatic activities (Tokuriki et al., 2008).

Indels (insertions and deletions) introduce or remove amino acids, significantly impacting protein structure and function. Insertions can add new amino acid sequences, creating new functional domains or motifs, while deletions can remove regions that regulate activity or affect structural integrity (Lynch, 2007). Functionally, indels can lead to the creation of new functional domains, such as the evolution of antifreeze proteins in Antarctic fish, where repetitive sequence insertions conferred ice-binding properties (Chen et al., 1997). Indels can also result in structural domain acquisition or loss, enabling new functional capabilities. For example, domain shuffling through indels has generated hybrid proteins with unique regulatory functions (Buljan et al., 2010). Additionally, deletions can promote new protein conformations, allowing interactions with different cellular components and contributing to the evolution of novel pathways or regulatory mechanisms (Conant & Wolfe, 2008). Domain shuffling involves the rearrangement of existing protein domains through recombination events, while exon shuffling facilitates the modular assembly of proteins (Buljan et al., 2010, Patthy, 1999; Long et al., 1995). Functionally, domain shuffling can lead to the development of novel protein functions by combining domains from different proteins. For example, the fusion of kinase and phosphatase domains results in proteins with unique regulatory capabilities. Additionally, hybrid proteins generated through domain shuffling can integrate multiple signaling pathways, thereby enhancing cellular adaptability and complexity.

Mutations in regulatory elements that affect gene expression can drive neofunctionalization by altering the timing and location of gene expression. Functionally, such mutations can lead to changes in tissue or spatial expression. For example, duplicated genes may be expressed in different tissues, developmental stages, or cellular compartments, allowing the resulting proteins to fulfill specialized roles. Escriva et al. demonstrated that retinoic acid receptor (*RAR*) paralogs in vertebrates evolved distinct expression patterns, with some paralogs

acquiring novel expression territories, such as the forebrain and caudal regions, contributing to their functional divergence (Escriva et al., 2006). A well-known case is duplicated opsin genes, which show tissue-specific expression and contribute to the evolution of color vision (Yokoyama, 2000).

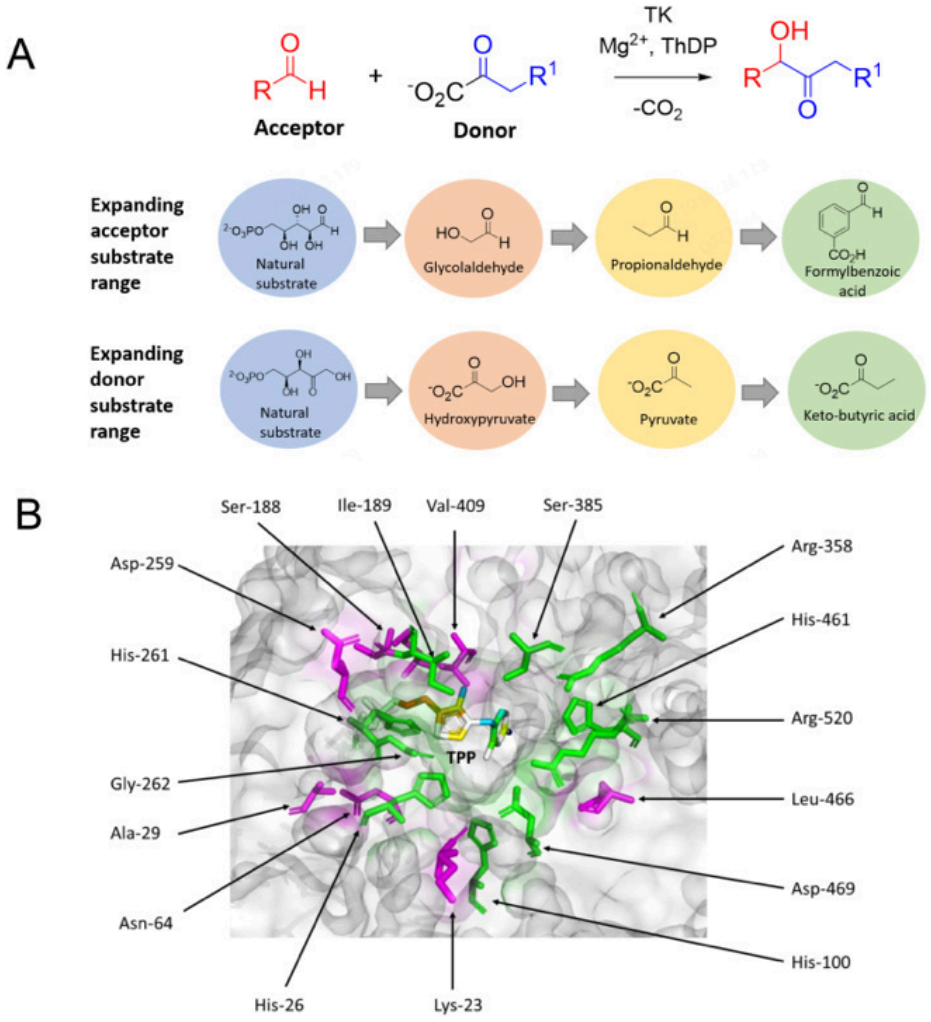


Figure 19. Engineering transketolase to expand substrate range by targeting mutations at specific active sites. (A) Range of acceptor and donor substrates accepted by engineered *E. coli* transketolase. (B) Phylogenetically (magenta) and structurally (green) defined mutation sites in *E. coli* transketolase with cofactor TPP (PDB ID: 1QGD). Structurally defined sites are within 4 Å of the docked substrate erythrose-4-phosphate, while phylogenetically diverse sites are within 10 Å of cofactor TPP, identified through analysis of 52 TK sequences across bacteria, fungi, plants, and trypanosomes. These sites were mutated to expand the substrate range of transketolase. Reproduced from Yu et al. (2022).

Additionally, mutations in regulatory sequences can cause subcellular localization shifts, affecting protein targeting signals (Warden et al., 2001). Mutations that affect amino acids

involved in post-translational modifications (PTMs) can significantly alter protein function, as PTMs act as molecular switches controlling protein activity, stability, localization, and interactions. Functionally, these mutations can lead to altered signaling pathways. For example, the introduction or elimination of serine, threonine, or tyrosine residues can modify phosphorylation patterns, impacting critical signaling processes such as growth and apoptosis (Pearlman et al., 2011).

Mutations affecting glycosylation sites, such as asparagine residues, can influence protein folding, stability, and cell-surface expression, thereby altering cellular communication (Helenius et al., 2001). Mutations that affect protein-protein interaction interfaces can either create new binding partners or disrupt existing ones. Functionally, such mutations can lead to the formation of novel interaction partners, where amino acid changes at interaction sites enable proteins to engage with new partners, allowing integration into different protein complexes or signaling pathways (Mintseris & Weng, 2005). Additionally, these alterations can modify interaction networks by introducing new protein complexes or altering existing ones, which can significantly impact cellular functions, including processes like signal transduction and metabolism (Samama et al., 1997). For example, specific mutations in the beta-2 adrenergic receptor were shown to stabilize an active state, allowing it to interact differently with G proteins and facilitating new signaling behaviors.

3.2. Analytical Approaches to Studying Neofunctionalization

Sequence analysis is a fundamental technique for detecting neofunctionalization. By aligning protein or gene sequences from different species or paralogous genes within the same genome, we can observe evolutionary changes that indicate functional divergence.

A key indicator of neofunctionalization is the presence of positive selection acting on a gene duplicate, which can be assessed through the ratio of nonsynonymous (dN) to synonymous (dS) substitution rates, commonly known as the dN/dS ratio. A ratio greater than one suggests that positive selection has favored amino acid changes, potentially leading to new or specialized functions (Yang & Nielsen, 2000). Advanced models using maximum likelihood estimation allow for more refined analyses, such as site-specific and branch-specific assessments, offering deeper insights into selective pressures acting on particular regions of the gene or across different evolutionary lineages (Yang, 2007). In particular, site-specific methods like the Mixed Effects Model of Evolution (MEME) are effective in detecting positive

selection at individual codon sites. MEME accounts for both pervasive and episodic selection by combining fixed effects for the proportion of selected sites with random effects for selection intensity, thereby increasing the sensitivity for identifying adaptive evolution (**Fig. 20**) (Murrell et al., 2012). This approach is crucial for understanding how gene duplicates can evolve new functions over time.

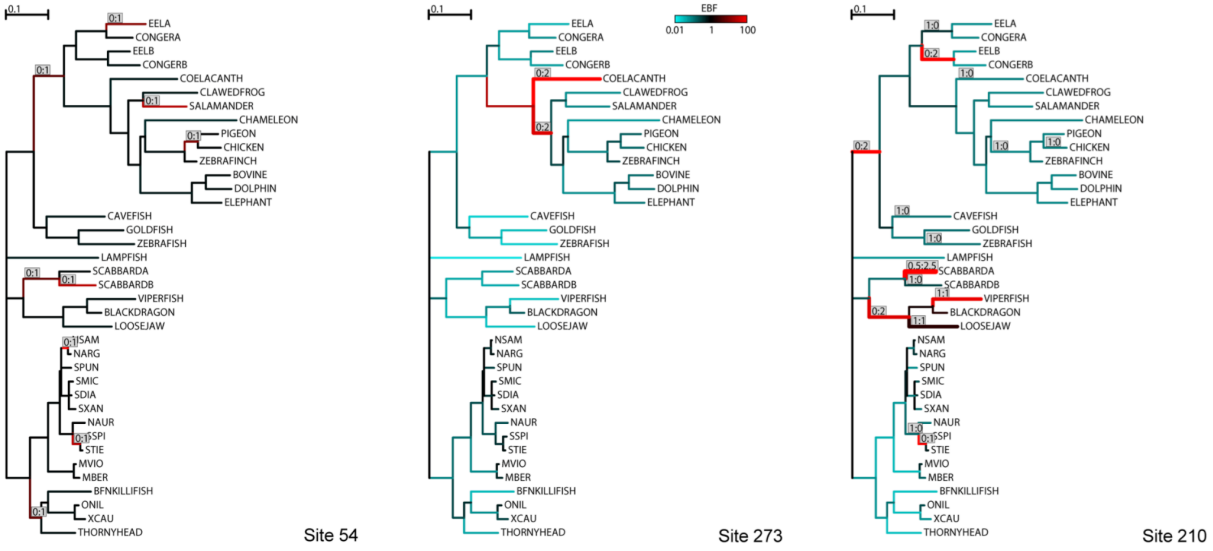


Figure 20. Comparison of FEL and MEME in detecting positive selection at individual codon sites in a vertebrate rhodopsin alignment. Branches are labeled with the count of synonymous substitutions, and their thickness reflects the minimal number of nucleotide substitutions. Branches are colored by the empirical Bayes factor (EBF) for positive selection: red indicates positive selection, teal indicates neutral or negative selection, and black indicates no information. MEME identified all three sites as under positive selection, while FEL reported varying selection pressures across sites. Reproduced from Murrell et al. (2012).

Structural and functional analyses provide insights into neofunctionalization by directly demonstrating how sequence changes influence protein structure and activity. One powerful approach is protein modeling, where computational methods predict the three-dimensional structure of proteins based on their amino acid sequences. Homology modeling, which infers an unknown structure using a known template, is particularly useful for visualizing how mutations may alter critical features such as active sites or binding interfaces (Webb & Sali, 2016). Additionally, molecular dynamics simulations offer a dynamic view, helping to understand how mutations affect protein stability and conformational changes over time (Hollingsworth & Dror, 2018).

Complementing these computational techniques, biochemical assays provide experimental validation of neofunctionalization by measuring functional changes in proteins. For instance,

enzyme kinetics studies can detect shifts in catalytic efficiency or substrate specificity, both of which are hallmark indicators of neofunctionalization (Aharoni et al., 2005). Similarly, binding assays can measure changes in ligand affinity, often linked to structural modifications in binding sites. Together, these structural and functional analyses offer a view of how sequence changes drive the evolution of new protein functions.

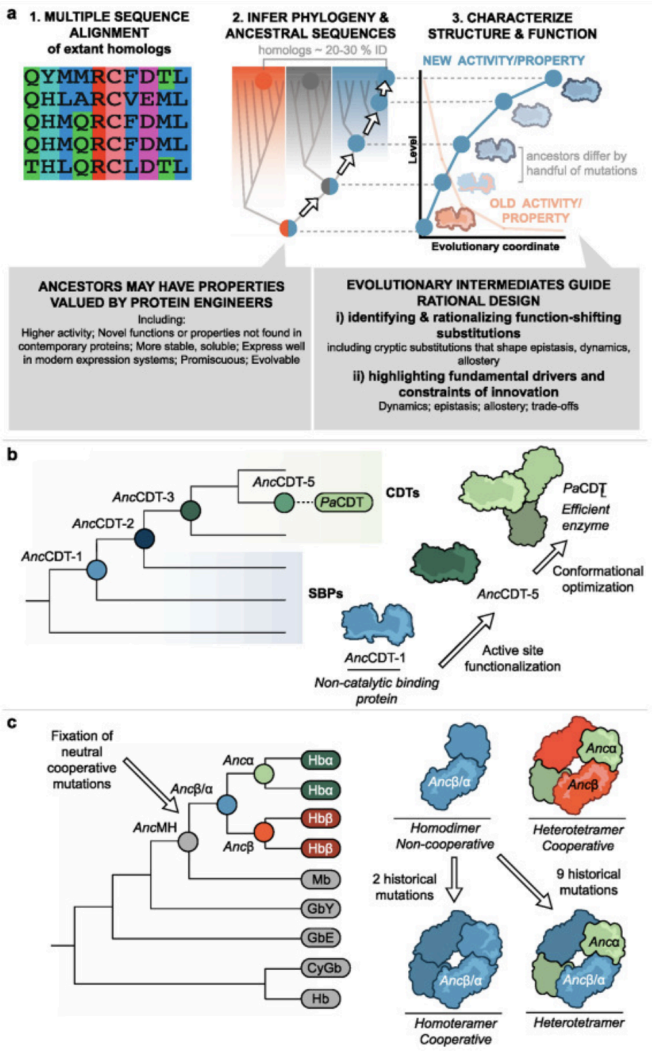


Figure 21. Overview and examples of ancestral sequence reconstruction (ASR). (a) General workflow of ASR, involving sequence collection, alignment, phylogeny calculation, ancestral sequence inference, and experimental characterization. (b) ASR revealed that contemporary cyclohexadienyl dehydratases (CDTs) evolved from a noncatalytic ancestral protein (AncCDT-1) through active site functionalization and remote mutations that altered the oligomeric state, favoring catalytically relevant states in the modern *Pseudomonas aeruginosa* CDT. (c) ASR demonstrated that modern heteromeric hemoglobin (Hb) evolved from a monomeric ancestor (AncMH) via a noncooperative homodimer (Ancβ/α), with two substitutions post-duplication conferring tetramerization and altered oxygen affinity. Reproduced from Spence (2021).

Differences in gene expression levels and patterns between gene duplicates can provide key evidence of functional divergence, a process that often precedes or accompanies neofunctionalization. Gene expression analysis techniques, such as quantitative real-time PCR and RNA sequencing, allow for quantification of gene expression across various tissues, developmental stages, or environmental conditions. Divergent expression patterns between duplicates suggest that each gene may have acquired specialized roles (Gu et al., 2002). This can translate in subfunctionalization, where one gene copy retains the original function while the other adapts to a different expression domain, which can later lead to neofunctionalization as the gene develops a new function. Proteomics, particularly mass spectrometry-based approaches, further complements gene expression studies by identifying and quantifying proteins in complex mixtures. Differences in protein abundance or post-translational modifications between gene duplicates can reveal additional layers of functional divergence, reflecting how the proteins produced by each gene copy may vary in their roles or regulatory mechanisms (Aebersold & Mann, 2016).

Phylogenetic methods are essential for tracing the evolutionary history of gene duplicates and inferring functional changes that may signal neofunctionalization. One powerful technique is ancestral sequence reconstruction (**Fig. 21**), which allows to infer and reconstruct the sequences of ancient proteins. These reconstructed ancestral proteins can then be experimentally characterized to determine their functions and compared with their modern descendants (Pauling et al., 1963). By performing functional assays on these "resurrected" proteins, we can pinpoint how specific mutations contributed to the development of new or altered functions over evolutionary time. Another key phylogenetic approach involves analyzing lineage-specific diversification. By constructing phylogenetic trees, we can identify lineages that have undergone rapid diversification or adaptive radiation, which often suggests periods of neofunctionalization (Innan & Kondrashov, 2010).

By combining comparative sequence analysis, structural and functional studies, gene expression profiling, and phylogenetic approaches, we can study the molecular mechanisms of neofunctionalization. These traditional techniques, when integrated with computational methods, such as machine learning, provide an even more powerful framework for understanding how gene duplicates acquire new functions. Machine learning has greatly enhanced the ability to analyze big datasets, predict functional outcomes, and automate complex tasks, offering new methods for studying evolutionary innovation and adaptation that were previously unreachable.

4. Machine Learning in Molecular Biology: Applications and Innovations

4.1. Introduction to Machine Learning in Biological Sciences

Machine learning (ML) is changing the way scientists analyze and interpret complex biological data. The integration of ML into biology began with early applications in sequence analysis and gene expression profiling, where algorithms were employed to identify patterns and correlations within large genomic datasets (Larrañaga et al., 2006).

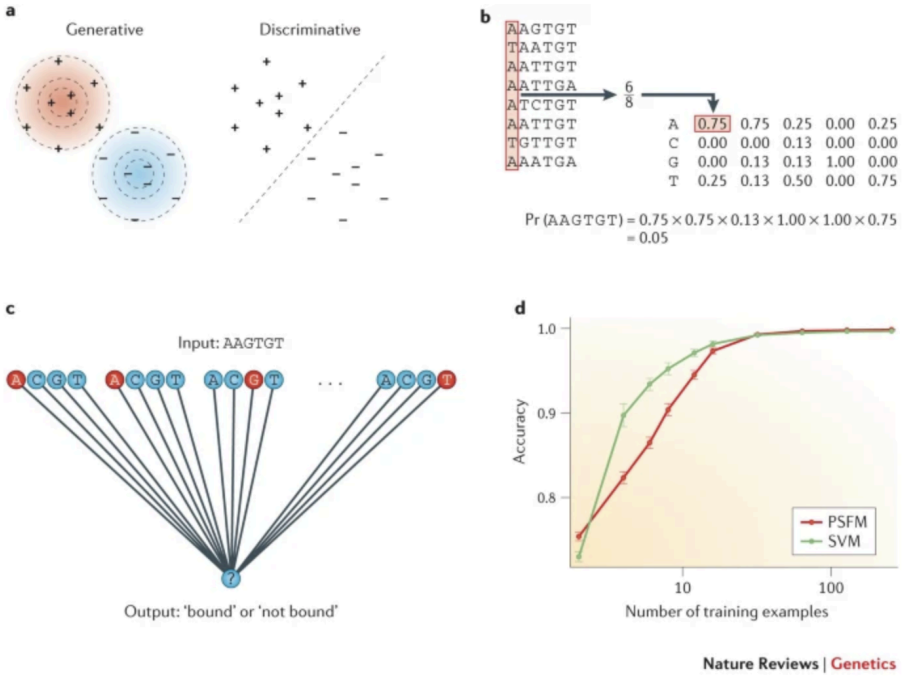


Figure 22. Two models of transcription factor binding. (a) Generative and discriminative models differ in interpretability and prediction accuracy, with generative models characterizing both classes completely and discriminative models focusing on the boundary between classes. (b) A position-specific frequency matrix (PSFM) model represents the frequency of each base at different positions, assuming independence across bases. (c) A linear support vector machine (SVM) model predicts transcription factor binding using labeled positive and negative examples. (d) The graph shows the mean accuracy (±95% confidence intervals) of PSFM and SVM in predicting transcription factor binding as a function of training set size. Reproduced from Libbrecht and Noble (2015).

These initial attempts demonstrated ML's potential to improve the accuracy and efficiency of biological analyses compared to traditional methods. As computational capabilities expanded and ML algorithms became more sophisticated, the influence of ML extended to diverse areas such as image analysis, where it aids in interpreting microscopic images; drug discovery, by predicting molecular interactions and potential drug candidates; and systems

biology, where it helps in modeling and understanding complex biological systems. The impact of ML in biology is underscored by its ability to handle vast datasets, uncover hidden patterns, and facilitate discoveries that were previously unreachable (**Fig. 22**) (Libbrecht and Noble, 2015). Historically, bioinformatics has relied heavily on statistical models and manual feature extraction to analyze biological data. These traditional methods, while effective to a certain extent, often required significant human intervention to identify relevant features and could struggle with the increasing complexity and volume of biological datasets.

Unlike traditional methods, AI-driven approaches enable models to learn hierarchical representations directly from raw data, eliminating the need for extensive manual feature engineering. This shift has not only enhanced predictive accuracy but also expanded the ability to manage and interpret unstructured data, such as genomic sequences and biological images. Consequently, AI-driven methodologies have facilitated breakthroughs in various biological domains (Ching et al., 2018; LeCun et al., 2015). Machine learning offers a number of tools for dealing with the complexities associated in the analysis of duplicated genes and their evolution. By leveraging large-scale datasets and sophisticated algorithms, ML can model the evolutionary patterns of proteins and predict functional residues with high precision. Machine learning models, particularly those based on deep learning, are adept at identifying patterns and correlations that may elude conventional analytical methods. These models can integrate various types of biological data, such as sequence information, structural data, and expression profiles (**Fig. 23**) (Almagro Armenteros et al., 2017; Hassanzadeh and Wang, 2016; Avsec et al., 2021). Furthermore, ML-driven approaches can facilitate the prediction of functional changes resulting from gene duplication, enabling researchers to identify potential new functions and evolutionary trajectories of duplicated genes. This capability not only enhances our understanding of protein evolution but also accelerates the discovery of novel biological functions and mechanisms.

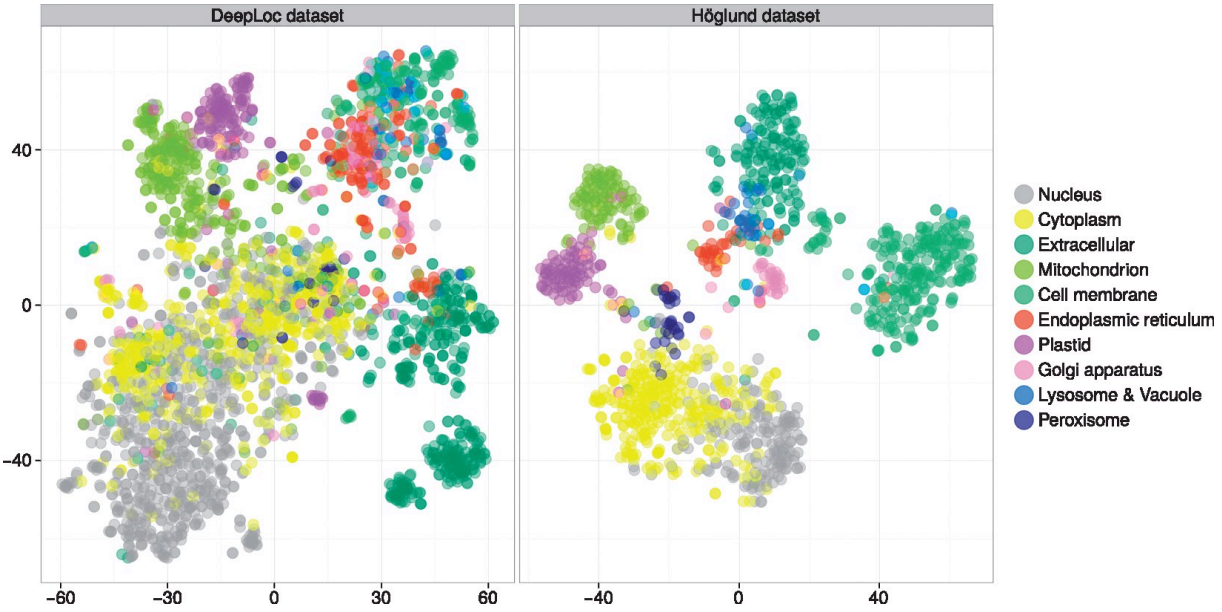


Figure 23. t-SNE representation of the context vector for a Conv A-BLSTM model trained on the DeepLoc and Höglund datasets, visualized for the respective test sets. Reproduced from Almagro Armenteros et al. (2017).

4.2. Machine Learning for Protein Structure Prediction

Proteins are fundamental macromolecules that perform numerous functions within biological systems, and their functionality is tied to their structural conformation. The structure of a protein is organized into four hierarchical levels: (1) Primary Structure: This refers to the linear sequence of amino acids linked together by peptide bonds. The specific order of amino acids determines the protein's unique characteristics and sets the foundation for its higher-level structures. (2) Secondary Structure: This level involves local folding patterns such as alpha-helices and beta-sheets, which are stabilized primarily by hydrogen bonds between the backbone atoms of the amino acids. These structures contribute to the overall stability and shape of the protein. (3) Tertiary Structure: This is the three-dimensional arrangement of the entire polypeptide chain, resulting from interactions among the side chains of amino acids. These interactions include hydrogen bonds, ionic bonds, hydrophobic interactions, and disulfide bridges, which collectively determine the protein's final shape and functional capabilities. (4) Quaternary Structure: Some proteins consist of multiple polypeptide chains, known as subunits, that assemble into a larger complex. The quaternary structure describes the spatial arrangement and interactions between these subunits (Branden and Tooze, 1999). Any alterations at the primary level can propagate through the

higher levels, potentially changing the protein's functionality and interactions within the cellular environment (Orengo et al., 1999).

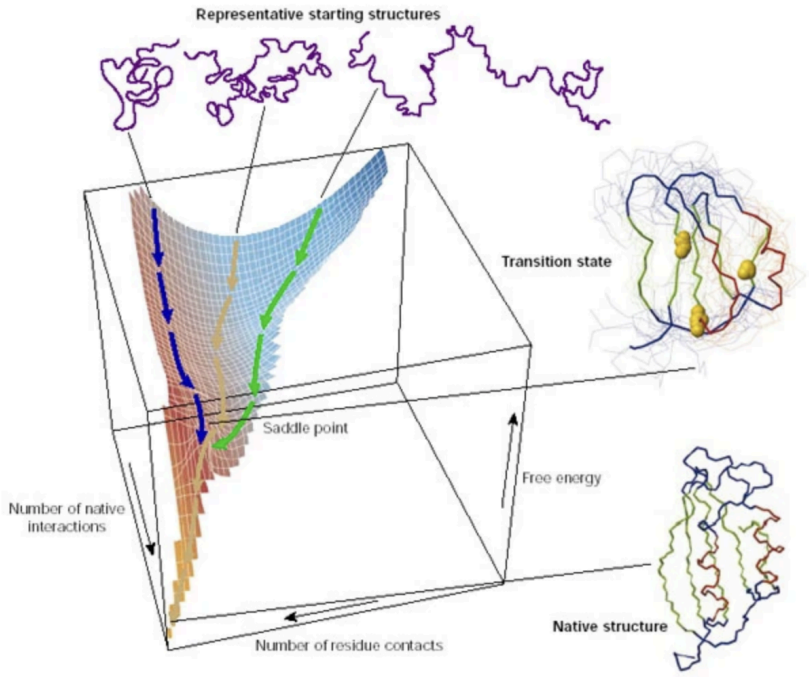


Figure 24. Schematic energy landscape for protein folding. The surface funnels various denatured conformations toward the unique native structure. The critical region is the saddle point representing the transition state, a barrier that must be crossed for folding to occur. Superimposed are ensembles of structures at different folding stages, with yellow spheres indicating 'key residues' essential for establishing the native fold. The native state is shown at the bottom, while unfolded species are represented at the top. Simplified folding trajectories are also indicated. Reproduced from Dobson (2003).

The precise arrangement of amino acids in space dictates how a protein interacts with other molecules, substrates, and cellular components. Structural features such as active sites, binding pockets, and conformational flexibility enable proteins to perform their specific biological roles effectively. For instance, enzymes rely on their structural configuration to facilitate catalytic activities, while structural proteins provide support and shape to cells and tissues. Moreover, even minor structural changes can lead to significant functional modifications, which is particularly relevant in the context of neofunctionalization following gene duplication. Understanding the relationship between protein structure and function allows to elucidate the mechanisms underlying protein activity, interactions, and regulation, thereby providing hints into biological processes and disease mechanisms (**Fig. 24**) (Anfinsen, 1973; Dobson, 2003). Homology modeling is a traditional method used to predict

the three-dimensional structure of a target protein based on its sequence similarity to one or more homologous proteins with known structures. This approach operates under the assumption that evolutionary related proteins share similar structures. The process involves aligning the target protein sequence with the sequences of homologous proteins, building a structural model based on this alignment, and refining the model to account for any discrepancies or insertions. Homology modeling has been widely adopted due to its relatively high accuracy when suitable homologs are available, making it a valuable tool for predicting protein structures in cases where experimental data is lacking (Schwede, 2003; Martí-Renom et al., 2000).

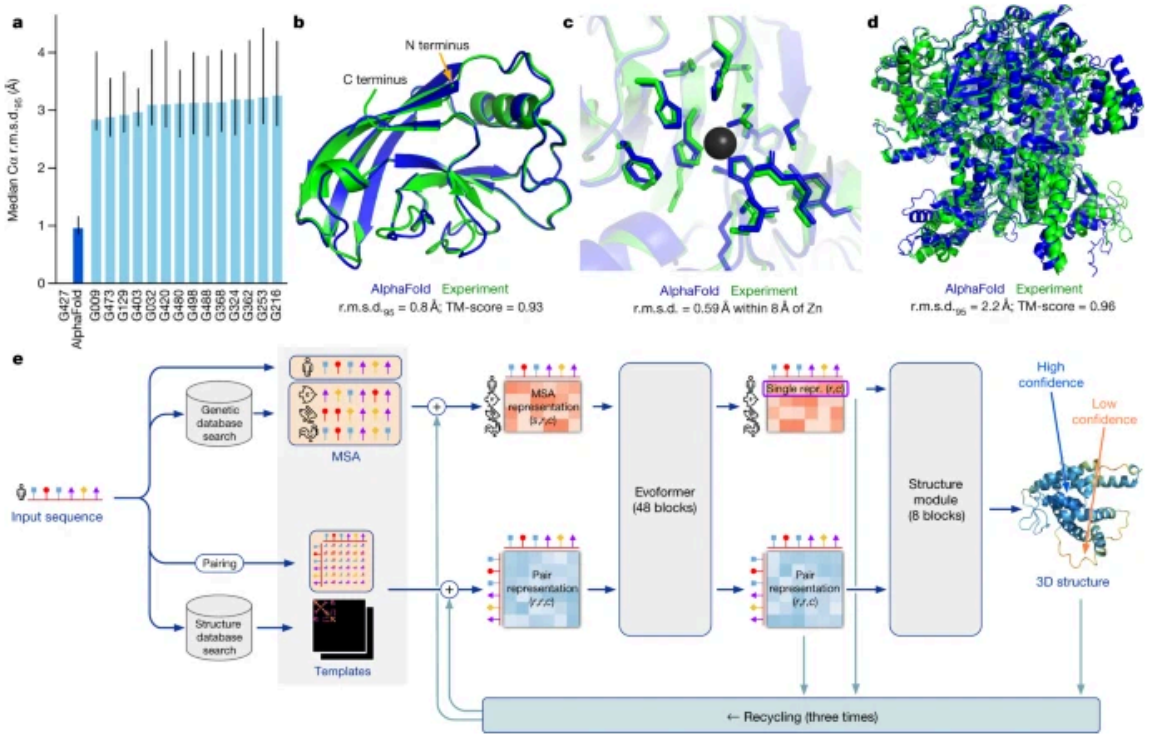


Figure 25. AlphaFold produces highly accurate protein structures. (a) Performance of AlphaFold on the CASP14 dataset compared to the top 15 entries, with median and 95% confidence intervals shown. (b) AlphaFold prediction for CASP14 target T1049 (blue) versus the true experimental structure (green). (c) Accurate prediction of a zinc-binding site in CASP14 target T1056. (d) Correct domain packing for CASP14 target T1044, a 2,180-residue protein, predicted post-CASP. (e) AlphaFold model architecture showing information flow between components. Reproduced from Jumper et al. (2021).

AlphaFold, developed by DeepMind, represents an advancement in the field of protein structure prediction. Introduced in 2018, AlphaFold gained attention by winning the Critical Assessment of Structure Prediction (CASP13) competition, where it demonstrated high

accuracy in predicting protein structures. This achievement showed the potential of deep learning techniques in handling complex biological problems. Building on this success, AlphaFold made substantial improvements in the subsequent CASP14 competition in 2020, achieving atomic-level accuracy. These successes not only highlighted AlphaFold's excellent predictive capabilities but also highlighted its potential to transform structural biology by providing reliable models for proteins that are difficult to study experimentally (**Fig. 25**) (Jumper et al., 2021; Senior et al., 2020).

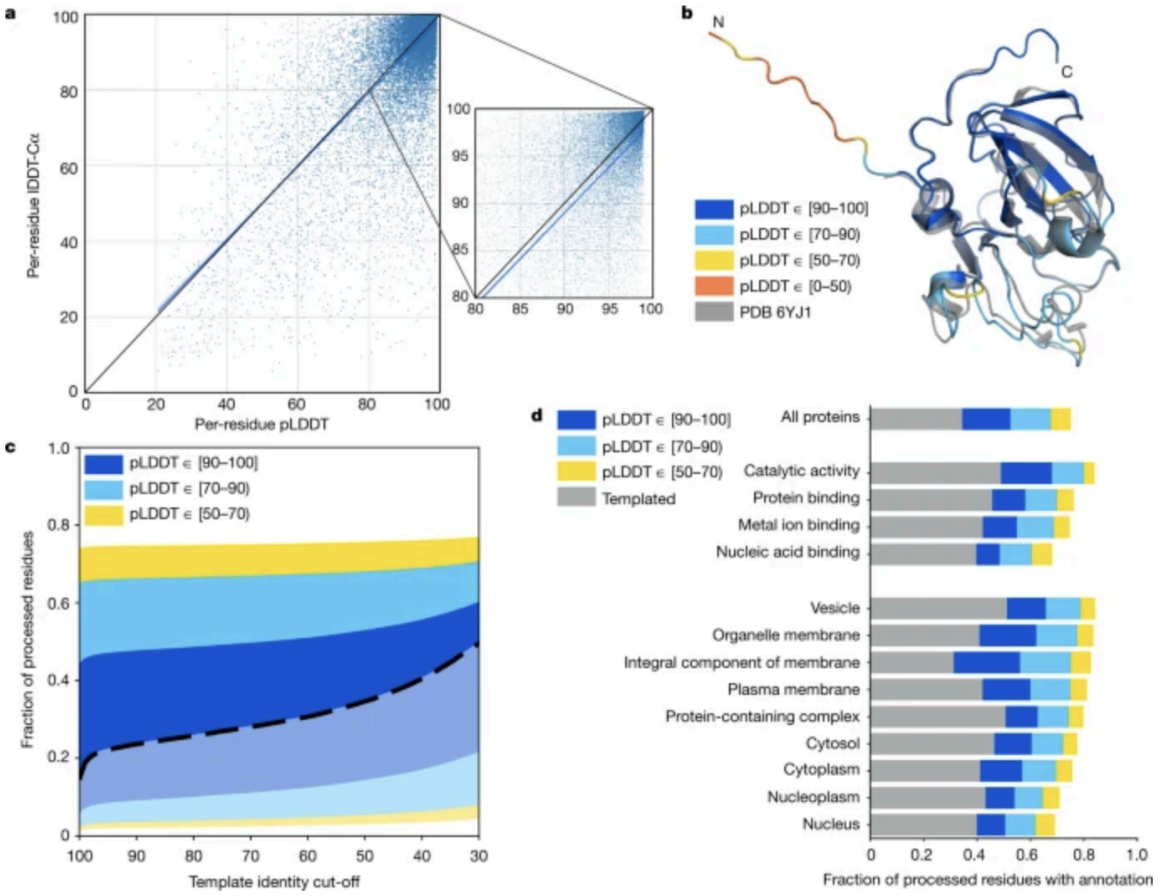


Figure 26. Model confidence and added coverage by AlphaFold. (a) Correlation between per-residue pLDDT and lDDT-Cα for a held-out set of recent PDB chains. The scatterplot shows a 1% subsample of residues, with the blue line indicating a least-squares fit. (b) AlphaFold prediction (colored by confidence) versus experimental structure for a CASP14 target (PDB: 6YJ1). (c) Model confidence for all predicted residues, with lighter colors indicating residues covered by templates of specified identity levels. (d) Added proteome coverage for high-level GO terms compared to template-based coverage with >50% sequence identity. Reproduced from Tunyasuvunakool et al. (2021).

At the core of AlphaFold's success are advanced deep learning techniques, including convolutional neural networks (CNNs) and attention mechanisms (**Fig. 28**) (Vaswani et al., 2017; Rao et al., 2020). These techniques enable AlphaFold to predict inter-residue distances and angles with high precision by effectively modeling the relationships within protein sequences. AlphaFold integrates evolutionary information derived from multiple sequence alignments, allowing it to capture patterns and co-evolutionary signals that are indicative of structural features. The architecture of AlphaFold is transformer-based, leveraging the power of attention mechanisms to focus on relevant parts of the protein sequence during prediction. This integration of deep learning methodologies allows AlphaFold to generate highly accurate structural models by learning hierarchical representations directly from raw sequence data. AlphaFold's accuracy in protein structure prediction has been reported as a major breakthrough, effectively addressing the long-standing protein folding problem. Its ability to predict structures with atomic-level precision has implications for various domains within structural biology. For instance, AlphaFold can accelerate drug discovery by providing accurate models of target proteins, thereby facilitating the identification of potential drug-binding sites and the design of therapeutic molecules. Additionally, AlphaFold enhances our understanding of diseases related to protein misfolding, such as Alzheimer's and Parkinson's, by enabling detailed structural analyses of pathogenic proteins. Moreover, AlphaFold contributes to filling gaps in structural databases, offering reliable models for proteins that have yet to be experimentally characterized. (**Fig. 26**) (Callaway, 2020; Tunyasuvunakool et al., 2021).

4.3. Protein Sequence Embeddings and Their Analytical Uses

Sequence embeddings represent protein sequences as numerical vectors that embed their underlying biochemical and biophysical characteristics. These embeddings transform the linear amino acid sequences into high-dimensional spaces, enabling machine learning models to process and analyze them effectively. The concept of sequence embeddings is inspired by word embeddings in natural language processing (NLP), where words are represented as vectors that capture semantic relationships (Asgari and Mofrad, 2015). Similarly, protein sequence embeddings aim to preserve the functional and evolutionary relationships between amino acids, allowing for the identification of patterns and similarities useful for understanding protein function and evolution. By converting sequences into continuous vector spaces, embeddings facilitate various downstream tasks such as classification, clustering, and prediction of protein properties (Yang et al., 2019). Protein

sequence embeddings are similar to word embeddings used in NLP, such as Word2Vec and GloVe, which capture the semantic relationships between words by positioning similar words close to each other in a vector space. In the context of proteins, sequence embeddings achieve a similar objective by representing amino acid sequences in a manner that reflects their functional and evolutionary relationships. Just as word embeddings allow NLP models to understand context and meaning, protein embeddings enable computational models to capture functional motifs and evolutionary conservation within protein sequences. This similarity facilitates the application of advanced NLP techniques to proteomics data. Consequently, protein sequence embeddings not only enhance the interpretability of protein data but also leverage the robustness of NLP models to advance the field of bioinformatics (Mikolov et al., 2013; Heinzinger et al., 2019).

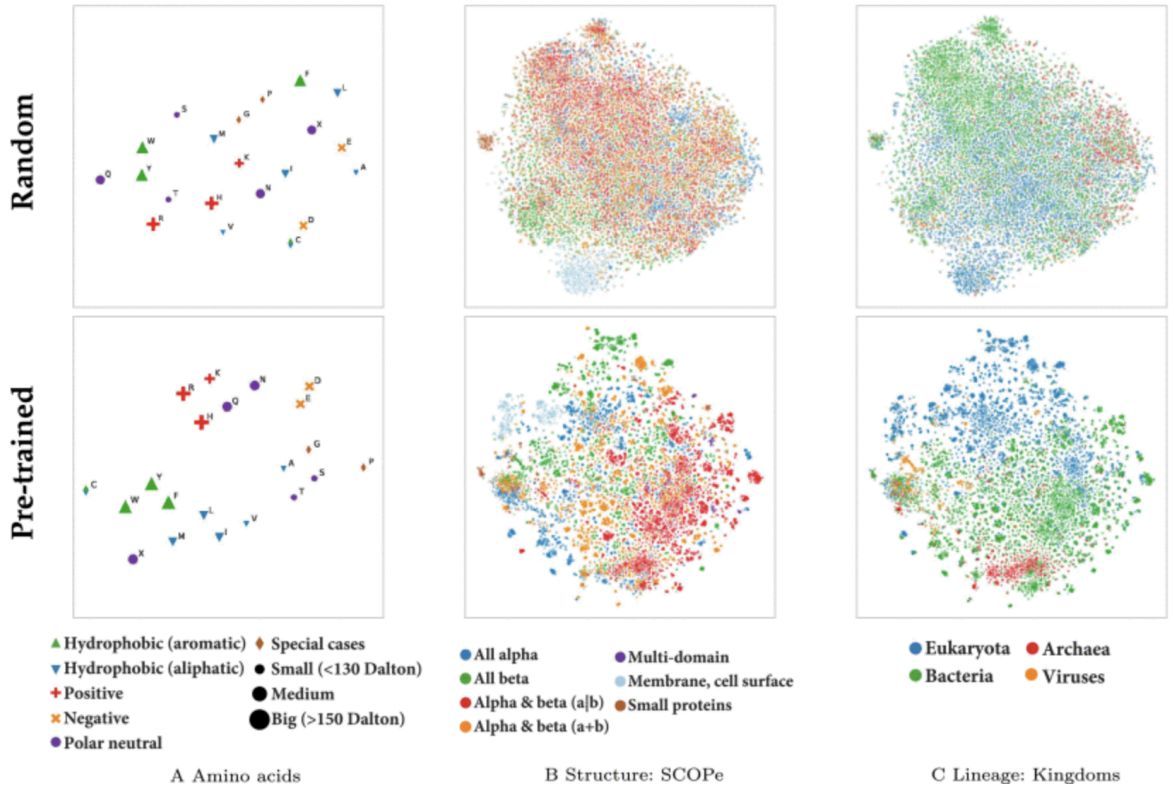


Figure 27. Protein language models (LMs) learned constraints. t-SNE projections of information extracted by the ProtT5-XL-U50 model. (Upper row) Random projections before training. (Lower row) Projections after pre-training on BFD & UniRef50. (A) Amino acids are highlighted by biophysical features. (B) Proteins are annotated by structural class (SCOPe). (C) Proteins are distinguished by their kingdom of life. The lower row demonstrates clear clustering patterns after training, with fine-grained distinctions in biophysical features, structural class, and organism origin. Reproduced from Elnaggar et al. (2021).

Protein language models have revolutionized computational biology by adapting advanced natural language processing (NLP) techniques to interpret and analyze protein sequences. ProtTrans T5, an extension of the Text-to-Text Transfer Transformer (T5) architecture, is specifically engineered to handle the unique complexities of amino acid sequences. Unlike traditional NLP models that operate on discrete word tokens, ProtTrans T5 incorporates a specialized tokenization strategy that integrates biochemical properties of amino acids, enabling the model to capture both local sequence motifs and long-range interactions essential for protein folding and function.

Pre-trained on extensive databases such as UniRef100, which encompasses a diverse array of protein sequences across various organisms, ProtTrans T5 leverages transfer learning to develop robust embeddings that reflect evolutionary conservation and structural dependencies. The model employs multi-head self-attention mechanisms, allowing it to simultaneously focus on multiple positions within a sequence, enhancing its ability to identify functionally critical residues and interaction sites. Additionally, ProtTrans T5 is trained using a combination of masked language modeling and next amino acid prediction tasks, which facilitate a deeper understanding of sequence context and evolutionary constraints. Comparative studies have demonstrated that ProtTrans T5 outperforms other protein language models, including ESM and BERT-based architectures, in key bioinformatics applications such as enzyme function prediction, subcellular localization classification, and protein-protein interaction mapping (**Fig. 27**) (Rao et al., 2021; Elnaggar et al., 2021). Furthermore, the model's interpretability is enhanced through attention weight visualization, providing valuable insights into the significance of specific amino acids in determining protein function and stability. This feature not only aids in hypothesis generation but also facilitates experimental validation, bridging the gap between computational predictions and laboratory research. By integrating domain-specific knowledge with state-of-the-art transformer architectures, ProtTrans T5 represents a significant advancement in the application of machine learning to protein biology, offering accuracy and biological relevance in the analysis of protein sequences.

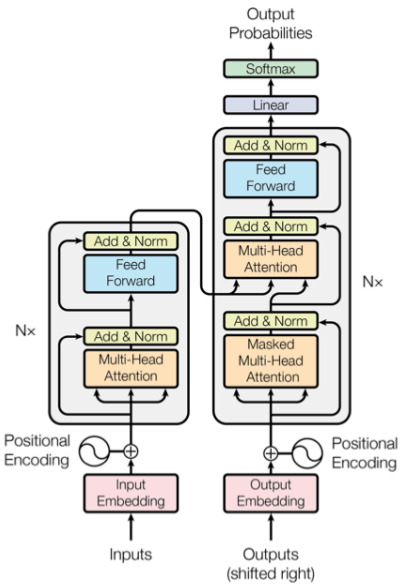


Figure 28. The Transformer model architecture, which utilizes self-attention mechanisms to process input sequences in parallel, enabling more efficient handling of long-range dependencies in data. Reproduced from Vaswani et al. (2017).

Through the extensive training, these models learn to recognize patterns that are indicative of evolutionary conservation and functional motifs within protein sequences. The ability to capture co-evolutionary signals, where mutations in one part of the sequence are compensated by changes in another, allows the models to identify residues that are critical for maintaining protein function and stability. Additionally, these embeddings can highlight regions of the protein that are involved in specific interactions or catalytic activities, thereby providing insights into functional residues without the need for explicit multiple sequence alignments.

By integrating different types of biological data and leveraging deep learning algorithms, embedding models facilitate the understanding of protein function and evolution, enabling researchers to make predictions about protein behavior and interactions (Rives et al., 2021). Embedding models play a key role in the identification of functional residues within proteins by leveraging the information encoded in the sequence embeddings. These models can assign importance scores or attention weights to specific amino acids, highlighting their significance in the protein's function. Techniques such as gradient-based saliency maps and layer-wise relevance propagation are employed to interpret the contributions of individual residues to the model's predictions. Gradient-based saliency maps involve calculating the gradient of the output with respect to the input residues, thereby identifying which residues have the most significant impact on the model's decision-making process. Layer-wise

relevance propagation, on the other hand, traces the relevance of each residue through the layers of the neural network, providing a detailed understanding of how each part of the sequence contributes to the overall prediction. These methods enable pinpoint key residues that are essential for the protein's activity, stability, and interaction with other molecules, thereby facilitating targeted studies and experimental validations (Shrikumar et al., 2017). Traditional sequence alignment methods primarily focus on identifying regions of similarity between protein sequences, relying on sequence homology to infer functional and evolutionary relationships. While effective for closely related sequences, these methods often struggle to detect remote homologs and functional residues in highly divergent sequences where sequence similarity is low. In contrast, sequence embeddings generated by deep learning models can capture complex patterns and relationships that extend beyond simple sequence similarity. Embeddings are capable of recognizing functional motifs and evolutionary constraints even in the absence of significant sequence identity, making them powerful tools for studying proteins with different backgrounds. This ability to detect remote homologs and identify functionally important residues in divergent sequences enhances the protein analysis, allowing for the exploration of a broader range of proteins and the discovery of novel functional insights that may be overlooked by traditional alignment-based approaches (Brandes et al., 2022).

Models such as ProtTrans T5 learn to recognize a wide array of sequence patterns, functional motifs, and evolutionary signatures. ProtTrans embeddings have proven to be highly effective in predicting various aspects of protein function, including the identification of functional sites, subcellular localization, and protein-protein interactions. By generating rich, context-aware representations of protein sequences, these embeddings enable machine learning models to discern subtle patterns and features that are indicative of specific functional properties. For instance, functional site prediction leverages the embeddings to identify regions of the protein that are critical for enzymatic activity or ligand binding, thereby highlighting potential active sites and binding interfaces (**Fig. 29**) (Littmann et al., 2021). Additionally, ProtTrans embeddings facilitate the prediction of subcellular localization by capturing sequence motifs and structural features that determine a protein's cellular compartmentalization (Stärk et al., 2021).

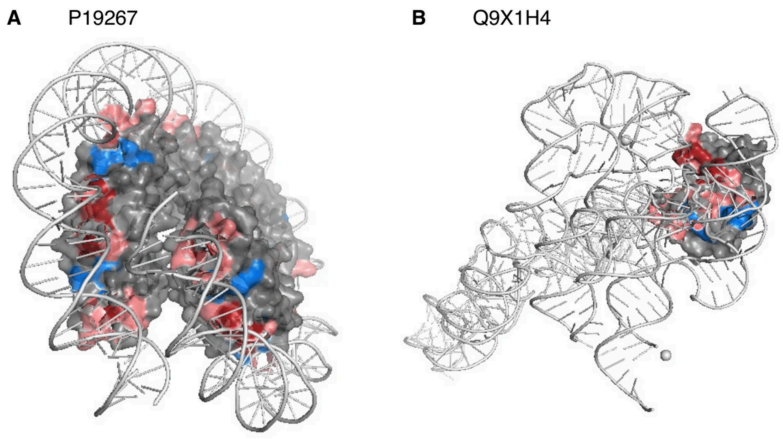


Figure 29. Annotations from low-resolution structures supported by reliable predictions. (A) For the DNA-binding protein HMf-2 (UniProt ID: P19267), bindEmbed21DL predicted four nucleic acid-binding residues with high confidence (dark red), matching 10 of 13 experimentally annotated DNA-binding residues in a lower-resolution structure (PDB: 5T5K). (B) For the ribonuclease P protein component (UniProt ID: Q9X1H4), bindEmbed21DL predicted four binding residues (dark red) not present in the PDB structure 6MAX but validated by two low-resolution structures (PDB: 3Q1Q and 3Q1R). Overall, 71% of binding residues were correctly predicted. Reproduced from Littmann et al. (2021).

The integration of machine learning into the study of gene duplication and protein evolution offers unparalleled insights into the mechanisms driving functional divergence, particularly in enzymes. Through advanced computational tools like AlphaFold and ProtTrans T5, alongside traditional sequence and phylogenetic analyses, we can study the complexities of how duplicated genes evolve and acquire new functions. In the context of vertebrates, this framework allows for investigation into how gene duplications lead to enzymatic diversification, contributing to adaptations and evolutionary novelty. By applying these methods, this work ultimately aims to advance our understanding of functional divergence in vertebrate enzymes, revealing the molecular underpinnings of evolutionary innovation.

5. Integrating Concepts: From Gene Duplication to Functional Divergence

Understanding gene duplication is fundamental to appreciating the mechanisms that drive genetic diversity and evolutionary change. This foundational knowledge not only elucidates how genetic variations arise but also sets the stage for exploring the subsequent processes of neofunctionalization and the evolution of new enzymatic functions. Through examination of methodologies and concepts involved in identifying duplicated genes, we have established

clear definitions and distinctions between homology, orthology, and paralogy. This classification is essential for accurately interpreting evolutionary and functional relationships among genes. The integration of computational tools, such as sequence alignments, substitution matrices, and conserved residue analysis, alongside biological insights, underscores the complexity of gene duplication and its evolutionary implications. Detecting neofunctionalization in proteins requires a multifaceted approach that combines comparative sequence analysis, structural and functional studies, expression profiling, phylogenetic analyses, and functional genomics. These methods collectively enhance our understanding of how gene duplication contributes to the emergence of new protein functions and, consequently, to biological diversity. Moreover, identifying the specific residues responsible for neofunctionalization involves a sophisticated strategy that integrates evolutionary conservation scores, embedding importance measures from machine learning models, and structural context. By focusing on highly conserved residues located in functionally relevant regions of proteins, we can pinpoint the molecular alterations that drive functional divergence following gene duplication. This approach not only identifies candidate residues but also elucidates the underlying mechanisms facilitating the acquisition of new functionalities. In summary, the exploration of gene duplication and its consequences through evolutionary analyses and functional predictions provides a synergistic framework for uncovering the molecular basis of biological diversity. By correlating evolutionary changes at the residue level with shifts in protein function, we gain profound insights into the processes that generate new functionalities and drive evolutionary innovation. This integrated perspective is essential for advancing our understanding of protein evolution and the diversification of biological functions, opening the way for future research into the dynamics of genetic and functional evolution.

Materials and Methods

1. Sequence Collection and Mapping

1.1. Homologous Sequence Collection

To systematically analyze gene duplications in humans, we developed a filtering pipeline starting with the UniProtKB/Swiss-Prot database (The UniProt Consortium et al., 2023), focusing exclusively on reviewed human protein sequences to ensure high-quality annotations. BLASTP (version 2.15.0) (Altschul et al., 1990) was employed using the human proteome dataset (UniProt UP000005640), which includes 20,434 reviewed uniprot sequences. The BLASTP search was configured with the following parameters: maximum number of target sequences per query set to 50, maximum number of high-scoring segment pairs (HSPs) limited to 1, and an e-value threshold of 1e-3. From the BLASTP results, we extracted intraspecies Best Reciprocal Hits (BRHs) to identify high confidence homologous pairs. BRHs were defined as pairs where each sequence is the best match for the other in both directions, minimizing the inclusion of distant homologous and isoforms from alternative splicing events. We further refined our dataset by filtering the BRH pairs based on their Pfam domain architectures. Pfam provides a collection of protein families, each represented by multiple sequence alignments and hidden Markov models. Gene pairs with identical Pfam architectures were selected to ensure conserved structural and functional domains and to avoid subfunctionalization events (Mistry et al., 2021). Additionally, we incorporated enzymatic function as a filter by selecting gene pairs for which at least one member possessed a Rhea or Enzyme Commission (EC) number. Rhea is a curated resource of biochemical reactions (Bansal et al., 2022), while EC numbers provide a hierarchical classification of enzyme-catalyzed reactions. Including gene pairs with such annotations allowed us to focus on proteins with characterized or predicted enzymatic activities, facilitating the exploration of functional divergence post-duplication.

1.2. Chromosomal Mapping and Gene Pair Visualization

To analyze the chromosomal distribution of duplicated gene pairs, we first mapped each gene to its respective chromosome using the UniprotKB/Swiss-Prot annotations.

Chromosomes were designated as c1 through c22, along the sex chromosomes X and Y. Circle plot facilitates the representation of complex genomic data in a circular layout, allowing for the clear depiction of relationships and duplications across chromosomes. The plots illustrate duplication events identified immediately after the BRH filtering step, and the distribution of the refined set of gene pairs with enzymatic activity. In each plot, chromosomes are arranged around the circle. Duplication events are represented by lines connecting the corresponding chromosomal locations of each gene pair. Genes located on the same chromosome are connected internally, while those on different chromosomes are linked externally, highlighting both intra- and inter-chromosomal duplications.

2. Orthologous Sequence Collection from Ensembl Compara

In order to better understand the evolutionary trajectories and functional divergence of the identified human duplicated gene pairs, we extended our analysis to include orthologous sequences from a broad range of vertebrate species. Utilizing the Ensembl Compara database (Release 112, Ensembl Genome Browser) (Harrison et al., 2024), which provides phylogenetic relationship and homology annotations across species, we aimed to reconstruct the duplication events along the vertebrate lineage and enrich our sequence dataset for downstream analyses. We queried Ensembl Compara to obtain orthologs of each gene of each gene in the duplicate pair across all available vertebrate species. The orthology relationships in Ensembl Compara are classified based on phylogenetic and syntenic evidence. We specifically filtered for genes with a “one-to-one” orthology relationship. To maintain the quality and reliability of the sequence alignments, we excluded any orthologous sequences containing ambiguous amino acids, denoted by the character ‘X’. Such ambiguities could arise from sequencing errors, assembly issues, or unresolved regions in the genome and could potentially disrupt alignment accuracy. By removing these sequences, we ensured that the subsequent multiple alignments would be based on high confidence amino acid residues, enhancing the precision of conservation and divergence analyses.

For each species, we assessed the presence or absence of both genes in each human duplicated gene pair. We divided our set in “both genes present” if both orthologs of the human gene pair are found in species, “single gene present” if only one gene of the pair has a corresponding ortholog, “neither gene present” if no orthologs for either gene are identified in the species. Duplicated gene pairs were sorted in descending order based on the number of species in which both genes are present. We visualized the presence and absence of gene pairs across species using a heatmap representation, where gene pairs were sorted

based on their distribution across vertebrate taxa. A phylogenetic dendrogram was used to classify species into major taxonomic classes: Mammalia, Sauropsida, and Actinopterygii. Cells in the heatmap were colored based on the presence status: dark color (high alpha) if both genes are present, medium color (medium alpha) if only one gene is present, light color (zero alpha) if neither gene is present. We quantified the number of duplicate gene pairs shared among the three taxonomic classes. Categories were established based on the combination of classes in which the duplicates are present (e.g., gene pairs were considered “only in mammals” if duplicated genes were present in at least ten mammals species and either absent or found in fewer than ten species in both sauropsids and fishes). We analyzed how gene duplication events are distributed across taxonomic classes using a Sankey diagram. This visualization highlights how gene pairs from distinct duplication categories are assigned to major vertebrate groups based on their presence across species. The flow diagram illustrates the movement of duplicated gene pairs, showing their contribution to different taxonomic classes and the degree of overlap between groups.

3. Multiple Sequence Alignment and Residue Classification

3.1. MSA Generation and Cleaning

We conducted multiple alignments for each duplicated gene pair and their respective orthologs to accurately assess the sequence conservation and facilitate downstream analyses. The alignment process was executed using Clustal Omega (CLUSTALO) (Version 1.2.4) (Sievers et al., 2011) on our University's High Performance Computing (HPC) cluster, ensuring efficient handling of the large dataset comprising all pairs containing gene A and gene B with their respective orthologous whose sequence number can potentially reach up to 354 sequences. Post-alignment, the resulting multiple sequence alignments (MSAs) underwent a rigorous cleaning process to enhance alignment quality and reliability. The cleaning was accomplished through a custom Python script, which employed the following steps. (1) Removal of sequences with ambiguous residues: sequences containing ambiguous amino acids (e.g., ‘X’) were excluded to prevent inaccuracies in downstream analyses. (2) Column filtering based on gap thresholds: columns with a high proportion of gaps were identified using a threshold of 0.3, meaning that if more than 30% of sequences in a column were gaps, the column was flagged for removal; columns exhibiting close gaps within a windows of 4 residues were also targeted to eliminate regions with excessive indels that could disrupt alignment integrity. (3) Exclusion of redundant or erroneous sequences:

sequences that deviated significantly in length or possessed additional sequence regions compared to their orthologs were considered potential errors and removed. (4) group based cleaning: alignments were processed in groups based on sequence identifiers to ensure consistent cleaning across related sequences. (5) Realignment: realignment was performed using Clustal Omega to optimize MSA after initial cleaning. (6) Dropping gapped columns: columns composed entirely of gaps were removed to streamline the alignment.

3.2. Alignment Scoring Metrics: ESPript and Custom Approach

To quantify the conservation and divergence within multiple sequence alignments (MSAs) of duplicated gene pairs and their orthologs, we developed a set of scoring metrics inspired by the ESPript tool (Gouet et al., 1999). These metrics facilitate the identification of conserved and variable regions across aligned sequences, enabling a detailed analysis of functional and evolutionary dynamics. Our scoring framework comprises three primary metrics: In-Group Score (ISc), Cross-Group Score (XSc) and Total Score (TSc). The metrics collectively assess the similarity and divergence within and between predefined groups of sequences in each column of the MSA.

The In-Group score (ISc) measures the average similarity within predefined groups of sequences. This classical similarity score assesses the conservation of residues among sequences belonging to the same functional or evolutionary group (different orthologous groups in our case). For a given column in the MSA comprising residues from a single group, the ISc is calculated as the average of all possible pairwise similarity scores within that group. Mathematically, for a column with residue R = {R1, R2, …, Rn}:

In-Group Score (ISc):

$$ISc = \frac{\sum_{i<j} S(R_i, R_j)}{\frac{n(n-1)}{2}} \quad (1)$$

Where S(Ri, Rj) represents the substitution score between residues Ri and Rj based on a chosen substitution matrix, in our case a BLOSUM62 (Henikoff and Henikoff, 1992). For example, for a column with residues A, C and D:

Example (ISc):

$$ISc = \frac{S(A,C) + S(A,D) + S(C,D)}{3}$$

The Cross-Group Score (XSc) evaluates the average similarity across different groups of sequences. This metric captures the conservation of residues between sequences from distinct functional or evolutionary groups. For a column containing residues from multiple groups, the XSc is computed as the average of all pairwise similarity scores between residues from different groups. If the column is divided into k groups, the XSc is the mean of the average similarities between each pair of groups. For example, for a column with residues divided into three groups: (A, C, D), (D, E), and (G):

Cross-Group Score (XSc):

$$XSc = \frac{\left(\frac{S(A,D)+S(A,E)+S(C,D)+S(C,E)+S(D,D)+S(D,E)}{6}\right) + \left(\frac{S(A,G)+S(C,G)+S(D,G)}{3}\right) + \left(\frac{S(D,G)+S(E,G)}{2}\right)}{3} \quad (2)$$

The total score (TSc) represents the mean of the In-Group Score and the Cross-Group Score, providing an overall measure of conservation within and between groups.

Total Score (TSc):

$$TSc = \frac{ISc + XSc}{2} \quad (3)$$

The Differential score (Dsc) represents the difference between in-group conservation and cross-group conservation within each alignment column. This score helps highlight columns where intra-group conservation is strong and inter-group differences are significant. In our custom implementation, DSc is calculated follows:

Custom Differential Score (DSc):

$$DSc = \min(ISc) - XSc \quad (4)$$

ESPript calculates the DSc as a normalized difference *((ISc-XSc)/2)*, providing a balanced

measure of differential conservation across all groups. By using the minimum ISc across all groups, our method emphasizes the lowest in-group conservation relative to cross-group conservation. This approach ensures that all groups maintain at least one high level of intra-group conservation while simultaneously exhibiting low cross-group conservation. Consequently, it prevents scenarios where a single highly conserved group inflates the average ISc, masking lower conservation in other groups. High custom Dsc indicates that all groups have high intra-group conservation and that cross-group conservation is low, suggesting significant divergence between groups.

3.3. Residue Classification: Robust, Adaptive, Plastic

To explore the functional and evolutionary dynamics of residues within duplicated gene pairs and their orthologs, we classified each residue in the multiple sequence alignment (MSAs) into three distinct categories: Robust, Adaptive and Plastic. This classification facilitates the identification of residues with varying degrees of conservation and divergence, which are important for understanding protein evolution and functional divergence. We began by aggregating the scoring data generated from our alignment scoring metrics (ISc, Xsc, Tsc and DSc). The scores were stored in individual files and residues corresponding to gaps in at least one of the duplicated genes were excluded to ensure that only fully aligned and valid residues were considered for classification. Thresholds were established to differentiate between Robust, Adaptive and Plastic residues based on the distribution of DSc and TSc scores. The adaptive threshold was defined as the median of the DSc scores plus one standard deviation (2.09). This threshold ensures that only residues with high differential conservation are classified as adaptive. The robust threshold was defined as the 75th percentile of the TSc scores (5.00). This threshold ensures that only the top quartile of conserved residues across groups are classified as robust. Based on the determined thresholds, each residue was classified into one of the three categories using the following criteria: robust residue if (Tsc >= adaptive threshold and DSc < adaptive threshold), representing the highly conserved residues across all groups, indicating a probable essential functional or structural role; adaptive residue if (DSc >= adaptive threshold), indicating residues that are conserved within specific groups but divergent between groups, suggesting roles in functional diversification post-gene duplication; plastic residue where all the residues that did not meet the criteria for robust or adaptive categories, reflecting variable residues that may contribute to structural flexibility or less critical functional roles.

4. Divergence Scoring and Functional Analysis

4.1. Residue-Level Divergence Scoring

To assess the functional divergence of amino acid residues within duplicated gene pairs and their orthologs, we employed a combination of conservation-based and embedding-based metrics. These metrics enable the identification of conserved residues indicative of functional importance and residues that signify functional divergence between protein groups.

Conservation-Based Metrics

Conserved residues with high conservation scores serve as indicators of potential functional residues, particularly when they are surrounded by regions with lower conservation scores. These conserved residues are crucial for maintaining the structural integrity and fundamental functions of proteins. The differential score (DSc) quantifies the difference between in-group conservation and cross-group conservation within each alignment column. High DSc values denote significant divergence between groups, highlighting Adaptive residues that have evolved to confer specialized functions post-gene duplication. To quantify the extent of functional divergence within protein pairs, we introduced the DSc Length (dsclen) metric. This metric is calculated as the number of DSc residues above threshold / max (protein length 1, protein length 2). A threshold for DSc (2.09) was established based on preliminary analyses to identify residues with significant differential conservation. For each protein pair, the number of residues with DSc values exceeding the defined threshold was counted. The count was normalized by dividing by the maximum length of the protein in the pair, yielding the dsclen score. This normalization accounts for variations in protein lengths across different pairs. A higher dsclen score indicates a greater extent of functional divergence, suggesting that a substantial proportion of residues have adapted to fulfill specialized roles in the respective protein groups.

Embedding-Based Metrics

We utilized sequence embeddings derived from the ProtTrans T5 model to capture the biochemical properties and contextual information of amino acid residues. Sequence embeddings transform amino acid sequence into high-dimensional vectors, encapsulating information about residue characteristics and their evolutionary context. We employed the ProtTrans T5 model, a transformer-based architecture pretrained on large-scale protein sequence data, to generate embeddings for each amino acid residue (Elnaggar et al., 2022). For each residue pair in the alignment, the corresponding embedding vectors were extracted

from the ProtTrans T5 model. For each amino acid residue pair, the delta score was calculated as the Euclidean distance between their respective embedding vectors. Embedding-based scores are context-dependent, meaning the similarity score between two amino acids is influenced by the neighboring residues and the overall sequence. The context can alter the embedding vector of a residue, thus changing its similarity score with other residues. The global delta score for each protein pair was computed as the average of the Euclidean distances across all residue pairs within the alignment. A higher delta score signifies greater dissimilarity between residue pairs in the embedding space, suggesting functional divergence. Conversely, lower delta scores indicate similarity and potential conservation of function. Both the dsclen and global delta scores serve as complementary indicators of possible functional divergence: dsclen score focuses on the proportion of residues exhibiting significant differential conservation, highlighting areas of the protein that have adapted post-duplication; global delta score provides a holistic measure of residue-level dissimilarity based on embedding representations, capturing both conservation and contextual divergence.

Structure-Based Metrics

Recognizing that functional divergence extends beyond changes in functional residues – where protein may maintain core functions while processing different substrates – we incorporated structural-based predictions to enhance our residue classification framework. To achieve this, we utilized P2Rank, a machine learning-based tool designed for rapid and accurate prediction of ligand binding sites on protein structures. P2Rank assesses the likelihood of residue being part of functional pockets by evaluating local chemical environments on the protein's solvent-accessible surface. P2Rank operates by predicting sites without relying on template structures, making it suitable for high-throughput analyses. Its efficiency and accuracy are particularly advantageous for processing large dataset, aligning with our study's requirements (Krivák and Hoksza, 2018). Protein structures corresponding to the humans sequences in our MSAs were obtained from Alphafold (Jumper et al., 2021). P2Rank was executed on each protein structure to identify potential ligand binding sites. For each residue, P2Rank provides a probability score indicating its likelihood of being part of a functional pocket. To classify residues as part of the functional pocket, we applied a threshold defined as the median probability score plus two standard deviations of all predicted scores. This threshold ensures that only residues with significantly high probabilities are considered functional hotspots.

Mixed Metrics

To enhance the identification of functionally significant residues within duplicated gene pairs and their orthologs, we integrated sequence embedding-based metrics with advanced functional residue prediction tools. Specifically, we employed BindEmbed21, a tool designed to predict the likelihood of residues participating in binding interactions with small molecules, metals, or nucleotides (Littmann et al., 2021). This integration aimed to refine our classification of residues by corroborating divergence scores with experimentally informed predictions of functional hotspots. BindEmbed21 leverages sequence embeddings to predict the probability that each residue within a protein sequence is involved in binding interactions. It provides three distinct probability scores for each residue, corresponding to binding molecules, metals and nucleotides. These predictions are grounded in extensive training on large experimental datasets, rendering BindEmbed21 a reliable tool for identifying functional residues beyond purely computational divergence metrics. Protein sequences from our MSAs were prepared in FASTA format to serve as input for BindEmbed21. The tool processes each residue in the sequence, outputting three probability values indicating the likelihood of the residue binding to small molecules, metals or nucleotides.

Following authors recommendations, residues with probability scores exceeding 0.5 for any of the three binding categories were classified as hotspots. Employing BindEmbed21 allows for the incorporation of experimentally derived functional predictions into our analyses. By identifying residues with high binding probabilities, we can prioritize these positions that are more likely to contribute to protein function and, consequently, to functional divergence following gene duplication. We introduced two new metrics, Hot Differential Score (hdsc) and Delta Differential Score (hdelta), which combines the functional significance of residues with their divergence score. A residue is defined as **hdsc** if it is both a hotspot and is associated with a DSc above the threshold. Hdsc serves as a composite metric, identifying residues that are not only predicted to be functionally significant (hotspot) but also exhibit high divergence scores. These residues are prime candidates for being Adaptive residues, playing key roles in functional diversification. This hybrid approach ensures that the identified residues are both functionally relevant and evolutionarily divergent, enhancing the reliability of our classification system. Residues identified as hotspots by BindEmbed21 and having Delta scores above the median value were defined as **hdelta**, these residues are both functionally significant (hotspots) and have high delta scores, indicating substantial dissimilarity in the embedding space and reinforcing their role in functional divergence. To refine our residue classification, we integrated the structural predictions from P2Rank with our existing divergence score metric DSc. We also defined as **pdsc** residues associated with a high DSc

and a high P2Rank probability. Integrating structural-based functional predictions with conservation and embedding-based divergence scores leverages complementary strengths of different analytical approaches.

4.2. Global Metrics Summarization

A systematic approach was developed to aggregate individual residue-level scores into global metrics for protein pairs, enabling the assessment of the effectiveness of divergence metrics in identifying cases of known functional divergence. This methodology enables the evaluation of how well our metrics correlate with experimentally validated instances of functional divergence. For each protein pair we synthesized the four residue-level metrics – DSc, pdsc, hdsc, and hdelta – into global scores. These global metrics facilitate protein-level assessments of functional divergence, enabling direct comparisons with known functional changes. The global metrics were calculated as follows:

Global Hotspot Differential Score (hdsc_r):

$$hdsc\_r = \frac{\text{Number of hdsc residues above threshold}}{\text{Number of hotspots above threshold}} \tag{5}$$

This ratio represents the proportion of residues classified as hdsc (i.e., residues with both DSc and hotspot probabilities above their respective thresholds) relative to the total number of hotspots identified by BindEmbed.

Global Pocket Differential Score (pdsc_r):

$$pdsc\_r = \frac{\text{Number of pdsc residues above threshold}}{\text{Number of hotspots above threshold}} \tag{6}$$

This metric quantifies the proportion of residues classified as pdsc (i.e., functional hotspots with significant pocket probability identified by P2Rank) relative to all identified hotspots.

Global Hotspot Delta Score (hdelta_r):

$$hdelta\_r = \frac{\text{Number of hdelta residues above threshold}}{\text{Number of hotspots above threshold}} \tag{7}$$

This ratio measures the proportion of residues classified as hdelta (i.e., functional hotspots with high embedding-based divergence scores) among all hotspots.

Global DSc Length (lendsc_r):

$$lendsc\_r = \frac{\text{Number of residues with DSc above threshold}}{\max(\text{Protein Length}_1, \text{Protein Length}_2)} \tag{8}$$

This normalized measure represents the density of residues exhibiting significant evolutionary divergence relative to the length of the longer protein in each pair.

To integrate multiple aspects of divergence, we derived additional global metrics by combining the existing scores. These combined metrics capture both average and peak performance across multiple divergence indicators, providing robust measures for functional divergence assessment.

hp_r_mean:

$$hp\_r\_mean = \text{Mean}(hdsc\_r, pdsc\_r) \tag{9}$$

hpd_r_mean:

$$hpd\_r\_mean = \text{Mean}(hdsc\_r, pdsc\_r, hdelta\_r) \tag{10}$$

hp_r_max:

$$hp\_r\_max = \max(hdsc\_r, pdsc\_r) \tag{11}$$

hpd_r_max:

$$hpd\_r\_max = \max(hdsc\_r, pdsc\_r, hdelta\_r) \tag{12}$$

4.3. RHEA Truth Set Construction

To evaluate the predictive accuracy of our global metrics, we established a truth set of protein pairs with experimentally validated functional divergence using the RHEA database. We extracted reaction entries from RHEA corresponding to the protein pairs under study. Each protein pair was associated with their respective RHEA entries detailing their enzymatic

functions. For each protein pair, we assessed the similarity of their RHEA entries: Truth Value = 0 assigned to protein pairs where all experimental RHEA entries for both protein were identical, indicating no functional divergence; Truth Value = 1 assigned to protein pairs where all experimental RHEA entries for both proteins were different, signifying functional divergence. Proteine with partial overlaps or inconclusive RHEA entries were excluded from the truth set to maintain data integrity.

4.4. Functional Divergence Identification Framework

The determination of optimal thresholds for our divergence metrics is a cornerstone of this study, as it directly impacts the accuracy and reliability of identifying functionally divergent protein pairs. Threshold optimization ensures that the classification of residues and, consequently, protein pairs, is both statistically robust and biologically meaningful. By employing Kernel Density Estimate (KDE) curves intersections to establish these thresholds (Hastie et al., 2001), we adopted a data-driven approach that minimizes bias and maximizes the discriminative power of each metric. This method allows for the objective delineation between true positives and true negatives based on the natural distribution of our metrics within the dataset. With the truth set established, we evaluated the predictive performance of each global metric using standard classification metrics, including Receiver Operating Characteristic (ROC) curves, F1 scores, recall, and precision (Powers, 2020). While ROC curves were generated using the continuous values of the metrics, the calculation of F1 Score, Precision, and Recall necessitated the conversion of these continuous scores into binary classifications. To establish appropriate thresholds for this binarization, we utilized the intersection points of Kernel Density Estimate (KDE) curves representing the true positive (truth value = 1) and true negative (truth value = 0) classifications. The selection of thresholds for converting continuous metric scores into binary classification is critical for optimizing the balance between Precision and Recall. To achieve this, we implemented the following procedure: for each global metric, we generated separate KDEs for residues classified as true positives (functional divergence present) and true negatives (no functional divergence) based on a truth set derived from the RHEA database. The point at which the KDE curves of true positive and true negative intersect were identified as potential thresholds. These intersection points represent score values where the probability density of true positives equals that of true negatives, providing a natural cutoff between the two classes. We computed the following performance metrics for each global score: ROC Curves: We plotted True Positive Rate (TPR) against False Positive Rate (FPR) to evaluate

the trade-off between sensitivity and specificity; F1 Scores: Harmonizing precision and recall to assess the balance between identifying true positive and minimizing false positives; Recall: Measuring the proportion of true positives among all positive predictions; Precision: Assessing the proportion of true positive among all actual positives. Performance metrics across all global scores were compared to determine which metrics most effectively distinguish between functionally divergent and non-divergent protein pairs. Metrics demonstrating higher Area Under the ROC Curve (AUC) and balanced F1 scores were considered more reliable indicators of functional divergence.

4.5. Probability and Permutation Analysis for Validation

To convert our scores into probabilities between 0 and 1, we applied a logistic transformation. This transformation allows us to interpret the scores probabilistically, where values closer to 1 indicate higher confidence and values closer to 0 suggest lower confidence (Hastie et al., 2001). We used the threshold derived from previous analyses, as the central point of the logistic function, with a steepness parameter of k = 15. The logistic function used is:

Probability Function:

$$P = \frac{1}{1 + \exp(-15 \times (\text{scores} - \text{threshold}))} \tag{14}$$

This transformation ensures that raw scores are smoothly scaled to a probability range, aiding in the interpretation of the results. We further validated our scores using a permutation test analysis (Ernst, 2004). The purpose of this analysis was to assess the stability and significance of the scores. For each analysis, we performed 1000 random resamplings, obtaining a distribution of scores by randomizing both the DSc and the functional prediction scores of the entire duplications set. The empirical p-value was calculated by comparing the true score to the distribution of randomly obtained scores. The p-value was computed as:

P-Value Calculation:

$$p\_value = \frac{1}{n} \sum_{i=1}^{n} \left[\left| \text{random\_score}_i - \overline{\text{random\_scores}} \right| \geq \left| \text{true\_score} - \overline{\text{random\_scores}} \right| \right] \tag{15}$$

This empirical p-value allows us to determine whether the observed score significantly deviates from what would be expected under random conditions.

5. Expression Analysis and Functional Enrichment

5.1. Tissue-Specific Expression Analysis

Functional divergence between protein pairs may be accompanied by alterations in their tissue-specific expression profiles. To investigate potential differences in tissue expression associated with divergent protein pairs, we utilized the Human Protein Atlas (HPA) database (Uhlén et al., 2015). The HPA is a resource that provides detailed information on the expression and localization of human proteins across a wide range of tissues and cell types. Specifically, we employed RNA-based transcriptomic data referred to as consensus normalized expression ("nTPM"), which summarizes transcript expression levels per gene across 50 tissues based on data from the Genotype-Tissue Expression (GTEx) project. The consensus nTPM values were calculated as the maximum nTPM value for each gene from the two data sources. For each protein in our study, we retrieved the nTPM values across all 50 tissues for both proteins. We then calculated the absolute expression difference (exp_diff) for each tissue by subtracting the nTPM values of protein B from that of protein A for each tissue. This approach enabled the assessment of tissue-specific expression changes that may underline the functional divergence between protein pairs.

5.2. Functional Enrichment: Gene Ontology and KEGG Pathways

Analyzing the roles of divergent and non-divergent protein pairs within different biological contexts involved retrieving and processing Gene Ontology (GO) annotations and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information (Ashburner et al., 2000; Kanehisa and Goto, 2000). The following methodology outlines the steps undertaken for this analysis. Each protein in the 1,112 protein pairs was annotated with its corresponding GO entries under the Biological Process and Molecular Function categories, as well as with KEGG pathway identifiers. Only proteins with at least one experimental GO annotation in any of the GO categories were included to ensure the relevance and reliability of the functional annotations. GO entries were obtained using the UniProt PI, which interfaces with the QuickGO database, maintained by the European Bioinformatics Institute (EBI) (Binns et al., 2009). This approach facilitated the efficient retrieval of up-to-date GO annotations for each protein. To streamline the analysis and focus on relevant biological functions, GO terms were slimmed using the goatools Python library (Klopfenstein et al., 2018). The slimming process employed the goslim_chembl.obo ontology file. This specific GO slim was chosen as it

strikes a balance between broad and specific terms, avoiding overly generic categories that could dilute functional insights while retaining sufficient specificity to capture meaningful biological processes and molecular functions. Functional enrichment analysis required ranking protein pairs based on their divergence scores derived from the hpd_r_mean metric. These scores were transformed to a standardized scale ranging from -2 to 4 to facilitate comparative analysis. The transformation was performed using the following formula:

$$\text{transform\_score}(\text{score}, \text{threshold}=0.152) = \begin{cases} 1 + 4 \times \left(1 - \frac{0.152}{\text{score} - 0.152}\right) & \text{if score} \geq 0.152 \\ -1 - 2 \times \left(\frac{0.152}{0.152 - \text{score}}\right) & \text{if score} < 0.152 \end{cases}$$

This transformation ensures that scores above the threshold are scaled positively, while those below the threshold are scaled negatively, with the threshold value itself mapped to zero. This scaling preserves the relative differences in divergence while normalizing the range of scores for downstream analysis. For KEGG pathway analysis, only metabolic KEGG terms prefixed with “00” were considered. The divergence scores for these pathways underwent the same transformation process as described above, ensuring consistency in the ranking methodology.

Enrichment analysis was conducted using the gseapy library, specifically utilizing the “prerank” function (Fang et al., 2023). This method leverages the transformed divergence scores to identify GO terms and KEGG pathways that are significantly associated with either divergent or non-divergent protein pairs. The input to the “prerank” function included a ranked list of protein pairs based on the transformed divergence scores, and a gene set including the slimmed GO terms and selected KEGG pathways relevant to the proteins under study. The prerank enrichment approach assesses whether predefined sets of GO terms or KEGG pathway are statistically overrepresented at the top or bottom of the ranked list, thereby highlighting biological processes and molecular functions that are potentially influenced by functional divergence.

6. Autonomous Functional Divergence Pipeline

An autonomous pipeline was implemented as a Google Colab notebook to ensure broad accessibility to functional divergence analyses. Google Colab, a cloud-based interactive coding environment, allows users to execute Python code directly within a web browser, eliminating the need for local installations or complex setup procedures. This platform supports collaborative work and provides robust computational resources, making it an ideal

choice for deploying bioinformatics workflows (Bisong, 2019). The pipeline is designed for ease of use, requiring users to input only two UniProt identifiers corresponding to any pair of homologous proteins, regardless of species or functional type, including non-enzymatic proteins. Once provided, the pipeline automatically retrieves orthologous sequences from the Compara database and performs multiple sequence alignments using CLUSTALO. The resulting alignments undergo a cleaning process to eliminate gaps and ambiguously aligned regions, ensuring high-quality data for subsequent analyses. Following alignment, the pipeline calculates various alignment scores, including TSC, ISCS, XSC, and DSC, which assess different aspects of sequence conservation and variability. Protein structures are automatically downloaded from AlphaFold and aligned using PyMOL to facilitate structural comparisons. The pipeline then generates embeddings using ProtTrans T5 and predicts functional hotspots and pocket residues utilizing BindEmbed21 and P2Rank, respectively. A phylogenetic tree is constructed with RAxML based on the curated orthologous sequences, providing insights into the evolutionary relationships among the protein pairs. The outputs of the pipeline include a detailed table listing aligned residues with their associated scores, UniProt features, GO gene ontology entries from QuickGO, and tissue-specific expression data from the Human Protein Atlas database. Residues are classified as robust, plastic, or adaptive based on their divergence profiles, integrating scores from various predictive tools alongside global divergence metrics (hdsc_r, pdsc_r, hdelta_r) and combined metrics that account for all three indicators. For visualization, the pipeline generates cleaned sequence alignments in both FASTA and PDF formats, styled similarly to ESPript for enhanced readability. Annotated PDB files are also produced, where the two structures downloaded from AlphaFold are aligned using PyMOL. Predicted binding pockets are depicted as grids, residues are color-coded with a red gradient based on their DSc values, and functionally important residues identified by BindEmbed21 are highlighted as sticks. Additionally, the pipeline outputs the phylogenetic tree generated by RAxML, providing a visual representation of the evolutionary relationships (Stamatakis, 2014). This pipeline streamlines complex functional divergence analyses, ensuring that researchers can efficiently conduct detailed studies with minimal computational expertise. In the **Results** section, we present an example protein pair to demonstrate the full range of outputs generated by this pipeline, illustrating its practical application and utility in real-world bioinformatics research.

7. Case Studies and Experimental Procedures

7.1. Case Study: AADAC-AADACL2 Docking and Evolutionary Insights

To evaluate the interaction between the catalytic triad serine residue of AADACL2 and the carbonyl carbon of hydroxy-epoxy ceramides, we employed a multi-step procedure integrating molecular docking, P2Rank-based binding site prediction, and clustering analysis. In the event that AADACL2 was not responsible for metabolizing the ceramide, we extended the analysis to include other proteins from the HPA “Skin - cornification” cluster, specifically focusing on those containing the serine residue characteristic of the catalytic triad found in lipases. AlphaFold PDB files served as the starting structures. These were processed to remove flexible or disordered regions based on B-factors, retaining only the most stable regions (B-factor < 90). Flexible N- and C-terminal regions were systematically trimmed using PyMOL scripting (Schrödinger, LLC, 2015), ensuring the core structure remained intact for subsequent docking. P2Rank was employed to predict potential binding sites on the prepared protein structures. For each protein, P2Rank generated a set of predicted binding pockets, which were then matched with the serine residue’s position to identify the pocket closest to the serine. The identified pocket was used to define the docking region in later steps. The docking process was conducted using AutoDock Vina (Eberhardt et al., 2021). The docking configuration included a search space centered on the predicted binding site, and the num_mode and exhaustiveness parameters were set to 200 to ensure thorough sampling of ligand conformations. Hydroxy-epoxy ceramide ligands were docked into the prepared protein structures, targeting the regions surrounding the serine residue. After docking, PyMOL was used to calculate the distance between the oxygen atom of the serine hydroxyl group and the carbonyl carbon of the hydroxy-epoxy ceramide for each docked conformation. This analysis was conducted across multiple ligand poses, and the minimum distance between the two atoms was recorded for further evaluation. To identify representative binding modes, the docked ligand conformations were first clustered based on their Root Mean Square Deviation (RMSD), with a threshold of 7 Å, grouping structurally similar conformations. To further refine the ranking of the ligand poses, we employed the Catalytically Favorable Conformations (CFC) and Close Contact-Catalytically Favorable Conformations (CC-CFC) metrics as described by Malatesta et al., (Malatesta et al., 2024). CFC was defined as any conformation where the carbonyl carbon of the hydroxy-epoxy ceramide is located within 4 Å of the serine hydroxyl group, indicating catalytic relevance. CC-CFC, in turn, refers to CFCs that are part of clusters identified through RMSD analysis, ensuring both catalytic favorability and structural consistency within the clustered poses.

These metrics ensured that only the most catalytically promising conformations were retained. The combination of CFC and CC-CFC rankings provided critical insights into the interaction between hydroxy-epoxy ceramides and these proteins, helping to prioritize the most likely binding modes. For each docking interaction, we visualized the ranking of the ligand conformations based on their clustering results (CFC, CC-CFC counts).

7.2. AADACL2 Expression and Solubility in *E. coli*

The coding sequence (CDS) of *Homo sapiens* arylacetamide deacetylase like 2 (*AADACL2*) (NM_207365.4:110-1315) was cloned into the pET-28a(+)-TEV plasmid obtained from GenScript Biotech. These expression vectors include an additional N-terminal 6xHisTag site to facilitate easier purification of the recombinant proteins (**Fig. 30**).

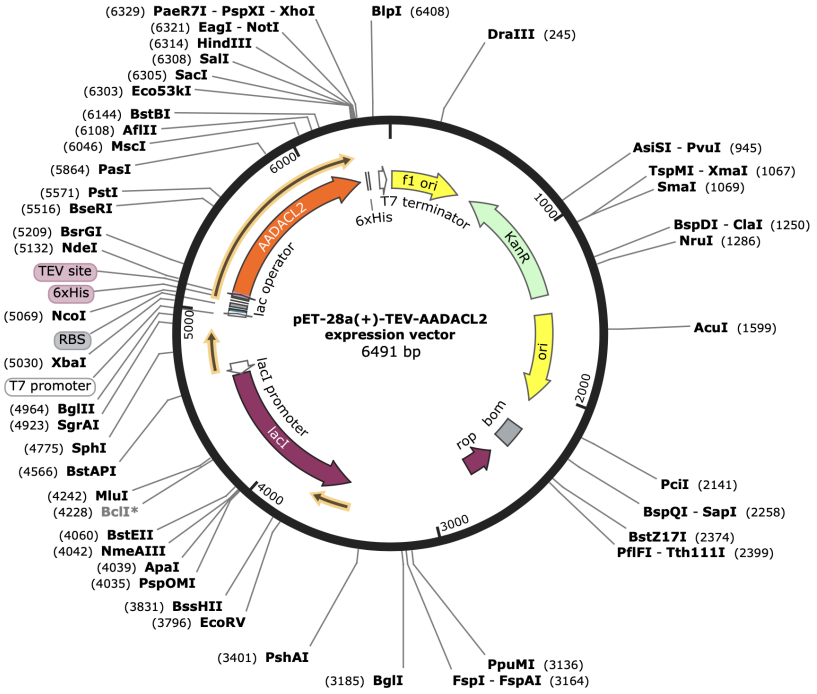


Figure 30: SnapGene-generated diagram of the pET-28a(+)-AADACL2 vector. The modified vector includes: (1) Lac operator for induction with IPTG; (2) A multiple cloning site containing the *AADACL2* CDS (highlighted in orange) cloned in frame with the N-terminal 6xHisTag and the TEV protease cleavage site; (3) Kanamycin antibiotic resistance gene for selection.

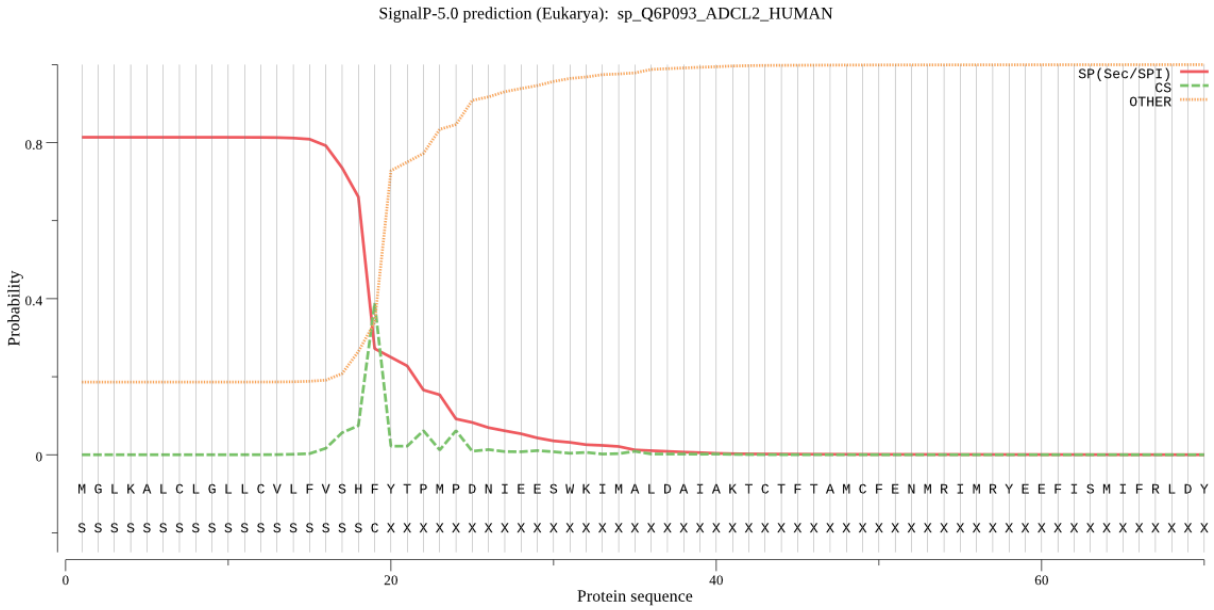


Figure 31: SignalP 5.0 prediction of the N-terminal signal peptide in AADACL2. The prediction indicates a Signal peptide (Sec/SPI) with a cleavage site between positions 19 and 20 (SHF-YT) and a probability score of 0.3887

The vector was further mutagenized by GenScript Biotech to exclude 18 residues from the N-terminus, thereby eliminating the predicted signal peptide as determined by SignalP 5.0 (Signal peptide (Sec/SPI) - Cleavage site between pos. 19 and 20: SHF-YT. Probability: 0.3887) (Almagro Armenteros et al., 2019). The mutagenesis aimed to enhance protein solubility by initiating the sequence with “FYTP” (**Fig. 31**).

Escherichia coli BL21-CodonPlus DE3 strain (Novagen) cells were cultured overnight at 37° in LB broth (comprising 1% NaCl, 1% tryptone, and 0.5% yeast extract) supplemented with 34 µg/mL chloramphenicol to serve as the host. Following incubation, the host cells were subjected to three centrifugation steps (4°C, 8000 g for 30 seconds each) and subsequently resuspended in sterile water under a laminar flow hood. Genscript provided 4 µg of DNA, which was diluted in 40 µL of water. From this solution, 1 µL was further diluted by adding 9 µL of water. The AADACL2 constructs were then electroporated, with 1 µL of the DNA solution, into the bacterial host cells and plated on LB agar medium containing the appropriate antibiotic (50 µg/mL kanamycin for plasmid selection and 34 µg/mL of chloramphenicol for host resistance). Subsequently, expression and solubility assays were performed. For expression testing, multiple kanamycin-resistant colonies were individually inoculated into two 50 mL centrifuge tubes, each containing 10 mL of LB broth with the appropriate antibiotics concentrations. One of the two cultures from each colony was induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.3 mM once

the cells reached an optical density at 600 nm (OD_{600}) of 0.6 at 37°C. The cultures were further incubated until they attained an OD_{600} of 6 a 37°C, after which the cells were harvested by centrifugation at 7000 g for 10 minutes at 4°C. Each pellet was resuspended in a volume calculated as x = (500 x OD measured of the less concentrated sample) / (OD measured of the sample) microliters of 1X Phosphate-Buffered Saline (PBS) and subjected to sonication (35 W, 30 seconds on and 30 seconds off on ice, for six cycles). The lysate was then separated by centrifugation, isolating the supernatant from the pellet. The pellet was subsequently resuspended in an equal volume of 1X PBS. Each resulting sample was analyzed using 12% SDS-PAGE to assess induction efficiency, expression levels, and to verify the presence of the protein within the soluble fraction.

7.3. AADACL2 Resuspension from Inclusion Bodies in *E. coli*

Despite utilizing various *Escherichia coli* strains, including BL21-CodonPlus, pGro7, pG-KJE8, TUNER, and different combinations of growth conditions, we were unable to achieve soluble expression of the AADACL2 protein. Consequently, we opted to attempt renaturation of the protein from the inclusion pellet through column purification. To this end, the transformed BL21-CodonPlus colony exhibiting the highest protein expression was inoculated into 1 liter of LB medium in a 5-liter flask and cultured under the same conditions described previously. The renaturation protocol encompassed the following steps: (1) The cell pellet was resuspended in 40 mL of resuspension buffer (20 Mm Tris-HCl, pH 7.8), sonicated on ice, subjected to centrifugation, and the supernatant was discarded; (2) the resulting pellet was resuspended in 30 mL of cold isolation buffer (20 mM Tris-HCl, 2 M Urea, 0.5 M NaCl, 2% Tween20), sonicated, centrifuged (this step was repeated twice to ensure through lysis and removal of insoluble materials); (3) the pellet was then resuspended in 50 mL of binding buffer (20 mM TrisHCl, 6M Urea, 0.5 M NaCl, 5 mM Imidazole, 1 mM β-mercaptoethanol) and agitated for 60 minutes at room temperature to facilitate protein refolding, and following the incubation, the mixture was centrifuged, and the supernatant was filtered through 0.45 µm filters. Purification was carried out using a Fast Protein Liquid Chromatography (FPLC) system (Akta Pure 25M, GE Healthcare) with a HisTrap HP 5 mL (GE Healthcare) cobalt affinity column, exploiting the N-terminal 6xHisTag. The column was equilibrated with binding buffer, after which a gradient was initiated at a flow rate of 1 mL/min using refolding buffer (20 mM Tris-HCl, 0.5 M NaCl, 5 mM Imidazole, 1 mM β-mercaptoethanol). Upon completion of the initial gradient, an elution gradient was applied at the same flow rate using an elution buffer (20 mM Tris-HCl, 0.5 M NaCl, 0.5 M Imidazole, 1

mM β-mercaptoethanol). To exchange the buffer containing the imidazole together with the purified protein, a VivaSpin Cytiva concentrator (Merck) with a 40 kDa molecular weight cutoff (MWCO) was utilized.

7.4. AADACL2 Activity Assay (4-Nitrophenyl Acetate)

To assess the enzymatic activity of the freshly purified AADACL2 protein, which is hypothesized to function as a lipase, we performed a 4-nitrophenyl acetate (PNPA) assay. The assay measured the absorbance at 405 nm, corresponding to the formation of 4-nitrophenol, using a spectrophotometer over time. The reaction mixture was prepared by combining 152 µL of 20 mM Tris-HCl (pH 7.8) and 150 mM NaCl. A blank was first measured, followed by the addition of 38 µL of 50 mM PNPA to achieve a final substrate concentration of 10 mM. The absorbance spectrum was recorded for 30 seconds before and after introducing 10 µL of the fresh AADACL2. The enzymatic activity was determined by comparing the slope of the absorbance increase immediately after PNPA addition with the slope following enzyme addition. A significant change in the slope indicated active enzymatic cleavage of PNPA, thereby confirming the lipase activity of the AADACL2 protein.

7.5. AADACL2 Expression in *Pichia pastoris*: Propagation and Transformation

The same coding sequence used for *Escherichia coli* was subcloned into the pPICZ A vector of *Pichia pastoris* from GenScript, which includes the inducible AOX1 promoter for recombinant protein expression, a c-myc epitope tag for the detection with anti-myc antibodies, a C-terminal 6xHisTag for protein purification, and zeocin resistance for selection of integrants (**Fig. 32**). We analyzed our protein with glycosylation prediction tools NetOGlyc 4.0 and NetNGlyc 1.0 (Steentoft et al., 2013). While NetOGlyc 4.0 did not predict any glycosylation sites, NetNGlyc 1.0 identified a potential glycosylation site at position N282 with a score just above the threshold: sp_Q6P093_ADCL2_HUMAN 282 NWSI 0.5009 (5/9) + (**Fig. 33**). Given our suspicion of glycosylations, we opted to use this vector for intracellular expression rather than the one with the alpha factor for secretion into the medium. Secretion pathways in yeast often lead to hyperglycosylation or the addition of heterogeneous glycan structures which can potentially interfere with protein folding, stability, and activity. Intracellular expression allows for better control over glycosylation patterns and enhances protein stability, which is advantageous for downstream purification and functional assays.

The procedures described in the next sections were partially performed following the protocols provided by Invitrogen™ in the “EasySelect™ Pichia Expression Kit For Expression of Recombinant Proteins Using pPICZ and pPICZa in *Pichia pastoris*” and “pPICZA, B, and C Pichia expression vectors for selection on Zeocin™ and purification of recombinant proteins”. XL1-Blue cells (from Agilent) were inoculated from a tetracycline-resistant glycerol stock (35 µg/mL) into 1 mL of LB broth and incubated overnight at 37°C. Genscript provided 4 µg of DNA, which was diluted in 40 µL of water. From this solution, 1 µL was further diluted by adding 9 µL of water. To prepare the cells for electroporation, 1 mL of the overnight culture was centrifuged at 8,000 g for 30 seconds at 4°C. The supernatant was discarded, and the pellet was washed by resuspending it in water, followed by centrifugation under the same conditions. This washing step was repeated three times. After the final wash, approximately 50 µL of the cell suspension remained, to which 1 µL of the prepared DNA solution was added. The mixture was subjected to electroporation, and the cells were incubated at 37°C for 1 hour. Following the recovery period, cells were plated on LB agar supplemented with tetracycline (35 µg/mL) and zeocin (25 µg/mL) and incubated overnight at 37°C. A single transformed colony was selected and inoculated into 15 mL of LB broth, then incubated overnight at 37°C. In order to isolate plasmid DNA from bacterial cultures, MiniPrep was performed using the following steps: (1) 1.5 mL of the overnight culture was transferred to a centrifuge tube and centrifuged at 14,000 g for 2 minutes. The supernatant was discarded, and an additional 1.5 mL of the culture was added to the same tube, followed by another centrifugation step. This process was repeated until a total of 12 mL was processed, resulting in four centrifuge tubes. (2) 300 µL of Solution 1 (50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA, 100 µg/mL RNase A) was added to each tube, and the suspension was vortexed. (3) 300 µL of Solution 2 (0.2N NaOH, 1% SDS) was added, and the contents were gently mixed by inversion. (4) 300 µL of Solution 3 (3M potassium acetate, 5M acetic acid) was then added, and the tubes were mixed by shaking. (5) The mixture was placed on ice for 5 minutes and subsequently centrifuged at 14,000 g for 10 minutes at 4°C. (6) 830 µL of the supernatant was carefully transferred to a new centrifuge tube. (7) 500 µL of isopropanol (0.6 volumes) was added to the supernatant, followed by vortexing and centrifugation at 14,000 g for 10 minutes at room temperature. (8) The supernatant was discarded, and the DNA pellet was washed with 500 µL of 70% ethanol, then left to evaporate. (9) The pellet was centrifuged at 14,000 g for 5 minutes at 4°C and resuspended in 50 µL of water. To prepare the plasmid for transformation and integration into *Pichia pastoris*, the pPICZ A-AADACL2 plasmid was linearized using the restriction enzyme SAC I (Termo Fisher Scientific). According to the manufacturer’s instructions, 2 units of SAC I were added to 200 µL of plasmid DNA containing approximately 20 µg, along with 10X MULTICORE buffer. The

reaction mixture was incubated for 2 hours at 37°C, followed by enzyme inactivation through heating at 65°C for 20 minutes. Successful digestion was confirmed by agarose gel electrophoresis, where both uncut and digested plasmid DNA were loaded to assess size shifts and verify the integrity of the digestion process. Upon confirmation of complete plasmid linearization, the linearized plasmid was prepared for transformation into *Pichia pastoris* GS115 cells. *Pichia pastoris* GS115 cells were inoculated into 5 mL of sterile YPD medium (1% yeast extract, 2% peptone, 2% dextrose) and incubated at 30°C with shaking overnight. The overnight culture (0.5 mL) was then transferred to 50 mL of fresh YPD in a 250 mL flask and grown at 30°C until the optical density at 600 nm reached between 1.3 and 1.5. Cells were harvested by centrifugation at 1500g for 5 minutes at 4°C, followed by three washes with sterile cold 1 M sorbitol. The prepared cells (80 µL) were mixed with the linearized plasmid DNA and electroporated using parameters recommended by the manufacturer. Following electroporation, cells were plated on YPDS agar containing zeocin (1% yeast extract, 2% peptone, 2% dextrose, 2% agar, and 100 µg/mL zeocin) and incubated at 30°C for 48 hours.

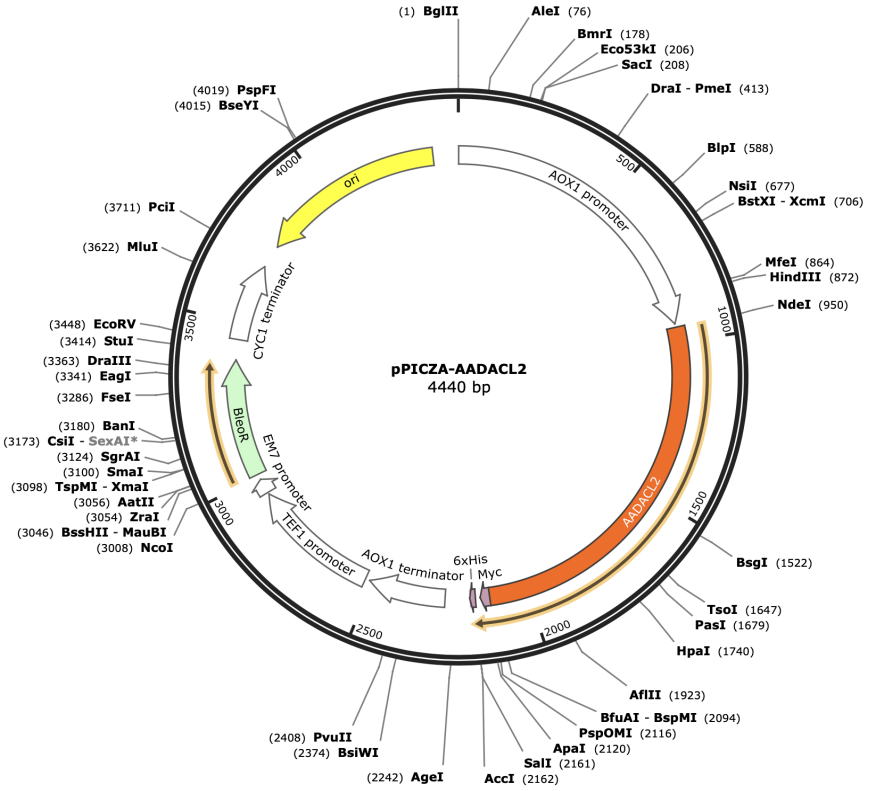


Figure 32: SnapGene-generated diagram of the pPICZA-AADACL2 vector. The modified vector includes: (1) AOX1 promoter for inducible expression; (2) a multiple cloning site containing the *AADACL2* CDS (highlighted in orange) cloned in frame with the C-terminal 6xHisTag and the c-myc epitope; (3) Zeocin antibiotic resistance gene for selection.

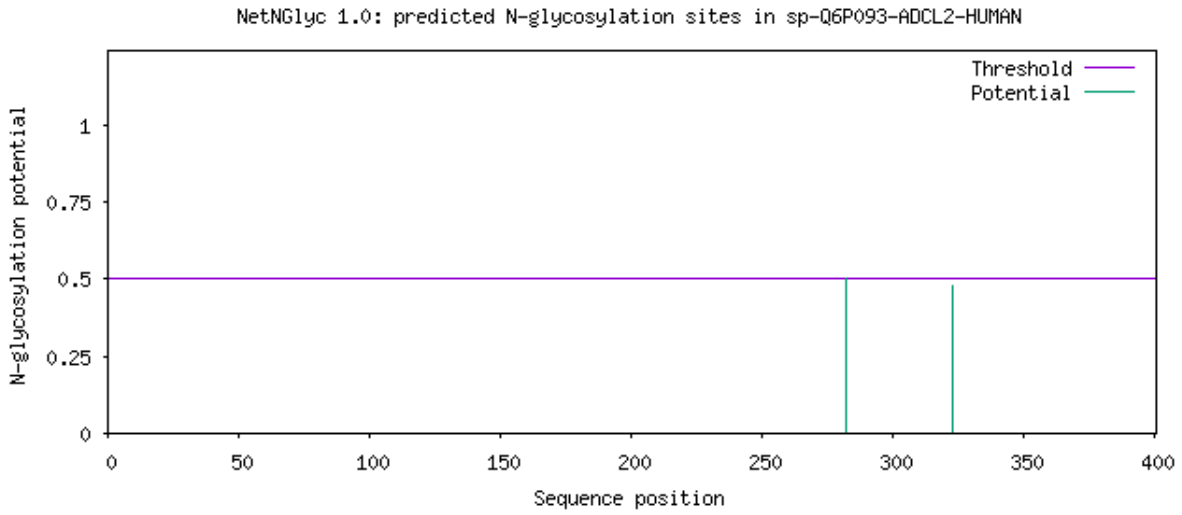


Figure 33: NetNGlyc 1.0 prediction of N-glycosylation sites in AADACL2. A possible glycosylation site is indicated at position N282 (sp_Q6P093_ADCL2_HUMAN 282 NWSI 0.5009 (5/9) +).

7.6. AADACL2 Expression and Solubility in *Pichia pastoris*

To confirm the ability of the transformed *Pichia pastoris* cells to utilize methanol, individual colonies were patched onto MDH and MMH agar plates containing 1.34% YNB, 4×10^{-5}% biotin, 1% agar, and either 2% dextrose for MDH or 0.5% methanol for MMH. These plates were incubated at 30°C for 48 hours. To evaluate the expression, solubility, and activity of the AADACL2 protein, multiple colonies were each inoculated into 10 mL of BMGH medium (100 mM potassium phosphate, pH 6.0; 1.34% YNB; 4×10^{-5}% biotin; 1% glycerol; 0.004% histidine) in 50 mL centrifuge tubes and cultured at 30°C with shaking at 250 rpm until the optical density at 600 nm (OD_{600}) reached ~3. The cells were then pelleted by centrifugation at 2500 g for 5 minutes at room temperature and resuspended in BMMH medium (100 mM potassium phosphate, pH 6.0; 1.34% YNB; 4×10^{-5}% biotin; 0.5% methanol; 0.004% histidine) to induce protein expression, adjusting the culture to an OD_{600} of 1. A portion of the resuspended cells was further diluted back into BMGH medium at the same OD_{600} to verify induction via SDS-PAGE analysis. At 2, 4, 6, 8, and 24 hours post-induction, 1 mL of culture was collected and transferred to 1.5 mL centrifuge tubes for processing. Each sample underwent the following processing steps: (1) centrifugation at 20,000 g for 3 minutes at room temperature; (2) addition of 100 µL lysis buffer (50 mM sodium phosphate, pH 7.4; 300 mM NaCl) to the pellet; (3) addition of an equal volume of acid-washed glass beads (0.5 mm in size) to facilitate mechanical disruption; (4) vortexing for 30 seconds followed by incubation on ice for 30 seconds, repeated eight times; (5) addition of 1% Tween20 and

subsequent rest for 20 minutes; (6) centrifugation at 20,000 g for 10 minutes at 4°C to separate the pellet and supernatant; and (7) resuspension of the pellet. A portion of both the pellet and the supernatant was reserved for SDS-PAGE analysis after all time points, while the remaining supernatant was used for subsequent enzymatic activity assays as described in the relevant methods section. This approach allowed for the assessment of protein expression levels, solubility, and functional activity over the induction period. For large-scale expression of AADACL2, 12.5 mL of culture was inoculated and grown at 30°C until an OD_{600} of 2.5 was reached. The culture was then transferred into 500 mL of BMGH medium and allowed to grow at 30°C until an OD_{600} of 3.5 was achieved. Cells were harvested by centrifugation and resuspended in 1,750 mL of BMMH medium. After an 8-hour incubation period, 4 grams of pellet were obtained and resuspended in 12 mL of the previously described lysis buffer. For cell lysis, two cycles of French press at 30 kpsi were performed, maintaining the samples on ice to prevent protein denaturation. Following lysis, 1% Tween20 was added, and the mixture was allowed to rest for 20 minutes. The lysate was then centrifuged at 14,000 g for 30 minutes, and the supernatant was purified using the same affinity purification method and equipment as described for *Escherichia coli*, eliminating the need for column refolding.

7.7. Lysis Protocols in *Pichia pastoris*

Prior to scaling up protein expression, we evaluated several lysis systems to optimize extraction of AADACL2 from *Pichia pastoris*. To achieve this, we followed the same inoculation, growth, and induction procedures detailed in the preceding section. After induction, the cell pellet was resuspended in a lysis buffer and subsequently divided into three aliquots. Each aliquot was subjected to a different lysis method: sonication, French press, or glass beads disruption. For the sonication method, samples were processed at 50% of power with alternating cycles of 5 minutes on and 5 minutes off, repeated three times. Throughout the sonication process, all samples were maintained on ice to prevent overheating and protein denaturation. The French press method utilized a Multi Cycle Cell disruptor from Constant Systems, set to an operating pressure of 30 kpsi. This procedure involved three cycles of high-pressure disruption, with each cycle followed by a 5-minute rest period on ice to prevent thermal damage. The glass bead disruption was performed as previously described, involving mechanical agitation with acid-washed glass beads to facilitate cell breakage. These comparative lysis protocols were implemented to determine the most effective method for achieving soluble AADACL2 protein in *Pichia pastoris*. The

efficiency of each lysis technique was subsequently assessed through SDS-PAGE analysis to evaluate protein yield and solubility, ensuring the selection of the optimal method for large-scale protein purification.

8. Pipeline Overview

The analysis pipeline (**Fig. 34**) integrates three major sources of information to systematically investigate gene duplication and functional divergence.

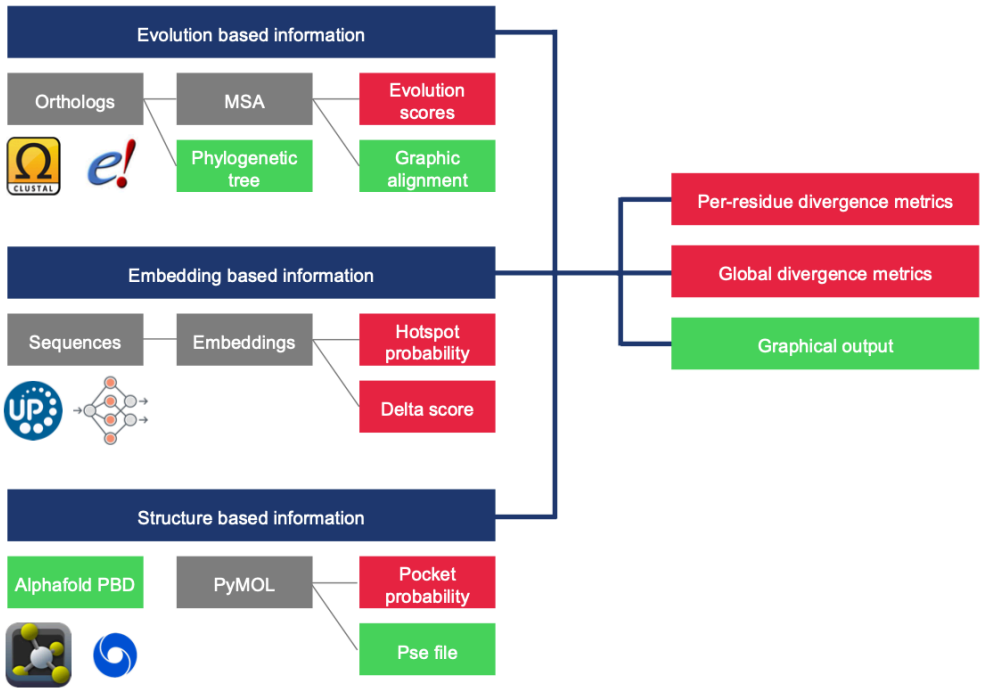


Figure 34. Overview of the analysis pipeline. The figure highlights the workflow with four distinct color-coded elements: in blue, the three main sections of the pipeline (evolution-based, embedding-based, and structure-based information); in gray, the tools and data sources used or retrieved throughout the process; in red, the generated data in terms of functional residues and evolutionary scores; and in green, the graphical outputs, including alignments, phylogenetic trees, and structural models.

First, the evolution-based information section collects orthologous sequences from a broad set of species, constructs multiple sequence alignments (MSAs), and generates phylogenetic trees. This step allows the calculation of evolutionary scores that quantify both in-group conservation and cross-group divergence, leading to differential scores (ISc, XSc, DSc). The embedding-based information section then leverages state-of-the-art deep learning models,

such as ProtTrans T5, to compute sequence embeddings for individual residues, capturing their biochemical properties, and uses BindEmbed21 and embeddings to predict functional hotspots. Embedding-based metrics, like the delta score, highlight functionally and context-divergent residues across gene pairs. Finally, the structure-based information section assesses protein structures using AlphaFold models and tools such as P2Rank to predict binding pockets. These predictions are integrated with divergence scores to refine the identification of adaptive or functionally significant residues. All three sections converge into an output with metrics that summarize residue-level divergence and those condensed into global divergence per gene pair, facilitating the identification of evolutionary trends and functional changes within gene pairs. The graphical outputs provide a visual representation of sequence, embedding, and structural divergence, including alignments, structures with calculated and predicted evolutionary annotations, phylogenetic trees, and expression profiles. The entire analysis was conducted using Python version 3.11, with various libraries to support data processing, machine learning, and visualization. Key Python libraries include torch, numpy, pandas, scikit-learn, matplotlib, plotly, biopython, statistics, pymol, transformers, seaborn, scipy, goatools, gseapy, and ete3. In addition to these, command-line tools such as clustalo, blastp, raxml, bindembed21, p2rank, and autosite were employed for sequence alignment, phylogenetic analysis, and structural predictions. The computationally intensive tasks, including the multiple sequence alignments and phylogenetic tree constructions, were carried out on the High Performance Computing (HPC) infrastructure of the University of Parma, provided in collaboration with INFN (Istituto Nazionale di Fisica Nucleare). The HPC facilities (HPC.unipr.it) enabled the efficient handling and processing of large datasets through advanced computational resources.

Results

1. Duplicated Gene Pairs: Numbers and Chromosomal Distribution

The all-against-all BLASTP search of the human proteome identified 16,261 homologous sequences from the initial 20,434 reviewed human protein sequences. From these homologous sequences, we extracted 10,130 intraspecies Best Reciprocal Hits (BRHs), likely representing the more closely related gene pairs. Filtering the BRH pairs based on identical Pfam domain architectures resulted in 4,083 gene pairs. This step ensured that only gene pairs with the same structural and functional domains were included. By incorporating enzymatic annotations, we further refined the dataset to 1,184 gene pairs where at least one member possessed a Rhea reaction or an EC number. Through this filtering approach, leveraging sequence similarity, reciprocal best hits, conserved domain architecture, and enzymatic annotations, we curated a high confidence set of duplicated gene pairs in humans. This refined dataset serves for subsequent analyses of gene duplication events and functional divergence.

To show the chromosomal distribution of duplicated gene pairs, we generated two radial plots (**Fig. 35**). The upper radial plot represents the 10,130 gene pairs identified immediately after the BRH filtering step. The distribution of duplication events spans across all chromosomes (c1-c22, X, Y), with a notable concentration on chromosome 19 as opposed to chromosome Y. The majority of duplication (~78%) events involve genes located on different chromosomes, indicating a prevalence of inter-chromosomal duplications in the initial homologous set. The lower radial plot depicts the final set of 1,184 gene pairs possessing enzymatic activity. Similar to the broader BRH dataset, these gene pairs are distributed across all chromosomes with ~80% events involving inter chromosome duplication. However there is a discernible pattern where certain chromosomes exhibit distinct duplication characteristics. Chromosome 19 emerges as the richest in terms of concentration of duplicated gene pairs. Chromosomes 11 and 12 display prominent clusters of intra-chromosomal duplications, suggesting localized duplication events. Chromosome 9 is notable for a cluster of inter-chromosomal duplications that predominantly connect to chromosome 1. The Y chromosome stands out as the poorest in duplicated gene pairs.

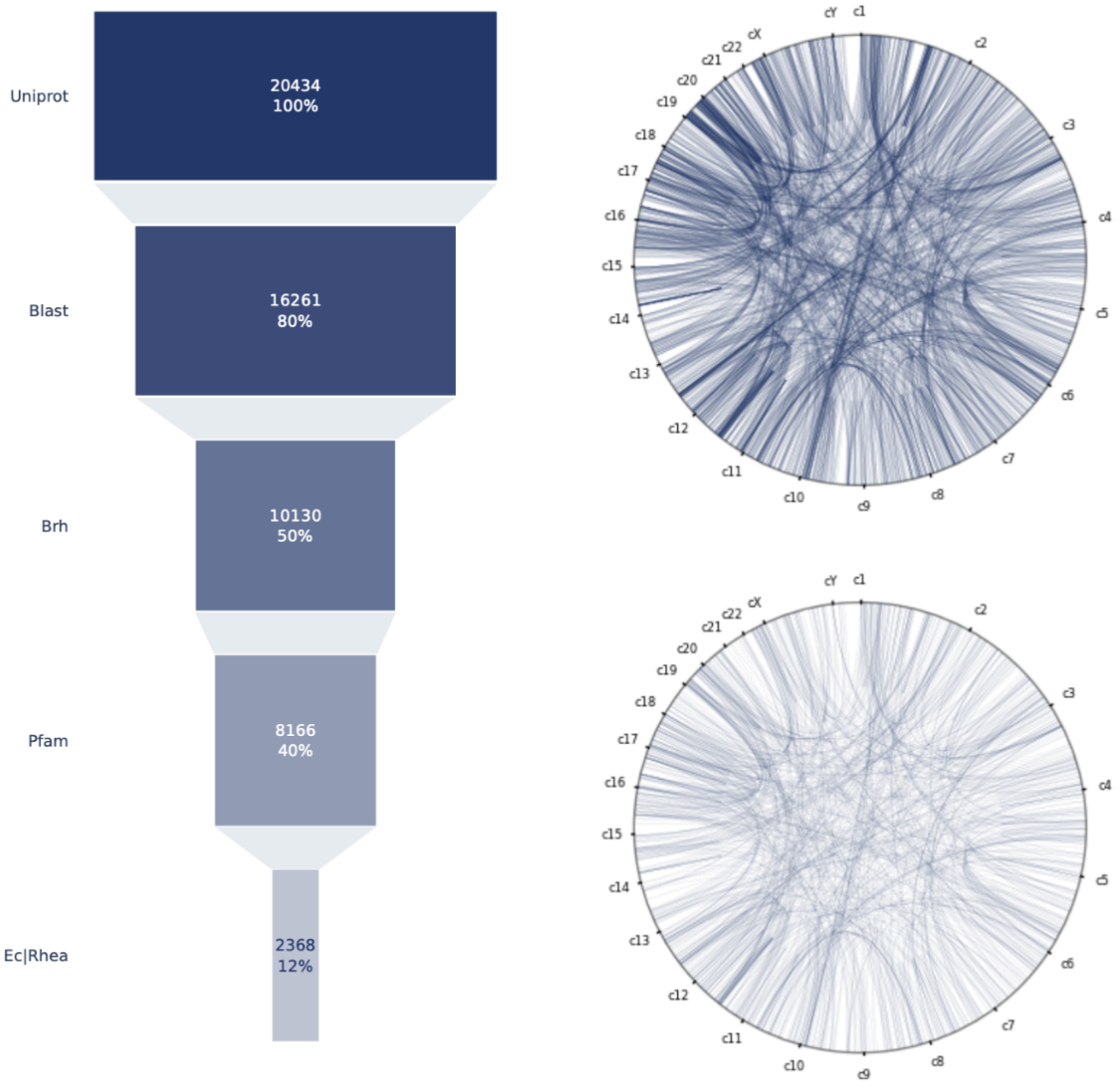


Figure 35: Overview of filtering and chromosomal distribution of human duplicated gene pairs. *Left panel*: A funnel plot showing the progressive reduction in the number of human duplicated gene pairs at each filtering step. Starting from 20,434 proteins in the UniProt database, subsequent steps reduced this number through BLAST, BRH, Pfam, EC/Rhea filtering, reaching a final set of 2,368 proteins with enzymatic activity. Percentages represent the proportion of proteins relative to the starting set. *Right panel*: Radial plots representing chromosomal locations and duplication events. Each section corresponds to a chromosome (c1-c22, X, Y). The top radial plot shows duplication events for gene pairs identified after the BRH filtering step. The bottom radial plot displays the distribution of the final set of gene pairs with probable enzymatic activity. Lines indicate duplications, with many involving genes located on different chromosomes.

2. Duplication Events Across Taxonomic Classes

A total of 458,028 orthologous sequences were retrieved across 177 vertebrate species for 1,147 human duplicate gene pairs with Rhea/EC annotation. Of these, 280,906 genes were classified as ortholog “one-to-one”. 17,492 sequences were excluded due to the presence of ambiguous amino acids (‘X’). On average, each gene had orthologs in 113 species.

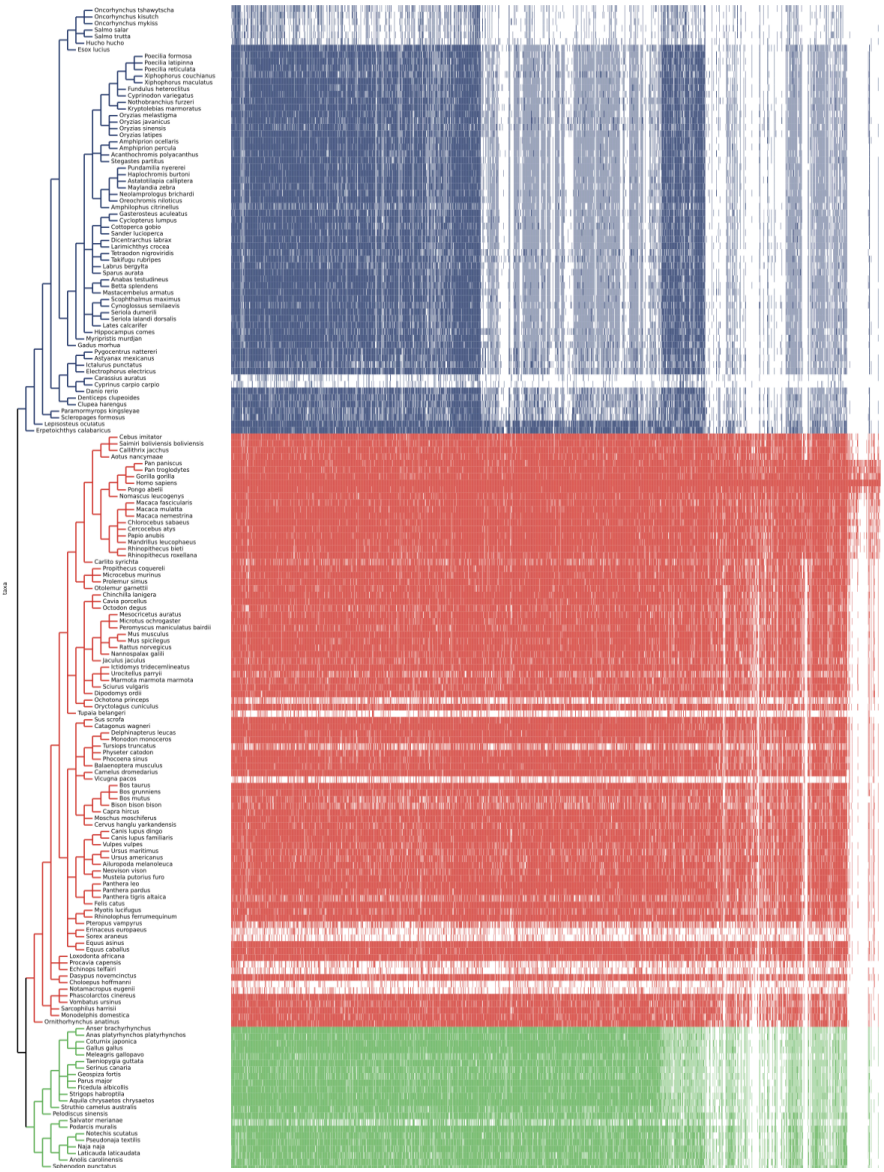


Figure 36: Heatmap of duplicated gene pairs across vertebrate species. This heat map displays the presence and the absence of duplicated gene pairs with Rhea/EC annotation across species from three taxonomic classes: Actinopterygii (blue), Mammalia (red) and Sauropsida (Green). The species are listed on the Y-axis, and their phylogenetic relationships are depicted by the tree on the left. Each tick on the X-axis represents a gene pair, sorted by the number of species in which the pair is present. Heatmap cells are color-coded based on the species taxonomic class and the alpha value indicates

whether the gene pair is present (maximum alpha), only one gene of the pair is present (intermediate alpha), or neither gene is present (zero alpha). This visualization highlights the distribution of gene duplications across species, with patterns of presence and absence visible across the taxonomic group.

Through the selecting only pairs in which both members find at least one orthologous gene, the number of pairs dropped to 1,121. The heatmap in **Figure 36** summarizes the presence of duplicate gene pairs across the 177 vertebrate species. Gene pairs are plotted along the X-axis and species along the Y-axis, with the associated phylogenetic tree. The X-axis is ordered to prioritize gene pairs present in the highest number of species. This sorting helps reveal clusters of widely conserved pairs, as well as those specific to certain taxonomic groups. The heatmap reveals a distinct pattern of gene duplications. A significant number (412, **Fig. 37**) of retained gene pairs can be observed across all three taxonomic classes: Mammalia, Sauropsida and Actinopterygii.

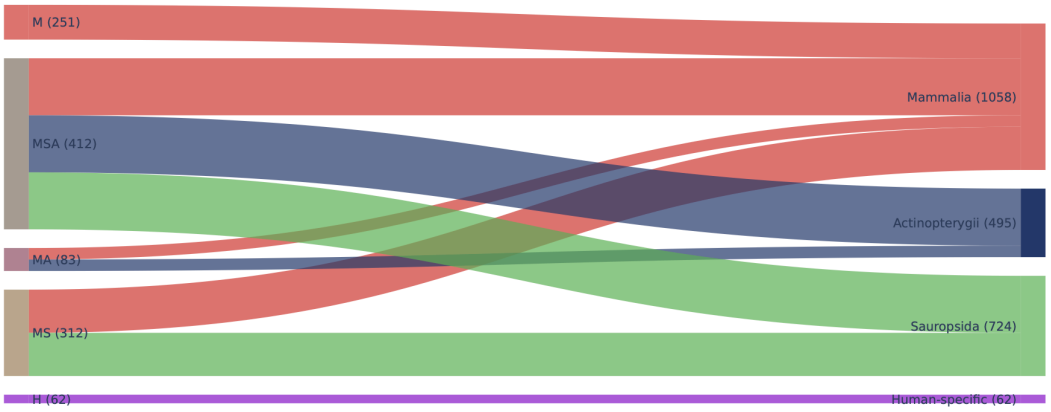


Figure 37. Sankey plot depicting the distribution of duplicated gene pairs among taxonomic classes. The Sankey plot visualizes the flow of duplicated gene pairs among Mammalia, Sauropsida and Actinopterygii. The left nodes represent categories of duplication presence (e.g., duplicates shared between mammals and fishes 'MA'). The flows indicates the number of duplicated gene pairs moving from these categories to the taxonomic classes on the right. The width of each flow is proportional to the number of gene pairs. Key observations include 251 duplicates exclusives to mammals ('M'), 83 exclusive to mammals and fishes ('MA'), 412 shared between the three groups ('MSA'), 312 exclusive to mammals and sauropsids ('MS'), 62 human exclusive ('H').

Certain species, such as *Carassius auratus* (goldfish) and *Vicugna pacos* (alpaca), exhibit rows with predominantly light colored cells, indicating a high number of missing genes. This pattern may result from incomplete genome assemblies or sequencing errors, leading to

underrepresentation of gene content in these species. We acknowledged that variations in genome assembly completeness across species could influence the presence or absence of gene pairs. Species with incomplete or low quality genomes were noted, and interpretations were made cautiously. *Erpetoichthys calabaricus* (reedfish) and *Lepisosteus oculatus* (spotted gar) show a notable presence of duplicated gene pairs that are absent in other fish species. The exclusive retention of these duplicates in these species may indicate lineage-specific duplication events or unique evolutionary pressure acting on their genomes. The middle right section of the heatmap reveals duplicate gene pairs present exclusively in primate species.

A Sankey plot quantifying the distribution of duplicated gene pairs among the three taxonomic groups is presented in **Figure 37**. Among duplicated gene pairs, 251 were present exclusively in mammals, 83 were shared between mammals and fishes, 312 were exclusive to mammals-sauropsids group, 62 were found to be primates or humans exclusive, and 412 were common to the three taxonomic groups. The cumulative count of duplicate gene pairs across all mammals, fishes and sauropsids species sums to 1,058, 495, 724 respectively.

3. Alignment Quality Evaluation

3.1. Visual Assessment of Alignment Quality Improvement

The alignment and subsequent cleaning processes significantly enhanced the quality and reliability of the multiple sequence alignments (MSAs) for duplicated gene pairs and their orthologs. Initially, alignments exhibited variability in sequence lengths, excessive gaps, and inconsistent residue conservation, which could impede accurate functional and evolutionary analyses.

Figure 38 illustrates a representative section of an MSA before and after the cleaning procedure. The pre-cleaning alignment on the left displays numerous gaps and misaligned residues, including sequences with additional regions compared to their orthologs. Post-cleaning, the alignment is improved, showing reduced gaps and more consistent residue positioning, indicative of higher alignment fidelity.

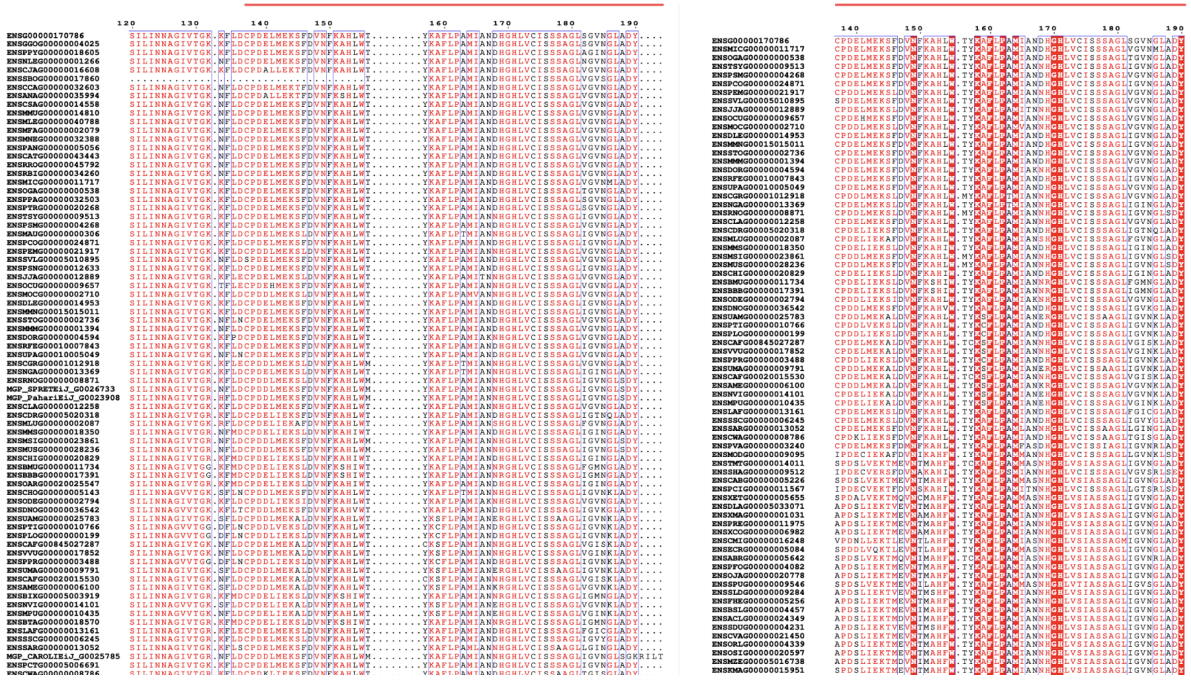


Figure 38: Comparison of Multiple Sequence Alignment before and after cleaning. A representative section of a multiple sequence alignment (MSA) generated with ESPript is shown before (left panel) and after (right panel) the cleaning procedure. The pre-cleaning alignment exhibits numerous gaps and misaligned residues, including sequences with additional regions inconsistent with orthologous counterparts. Post-cleaning, the alignment demonstrates reduced gaps and improved residue alignment, highlighting the effectiveness of the cleaning process in enhancing alignment quality.

To quantitatively assess the impact of the cleaning process, we compared several alignment metrics before and after cleaning across all processed alignments. These metrics include the number of sequences per alignment, the total number of gaps, and the percentage identity among aligned sequences. **Figure 39** presents three scatter plots accompanied by histograms, illustrating the distribution of these metrics before and after cleaning. The histogram above each scatter plot highlights the concentration of data points along the X-axis. After cleaning (Dots) the number of sequences per alignment decreases uniformly, reflecting the removal of redundant or erroneous sequences. Alignments showed a concentration of approximately 300 sequences. After cleaning a significant reduction in the number of gaps was observed, indicating the removal of poorly aligned regions and excessive indels. After cleaning, alignments contained an average of ~200,000 gaps. After cleaning there was an overall increase in percentage identity, signifying improved conservation across aligned sequences. A few rare instances showed a slight decrease, potentially due to the exclusion of highly divergent sequences that skewed the average. Average identity after cleaning was around 60%.

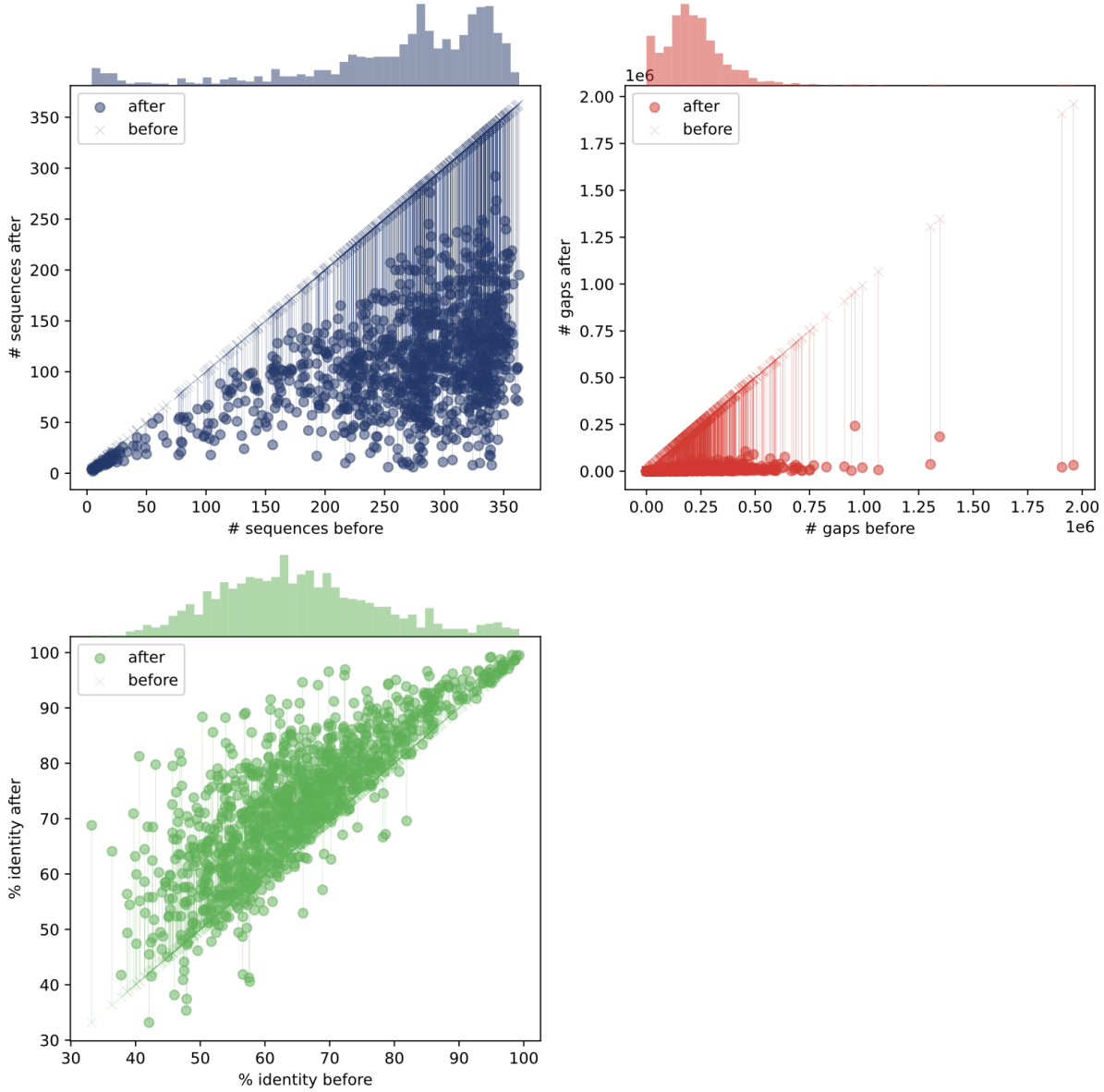


Figure 39: Impact of cleaning on alignment metrics across all gene pairs. The figure comprises three scatter plots with corresponding histograms, comparing alignment metrics before (X symbols) and after (dots) the cleaning process. *Top left panel*: each vertical line connects the number of sequences in alignment before (X) and after (dot) cleaning; histogram displays the distribution of sequence counts, with a concentration around 300 sequence post-cleaning. *Top right panel*: vertical lines indicate the number of gaps before and after cleaning for each alignment; histogram shows a high concentration of gaps (~0.2 to 1e6) after cleaning. *Bottom panel*: each line connects the percentage identity before (X) and after (dot) cleaning; histogram illustrates an increase in average identity to approximately 60% post-cleaning, with some rare alignments showing a slight decrease.

3.2. Quantitative Analysis of Alignment Metrics

We classified each residue in the multiple sequence alignments (MSAs) into three distinct categories: Robust, Adaptive and Plastic. This classification was based on the calculated Differential Score (DSc) and Total Score (TSc), as defined in the Methods section. **Figure 40** illustrates the distribution of amino acid residues classified into Robust, Adaptive, and Plastic categories based on their Differential Score (DSc) and Total Score (TSc). The classification identified approximately 260,000 Plastic residues, 160,000 Adaptive residues, and 175,000 Robust residues.

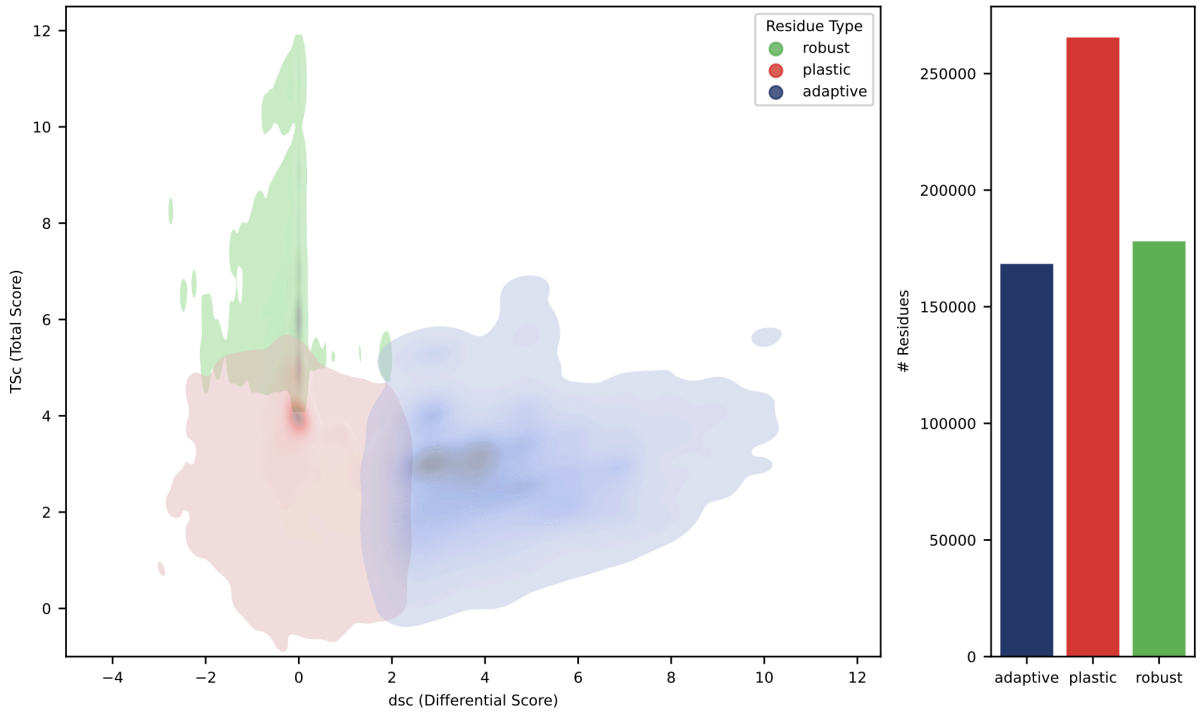


Figure 40: Distribution of amino acid residues types based on differential and total scores. *Left panel*: A kernel density estimate (KDE) plot showing the density distribution of amino acid residues identified in the alignments. The X-axis represents the Differential Score (DSc), while the Y-axis represents the Total Score (TSc). Residue types are colored based on thresholds defined in the methods section: robust residues are shown in green, plastic residues in red, and adaptive residues in blue. The KDE plot illustrates the overlap and separation of residue types across these scoring metrics, with robust residues clustering near zero DSc and higher TSc values, while adaptive residues show broader dispersion across higher DSc values. *Right panel*: A bar plot showing the absolute counts of the three residue types. The plastic residues are the most abundant, followed by robust and adaptive residues, reflecting the distribution of residue classifications based on the scoring thresholds.

Notably, our scoring system, grounded in the BLOSUM62 substitution matrix, imposes inherent mathematical constraints on the achievable combinations of DSc and TSc values.

For instance, the maximum attainable DSc is 11, and the corresponding TSc could be at maximum 6, as demonstrated by the residue pair C-W with ISc = 9, ISc2 = 11, and XSc = -2. This results in DSc = min(9, 11) - (-2) = 11 and TSc = (9 + 11 + (-2)) / 3 = 6. Consequently, regions in the DSc-TSc space, such as DSc = 10 and TSc = 6.5, are mathematically impossible to achieve within our framework, validating the absence of residues in these areas on the KDE plot. Additionally, Robust residues consistently exhibit XSc values below zero, reaffirming their high in-group conservation without significant cross-group similarity. Adaptive residues occasionally exceed the TSc threshold of 5, although they predominantly remain below, indicating a balance between substantial differential conservation and moderate overall conservation across groups. Conversely, Plastic residues are reliably located within the range of approximately 0 <= TSc <= 5 and -2 <= Dsc <= 2, aligning with their classification as variable residues.

4. Divergence Metrics and Functional Insights

4.1. Residue-Level Divergence Metrics Analysis

The relationship between our base divergence score metrics was evaluated by calculating Pearson correlation coefficients between the embedding-based delta score, which quantifies the average Euclidean distance between residue embedding vectors, and the conservation-based differential score (DSc), as well as between delta and the total score across all amino acid residues (TSc). The analysis revealed a moderate positive correlation between delta and DSc (Pearson's $r = 0.62$, $p < 0.001$) (Kirch, 2008). Additionally, the correlation between delta and TSc was found to be moderately negative (Pearson's $r = -0.36$, $p < 0.001$). **Figure 41** visualizes these relationships through scatter plots accompanied by regression lines. The left panel displays the positive trend between delta and DSc, illustrating that as delta scores increase, DScs also tend to rise. The right panel depicts the negative trend between delta and TSc, showing that higher delta scores are generally associated with lower TScs.

In total, our analysis encompassed 609,927 amino acid residues, each evaluated using four distinct metrics: Differential Score (DSc), Pocket Differential Score (pdsc), Hotspot Differential Score (hdsc), Hotspot Delta Score (hdelta). These metrics collectively integrate evolutionary information, structural insights, and sequence embedding data to classify residues into functionally significant categories.

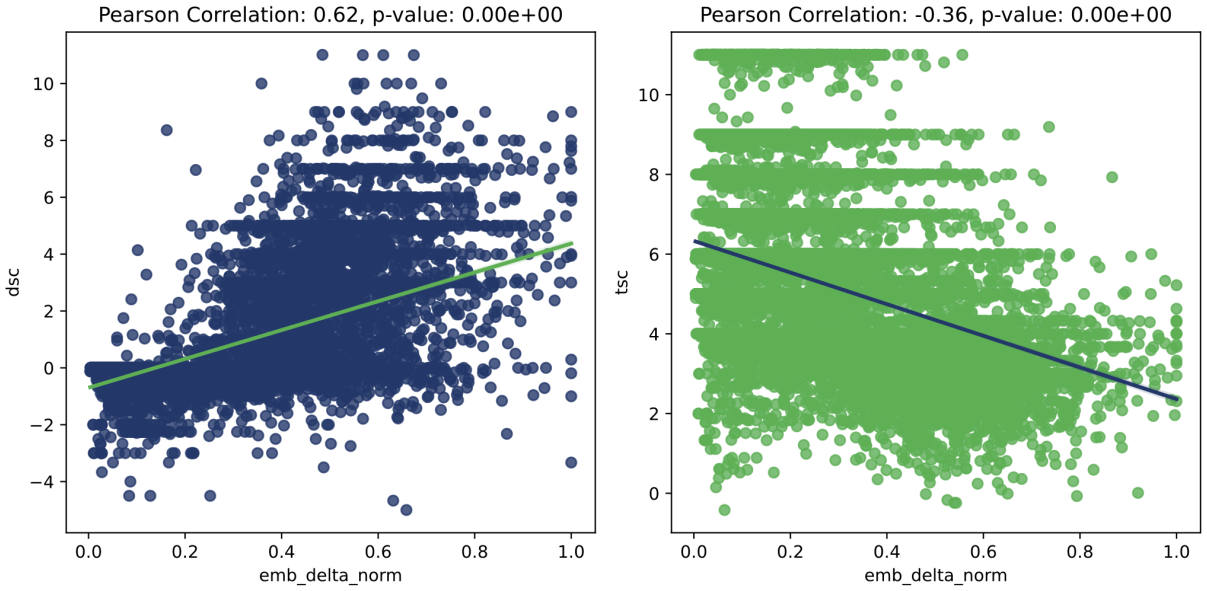


Figure 41: Scatter plots of the relationship between embedding-based delta score and conservation-based metrics (DSc and TSc). The scatter plots illustrate the relationship between the embedding-based delta score (emb_delta_norm) and two conservation metrics: the Differential Score (DSc, left panel) and the Total Score (TSc, right panel). Pearson correlation coefficients are shown above each plot. In the left panel, a moderate positive correlation ($r = 0.62$, $p < 0.001$) is observed between delta and DSc, indicating that residues with higher delta scores tend to exhibit higher differential conservation. In contrast, the right panel shows a moderate negative correlation ($r = -0.36$, $p < 0.001$) between delta and TSc, suggesting that residues with higher delta scores are generally less conserved overall. Regression lines in both plots highlight these trends.

Figure 42 displays the distribution of residues across these metrics through bar plots, with the respective thresholds indicated by vertical lines. The distribution highlights the following key findings: 167,333 residues exhibited a DSc above the predefined threshold, indicating significant evolutionary divergence; 243,971 residues exceeded the delta score threshold, reflecting considerable dissimilarity in embedding space; among these, 142,817 residues concurrently exceeded both DSc and delta thresholds, suggesting a strong overlap between evolutionary divergence ad embedding-based dissimilarity; 36,633 residues were identified as hotspot based on a probability score greater than the threshold, as predicted by BindEmbed21; additionally, 26,742 residues were classified as part of functional pockets using P2Rank, based on a probability exceeding the predefined threshold.

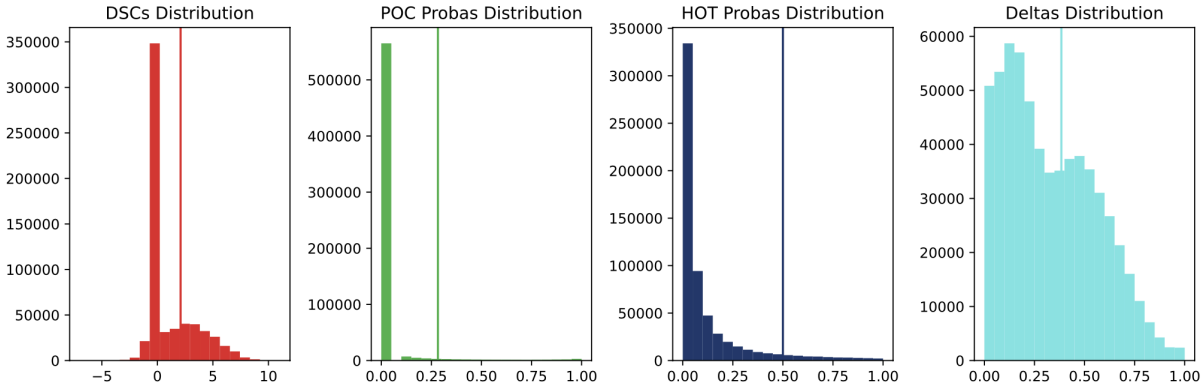


Figure 42: Distribution of residue metrics with threshold values. This bar plot illustrates the distribution of the 609,927 residues based on four metrics: DSc, delta, hotspot probability, and pocket probability. Of these, 167,333 residues have DSc above the threshold, 243,971 residues exceed the delta threshold, and 142,817 residues are above the delta threshold, 36,633 residues have a hotspot probability greater than the threshold, while 26,742 residues exceed the threshold for pocket probability. The vertical lines indicate the thresholds used for each metric, providing a reference for the classification of residues across evolutionary and structural criteria.

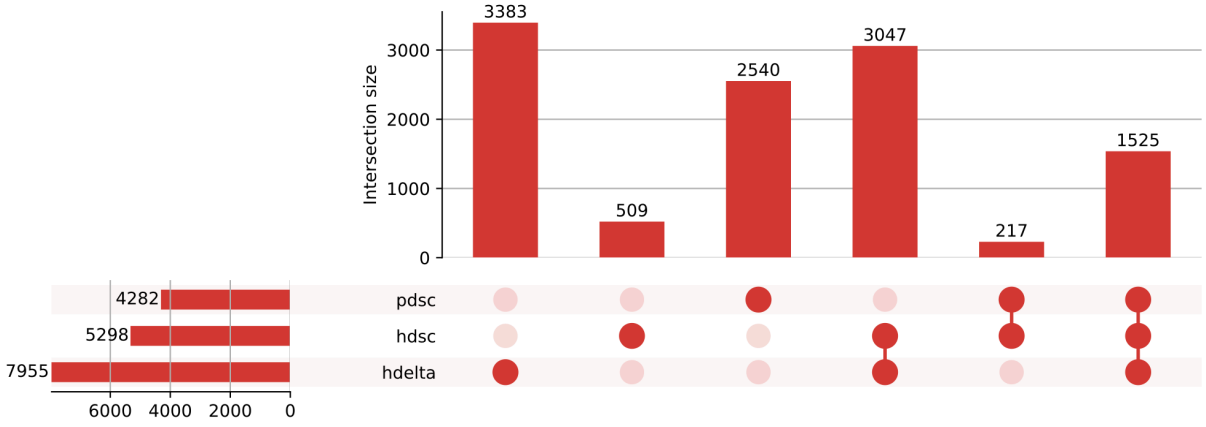


Figure 43: Upset plot showing intersections between residue classifications. This upset plot visualizes the qualitative intersections among residues classified by the delta, hdsc, and pdsc metrics. The horizontal bars represent the absolute counts of residues classified by each individual metric, while the vertical bars display the number of residues shared across various combinations of the metrics. Specifically, 7,955 residues are classified as delta, 5,298 as hdsc, and 5,282 as pdsc. A total of 1,525 residues exceed the thresholds for all three metrics.

Figure 43 presents an UpSet plot, illustrating the intersections among the four classification metrics: 7,955 residues classified as hdelta, integrating embedding information with high delta scores; 5,298 residues were classified as hdsc, combining evolutionary divergence (DSc) with embedding embedding-based information; 5,282 residues were classified as pdsc, merging evolutionary divergence with structural functional pocket data. Combining the

classifications, 1,525 residues met all four classification criteria, indicating high significance across all metrics; 217 residues were identified as both pdsc and hdsc but did not meet the hdelta threshold; 509 residues were classified solely as hdsc, without meeting pdsc or hdelta threshold.

4.2. Validation of Divergence Metrics with RHEA Truth Set

We constructed a truth set using the RHEA database, which provides experimentally validated biochemical reactions, to validate the effectiveness of our divergence metrics in identifying functionally divergent protein pairs. **Figure 44** presents the distribution of truth scores for our protein pairs through two histograms.

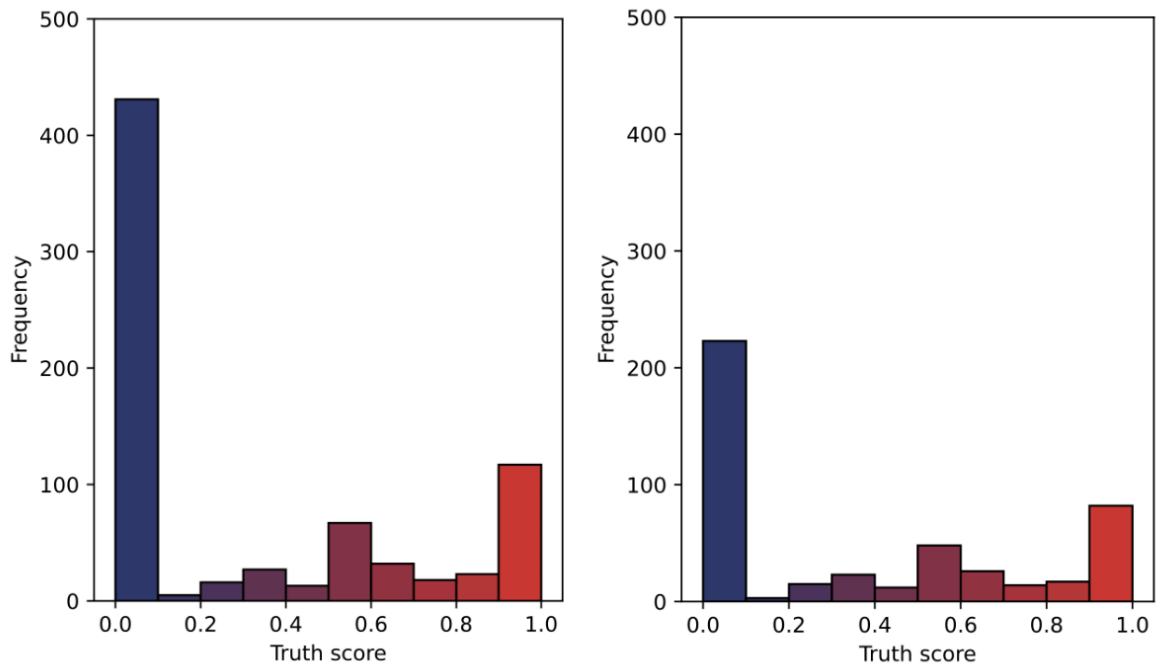


Figure 44: Distribution of Truth Scores in protein pairs using the RHEA database. Histograms depicting the distribution of truth scores for all analyzed protein pairs, including both experimental and non-experimental entries. The x-axis represents the truth score ranging from 0 to 1, while the y-axis indicates the frequency of protein pairs. The distribution shows peaks at approximately X = 0 and X = 1. A color gradient from blue to red highlights the transition from low to high truth scores.

The left histogram illustrates the overall distribution, encompassing both experimental and non-experimental entries, with approximate peaks at X = 0 (430 pairs) and X = 1 (106 pairs).

The right histogram focuses exclusively on protein pairs with experimental RHEA entries, revealing distinct peaks at X = 0 (222 pairs) and X = 1 (73 pairs). These peaks correspond to non-divergent and divergent protein pairs, respectively, confirming the bimodal nature of our truth set. The color gradient from blue (low sores) to red (high scores) enhances the visual distinction between the two categories.

4.3. Impact of KDE-Derived Thresholds on Metrics Performance

The determination of optimal thresholds for our divergence metrics using Kernel Density Estimate (KDE) curve intersections was useful in enhancing the accuracy and reliability of identifying functionally divergent protein pairs. These thresholds were instrumental in converting continuous metric scores into binary classification, thereby enabling the calculation of discrete performance metrics such as Precision, Recall, and F1 Score. By analyzing the KDE curves for true positives (functional divergence present) and true negatives (no functional divergence), we identified intersection points that served as natural cutoffs between the two classes. Notably, thresholds determined via ROC curve analysis using Youden's J statistic (sensitivity+specificity-1) fell approximately at the same points as the KDE intersections (Youden, 1950). Given this concordance, we opted to use the KDE intersections thresholds. These intersection thresholds are detailed on **Table 1** and are visually represented in **Figure 45**, where selected KDE intersections are highlighted to illustrate their role in threshold setting.

Upon analyzing the ROC curves and the corresponding heatmaps for F1, precision, and Recall (**Fig. 46**), distinct performance patterns emerged across the various divergence metrics. Evolutionary-only metric (ldsc_r), which relies solely on evolutionary information, demonstrated the poorest performance with an Area Under the ROC Curve (AUC) of 0.747 and an F1 Score of 0.5. In contrast, metrics that integrated evolutionary information with additional data sources exhibited significantly improved performance. For instance, hdsc_r, which integrates evolutionary information with functional residues predicted by neural networks, achieved an AUC of 0.909 and an F1 Score of 0.75, marking it as the top-performing non-combined metric. Similarly, pdsc_r, which combines evolutionary and structural information from functional pockets, is closely followed with an AUC of 0.892 and an F1 Score of 0.74. Among the combined metrics, hdelta_r, which incorporates only embedding information, yielded an AUC of 0.872 and an F1 Score of 0.68, indicating moderate performance relative to mixed metrics.

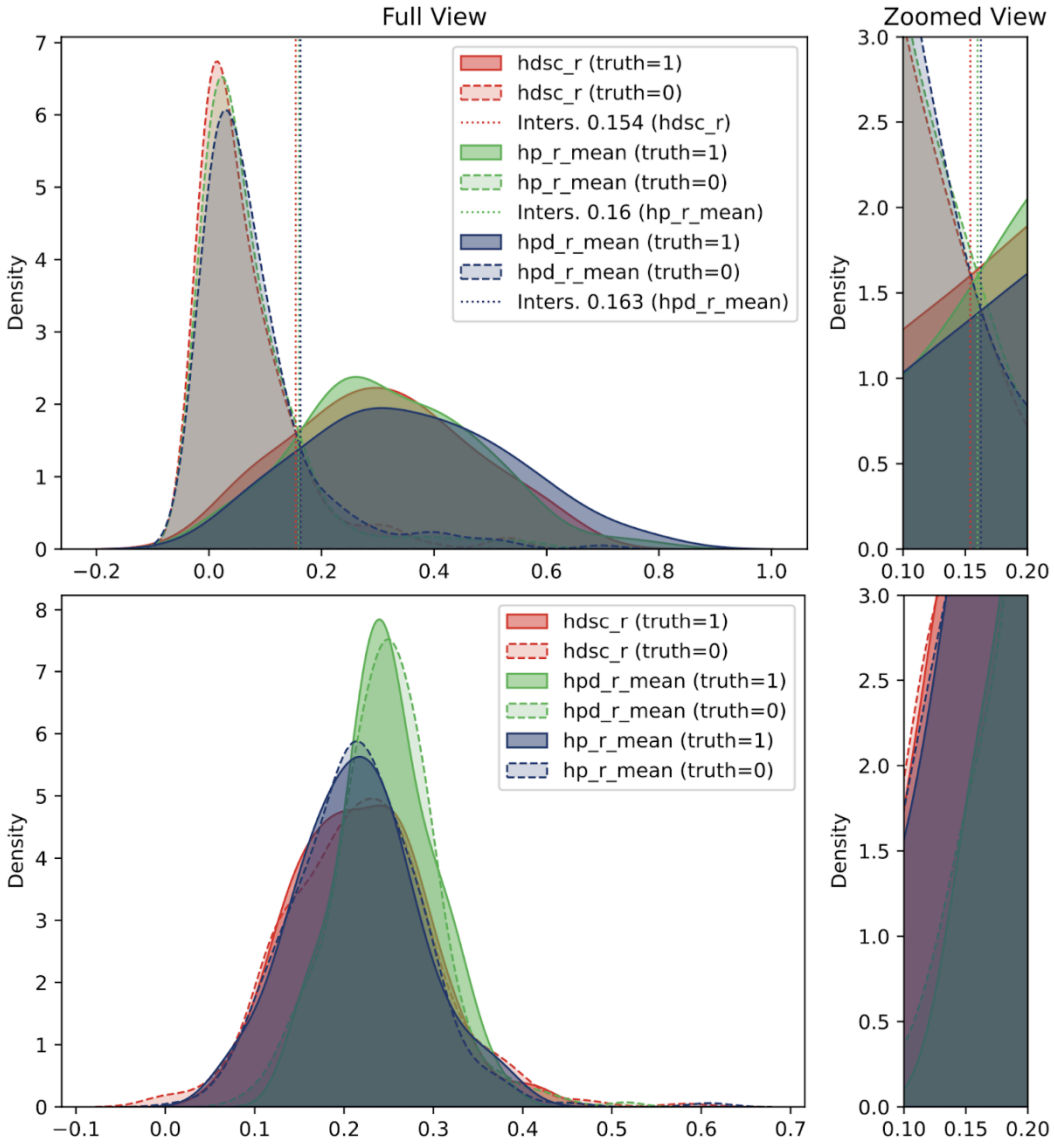


Figure 45. KDE curve intersections for divergence metrics. *Top panels*: KDE curves for three divergence metrics – hdsc_r (red), hpd_r_mean (green) and hp_r_mean (blue) – depicting the distribution of predicted scores for truth = 0 (non-divergent and highlighted as dotted lines) and truth = 1 (divergent and highlighted as continue lines) protein pairs. The intersection points, where the curves for truth = 0 and truth = 1 cross, are marked with vertical dotted lines at thresholds of 0.154, 0.160, and 0.163, respectively. These intersections denote the optimal thresholds used for binary classification, highlighting regions where the model transitions from predicting non-divergent to divergent protein pairs. *Top right panel*: A zoomed-in view focusing on a narrower range of predicted scores (between 0.1 e 0.2) provides a closer examination of the overlapping areas around the intersection thresholds. *Bottom panels*: The same KDE curves are plotted for randomly generated data, used to test the functionality and robustness of predictive scores. Similar to the real data, the left bottom panel shows the full range of scores, and the right bottom panel zooms in on the same intersection region as above.

However, when combining multiple metrics, performance further improved. Hpd_r_mean (which averages hdsc_r, pdsc_r, and hdelta_r) attained the highest AUC of 0.919 and an F1 Score of 0.76, underscoring the advantage of averaging multiple metrics to enhance predictive power. Additionally, hp_r_mean, which averages hdsc_r and pdsc_r, achieved an AUC of 0.917 and the highest F1 Score of 0.79, primarily driven by improved precision. Additional metrics, such as pdsc_r_norm (pdsc_r divided by the greater length of the two proteins in the pair), were also evaluated. However, these normalized counterparts did not outperform their non-normalized versions, indicating that normalization in this context did not confer additional predictive benefits.

To contextualize these findings, we generated random prediction values to serve as baseline comparison. In order to randomize data, either DSc or Delta values were randomized across the entire dataset, while keeping other features intact. The ROC curves for these random values yielded AUCs around 0.5, and the corresponding F1 Scores hovered near 0.35 across all metrics, indicating no discriminative power. The recall of ~1 indicates that the scoring method successfully identified all true divergent protein pairs (i.e., no false negatives) but the precision of 0.2 suggests that the majority of the predictions were false positives, with only 20% of the predicted pairs being correct. In the KDE curves, the truth=0 and truth=1 distributions for the random predictions exhibited significant overlap, further underscoring the lack of effective separation and validating the superior performance of our optimized metrics. The numerical outcomes clearly demonstrate that metrics integrating multiple sources of information – particularly evolutionary and functional data – outperform those based solely on a single type of data. The high AUC values and balanced F1 Scores of the combined metrics validate the effectiveness of our threshold optimization approach using KDE curve intersections. This optimization ensures that our divergence metrics are finely tuned to distinguish between functionally divergent and non-divergent protein pairs, thereby enhancing the reliability of our classification framework. While the high AUC values and balanced F1 Scores affirm the model's effectiveness, the overlapping regions near the threshold values of approximately 0.15 indicate areas where predictive reliability may decrease. This suggests potential classification errors, such as false positives and false negatives, in these regions.

The overlapping regions near the threshold values of approximately 0.16 (**Fig. 45**) indicate areas where the model's predictive reliability may decrease, suggesting potential for classification errors such as false positives and false negatives.

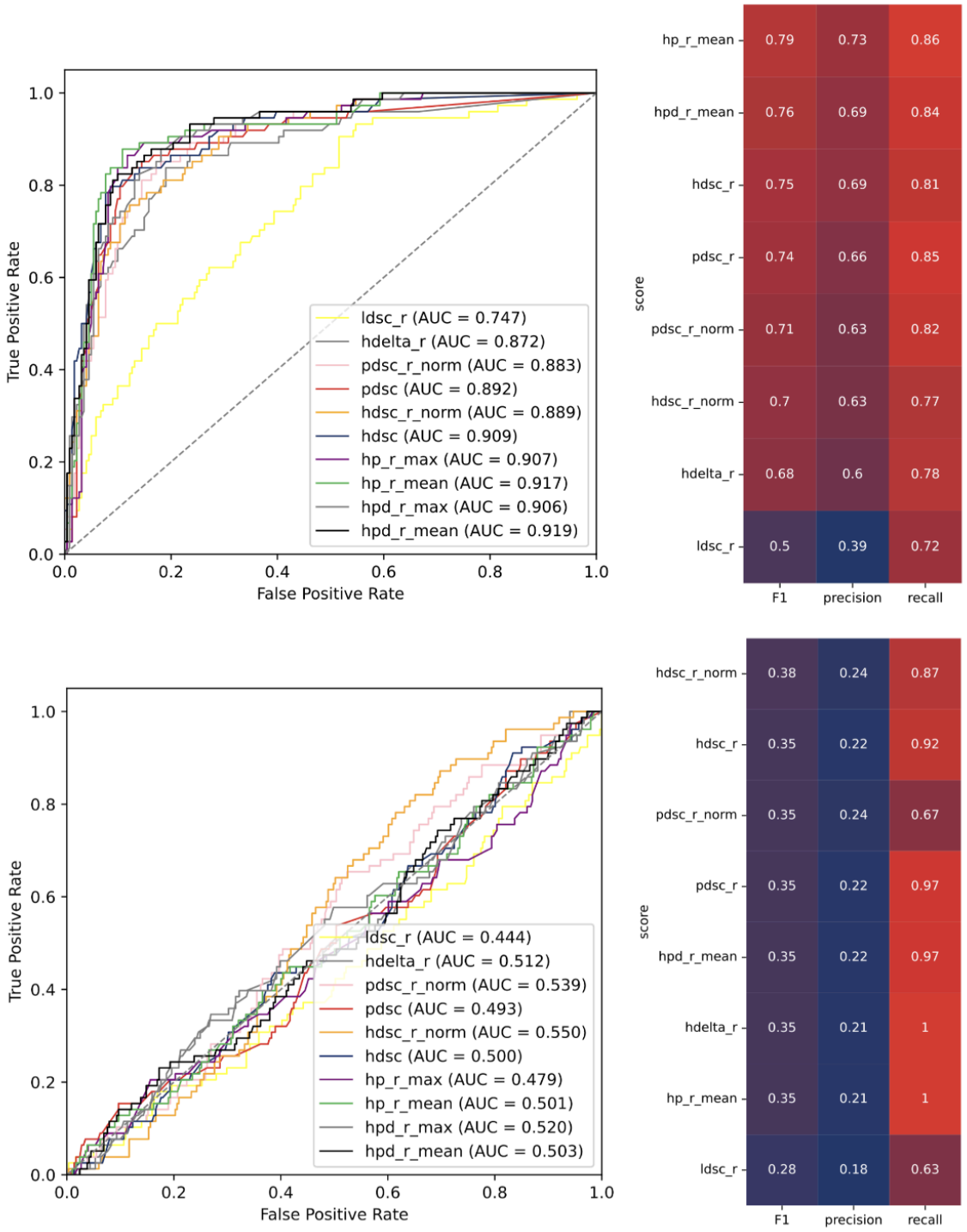


Figure 46: ROC curves and performance heatmaps for real and randomly generated data. *Top panels*: The ROC curve on the left shows the performance of 10 different metrics evaluated on real data, while the heatmap on the right displays F1 score, precision, and recall for the same metrics. These panels highlight the effectiveness of the scoring methods in distinguishing between divergent and non-divergent residues. *Bottom panels*: The ROC curve on the left and the heatmap on the right present the corresponding results for randomly generated data, serving as control to assess the robustness of the metrics.

score	pos	precision	recall	F1	AUC	threshold
ldsc_r	510	0.39	0.72	0.50	0.75	0.237
hdelta_r	400	0.60	0.78	0.68	0.87	0.187
pdsc_r_norm	368	0.63	0.82	0.71	0.88	0.00030
pdsc_r	349	0.66	0.85	0.74	0.89	0.163
hdsc_r_norm	358	0.63	0.77	0.70	0.89	0.00031
hdsc_r	349	0.69	0.81	0.75	0.91	0.154
hp_r_mean	343	0.73	0.86	0.79	0.92	0.160
hpd_r_mean	382	0.69	0.84	0.76	0.92	0.163

Table 1: Performance metrics of predictive scoring methods based on divergence score. The table displays performance metrics for eight predictive scoring methods evaluated using various scoring criteria. Each method is assessed by its precision, recall, F1 score, AUC, and optimal threshold. The “pos” column refers to the number of positive predictions above the threshold for each method. The thresholds are used to distinguish between divergent and non-divergent protein pairs.

4.4. Identification of Functionally Divergent Protein Pairs

Applying the established threshold to our dataset, we predicted approximately 35% of protein pairs as functionally divergent. Specifically, utilizing the hpd_r metric threshold, the dataset was divided into 382 positive pairs (divergent proteins) and 730 negative pairs (non-divergent proteins), representing roughly 34% and 66% of the total respectively. **Figure 47** presents an UpSet plot illustrating the intersections among the three divergence metrics: pdsc_r (pocket divergence), hdsc_r (functional residue divergence), and hdelta_r (embedding space divergence). The plot reveals that 217 protein pairs were unanimously predicted as divergent by all three metrics (hdsc_r, pdsc_r, hdelta_r); 83 pairs were identified as divergent exclusively by the pdsc_r metric, indicating divergence drive solely by changes in the functional pockets without corresponding changes in functional residues or embedding space, likely reflecting changes in substrate processing while maintaining overall enzymatic function; 83 pairs were flagged as divergent only by the hdelta_r metric, signifying differences in functional residues not accompanied by evolutionary conserved differences; 574 pairs did not receive a suprathreshold score from any of the three metrics, remaining classified as non-divergent.

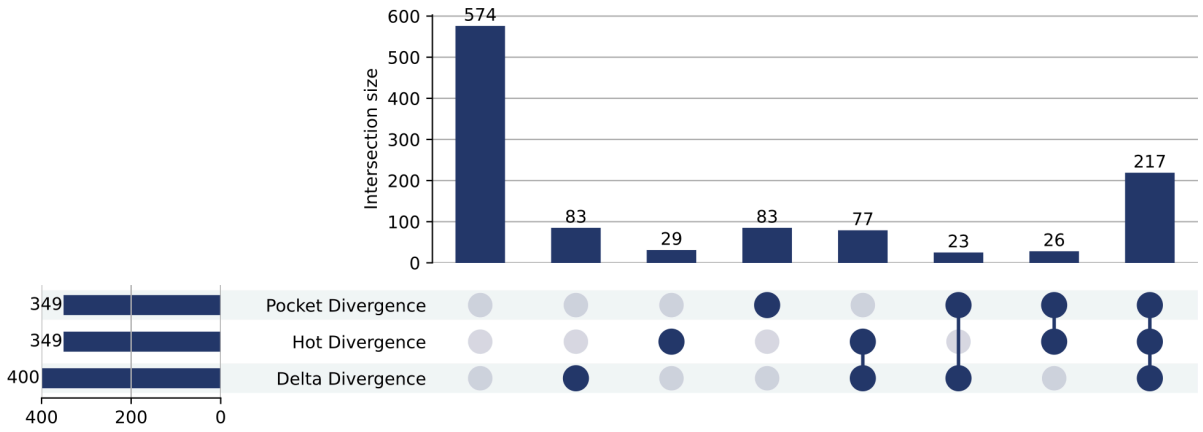


Figure 47: UpSet plot illustrating overlaps among divergence metrics. This figure presents an UpSet plot showing the intersections among three divergence metrics: pdsc_r (pocket divergence), hdsc_r (hotspot divergence), and hdelta_r (delta divergence). Dots represent the presence of protein pairs in specific metric combinations. Horizontal bars show the total number of protein pairs in each intersection. Vertical bars indicate the total number of protein pairs predicted as divergent for each individual metric.

4.5. Probability and Permutation Validation Analysis

To enhance the interpretability and reliability of our functional divergence scores, we applied a logistic transformation to convert the hpd_r scores into probabilities ranging from 0 to 1. This transformation facilitates a probabilistic interpretation of the scores, where values closer to 1 indicate higher confidence in functional divergence, and values nearer to 0 suggest lower confidence. By setting the threshold at the central point of the logistic function and using a steepness parameter of k=15, we ensured a sharp transition between divergent and non-divergent classifications. The resulting probability distribution revealed two distinct tails, effectively segregating our protein pairs into functionally divergent and non-divergent categories. Specifically, 237 protein pairs fell within the intermediate probability range of $0.3 < prob < 0.7$, 293 pairs were associated with high probabilities ($prob \geq 0.7$), and 582 pairs with low probabilities ($prob \leq 0.3$) (**Fig. 48**). To rigorously assess the statistical significance and stability of these probabilistic scores, we conducted a permutation test analysis comprising 1000 random resamplings. This method allowed us to evaluate whether the observed scores significantly deviated from what would be expected under random conditions. Initially, we employed a one-tailed p-value approach to determine if the observed hpd_r scores were significantly higher than those obtained through randomization. However, this approach consistently yielded p-values equal to 1. This outcome indicates that the observed scores

were uniformly lower than those generated by random permutations. Recognizing the limitation of the one-tailed approach in capturing the full spectrum of our data, we transitioned to a two-tailed p-value methodology. The two-tailed analysis revealed that 840 gene pairs exhibited p-values below the significance threshold of 0.05, with 560 of these pairs obtaining p-values less than 1e−4.

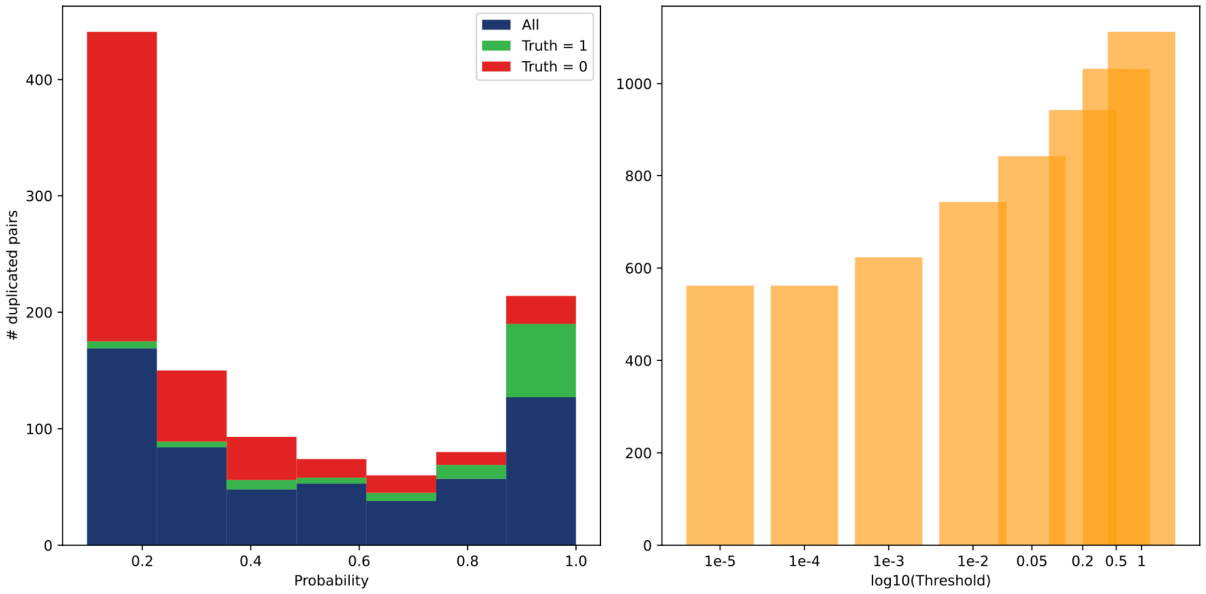


Figure 48: Histplot and barplot showing functional divergence probabilities and the p-values calculated for these probabilities. On the left, a stacked hist plot shows the probabilities obtained through the sigmoid function and the ration calculated for each pairs, with bars colored based on truth values (all predictions on unknown pairs in blue, probabilities associated with truth RHEA = 1 in green, and those associated with truth = 0 in red). The Y-axis represents the frequency of gene pairs. On the right, a barplot displays the log10 of the p-value thresholds applied on the X-axis, with the count of gene pairs with p-values below those thresholds on the Y-axis. There is no noticeable decrease in the number of pairs with a p-value less than 1e-4.

5. Gene Expression and Functional Enrichment

5.1. Tissue-Specific Expression Analysis

Expression data from the Human Protein Atlas (HPA) facilitated the analysis of tissue-specific differences among 434 protein pairs. For each pair, the expression difference (exp_diff) between two proteins was calculated across all 50 tissues available in the HPA dataset. By setting a threshold of 150 for the exp_diff value, we divided the dataset into 212 duplicate

pairs below the threshold and 222 pairs above it. The average hpdsc_r value is 0.150 for the first group and 0.159 for the second group, indicating only a slight difference between the two groups. This process also identified the tissue with the highest absolute exp_diff for each pair, along with the corresponding difference value.

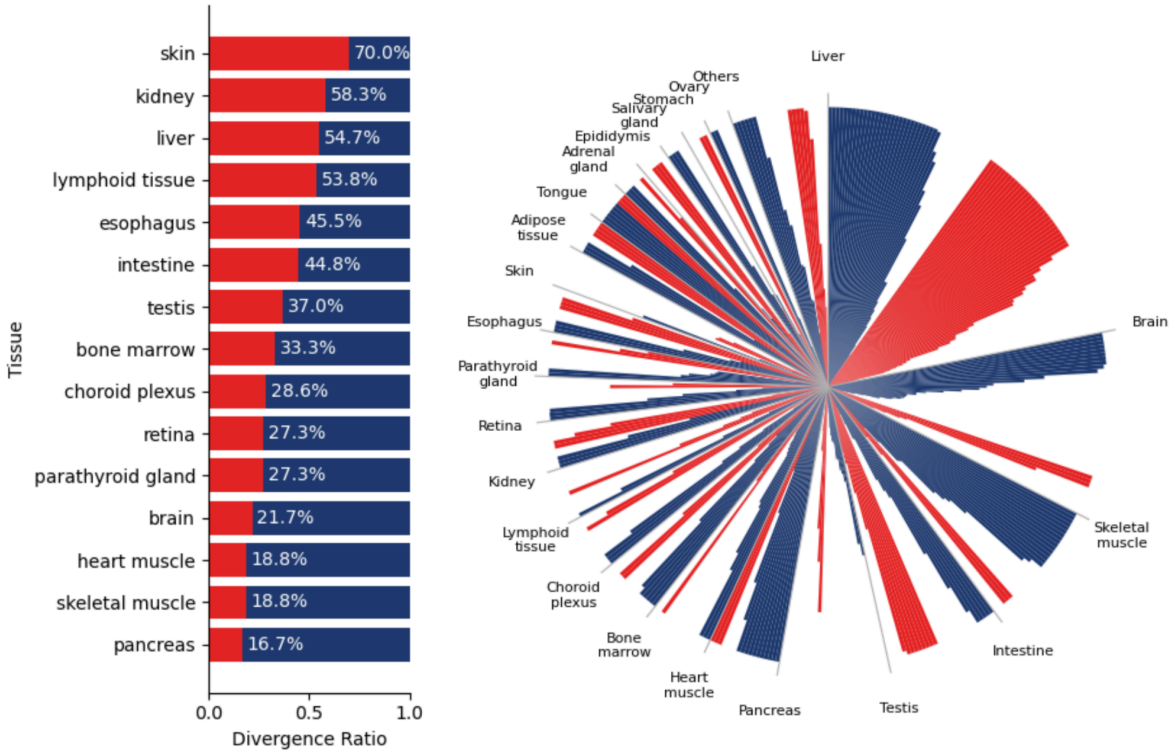


Figure 49: Combined radial and bar plot illustrating tissue-specific expression differences among protein pairs. The right panel shows a radial plot where each slice corresponds to a tissue with angular size proportional to the number of pairs, capped at 200 exp_diff values, and low-count tissues grouped under “Others”. Bars represent individual protein pairs, colored by divergence status (red for divergent, blue for non-divergent pairs). The left panel presents a bar plot of the percentage of divergent protein pairs (red) per tissue, ordered from highest to lowest percentage.

A radial plot (**Fig. 49**) illustrates the distribution of tissues identified as having the highest expression differences: slices represent the 50 tissue, with their angular size proportional to the number of protein pairs for which they were the tissue of maximum expression difference; bars correspond to individual protein pairs, where the height indicates the magnitude of exp_diff (capped at 200 for values exceeding this threshold). Bars are ordered based on their exp_diff values and color-coded to reflect functional divergence. The liver emerged as the most frequently identified tissue, accounting for 95 protein pairs (~22%). This was followed by the brain and skeletal muscle. Conversely, tissues such as the ovaries and salivary glands exhibited fewer instances of significant expression differences. A bar plot (**Fig. 49**) displays

the percentage of protein pairs classified as divergent within each tissue, ordered from highest to lowest: skin exhibited the highest proportion of divergent pairs at 70%, indicating substantial functional innovation in this tissue; the kidney followed with 58.3% of pairs classified as divergent; the liver maintained a significant proportion of 54.7% divergent pairs.

5.2. Functional Enrichment: Gene Ontology and KEGG Pathways

The functional enrichment analysis elucidated distinct biological functions and processes associated with divergent and non-divergent protein pairs (**Fig. 50**). By examining the enriched GO terms within the categories of Molecular Function (MF) and Biological Process (BP), as well as KEGG pathways, meaningful patterns emerged that highlight the potential roles of these protein pairs in various cellular contexts. Of the 1,112 protein pairs analyzed, 1,089 pairs were annotated with at least one GO BP entry, and 1,106 with at least one GO MF entry. A total of 963 pairs possessed both GO BP and GO MF annotations. THese annotations encompassed 4,033 unique GO BP entries, which were slimmed down to 94 GO BP terms, and 1,639 unique GO MF entries, reduced to 121 GO MF terms. Additionally, 384 pairs were associated with at least one KEGG pathway entry, with 324 pairs annotated with KEGG pathways for both proteins in the pair, covering a total of 83 unique KEGG pathways.

The divergent protein pairs exhibited significant enrichment in several catalytic activities, including transferase activity, oxidoreductase activity, hydrolase activity and lyase activity. Conversely, non-divergent pairs showed enrichment in binding-related molecular functions including nucleotide, RNA, and DNA Binding, protein binding and enzyme regulator activity, helicase and kinase activity, chromatin and transcription factor binding. The biological processes enriched among divergent protein pairs include metabolic processes such as amine, lipid, xenobiotic, catabolic, small molecule, and amino acid metabolism, and biosynthetic processes. Non-divergent pairs were enriched in biological processes fundamental to cellular integrity and function, including DNA metabolic process and chromosome organization, mitotic cell cycle and cell division, cytoskeleton organization. Divergent protein pairs showed significant enrichment in various specialized metabolic pathways, including arachidonic acid metabolism. alpha-linoleic acid metabolism, fatty acid elongation, steroid biosynthesis, tyrosine and histidine metabolisms. These enriched pathways indicate that divergent protein pairs are involved in a wide range of metabolic processes. Non-divergent pairs were enriched in central and fundamental metabolic pathways, including pyruvate metabolism, fructose and mannose metabolism, mucin type O-glycan biosynthesis, purine metabolism, and pentose phosphate pathways.

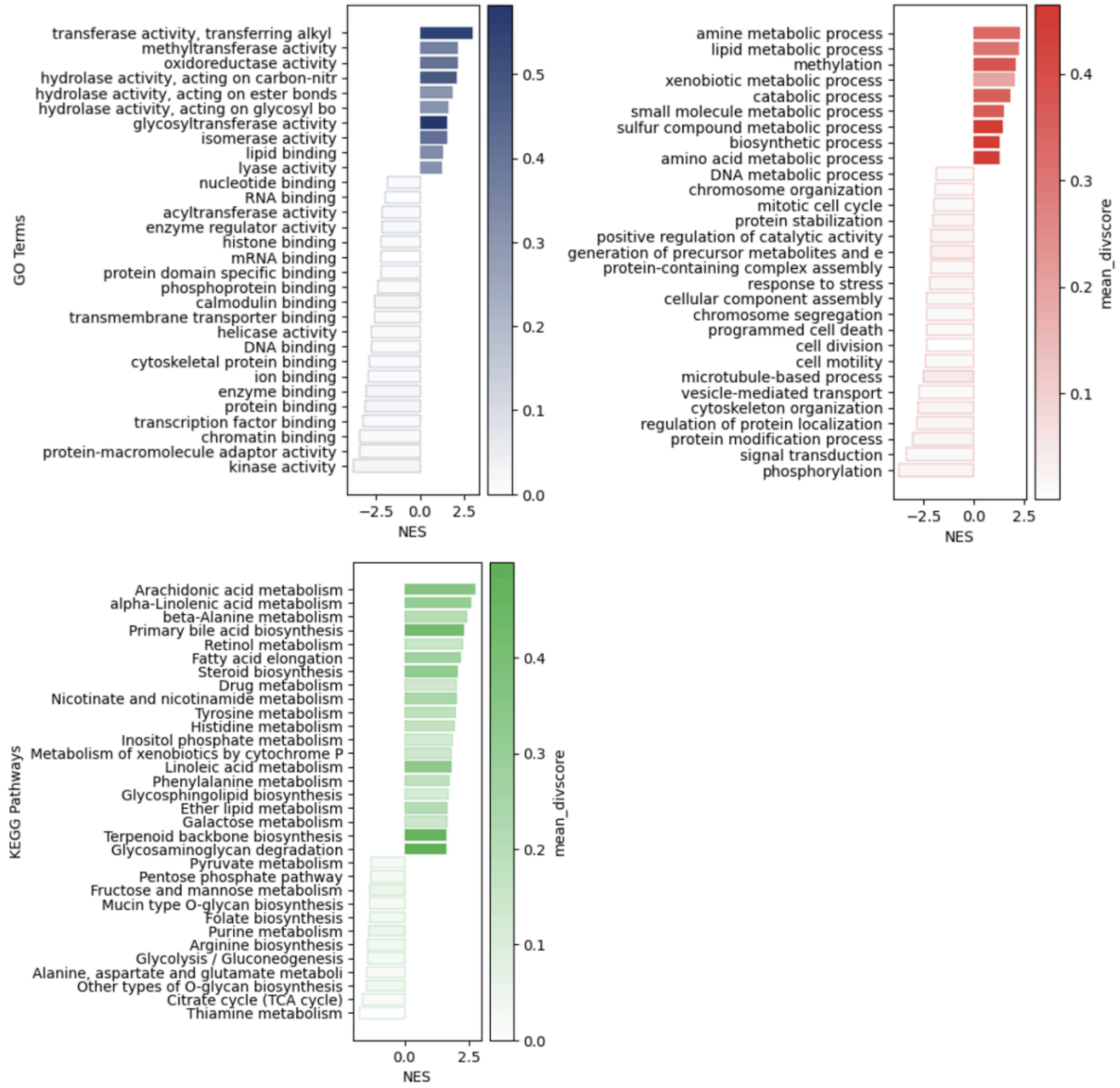


Figure 50: Functional and biological pathway enrichment for divergent and non-divergent protein pairs. This figure displays the enrichment of GO terms and KEGG pathways as horizontal bar plots. On the Y-axis, the GO terms and KEGG pathway descriptions are listed, sorted by the highest Normalized Enrichment Scores (NES). Positive NES values indicate that the GO terms and KEGG pathways are enriched among protein pairs with higher divergence scores, while negative NES values indicate enrichment among protein pairs with lower divergence scores, reflecting conserved functions. The X-axis represents the NES scores, providing a quantitative measure of the enrichment for each category. The bars are color-coded based on the mean divergence scores of the core set of proteins that contributed to the enrichment. Blue bars correspond to MF GO terms, red bars represent BP GO terms, and green bars depict KEGG pathways. The intensity of the color reflects the average divergence scores of the protein pairs, with more intense colors indicating higher divergence.

6. Results from Autonomous Pipeline: Divergence of AOC2 and AOC3

Known protein pairs with conserved functional residues exhibiting the same enzymatic activity but likely acting on different substrates were identified by ranking based on shared EC numbers and a pdsc_r score above the established threshold, while excluding channel proteins (**Table 2**).

gene1	gene2	len1	len2	h	hdsc	hdsc_r	hdelta	hdelta_r	p	pdsc	pdsc_r	hpd_r	ediff	truth	name1	name2
DTD1	DTD2	209	168	9	3	0.33	9	1.00	17	12	0.71	0.68	39	0.5	D-aminoacyl-tRNA deacylase 1	D-aminoacyl-tRNA deacylase 2
TRUB2	TRUB1	331	349	14	6	0.43	9	0.64	32	18	0.56	0.54	20	0.3	Pseudouridylate synthase TRUI	Pseudouridylate synthase TRUI
XKR9	XKR8	373	395	9	3	0.33	3	0.33	97	54	0.56	0.41	24	0.0	XK-related protein 9 (hXKR9) [C	XK-related protein 8 (hXkr8) [Cl
PDSS1	PDSS2	415	399	42	23	0.55	42	1.00	53	29	0.55	0.70	23	0.0	All trans-polyprenyl-diphosphate	All trans-polyprenyl-diphosphate
JMJD6	JMJD4	403	417	39	22	0.56	32	0.82	39	20	0.51	0.63	117	1.0	Bifunctional arginine demethyla	2-oxoglutarate and iron-depend
RPUSD3	RPUSD4	343	377	12	6	0.50	6	0.50	22	11	0.50	0.50	60	0.7	Mitochondrial mRNA pseudouri	Pseudouridylate synthase RPUS
PTDSS1	PTDSS2	473	487	29	13	0.45	13	0.45	102	46	0.45	0.45	44	0.9	Phosphatidylserine synthase 1 (	Phosphatidylserine synthase 2 (
MEPCE	BCDIN3D	689	292	27	9	0.33	21	0.78	30	13	0.43	0.51	75	1.0	7SK snRNA methylphosphate c	RNA 5'-monophosphate methylt
POMT2	POMT1	750	725	72	38	0.53	48	0.67	12	5	0.42	0.54	50	0.0	Protein O-mannosyl-transferase	Protein O-mannosyl-transferase
RASL10A	RASL10B	203	203	17	4	0.24	8	0.47	22	9	0.41	0.37	30	0.0	Ras-like protein family member	Ras-like protein family member
PUS7	PUS7L	661	701	13	3	0.23	3	0.23	32	13	0.41	0.29	9	0.7	Pseudouridylate synthase 7 hor	Pseudouridylate synthase PUS7
OSGEPL1	OSGEP	414	335	40	13	0.33	31	0.78	42	17	0.40	0.50	8	0.0	tRNA N6-adenosine threonylcar	tRNA N6-adenosine threonylcar
HACD4	HACD3	232	362	42	12	0.29	20	0.48	38	14	0.37	0.38	114	0.0	Very-long-chain (3R)-3-hydroxy	Very-long-chain (3R)-3-hydroxy
AOC2	AOC3	756	763	41	8	0.20	6	0.15	35	12	0.34	0.23	403	0.7	Amine oxidase [copper-containi	Amine oxidase [copper-containi
ART1	ART5	327	291	43	17	0.40	14	0.33	30	10	0.33	0.35	33	0.0	GPI-linked NAD(P)(+)--arginine	Ecto-ADP-ribosyltransferase 5 (

Table 2: Top 15 protein known pairs sorted by pdsc_r score. The table shows the top 15 protein pairs ranked by their pdsc_r score, identifying known pairs with conserved functional residues that may exhibit the same enzymatic activity but likely act on different substrates. These pairs were selected based on shared EC numbers and a pdsc_r score exceeding the established threshold, while excluding channel proteins. The table includes columns for the gene names of both proteins, their respective lengths, the number of identified hotspots, hdsc value, hdsc_r value, hdelta value, hdelta_r value, the number of predicted pocket residues, pdsc and pdsc_r values, the mean hpd_r value, expression difference (exp_diff), the truth ratio from the RHEA analysis, and a description of each gene.

Among these, the pair comprising AOC2 (UniProt ID: O75106) and AOC3 (UniProt ID: Q16853) stood out due to their relatively high pdsc_r scores despite having lower hdelta_r and hdsc_r values but still above the respective thresholds, and for its high exp_diff value. Both AOC2 and AOC3 belong to the semicarbazide-sensitive amine oxidase (SSAO) family and display distinct functional and structural characteristics. AOC2 primarily catalyzes the oxidative deamination of larger aromatic monoamines such as 2-phenylethylamine, tryptamine, and p-tyramine, and is predominantly active in the human retina. In contrast, AOC3 functions both as an adhesion molecule and an amine oxidase, with a preference for

smaller amines like methylamine and benzylamine, and is widely expressed across various tissues, particularly in vascular contexts related to leukocyte adhesion and recruitment (Kaitaniemi et al., 2009).

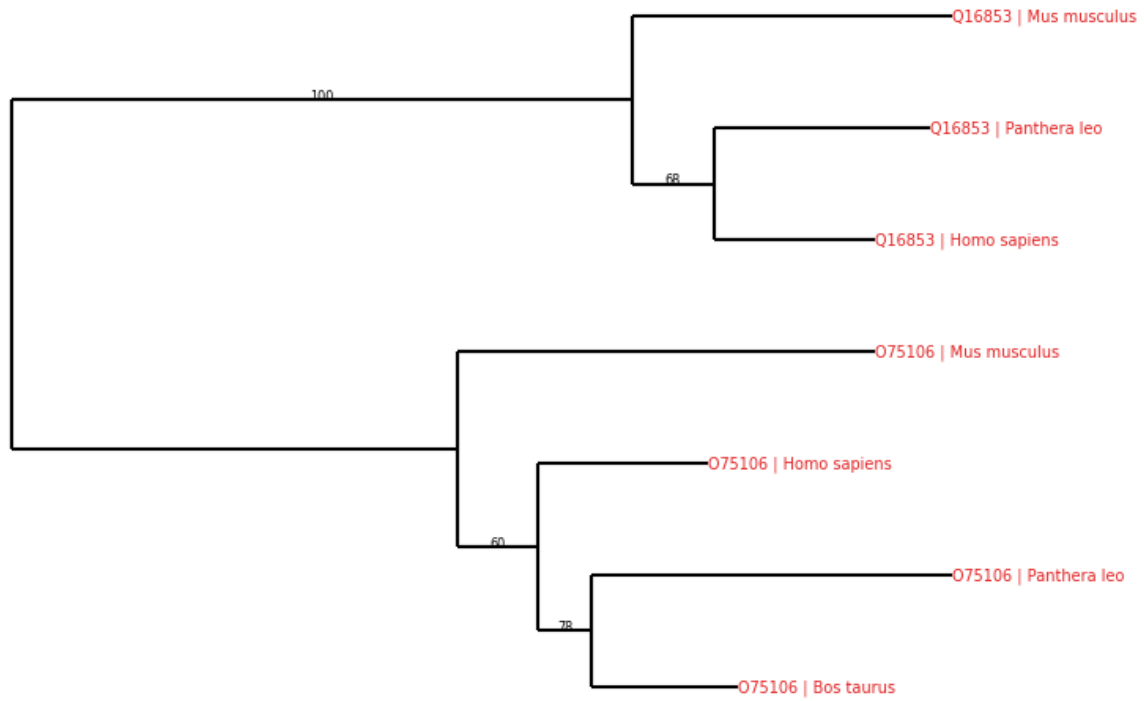


Figure 51: Phylogenetic tree generated with RAxML based on CLUSTALO cleaned alignments. The tree illustrates the evolutionary relationship of AOC2 and AOC3 orthologs across selected vertebrate species. The duplication event is present as one-to-one orthologous relationships exclusively in mammals, with *Bos taurus* missing for AOC3 because showing a one-to-many relationship with AOC3. The branches are labeled with Bootstrap support values.

The AOC2 and AOC3 pair using the main pipeline analysis or using the autonomous Google Colab pipeline, generated complete outputs that illustrate the functional divergence between these proteins. One of the initial outputs was a phylogenetic tree constructed with RAxML (**Fig. 51**), which included orthologs from multiple vertebrate species. This tree revealed that the gene duplication event leading to *AOC2* and *AOC3* is present with one-to-one orthologous relationships exclusively in mammals. Notably, in *Bos taurus*, the relationship is annotated as one-to-many with *AOC3*, resulting in its absence from the tree and potentially highlighting species-specific duplications and evolutionary divergence within this protein family.

Kaitaniemi et al. identified several key residues responsible for the differing substrate preferences between AOC2 and AOC3. In AOC2, residues Asn388 and Gly463 contribute to

a larger and more accessible active site, facilitating the accommodation of larger substrates such as aromatic amines. Conversely, AOC3 possesses residues Tyr394 and Leu469, which create a smaller active site better suited for smaller substrates like methylamine. Notably, Leu469 in AOC3 serves a dual function by restricting substrate access and stabilizing binding through hydrophobic interactions, whereas in AOC2, this residue is substituted by Gly463, allowing larger substrates to enter the active site (Kaitaniemi et al., 2009).

uni1	uni2	aln	pos1	pos2	res1	res2	iscs1	iscs2	xsc	tsc	dsc	p1	p2	h1	h2	delta	act	bin
O75106	Q16853	214	206	212	H	T	6.28	4.06	-1.89	2.82	5.94	0.65	0.24	0.33	0.16	0.50	F	F
O75106	Q16853	327	317	323	Y	F	6.75	6	3.1	5.28	2.9	0.34	NaN	0.10	0.12	0.33	F	F
O75106	Q16853	347	337	343	V	A	3.11	4	0.43	2.51	2.68	NaN	NaN	0.60	0.58	0.38	F	F
O75106	Q16853	390	380	386	D	D	6	6	6	6	0	0.53	0.20	0.52	0.59	0.21	T	F
O75106	Q16853	400	390	396	R	T	4.75	4.35	-1	2.7	5.35	0.13	0.39	0.46	0.18	0.66	F	F
O75106	Q16853	401	391	397	G	P	4.93	6.1	-1.84	3.06	6.78	NaN	0.30	0.05	0.05	0.48	F	F
O75106	Q16853	462	452	458	S	T	4	4.66	0.94	3.2	3.06	0.29	0.29	0.05	0.05	0.35	F	F
O75106	Q16853	463	453	459	A	V	3.44	4	-0.21	2.41	3.64	0.32	NaN	0.04	0.05	0.40	F	F
O75106	Q16853	471	461	467	S	T	3.81	5	1.13	3.31	2.69	0.53	0.49	0.67	0.62	0.38	F	F
O75106	Q16853	472	462	468	V	L	3.97	3.59	1.01	2.86	2.58	0.70	0.12	0.64	0.68	0.37	F	F
O75106	Q16853	473	463	469	G	L	6	3.77	-3.97	1.93	7.74	0.71	0.45	0.61	0.83	0.51	F	F
O75106	Q16853	475	465	471	Y	Y	7	7	7	7	0	0.33	NaN	0.91	0.93	0.18	T	F
O75106	Q16853	483	473	479	L	F	3.42	6	0.48	3.3	2.94	0.32	0.16	0.02	0.02	0.38	F	F
O75106	Q16853	500	490	496	N	S	6	4	1	3.67	3	0.32	0.19	0.54	0.52	0.41	F	F
O75106	Q16853	520	510	514	R	H	3.54	8	0.63	4.06	2.91	NaN	NaN	0.53	0.46	0.42	F	F
O75106	Q16853	521	511	515	V	T	4	4.33	0.06	2.8	3.94	NaN	NaN	0.49	0.52	0.36	F	F

Table 3: Alignment details and scores of AOC2 and AOC3. The table includes the indices of positions in the alignment and in the two proteins. For each aligned position in the two proteins, the consensus residue is shown using one-letter code. The alignment scores include ISCs, XSc, TSc, and DSc. Scores used for divergence assessment are also presented, with pocket and hotspot predictions indicated as p1, p2, h1, and h2 respectively. Additionally, the table shows the delta of the embeddings. Columns indicating UniProt features are provided, with active sites marked as act and binding sites as bin. Values above the respective thresholds are highlighted according to their significance: green for values indicating conservation or functional metrics, and red for values representing divergence.

Table 3 presents all residues with Dsc scores above the threshold or with at least one functional prediction (pdsc and hdsc) exceeding the threshold, alongside functional residues annotated in UniProt features columns act and bin (where F indicates False and T indicates True). Residues with functional predictions above the threshold are highlighted in green, while those with divergence scores (DSc and Delta) above the threshold are marked in red. Notably, the G463 residue showed a Dsc of 7.74, a pocket probability of 0.71, a hotspot probability of 0.61, and a delta of 0.51. This particular residue is located near residue 465, an annotated active site evolutionarily conserved, evinced by a TSc of 7.

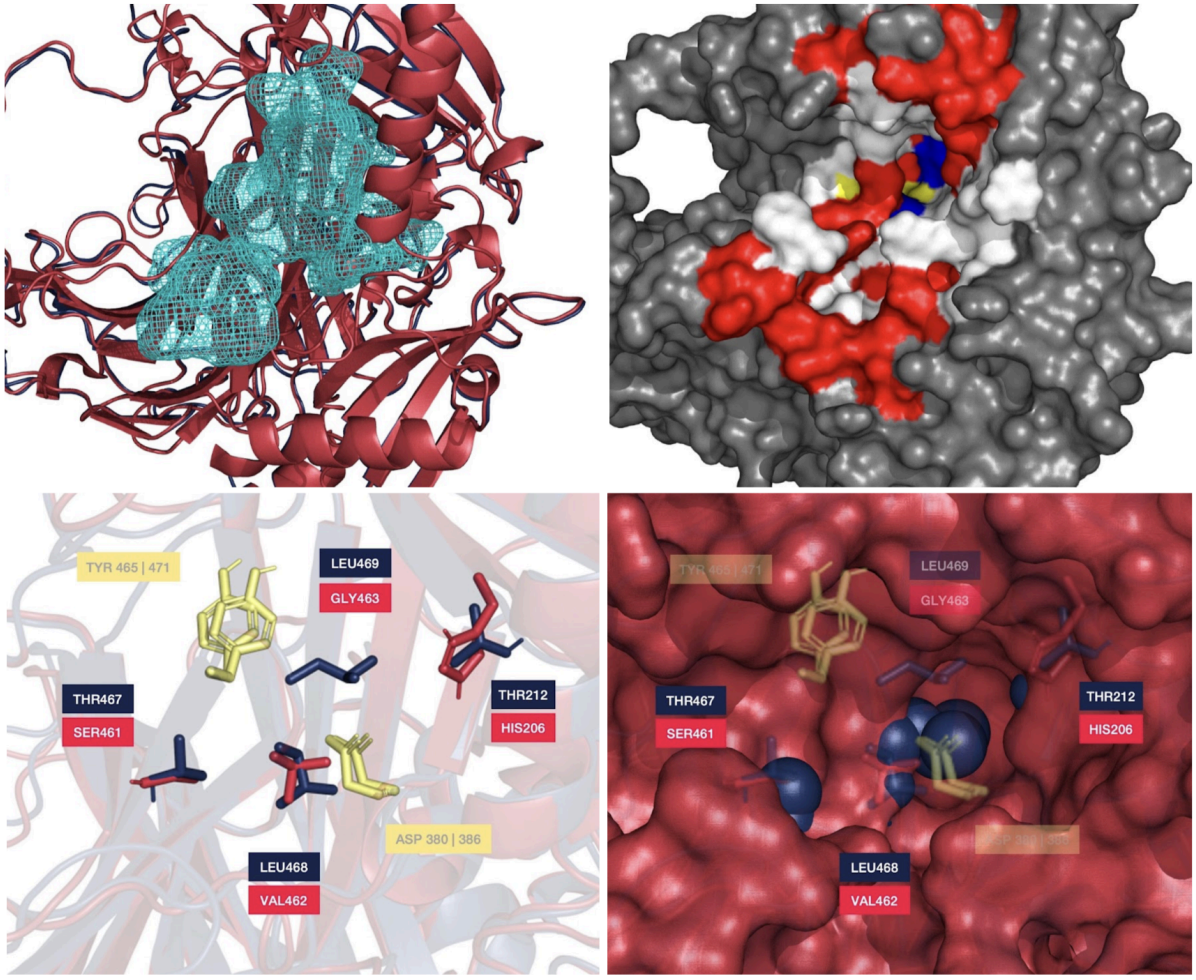


Figure 52: Four-panel structural output from the analysis pipeline. In the top-left panel, the aligned structures of AOC2 (red) and AOC3 (blue) are shown with the active site pocket highlighted as a blue grid. The top-right panel displays the same structures as surfaces, with active site residues in yellow, binding site residues in blue, pocket residues in white, and residues with high DSc in red. The bottom-left panel presents the structures in transparency, with functional and divergent residues highlighted as colored sticks labeled with their respective amino acids and position numbers (e.g., LEU469). The bottom-right panel mirrors the bottom-left visualization but includes surface overlays, clearly illustrating the hindrance introduced by the blue surface.

Furthermore, residues surrounding position 463 (462, 465, 473), which contribute to a smaller pocket, are less frequently recognized as part of the binding pocket compared to their counterparts, which are no longer identified as pocket residues in the corresponding protein. The analysis of AOC2 and AOC3 revealed several functional positions with high DSc, pinpointing key residues responsible for their differing substrate preference. Structural examination was conducted using AlphaFold-generated PDB files, which were automatically processed through the analysis pipeline. The protein structures were aligned using and

analyzed using PyMOL, and the relevant information derived from the analysis was incorporated into the structural files. In the visual representation (**Fig. 52**, top left panel), AOC2 is depicted in red and AOC3 in blue, with the active site pocket highlighted as a grid.

Residues within the pocket were color-coded based on their functional significance (top right panel): red indicates functional positions with DSc above the established cutoff, yellow denotes active site residues, and blue marks binding site residues. The visualizations show that the active site itself remains conserved, while the surrounding residues have undergone mutations, as shown in the sticks visualization (bottom left panel). These mutations introduce hindrances that alter substrate accessibility and specificity (bottom right panel). Specifically, the altered residues near the active site restrict the entry of larger substrates in AOC3.

7. Case Studies and Experimental Results

7.1 Case Study: AADACL2 Functional Divergence and Evolutionary Insights

We employed a custom ranking system to systematically prioritize protein pairs, with the aim of identifying cases where at least one protein in the pair was functionally characterized (**Table 4**).

The selection criteria were as follows: (1) at least one protein in the pair must have a RHEA annotation, while the other lacks such annotation, thereby excluding pairs previously used for defining score thresholds or those where both proteins are characterized; (2) at least one protein must possess an assigned Enzyme Commission (EC) number, while the other either lacks an EC number or has an incomplete designation, indicated by a '–' in the identification code; and (3) the expression difference (exp_diff) between the two proteins must be greater than 70. These criteria were designed to focus on pairs in which at least one partner is uncharacterized, thereby facilitating the discovery of its potential functional role.

Among these pairs, AADAC (P22760) and AADACL2 (Q6P093) stand out, particularly due to their high expression difference (exp_diff, yellow in the table) derived from tissue expression profiles. AADAC (Arylacetamide deacetylase) is annotated with the EC number 3.1.1.3, while AADACL2 (Arylacetamide deacetylase-like 2) has an incomplete EC number 3.1.1.-, suggesting a potential variation in their substrate specificity. Notably, only four other protein pairs within the top 30 ranked pairs exhibit this pattern of EC numbers before AADAC and AADACL2. Alphafold structures and UniProt features showed that AADAC and AADACL2

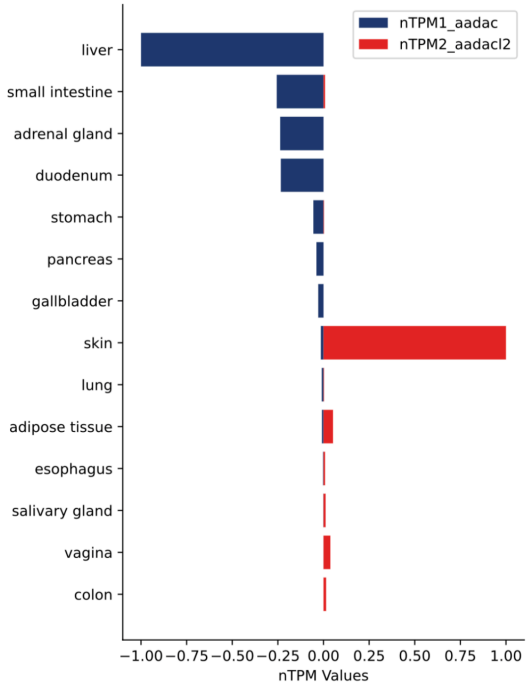
share a common α/β-hydrolase fold and a catalytic triad essential for hydrolysis. The catalytic triad in AADAC is composed of Ser189, Asp343, His373; we have identified the conserved catalytic triad in AADACL2 as well (Ser189, Asp343, His371, **Table 5**). In addition to the catalytic triad, these enzymes also share residues responsible for forming the oxyanion hole, which stabilizes the transition state during hydrolysis. In AADAC, the oxyanion hole is formed by His111, Gly112, and Gly113, and we observe the same residues forming the oxyanion hole in AADACL2 (His111, Gly112, and Gly113).

gene1	gene2	len1	len2	h	hdsc	hdsc_r	hdelta	hdelta_r	p	pdsc	pdsc_r	hpd_r	prob	pval	ediff	name1	name2
RTCA	RCL1	366	373	27	18	0.67	24	0.89	25	17	0.68	0.75	1.00	0.00	216	RNA 3'-terminal phosphate (	RNA 3'-terminal phosphate (
PPT1	PPT2	306	302	51	30	0.59	44	0.86	18	9	0.50	0.65	1.00	0.00	160	Palmitoyl-protein thioesteras	Lysosomal thioesterase PPT
PHYH	PHYHD1	338	291	31	14	0.45	27	0.87	32	17	0.53	0.62	1.00	0.00	571	Phytanoyl-CoA dioxygenase	Phytanoyl-CoA dioxygenase
PRXL2B	PRXL2A	198	229	37	24	0.65	23	0.62	11	6	0.55	0.61	1.00	0.00	383	Prostamide/prostaglandin F	Peroxiredoxin-like 2A (Perox
HHAT	HHATL	493	504	113	60	0.53	83	0.73	91	46	0.51	0.59	1.00	0.00	500	Protein-cysteine N-palmitoyl	Protein-cysteine N-palmitoyl
DHRS7C	DHRS7B	311	325	42	21	0.50	35	0.83	31	13	0.42	0.58	1.00	0.00	316	Dehydrogenase/reductase S	Dehydrogenase/reductase S
ASRGL1	TASP1	308	420	25	12	0.48	21	0.84	13	4	0.31	0.54	1.00	0.00	121	Isoaspartyl peptidase/L-aspa	Threonine aspartase 1 (Tasp
ADPRH	ADPRHL1	357	1967	29	10	0.34	24	0.83	18	7	0.39	0.52	1.00	0.00	274	ADP-ribosylhydrolase ARH1	Inactive ADP-ribosyltransfer
YJEFN3	NAXE	299	288	23	10	0.43	14	0.61	29	12	0.41	0.49	0.99	0.00	132	YjeF N-terminal domain-con	NAD(P)H-hydrate epimerase
LHPP	HDHD2	270	259	34	13	0.38	21	0.62	17	6	0.35	0.45	0.99	0.00	337	Phospholysine phosphohisti	Haloacid dehalogenase-like
EBP	EBPL	230	206	43	17	0.40	22	0.51	33	13	0.39	0.43	0.98	0.00	222	3-beta-hydroxysteroid-Delta	Emopamil-binding protein-lik
TSTD3	TSTD1	97	115	8	3	0.38	7	0.88	0	0	0.00	0.42	0.98	0.02	169	Thiosulfate sulfurtransferase	Thiosulfate:glutathione sulfu
EPHX4	EPHX3	362	360	47	15	0.32	22	0.47	33	15	0.45	0.41	0.98	0.00	74	Epoxide hydrolase 4 (EC 3.3	Epoxide hydrolase 3 (EH3) (
IRAK2	IRAK1	625	712	34	11	0.32	23	0.68	26	5	0.19	0.40	0.97	0.00	81	Interleukin-1 receptor-assoc	Interleukin-1 receptor-assoc
TECRL	TECR	363	308	73	23	0.32	34	0.47	37	15	0.41	0.40	0.97	0.00	458	Trans-2,3-enoyl-CoA reducta	Very-long-chain enoyl-CoA r
A4GALT	A4GNT	353	340	37	11	0.30	17	0.46	27	8	0.30	0.35	0.94	0.00	75	Lactosylceramide 4-alpha-ga	Alpha-1,4-N-acetylglucosam
SUMF2	SUMF1	301	374	23	10	0.43	3	0.13	15	7	0.47	0.34	0.94	0.00	75	Inactive C-alpha-formylglycin	Formylglycine-generating en
FASTK	FASTKD3	549	662	7	6	0.86	1	0.14	0	0	0.00	0.33	0.93	0.00	113	Fas-activated serine/threonin	FAST kinase domain-contain
AADAC	AADACL2	399	401	43	8	0.19	18	0.42	37	12	0.32	0.31	0.90	0.01	917	Arylacetamide deacetylase (	Arylacetamide deacetylase-l
HSD11B1	HSD11B1L	292	286	54	5	0.09	38	0.70	33	4	0.12	0.31	0.90	0.03	933	11-beta-hydroxysteroid dehy	Hydroxysteroid 11-beta-dehy
SMPDL3A	SMPDL3B	453	455	26	6	0.23	7	0.27	25	10	0.40	0.30	0.89	0.00	77	Cyclic GMP-AMP phosphodi	Acid sphingomyelinase-like p
ABHD14B	ABHD14A	210	271	37	11	0.30	11	0.30	7	2	0.29	0.29	0.88	0.09	127	Putative protein-lysine deac	Protein ABHD14A (EC 3.-.-.-
ODC1	AZIN2	461	460	48	8	0.17	26	0.54	31	5	0.16	0.29	0.87	0.00	157	Ornithine decarboxylase (OI	Antizyme inhibitor 2 (Azl2) (
KLB	KL	1044	1012	7	2	0.29	2	0.29	12	3	0.25	0.27	0.84	0.00	131	Beta-klotho (BKL) (BetaKlot	Klotho (EC 3.2.1.31) [Cleave
TLCD3B	TLCD3A	274	257	43	12	0.28	11	0.26	60	17	0.28	0.27	0.84	0.10	90	Ceramide synthase (EC 2.3.1	TLC domain-containing prot
HSDL1	HSD17B12	330	312	51	14	0.27	20	0.39	34	4	0.12	0.26	0.81	0.14	115	Inactive hydroxysteroid dehy	Very-long-chain 3-oxoacyl-C
ENPP5	ENPP4	477	453	33	7	0.21	9	0.27	18	5	0.28	0.25	0.80	0.09	227	Ectonucleotide pyrophospha	Bis(5'-adenosyl)-triphosphat
ACO1	IREB2	889	963	43	7	0.16	17	0.40	30	6	0.20	0.25	0.79	0.03	108	Cytoplasmic aconitate hydra	Iron-responsive element-bin
LIPA	LIPM	399	423	70	12	0.17	21	0.30	35	8	0.23	0.23	0.74	0.34	388	Lysosomal acid lipase/chole	Lipase member M (EC 3.1.1
ST3GAL4	ST3GAL6	333	331	36	8	0.22	12	0.33	26	3	0.12	0.22	0.71	0.48	108	CMP-N-acetylneuraminate-b	Type 2 lactosamine alpha-2,

Table 4: Top 30 protein pairs with at least one known protein, sorted by hpd_r score. This table presents the top 30 protein pairs ranked according to their hpd_r score, focusing on pairs where at least one protein is known. The table includes columns for the gene names of both proteins, their respective lengths, the number of identified hotspots, hdsc value, hdsc_r value, hdelta value, hdelta_r value, the number of predicted pocket residues, pdsc and pdsc_r values, the mean hpd_r value, expression difference (exp_diff), the truth ratio from the RHEA analysis, and a description for each protein.

Further analysis of the tissue expression profiles generated using HPA data reveals a distinctive pattern (**Fig. 53**). *AADAC* is predominantly expressed in the liver, whereas *AADACL2* shows almost exclusive expression in the skin. Supporting this observation, HPA clustering assigns AADAC to cluster 83, labeled “Liver - Metabolism”, and AADACL2 to cluster 13, labeled “Skin - Cornification”. Among the top 15 neighboring genes in the AADACL2 cluster are proteins such as LCE1E (Late Cornified Envelope 1E), various keratins (KRT35, KRT28, KRT82), and filaggrin.

Several enzymes of our set have been identified in the ceramide processing pathway. ALOX12B and ALOXE3 are also parts of our duplicated protein pairs set with the following scores: probability of 0.96, p-value of 0.0, hdsc_r 0.17, pdsc_r 0.04, and hdelta_r 0.29. EPHX3 and EPHX4 also appear with these scores: probability of 0.98, p-value of 0.0, hdsc_r 0.32, pdsc_r 0.45, and hdelta_r 0.47. Similarly, TGM1 and F13A1 have these scores: probability=0.61, p-value=0.93, hdsc_r=0.21, pdsc_r=0.20, hdelta_r=0.17.



Neighbour1	Des1	Clu1	Neighbour2	Des2	Clu2
LCE1E	Late cornified envelope 1E	13	CYP3A4	Cytochrome P450 family 3 subfamily A member 4	83
PSORS1C2	Psoriasis susceptibility 1 candidate 2	13	APOB	Apolipoprotein B	83
KRTAP26-1	Keratin associated protein 26-1	13	OTC	Ornithine transcarbamylase	83
KRT35	Keratin 35	13	APOA1	Apolipoprotein A1	83
LCE1B	Late cornified envelope 1B	13	SMLR1	Small leucine rich protein 1	83
THEM5	Thioesterase superfamily member 5	13	CPS1	Carbamoyl-phosphate synthase 1	83
KRT77	Keratin 77	13	SULT2A1	Sulfotransferase family 2A member 1	83
LCE2C	Late cornified envelope 2C	13	TMEM220	Transmembrane protein 220	83
KRTAP17-1	Keratin associated protein 17-1	13	CYP2C19	Cytochrome P450 family 2 subfamily C member 19	83
LCE1A	Late cornified envelope 1A	13	NAT2	N-acetyltransferase 2	83
CASP14	Caspase 14	13	KDM8	Lysine demethylase 8	83
LCE5A	Late cornified envelope 5A	13	A1CF	APOBEC1 complementation factor	83
KRT28	Keratin 28	13	SLC2A2	Solute carrier family 2 member 2	83
KRT82	Keratin 82	13	GLYCTK	Glycerate kinase	83
FLG	Filaggrin	13	SLC41A2	Solute carrier family 41 member 2	83

Figure 53. Tissue expression analysis of *AADAC* and *AADACL2*. The left panel displays normalized tissue expression values for *AADAC* (blue) and *AADACL2* (red), obtained from HPA data. Expression levels were normalized for each protein by dividing by their maximum expression value. *AADAC* exhibits predominant expression in the liver, while *AADACL2* is almost exclusively expressed in the skin. The right panel presents a table summarizing the HPA clusters for the two proteins. The first three columns pertain to *AADACL2*, and the last three to *AADAC*. Each cluster is listed along with the names and descriptions of the top 15 neighboring genes, sorted by correlation as calculated by the HPA.

Upon closer examination of our dataset, we identified another pair of proteins from the same family that exceeded our thresholds and were included in the pool of 1,112 pairs: AADACL3 (Q5VUY0) and AADACL4 (Q5VUY2) with probability = 0.897, p-value = 0.004, hdsc_r = 0.15, pdsc_r = 0.23, hdelta_r = 0.54. While this section primarily focuses on AADAC and AADACL2, we include all four family members in the RAxML phylogenetic tree presented in **Figure 54**. In this tree, mammals are colored in red, sauropsids in green, and fish in blue. Notably, all four proteins have orthologs across these three major vertebrate groups. However, an intriguing pattern emerges: whereas AADACL3 and AADACL4 have orthologs in aquatic mammals (e.g., *Tursiops truncatus*), AADAC ad AADACL2 do not, although they remain duplicated in other terrestrial mammals.

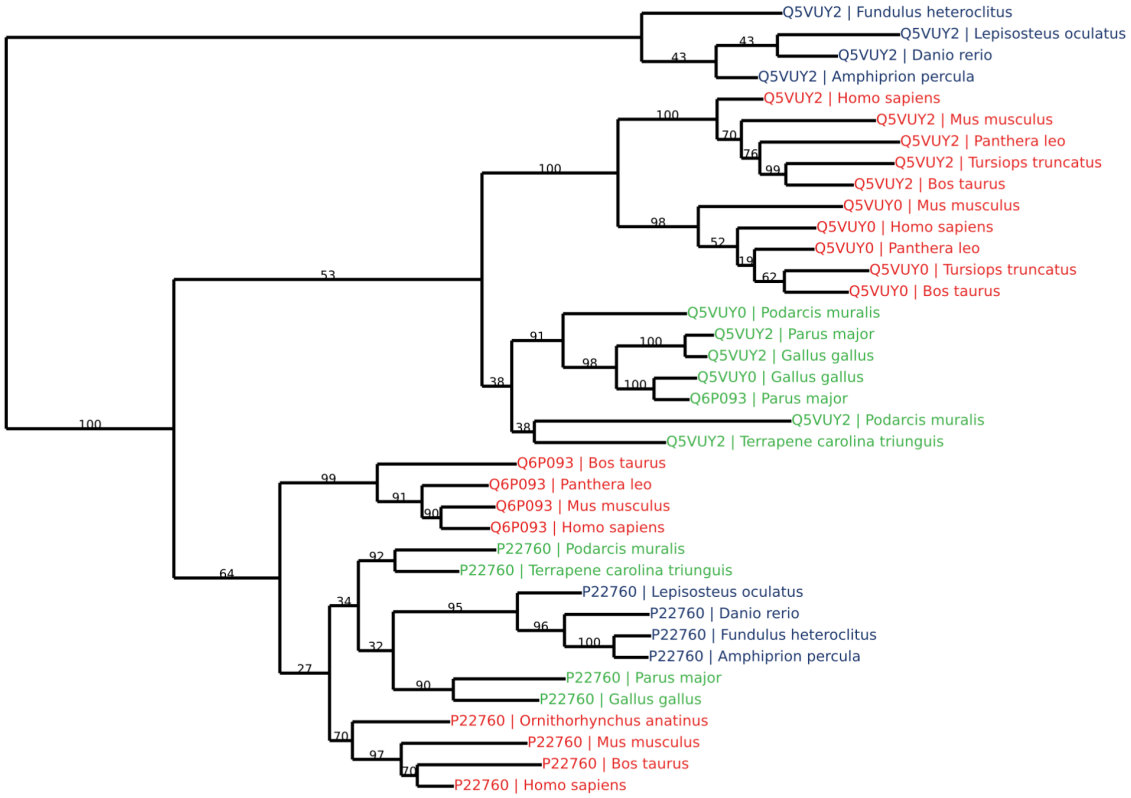


Figure 54: Phylogenetic tree of AADAC family proteins. This phylogenetic tree was generated using RAxML based on cleaned CLUSTALO alignments and illustrates evolutionary relationships among COMPARA orthologs of AADAC (P22760), AADACL2 (Q6P093), AADACL3 (Q5VUY0), and AADACL4 (Q5VUY2) across selected vertebrate species. The tree's leaves are color-coded according to vertebrate classes: red for mammals, green for sauropsids, and blue for fishes. Each leaf is labeled with the UniProt identifier of the human protein and the species name. Bootstrap support values are indicated on the branches. The tree reveals that all four genes are present in the three vertebrate classes; however, AADAC and AADACL2 are absent in the dolphin (*Tursiops truncatus*).

The alignment summary table (**Table 5**), which includes alignment scores, functional probability scores, and UniProt features, revealed several mutations with high Differential Conservation Scores (DSc) around the active site and predicted pocket residues. Pocket predictions between residues 200 and 270 show discrepancies between AADAC and AADACL2, with predictions in AADAC falling below the threshold level, suggesting a reduction in the pocket cavity size. Although no mutations were identified within the oxyanion hole region, a notable mutation from tryptophan (W) to phenylalanine (F) was observed at residue 115, adjacent to the oxyanion hole. While W115 is not predicted to be part of the pocket, F115 is, suggesting that the bulkier side chain of W115 may contribute to the narrower pocket observed in AADAC. To validate these structural differences, AlphaFold-predicted structures of AADAC and AADACL2 were aligned and analyzed using PyMOL. The structural alignment and surface visualization (**Figure 55**, bottom panels) revealed an elongated pocket in AADACL2, as clearly shown in the side view of the pockets in the bottom left panel.

uni1	uni2	aln	pos1	pos2	res1	res2	iscs1	iscs2	xsc	tsc	dsc	p1	p2	h1	h2	delta	act	bin
P22760	Q6P093	46	40	40	L	A	2.23	2.85	-1.29	1.27	3.52	0.71	0.92	0.31	0.40	0.48	F	F
P22760	Q6P093	49	43	43	I	C	1.38	5.01	-0.87	1.84	2.25	NaN	0.74	0.54	0.39	0.80	F	F
P22760	Q6P093	79	66	66	G	F	1.31	2.3	-1.01	0.86	2.32	0.33	1.00	0.23	0.54	0.62	F	F
P22760	Q6P093	130	115	115	W	F	11	5.64	1.06	5.9	4.57	NaN	0.86	0.85	0.84	0.62	F	F
P22760	Q6P093	132	117	117	V	F	2.98	4.82	0.27	2.69	2.71	0.92	1.00	0.79	0.85	0.46	F	F
P22760	Q6P093	137	122	122	L	Q	2.16	3.59	-0.63	1.71	2.79	0.71	0.81	0.29	0.19	0.55	F	F
P22760	Q6P093	204	189	189	S	S	4	4	4	4	0	0.84	0.99	0.95	0.95	0.28	T	F
P22760	Q6P093	205	190	190	A	S	3.89	3.74	0.99	2.87	2.75	NaN	0.92	0.92	0.95	0.64	F	F
P22760	Q6P093	237	221	221	A	G	1.89	2.11	-0.91	1.03	2.8	NaN	0.72	0.42	0.64	0.59	F	F
P22760	Q6P093	240	224	224	P	I	1.19	2.75	-0.95	0.99	2.14	NaN	0.43	0.12	0.40	0.62	F	F
P22760	Q6P093	262	246	246	M	A	4.59	2.78	-1.19	2.06	3.97	NaN	0.78	0.47	0.37	0.63	F	F
P22760	Q6P093	265	249	249	F	L	5.66	4	0.1	3.25	3.9	0.86	0.98	0.78	0.74	0.52	F	F
P22760	Q6P093	266	250	250	W	V	10.38	2.87	-2.8	3.48	5.67	NaN	0.84	0.62	0.65	0.62	F	F
P22760	Q6P093	368	343	341	D	D	6	6	6	6	0	NaN	NaN	0.86	0.80	0.11	T	F
P22760	Q6P093	398	373	371	H	H	8	8	8	8	0	0.90	0.98	0.93	0.91	0.28	T	F

Table 5: Alignment details and scores of AADAC and AADACL2. The table includes the indices of positions in the alignment and in the two proteins. For each aligned position in the two proteins, the consensus residue is shown using one-letter code. The alignment scores include ISCs, XSc, TSc, and DSc. Scores used for divergence assessment are also presented, with pocket and hotspot predictions indicated as p1, p2, h1, and h2 respectively. Additionally, the table shows the delta of the embeddings. Columns indicating UniProt features are provided, with active sites marked as act and binding sites as bin. Values above the respective thresholds are highlighted according to their significance: green for values indicating conservation or functional metrics, and red for values representing divergence.

In our investigation of the newly identified lipase AADACL2, we focused on selecting a substrate that accurately reflects its role within the ceramide metabolism pathway. Based on the established biochemical pathways of ceramide processing, particularly the P-O ceramide

pathway, we hypothesized that the lipase acts on epoxy-hydroxy ceramides. Given that epoxy-hydroxy ceramides represent the key intermediate in this process, we chose them as the substrate for our docking studies. In the top panel of **Figure 56**, we present the 3D structure of the epoxy-hydroxy ceramide molecule used for docking.

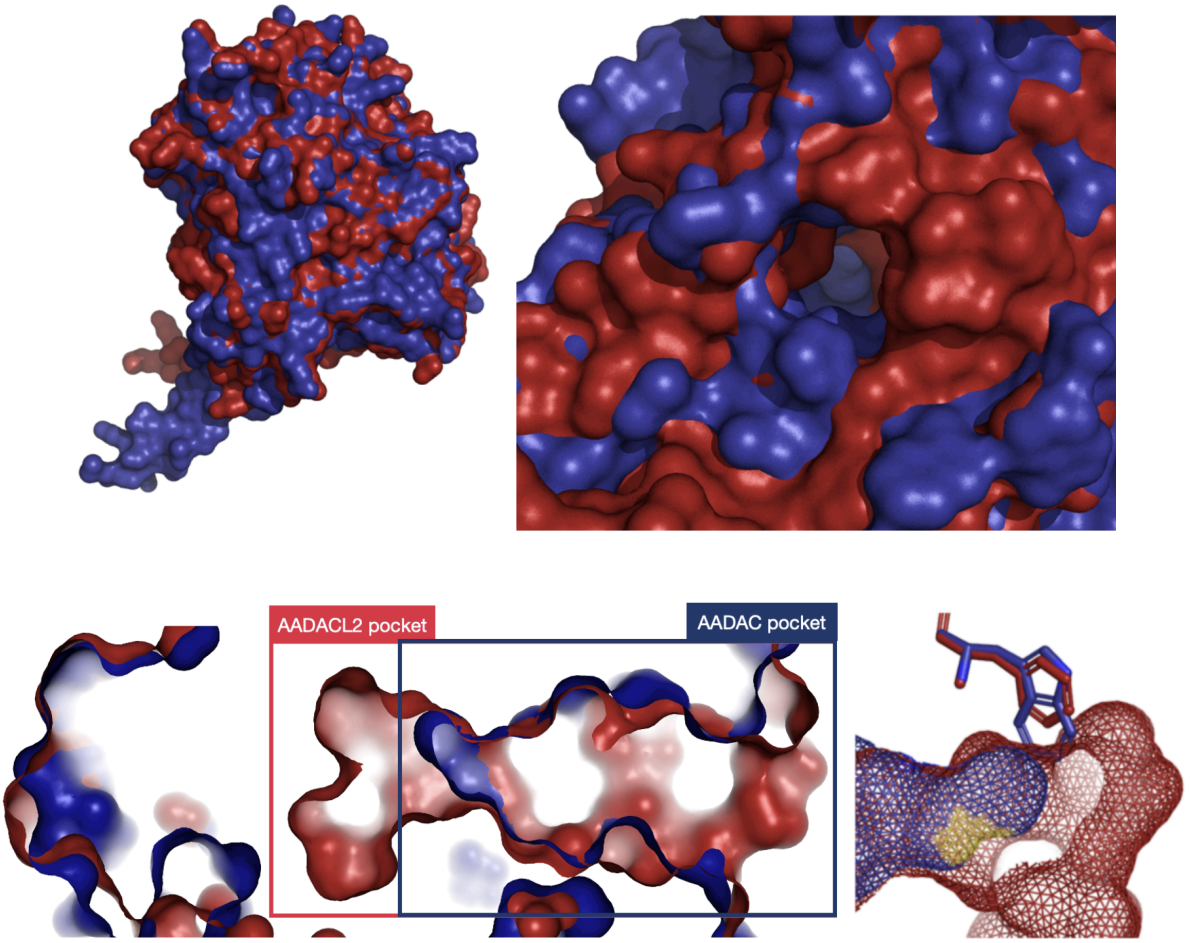


Figure 55: Structural comparison of AADAC and AADACL2 AlphaFold models. The figure presents the aligned AlphaFold-predicted structures of AADAC and AADACL2. The top left panel displays the overall structures aligned as surfaces, with AADAC colored in blue and AADACL2 in red. The top right panel provides a zoomed-in view of the functional pockets within the aligned structures. The bottom left panel shows the grid surfaces of the pockets from a side perspective, highlighting the elongation of the pocket in AADACL2. The bottom right panel focuses on the narrowing region, illustrating the W115 residue in AADAC and the corresponding F115 residue in AADACL2 as stick models. Additionally, serine S189 of the catalytic triad is depicted as yellow, emphasizing its proximity to the mutation site. This structural analysis underscores the potential for AADACL2 to interact with larger lipid substrates due to the expanded pocket.

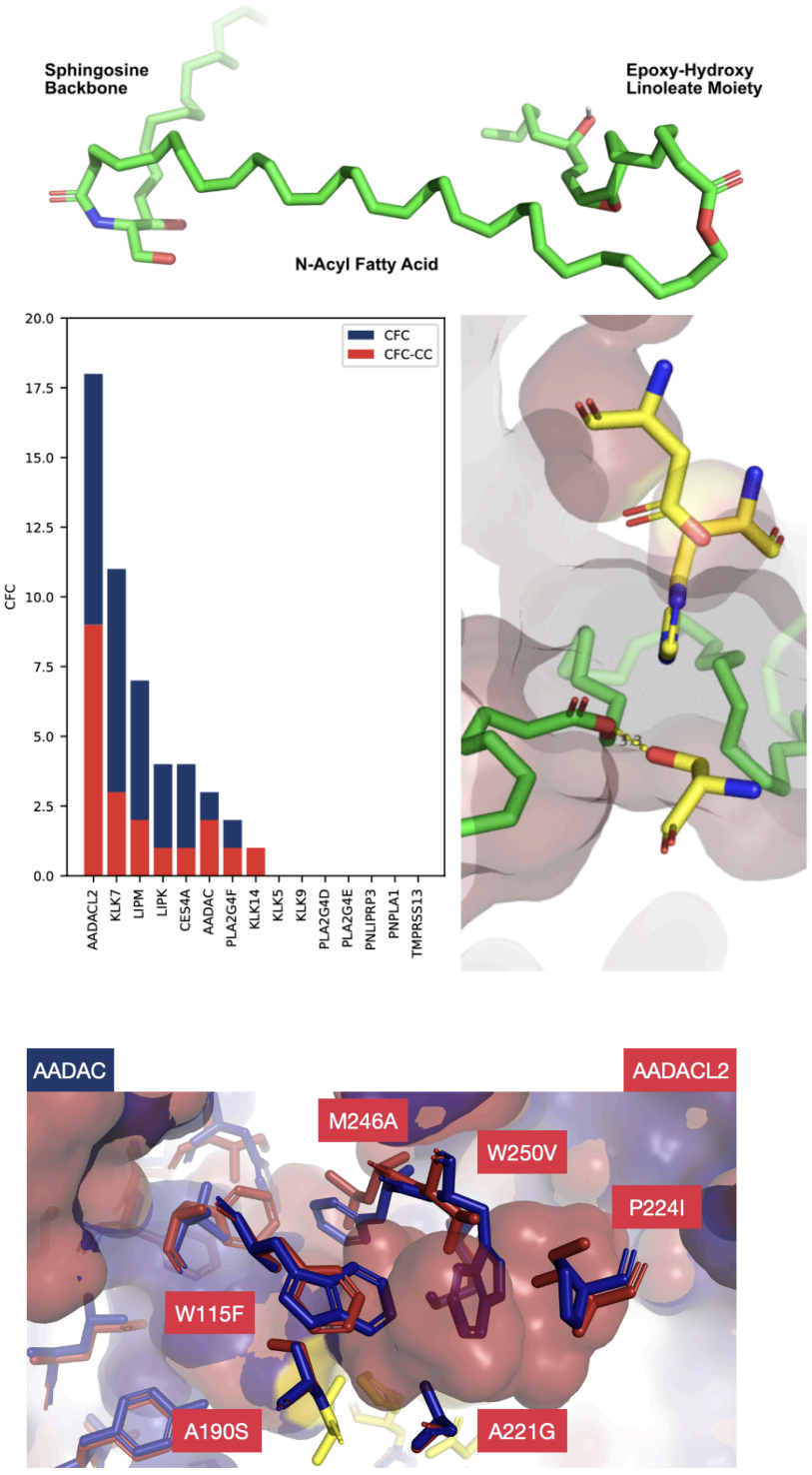


Figure 56: Docking analysis of epoxy-hydroxy ceramide with AADACL2. ***Top panel*****: displays the 3D structure of the epoxy-hydroxy ceramide molecule used for docking with labels**. *Central left panel*: shows the docking results for AADAC, AADACL2, and other proteins in the HPA "Skin - cornification" cluster with the catalytic serine of the catalytic triad. CFC refers to the most favorable conformations, whereas CC-CFC refers to the conformations in which the catalytic triad serine is within 4 angstroms from the carbonyl carbon of the substrate. *Central right panel*: illustrates the

docking-generate structure, highlighting the most favorable conformation where the catalytic serine (yellow) is positioned 3.3 angstroms from the substrate's carbonyl group (green). *Bottom panel*: presents the AADACL2 structure in a transparent surface view, showing the binding pocket. The catalytic triad residues are shown as yellow sticks, while functional residues with high DSc (hdsc) are in blue-red.

The molecule comprises a sphingoid base containing an amino group, a hydroxyl group, and a long hydrocarbon chain; an N-acyl fatty acid ranging from 30 to 36 carbons linked to the sphingoid base via an amide bond (30 in our case), which includes an omega-hydroxyl group (Akiyama, 2021). This omega-hydroxyl group is utilized for covalent attachment to proteins in the CLE (Nemes et al., 1999), remaining esterified with linoleic acid in the context of epoxy-hydroxy ceramides. The linoleic acid is esterified to the omega-hydroxyl group of the fatty acid, which has been oxidized by ALOX12 and ALOXE3, and contains two key reactive groups: an epoxide group, increasing its activity, and a hydroxyl group.

In the central left panel of **Figure 56**, we display the docking results for AADAC, AADACL2, and other proteins within the HPA "Skin - cornification" cluster that possesses the catalytic serine of the catalytic triad. The results indicate that nine CFC clusters exhibited favorable conformations, and nine CC-CFC clusters showed the catalytic serine within 4 angstroms of the carbonyl carbon of the substrate (3 and 2 clusters, respectively, for AADAC). The second highest number of CC-CFC clusters was observed in KLK7, with three such clusters. The central right panel shows the structures generated through docking, highlighting the most favorable conformation where the catalytic serine (depicted in yellow) is positioned 3.3 angstroms from the carbonyl carbon of the substrate (shown in green). The bottom panel presents the structure of AADACL2 in a surface representation with transparency, illustrating the entirety of the binding pocket. Residues comprising the catalytic triad are depicted as yellow sticks, while residues classified as functional and exhibiting high DSc (hdsc) are shown in blue-red. Notably, mutations are distributed along the entire length of the pocket, effectively outlining its contour.

7.2. AADACL2 Induction and Expression in *E. coli*

We conducted induction assays following the growth of *E. coli* BL21-CodonPlus cells transformed with the pET-28a(+)-TEV vector containing the coding sequence (CDS) of *Homo sapiens* arylacetamide deacetylase like 2 (*AADACL2*) (NM_207365.4:110-1315), which had been further mutagenized to exclude 18 residues from the N-terminus. The transformed

colonies were grown at 37°C and induced with 0.3 mM IPTG then the OD_{600} reached 0.6, with harvesting performed once the OD_{600} approached 6. Induction was analyzed using SDS-PAGE by comparing whole cells lysed with SDS from both induced and non-induced samples (**Fig. 57**). From left to right, the lanes included two non-induced colonies, two induced colonies, an induction test conducted at 20°C, another induced colony, and a molecular weight marker. As depicted in the figure, the induction band is clearly identifiable, marked by a red arrow at approximately 46.7 kDa (representing the 405 residues including the His tag and the TEV cleavage site, approximately 2.5 kDa).

Following the induction assays, we performed lysis experiments to assess the solubility of the expressed AADACL2 protein. After sonication, expression profiles were analyzed using SDS-PAGE. As shown in **Figure 58** (left panel), which displays the pellet, whole cell lysate, and supernatant from Colony 4 and Colony 6 (lanes 4, 5, 6 and 8, 9, 10 respectively), all of the AADACL2 protein remained in the pellet fraction, indicating poor solubility. To address this, we implemented a renaturation protocol directly in the column, creating a denaturation gradient with 6 M urea and a renaturation gradient (detailed in the Methods section). This approach yielded a minimal amount of soluble protein (**Fig. 58**, right panel). However, subsequent activity assays using the PNPA assay did not detect any enzymatic activity, suggesting that the renatured protein was either inactive or present in insufficient quantities.

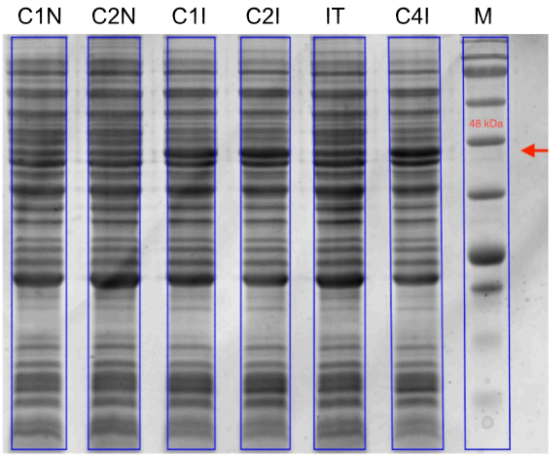


Figure 57: Induction test of *E. coli* BL21-CodonPlus colonies transformed with pET-28a(+)-TEV vector containing *AADACL2*. SDS-PAGE gel analysis showing the inducted (I) and non-inducted (N) colonies from 1 (C1) to 4 (C4). From left to right, the lanes include: C1N, C2N, C1I, C2I, induction test conducted at 20°C, C4I, Marker (Tris-Glycine). The red arrow points to the inducted protein band at approximately 47 kDa, located just below the 48 kDa marker band. Each culture (10 mL at OD_{600} = 3.5) was pelleted and resuspended in 500 µL of buffer. For each lane, 1 µL of the final lysate (~1/500 of the total) was loaded, corresponding to ~7×10^7 cells per lane.

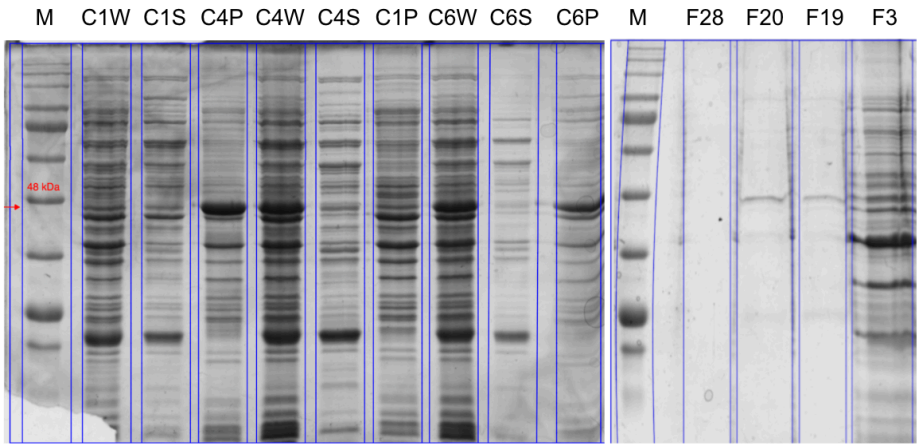


Figure 58: Solubility and purification analysis of AADACL2 from *E. coli* BL21-CodonPlus colonies transformed with pET-28a(+)-TEV-AADACL2. *Left panel*: SDS-PAGE analysis of solubility tests showing the distribution of AADACL2 protein in pellet (P), whole cell lysate (W), and supernatant (S) fractions from Colony 1 (C1) and Colony 4 (C4). The lanes are as follows: Marker (Tris-Glycine), C1W, C1S, C4P, C4W, C4S, C1P, C6W, C6S, C6P. The red arrow points to the inducted protein band at approximately 47 kDa, located just below the 48 kDa marker band. Each culture (10 mL at $OD_{600} = 4$) was pelleted and resuspended in 500 µL of buffer. For each fraction, 1 µL of lysate was loaded (~8×10^7 cells per lane). *Right panel*: SDS-PAGE analysis of the purification process for renatured AADACL2 using affinity chromatography. The lanes included: Marker (Tris-Glycine), F28 (Fraction 28), F20 (Fraction 20), F19 (Fraction 19), F3 (Fraction 3). For purified fractions, 5 µL of samples were loaded per lane corresponding to ~0.2 µg loaded for Fraction 20.

7.3. AADACL2 Induction and Activity Assay in *Pichia pastoris*

To transform *Pichia pastoris*, the pPICZ A vector containing the mutagenized *AADACL2* gene was propagated in *Escherichia coli* XL1-Blue cells. Plasmid extraction and purification from *E. coli* cells were performed using a miniprep procedure. Once a sufficient quantity of purified plasmid was obtained, it was digested with the restriction enzyme SAC I. Successful digestion was verified by agarose gel electrophoresis, as shown in **Figure 59** (left panel), where undigested plasmid (lanes 2 and 4) displayed bands at approximately 6 kb, and digested plasmid (lanes 3 and 5) showed a shift to approximately 3.5 kb, corresponding to the linearized plasmid size. The shift occurs due to the linearization of the plasmid from its circular form. Following confirmation of correct digestion, the concentration of the digested DNA was quantified using a Nanodrop spectrophotometer (**Fig. 59**, right panel).

The absorbance curves between 200 and 340 nm exhibited a peak at 260 nm with an average A260/A280 ratio of 1.76 and an A260/A230 ratio of 1.54, yielding approximately 30 µg of DNA from 24 mL of culture–sufficient for four transformation attempts. After electroporation of the digested DNA into *Pichia pastoris* GS115 cells, individual colonies were patched onto MDH and MMH agar plates to verify the MutS phenotype, as we transformed GS115 cells which are expected to exhibit MutS characteristics, allowing growth on both methanol-containing and methanol-free media (detailed in the Methods section). All colonies successfully grew on both types of media, confirming the MutS phenotype.

Upon obtaining *Pichia pastoris* transformants that successfully grew on zeocin and methanol-containing media, we proceeded to assess the expression, solubility, and activity of the AADACL2 protein. As depicted in **Figure 60**, SDS-PAGE analysis of induced and non-induced whole cell lysates revealed two distinct protein bands: one at approximately 47 kDa, corresponding to the expected size of AADACL2 (411 residues including the His tag and c-myc epitope, ~3.25 kDa), and another higher band at approximately 70 kDa. We hypothesize that the higher band represents a hyperglycosylated form of AADACL2, as glycosylation adds molecular weight, causing slower migration on SDS-PAGE. The hyperglycosylated version appears to be more abundant than the correctly sized protein.

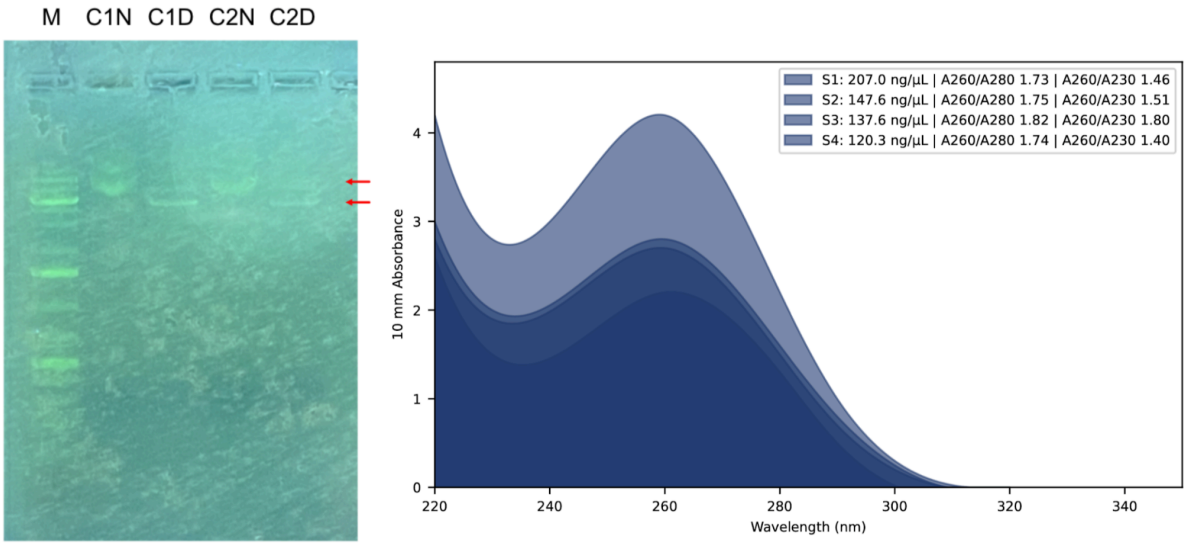


Figure 59: Agarose gel verification and DNA quantification of SAC I digestion of pPICZ A-AADACL2 plasmid. *Left panel*: agarose gel electrophoresis demonstrating the digestion of the pPICZ A-AADACL2 plasmid with SAC I. The lanes are as follows: marker, C1N (colony 1 undigested), C1D (colony 1 digested), C2N (colony 2 undigested), C2D (colony 2 digested). The digested samples (C1D and C2D) show a band shift from approximately 6 kb to ~3.5 kb. *Right panel*: absorbance curves obtained from Nanodrop quantification of digested plasmid DNA. The X-axis represents wavelength in

nm (220-340 nm), and the Y-axis represents absorbance (Abs). The legend indicates the concentration of individual samples along with their A260/A280 and A260/A230 ratios.

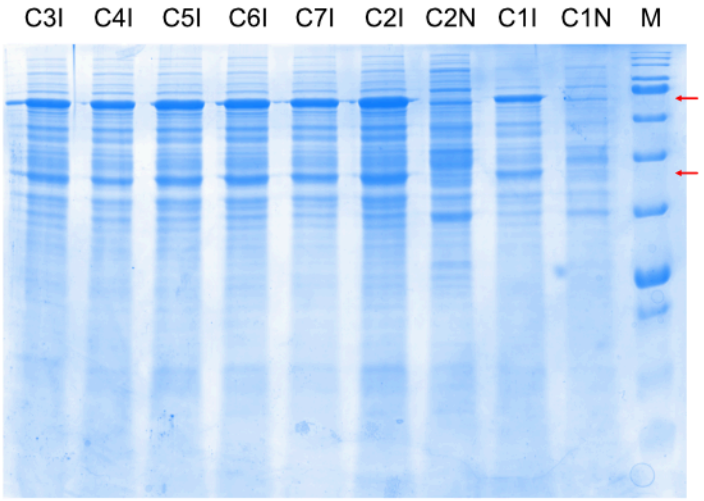


Figure 60: SDS-PAGE analysis of whole cell lysates from *Pichia pastoris* GS115 cells transformed with pPICZ A-AADACL2. From left to right the lanes include: C3I to C7I (Colonies from 3 to 7 Induced), C2I (Colony 2 Induced), C2N (Colony 2 Non-induced), C1I (Colony 1 Induced), C1N (Colony 1 Non-induced), Marker (Tris-Glycine). The red arrows point to the induced protein band at approximately 47 kDa, located just below the 48 kDa marker band, and the induced hyperglycosylated protein band at approximately 70 kDa, located just below the 75 kDa marker band. Each 10 mL culture at $OD_{600} = 3$ was pelleted and resuspended in 500 µL of buffer. For each sample, 1 µL of lysate (~6×10^5 cells) was loaded per lane.

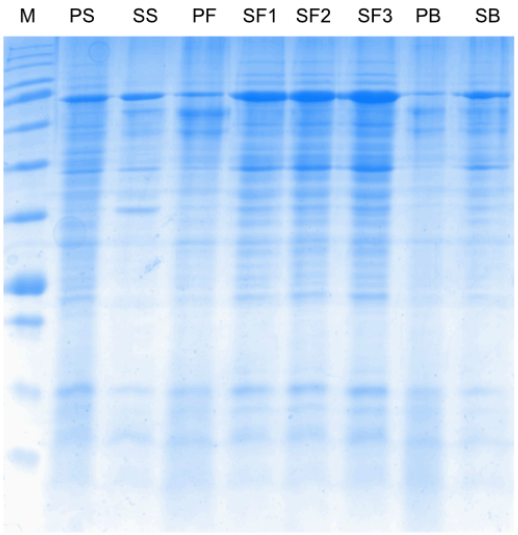


Figure 61: SDS-PAGE analysis of lysis methods for *Pichia pastoris* GS115 cells expressing AADACL2. This SDS-PAGE gel image compares the effectiveness of different lysis methods for *Pichia pastoris* cells transformed with pPICZ A-AADACL2. The lanes include: Marker (Tris-Glycine),

PS (Pellet after Sonication), SS (Supernatant after Sonication), PF (Pellet after first French press cycle), SF1 (Supernatant after first French press cycle), SF2 (Supernatant after second French press cycle), SF3 (Supernatant after third French press cycle), PB (Pellet after Glass bead disruption), SB (Supernatant after Glass bead disruption). Each 10 mL culture at $OD_{600} = 6$ was pelleted and resuspended in 500 µL of buffer. For each lane, 0.5 µL of lysate (~6×10^5 cells) was loaded.

To determine the most effective method for lysis of *Pichia pastoris* cells expressing AADACL2 using the equipment available in our laboratory, we tested three different protocols: sonication, French press, and glass bead disruption. Following the same inoculation, growth, and induction outlined earlier, the cell pellet was resuspended in lysis buffer and divided into three aliquots, each subjected to one of the lysis methods. Sonication was performed at 50% of power with cycles of 5 minutes on and 5 minutes off. The French press method utilized three iterations at 30 kpsi with 5-minute ice rests between cycles. Glass bead disruption involved mechanical agitation with acid-washed glass beads, as described in the Methods section. The efficiency of each lysis method was evaluated by analyzing the distribution of AADACL2 protein between the pellet and supernatant fractions using SDS-PAGE (**Fig. 61**).

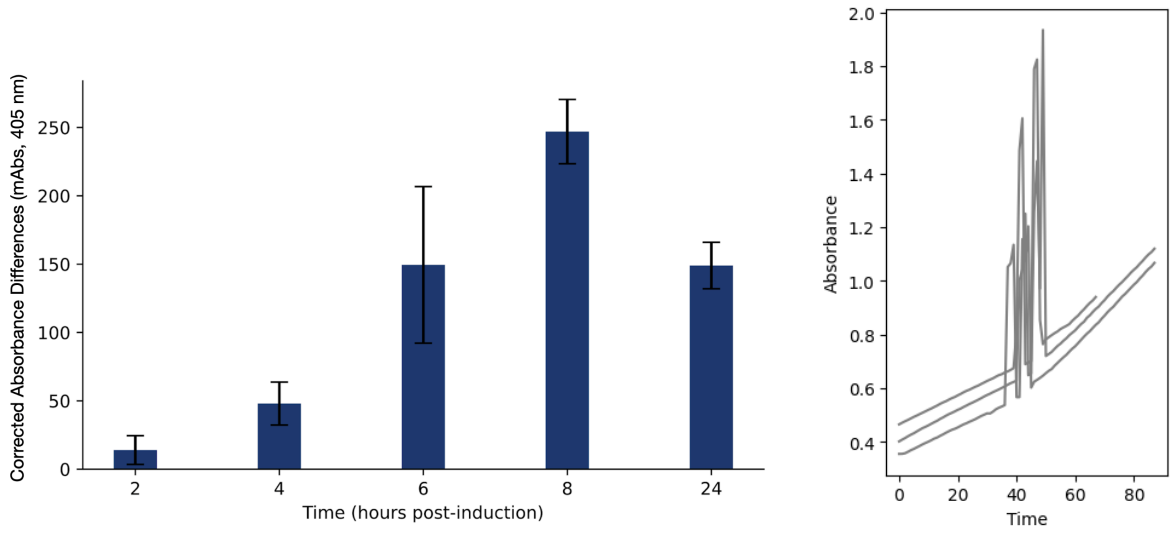


Figure 62. *Left panel:* Bar plot of mean and standard error of delta absorbance between only PNPA and PNPA+AADACL2. This bar graph displays the mean and standard error of delta absorbance between reactions containing only PNPA and those with PNPA plus AADACL2. The X-axis represents hours post-induction (2, 4, 6, 8, and 24 hours), and the Y-axis represents the mean delta absorbance expressed in milliAbsorbance units (mAbs) across three replicates. The peak activity was observed at 8 hours post-induction with a mean delta absorbance of 240 mAbs, followed by a decrease at 24 hours to a mean delta of 150 mAbs, similar to the value at 6 hours but with a higher standard error. *Right panel*: Line plot showing three replicates measured at 8 hours. The X-axis represents time

(seconds), and the Y-axis represents absorbance at 405 nm. The first 30 seconds were recorded without substrate, followed by the addition of substrate.

Comparing the protein amounts in the pellet and supernatant post-lysis, sonication was found to be the least effective, as a significant proportion of the protein remained in the pellet, with the ~47 kDa band still prominently visible. Glass bead disruption proved to be the most effective method, with the pellet band almost disappearing, indicating efficient protein release into the supernatant. The french press method did not show substantial improvement in protein yield between the second (lane 6) and third cycles (lane 7). Given the comparable performance between glass bead disruption and French press, and considered the simplicity and immediacy of the latter, we opted for two cycles of French press for subsequent analyses.

At 2, 4, 6, 8, and 24 hours post-induction, 1 mL of culture was collected and transferred to 1.5 mL centrifuge tubes for processing. Each sample underwent the processing steps described in the method sections. The cells collected at different time points were processed, and the supernatants were immediately analyzed following the activity assay protocol outlined in the Methods. The activity peak, determined by the delta absorbance of the sample compared to the average of the control sample (time point of 0), was observed at 8 hours with a mean delta absorbance of 240 mAbs across three replicates (**Fig. 62**). At 24 hours, a decrease in activity was observed with a mean delta absorbance of 150 mAbs, comparable to the value at 6 hours, although with a much higher standard error. We did not include a graph of values normalized to the total protein concentration measured with the Bradford assay because it did not alter the trend. These assays are not intended to quantify or characterize the activity of the protein but rather to provide an indication of when the protein may exhibit peak activity. Accurately assaying the protein's activity requires precise knowledge of the amount added to the reaction environment, which is not possible when measuring the activity of the supernatant. To further characterize the activity of AADACL2, we performed large-scale expression; however, we were unable to purify the protein as it did not bind to the column, and it was not detected in any of the fractions obtained after FPLC purification. Further studies are necessary to understand the underlying reasons.

Discussion

1. Duplicated Gene Pairs: Numbers and Chromosomal Distribution

Our analysis of homologous human gene sequences has yielded insights into the extent, distribution, and functional implications of gene duplications within the human genome. By employing a rigorous filtering pipeline that integrates sequence similarity and reciprocal best hits (BRHs), we first identified potential duplicated gene pairs. We then refined these pairs by considering conserved domain architectures and enzymatic annotations, resulting in a high-confidence set of 1,184 duplicated gene pairs. This methodological approach ensures the reliability of our findings by minimizing the inclusion of distant homologs and isoforms, thereby focusing on gene pairs with potential functional relevance. The stepwise reduction from 20,434 reviewed human protein sequences to 1,184 high-confidence duplicated gene pairs underscores the stringency of our filtering criteria. Unlike Babbi et al., who identified 3,428 enzymes from an initial set of 20,365 proteins in SwissProt (Babbi et al., 2020), we focused specifically on duplicated gene pairs to facilitate the analysis of functional divergence. Similarly, our filtering pipeline revealed 2,368 proteins associated with EC or RHEA annotations, emphasizing the presence of enzymatic activity. Starting with an all-against-all BLASTP search, the identification of 16,261 homologous sequences highlights the extensive duplication events within the human proteome, reflecting the wide presence of gene duplication within the human genome. This aligns with Lallemand et al., who emphasize the ubiquity of gene duplication across genomes, with an estimated 65.8% of human genes being duplicated (Lallemand et al., 2020). In our analysis, we obtained a slightly higher estimate of 80% due to the use of a more permissive threshold for filtering. The subsequent extraction of 10,130 BRHs emphasizes the presence of robust reciprocal relationships, indicative of true gene duplications rather than transient or spurious matches. Reciprocal best hits are significant because they help confirm that both genes in a pair are each other's closest homologs, reducing the likelihood of false positives in identifying true duplications. Our filtering approach was not only designed to identify duplicates but also to ensure that domain architectures were conserved in both members of the gene pair, thereby facilitating subsequent analyses of functional divergence. Further refinement through Pfam domain architecture filtering and enzymatic annotation ensured that the final dataset comprises gene pairs with conserved structural and functional domains, enhancing the biological relevance of

our findings. Identifying and maintaining duplicated genes can be challenging, as many undergo pseudogenization or subfunctionalization (Innan and Kondrashov, 2010). Pseudogenization refers to the process by which a gene loses its function, often due to mutations, while sub-functionalization involves the partitioning of the original functions between gene duplicates. Our stepwise refinement process, including filtering by domain architectures, helped to ensure that the 1,184 gene pairs in our final dataset were those most likely to have retained functional significance, avoiding many of the challenges associated with identifying functionally divergent or nonfunctional duplicates.

Intra-chromosomal duplications, though less frequent, exhibit distinct patterns on chromosomes 11 and 12. On chromosome 11, there are prominent clusters of duplicated gene pairs, which are the result of localized duplication events, such as the olfactory receptor gene family, which includes 369 genes clustered along the chromosome (Taylor et al., 2006). These localized duplication events, likely driven by mechanisms such as unequal crossing over or replication errors, have contributed to the expansion of specific gene families. Chromosome 9's clustering of inter-chromosomal duplications, predominantly linking to chromosome 1, indicates potential hotspots for genomic interactions and rearrangements. These block duplications likely occurred early in vertebrate evolution and may represent hotspots for genomic interactions and rearrangements, providing insights into chromosomal synteny and the evolutionary history of these gene families (Hughes, 1999). Conversely, the Y chromosome exhibits minimal duplication events, which corresponds with its reduced gene content and unique evolutionary trajectory. This chromosome is characterized by limited recombination and a higher rate of gene loss compared to autosomes, highlighting its distinct role and evolution within the human genome (Rhie et al., 2023).

2. Duplication Events Across Taxonomic Classes

By examining genes across different species, we can infer the timing and frequency of duplication events, as well as the evolutionary pressures acting on these genes. A larger and more diverse sequence database improves the quality of multiple sequence alignments. It allows for better identification of conserved residues and domains, which are indicative of functional and structural importance. This is crucial for subsequent analyses, such as identifying functionally divergent sites between duplicated genes.

In expanding our analysis beyond the human genome, we compiled a substantial dataset of 458,028 orthologous sequences across 177 vertebrate species for 1,147 human duplicated

gene pairs. We classified 280,906 genes as 'one-to-one' orthologs, highlighting their evolutionary conservation across diverse vertebrate lineages. The use of Ensembl Compara facilitated this classification by providing reliable 'one-to-one' ortholog identification, a feature crucial for assessing evolutionary conservation across species. This stringent criterion ensures that for each human gene, only a single orthologous gene from each species is included, reflecting a direct evolutionary counterpart without considering additional duplication events in that lineage. This approach minimizes the confounding effects of species-specific gene expansion and maintains the clarity of orthologous relationships. An average of 113 orthologs per gene across our dataset suggests significant retention, indicative of potential functional importance. Given the scope of our study and the need for accurate, large-scale ortholog mapping, Ensembl Compara was chosen for its established reliability and broad species coverage, although tools like OMA also represent viable alternatives for such analyses (Altenhoff et al., 2020).

Notably, we excluded 17,492 sequences containing ambiguous amino acids (‘X’), underscoring challenges with incomplete or error-prone genome assemblies—a common issue in comparative genomics studies (Rhie et al., 2021). Despite recent advancements in genome assembly technologies, significant challenges remain, particularly in dealing with incomplete and error-prone assemblies. Short-read sequencing methods, which were previously widely used, often result in fragmented genomes with many mis-joins and large numbers of unresolved gaps, especially in repetitive and GC-rich regions. These errors can lead to false gene duplications, which complicate the analysis of evolutionary events. For instance, the presence of ambiguous amino acids (‘X’) and incomplete gene annotations in species like *Carassius auratus* (goldfish) and *Vicugna pacos* (alpaca) may not solely reflect true biological variation, but rather result from the limitations of current genome assemblies. Even with long-read technologies, heterozygosity and repetitive sequences continue to pose significant problems, leading to errors such as false duplications or truncated genes, which can mislead functional and comparative genomic analyses (Rhie et al., 2021). Interestingly, *Erpetoichthys calabaricus* (reedfish) and *Lepisosteus oculatus* (spotted gar) show duplicated pairs that are not found in other fish species. Unlike many teleosts, these species belong to more basal lineages of ray-finned fishes, which may explain the retention of certain duplicated genes. This suggests that they might have experienced lineage-specific duplications or been subject to unique selective pressures, leading to specialized adaptations that distinguish them from other, more derived fish species (Clarke et al., 2016).

Our heatmap analysis (**Fig. 36**) provides an overview of the distribution of duplicated gene pairs across various vertebrate taxa. Notably, we identified 412 duplicate pairs conserved

across Mammalia, Sauropsida, and Actinopterygii. This widespread conservation suggests that these gene duplications originated early in vertebrate evolution and have been maintained due to their essential roles in fundamental biological processes shared among these lineages. These findings align with previous research highlighting the influence of both whole-genome duplications (WGDs) and smaller-scale duplications in shaping the vertebrate genome (Pervaiz et al., 2019; P. P. Singh et al., 2015). Such studies emphasize the retention of duplicated genes because of their critical functions in development and physiology, supporting our observation that these gene pairs are crucial for evolutionary innovation and key biological activities. Further illustrating the distribution of duplicated gene pairs, the Sankey plot (**Fig. 37**) shows that 83 gene pairs are shared between mammals and fishes ('MA'). These likely reflect ancient duplications that occurred before the divergence of these groups, with subsequent gene loss in sauropsids possibly due to selective pressures or genomic rearrangements. The divergence between Actinopterygii and the lineage leading to tetrapods (which includes Mammalia and Sauropsida) occurred earlier than the split between Mammalia and Sauropsida (Benton, 2014). Therefore, any gene duplications present in both fishes and mammals but absent in sauropsids suggest that the duplication may have occurred before the divergence of these groups, and the absence in sauropsids could be due to gene loss in that lineage. Duplications present across all three classes are ancient and likely occurred in a common ancestor of all vertebrates. In contrast, the 312 pairs exclusive to mammals and sauropsids ('MS') probably represent duplications that arose after the split from Actinopterygii but before the divergence between mammals and sauropsids. These duplications may have facilitated shared traits such as endothermy and advanced sensory systems. The evolutionary divergence between mammals and sauropsids (reptiles and birds) occurred approximately 310 million years ago (maximum constraint 370 MYA) (Benton, 2014), more recently than the separation between actinopterygians (ray-finned fishes) and the common ancestor of mammals and sauropsids around 425 million years ago (maximum constraint 495 MYA) (Donoghue et al., 2003). This closer evolutionary relationship explains why more gene duplications are shared between mammals and sauropsids than between mammals and actinopterygians. Additionally, the teleost-specific whole-genome duplication in actinopterygians, followed by gene loss or specialization, led to greater divergence in their genomic architecture, reducing the number of shared duplications with mammals (Ravi and Venkatesh, 2008). In contrast, mammals and sauropsids have retained more of these ancient gene duplications due to their more recent split and similar selective pressures. The presence of 251 gene pairs exclusive to mammals ('M') points to more recent duplications contributing to mammalian-specific traits. These pairs exclusive to mammals likely drive adaptations such as immune system functions, reproductive processes, and skin formation

(Luis Villanueva-Cañas et al., 2017). For example, the Late Cornified Envelope (LCE) genes help maintain the flexibility of mammalian skin (Strasser 2014), while caseins are essential for milk production (Kawasaki et al., 2011). Additionally, genes like neuronatin and IGIP contribute to brain development and immune responses (Luis Villanueva-Cañas et al., 2017). Finally, the 62 gene pairs exclusive to primates or humans ('H') highlight ongoing genomic evolution, potentially playing roles in uniquely human traits and disease susceptibilities. These primate-specific duplications likely support unique adaptations such as advanced cognitive functions, immune responses, and metabolic processes (Marques et al., 2005). The conservation and lineage-specific retention of duplicated gene pairs emphasize the dual role of gene duplication in fostering evolutionary innovation and maintaining essential biological functions. Conserved duplicates across vertebrates likely play roles in core physiological processes, while lineage-specific duplicates may offer adaptive advantages in response to environmental pressures or ecological niches.

3. Alignment Quality and Residue-Level Divergence Analysis

The rigorous alignment and cleaning processes in this study have significantly improved the quality and reliability of the multiple sequence alignments (MSAs) for duplicated gene pairs and their orthologs. Initially, MSAs showed variability in sequence lengths, excessive gaps, and inconsistent residue conservation, potentially compromising functional and evolutionary inferences (Carrillo and Lipman, 1988). As shown in **Figure 39**, the cleaning reduced gaps and improved residue alignment consistency. High-quality alignments are essential for reliable downstream analyses, including phylogenetic tree construction, detection of selection pressures, and identification of conserved functional motifs. The quantitative assessment of alignment metrics highlights the cleaning process's effectiveness. The decrease in sequence numbers per alignment indicates the removal of redundant or erroneous entries, streamlining the dataset (Just, 2001). The significant reduction in gaps, averaging 200,000 post-cleaning, eliminates poorly aligned regions and excessive indels. The observed increase in percentage identity, averaging 60% after cleaning, reflects enhanced conservation across aligned sequences, enabling more accurate detection of evolutionary constraints and functional residues. These improvements are crucial for robust downstream analyses. Accurate residue alignment is key for identifying conserved domains and understanding the structural and functional implications of gene duplications. Enhanced alignment fidelity allows for more precise phylogenetic reconstructions and protein function analyses, providing better insights into the evolutionary relationships of duplicated genes (Phillips et al., 2000). Overall, the

cleaning process has fortified the dataset, laying a solid foundation for analyzing duplicated gene pairs and their evolutionary trajectories. The quantitative analysis of alignment metrics offers insights into the conservation and divergence of amino acid residues within duplicated gene pairs. By classifying residues into Robust, Adaptive, and Plastic categories based on their Differential Score (DSc) and Total Score (TSc), we delineated regions of high conservation, moderate adaptability, and significant variability. As depicted in **Figure 40**, the distribution of these residue types reveals patterns indicative of their evolutionary and functional roles. Robust residues, approximately 175,000 positions, exhibit low variability (DSc) and high conservation (TSc). These residues are critical for maintaining structural integrity and fundamental functions, remaining conserved across diverse species and gene copies. The low XSc values suggest strong purifying selection. Single positions that are critical for a protein's activity—such as catalytic residues or binding site residues—often experience strong selective pressure to maintain their function. These positions are typically conserved across species or paralogs because even small changes in these residues can dramatically alter protein function or render the enzyme inactive (Lynch and Conery, 2000) (Lynch 2000). Adaptive residues, around 160,000 positions, have high DSc values, indicating they are evolutionarily important and conserved in both groups of duplicated genes but differ between the two groups. These residues, also referred to as plasticity residues, play a crucial role in determining enzyme specificity and function, providing the ability to adopt new or altered functions through minimal changes. Yoshikuni et al. demonstrated this through experiments on g-humulene synthase, where saturation mutagenesis identified residues such as W315, M447, S484, and Y566 as critical in shifting product distribution and enabling new enzyme functions (Yoshikuni et al., 2006). By facilitating functional divergence, these residues allow duplicated genes to acquire specialized roles. Occasional exceedance of the TSc threshold in plasticity residues suggests positive selection, contributing to species-specific adaptations. Plastic residues, the most abundant category with approximately 260,000 positions, exhibit low DSc and low TSc values, indicating that they can mutate without significantly affecting protein functionality. These residues are often found in structural regions, such as loops, or in regions that are not critical for protein function, allowing mutations without compromising protein stability and functionality. They are also characterized by a faster evolutionary rate compared to highly conserved and functional residues, which mutate more slowly due to selective pressure to maintain stability and function (Tóth-Petróczy and Tawfik, 2011). Classifying residues into these categories help understanding of the evolutionary pressures on duplicated genes. Robust residues highlight regions under strong functional constraints, adaptive residues point to functional diversification, and plastic residues suggest structural and functional flexibility. This

stratification helps interpret gene duplications' functional implications, offering avenues for further research into conserved and variable regions' roles in protein evolution.

The classification of residues based solely on DSc and TSc scores, while informative, may not capture the full complexity of evolutionary pressures acting on specific amino acid positions. The assessment of our base divergence score metrics revealed relationships between embedding-based and conservation-based measures of residue divergence. Specifically, the moderate positive correlation between the embedding-based delta score and the conservation-based Differential Score (DSc) (Pearson's $r = 0.62$, $p < 0.001$) indicates that residues exhibiting greater dissimilarity in the embedding space are also subject to significant differential conservation across protein groups. This alignment suggests that the embedding-based delta score effectively captures aspects of functional divergence that are corroborated by traditional conservation metrics. Embedding-based scores are context-dependent, meaning the similarity score between two amino acids is influenced by the neighboring residues and the overall sequence. The context can alter the embedding vector of a residue, thus changing its similarity score with other residues (Elnaggar et al., 2021). Additionally, a recent study attempted to compare the embedding cosine similarity to the BLOSUM matrices and obtained good results, further validating the utility of embedding-based approaches in capturing evolutionary relationships (Ashrafzadeh et al., 2024). Conversely, the moderately negative correlation between the delta score and the Total Score (TSc) (Pearson's $r = -0.36$, $p < 0.001$) implies that residues with higher delta scores tend to be less conserved overall, highlighting their potential role in adaptive or specialized functions. The scatter plots depicted in **Figure 41** visually affirm these correlations, with the positive trend between delta and DSc reinforcing the utility of delta scores in identifying functionally divergent residues. The negative trend between delta and TSc further underscores the inverse relationship between embedding-based dissimilarity and overall conservation, aligning with the notion that less conserved residues are more likely to contribute to functional diversification. However, the moderate strength of these correlations suggests that while delta and DSc are related, they capture distinct facets of residue divergence and should be used complementarily rather than interchangeably, especially because the DSc takes into account conservation along ortholog sequences. The integration of additional metrics—Pocket Differential Score (pdsc), Hotspot Differential Score (hdsc), and Hotspot Delta Score (hdelta)—provides a framework for identifying residues that are both evolutionarily divergent and functionally significant. Starting with the Differential Score (DSc), we pinpointed residues that exhibit significant evolutionary divergence, suggesting potential roles in functional specialization following gene duplication. To capture the context in which

these residues operate, we employed the embedding Delta Score (delta). This metric assesses changes in the physicochemical and structural environment of residues across homologous proteins, offering insights into how alterations in residue context contribute to functional divergence. Building upon this foundation, we identified hotspot residues (hdsc) and residues within functional pockets (pdsc) using structural prediction tools P2Rank and BindEmbed21. These tools focus on residues critical for protein function—such as binding sites and catalytic centers—highlighting positions where evolutionary divergence aligns with functional importance. The UpSet plot in **Figure 43** illustrates the intersections among these classifications, revealing that a subset of residues (1,525) meets all three high-significance criteria: evolutionary divergence (DSc), contextual change (delta), and functional significance (hdsc or pdsc). This convergence underscores the potential of these residues to contribute to neofunctionalization in duplicated genes. By focusing on these high-significance residues, we can more effectively analyze human neofunctionalized gene pairs. The alignment of evolutionary divergence with functional hotspots and pockets suggests that these residues may be important in acquiring new functions or modifying existing ones. Our multifaceted approach enables a detailed examination of the molecular mechanisms underlying neofunctionalization. By integrating evolutionary metrics with structural and functional context, we provide a robust framework for identifying residues that are likely to drive functional innovation in duplicated genes. In this context, we adopt the concept of 'multi-view data', to refer to heterogeneous data that provide complementary information for characterizing biological systems (Li et al., 2016). This concept aligns closely with our multi-dimensional analysis using various metrics. Such data can vary in type, source, statistical distribution, semantics, and levels of imprecision and uncertainty. In our study, we focused on integrating different metrics including: (1) pocket residues that highlight substrate binding, (2) active site residues that are critical for catalysis, (3) differential scores from evolutionary alignments that indicate conservation and divergence, and (4) embedding space distances that capture the similarity between residue contexts. By integrating these diverse metrics, we aimed to capture the complexity of evolutionary patterns and functional divergence among duplicated genes. In conclusion, the combined use of DSc, delta, hdsc, and pdsc metrics allows us to pinpoint residues that are evolutionarily divergent, contextually altered, and functionally significant within protein structures.

4. Validation of Divergence Metrics with RHEA: Thresholds and Probability Analysis

The development of global metrics from residue-level scores marks an advancement in our ability to systematically identify functionally divergent protein pairs. By aggregating individual residue-level metrics—Differential Score (DSc), Pocket Differential Score (pdsc), Hotspot Differential Score (hdsc), and Hotspot Delta Score (hdelta)—into global scores, we have established a robust framework that encapsulates multiple dimensions of functional divergence. This integrative approach not only enhances the sensitivity and specificity of divergence detection but also facilitates a nuanced understanding of the underlying evolutionary and structural mechanisms driving functional specialization. Our systematic aggregation methodology synthesizes the four residue-level metrics into global scores that provide protein-level assessments of functional divergence. This synthesis involves calculating ratios and normalized measures that reflect the proportion and density of significant divergence indicators within each protein pair. Specifically, metrics such as the proportion of residues classified as hdsc relative to total hotspots, the proportion of pdsc relative to all hotspots, and the proportion of hdelta among all hotspots offer views of how evolutionary and structural diversifications interplay within protein pairs.

To rigorously evaluate the predictive accuracy of our global metrics, we established a truth set of protein pairs with experimentally validated functional divergence using the RHEA database. This truth set serves as a gold standard, enabling an objective assessment of our metrics' ability to distinguish between functionally divergent and non-divergent protein pairs. The bimodal distribution of truth scores, with distinct peaks at 0 (non-divergent) and 1 (divergent), underscores the reliability of the RHEA-based classifications and validates the efficacy of our truth set construction methodology. Importantly, the dataset was imbalanced, with the negative group being more abundant than the positive group. This imbalance necessitated careful consideration during the analysis to ensure that the metrics were not biased towards the more prevalent class. Central to our approach was the optimization of thresholds for each global metric, achieved through the intersection of Kernel Density Estimate (KDE) curves representing true positives (functional divergence present) and true negatives (no functional divergence). This data-driven thresholding method ensures an unbiased and statistically robust delineation between divergent and non-divergent protein pairs. Using the intersection as a threshold is a common approach to minimize overlap between groups, often achieving a balanced trade-off between false positives and false negatives. The Youden Index is particularly helpful for identifying the 'optimal' threshold value, which maximizes the trade-off between sensitivity and specificity (Youden, 1950b).

This approach is especially valuable when prevalence rates and decision error costs are unknown, as it simplifies the cutoff selection by focusing solely on the test's discriminatory performance (Fluss et al., 2005). The concordance between KDE-derived thresholds and those identified via ROC curve analysis using Youden's J statistic reinforces the validity of our threshold optimization strategy. The performance evaluation of our global metrics, as detailed in **Table 1**, reveals a clear hierarchy in predictive accuracy. Metrics integrating evolutionary and functional data—such as hdsc_r and pdsc_r—demonstrate superior performance, with AUC values exceeding 0.89 and F1 scores approaching 0.76. These metrics leverage both evolutionary conservation and structural/functional insights, highlighting the importance of multi-dimensional data integration in accurately identifying functional divergence. In contrast, metrics based solely on evolutionary information (e.g., ldsc_r) exhibit lower performance, emphasizing the limitations of relying exclusively on sequence conservation without considering structural or functional context. The numerical outcomes clearly demonstrate that metrics integrating multiple sources of information – particularly evolutionary and functional data – outperform those based solely on a single type of data. The high AUC values and balanced F1 Scores of the combined metrics validate the effectiveness of our threshold optimization approach using KDE curve intersections. This optimization ensures that our divergence metrics are finely tuned to distinguish between functionally divergent and non-divergent protein pairs, thereby enhancing the reliability of our classification framework. While the high AUC values and balanced F1 Scores affirm the model's effectiveness, the overlapping regions near the threshold values of approximately 0.15 indicate areas where predictive reliability may decrease (**Fig. 45**). This suggests potential classification errors, such as false positives and false negatives, in these regions.

Applying the optimized thresholds to our dataset, we predicted approximately 35% of protein pairs as functionally divergent. This prediction rate aligns closely with the distribution observed in the RHEA-based truth set, where about 34% of protein pairs exhibited functional divergence. The UpSet plot analysis further elucidates the complementary strengths of different metrics, revealing that a significant subset of divergent pairs is consistently identified across multiple criteria—specifically, pdsc_r, hdsc_r, and hdelta_r. Notably, 217 protein pairs were unanimously predicted as divergent by all three metrics, underscoring the robustness of our classification framework. Additionally, the exclusive identification of divergent pairs by individual metrics—83 pairs by pdsc_r and 83 pairs by hdelta_r—highlights distinct pathways of functional divergence, such as pocket-specific alterations and embedding space dissimilarities that do not necessarily coincide with evolutionary conservation. The high concordance of predictions across multiple metrics reinforces the validity of our integrative

approach and provides compelling evidence for the multifaceted nature of functional divergence. The combination of low hdsc_r and pdsc_r suggests that while the functional and pocket residues are conserved in their roles, there is significant variability in their contextual or structural environments as captured by hdelta_r. For example, protein pairs such as UGT2A1 and UGT2A2 or SULT1C2 and SULT1C4, which have different substrates and share the same EC but exhibit slight differences in RHEA annotations, clearly illustrate this phenomenon. In both pairs, there are less than three hotspots with high DSc, while approximately more than twenty positions show a DSc below the threshold but a high delta. This indicates that the residues maintain their essential functions but operate within different structural contexts or interaction networks. High delta and low DSc can mean that the residues could also be the same but the context changes. Furthermore, it is possible that the changes to these important residues are present but not evolutionarily conserved, likely because they are recent or human-specific alterations. This is attributable to our methodology, where hotspots and pocket residues probabilities are calculated exclusively for humans, whereas the Differential Score (DSc) accounts for conservation across orthologs. Consequently, recent or lineage-specific changes may not be captured by DSc but are reflected in the high hdelta_r scores.

In addition to analyzing protein pairs with low hdsc_r and pdsc_r alongside high hdelta_r, it is equally important to investigate and interpret the patterns exhibited by protein pairs with high pdsc_r but low hdsc_r, as well as those with high hdsc_r but low pdsc_r. The Pocket Differential Score (pdsc_r) quantifies the proportion of residues classified as functional pockets with significant pocket probability identified by P2Rank, relative to all identified pocket residues. A high pdsc_r indicates substantial divergence in the structural or functional characteristics of the protein pockets. Conversely, the Hotspot Differential Score normalized (hdsc_r) represents the proportion of residues classified as hdsc—residues with both DSc and hotspot probabilities above their respective thresholds—relative to the total number of hotspots identified. A low hdsc_r suggests minimal differential conservation of these functional residues between protein pairs. Protein pairs exhibiting high pdsc_r and low hdsc_r signify that while the core functional residues remain conserved (low hdsc_r), there is significant divergence in the surrounding pocket regions. This pattern indicates that the primary catalytic or binding residues are preserved to maintain essential biochemical functions, while the structural context or the environment of these residues within the functional pockets has undergone substantial variation. Divergence in pocket regions can lead to altered substrate specificity or binding affinities, enabling duplicated proteins to interact with different substrates or regulatory molecules. This facilitates the expansion of

metabolic pathways and the specialization of enzymatic functions without compromising the core activity. Changes in pocket regions may also influence the interaction with allosteric regulators or inhibitors, allowing for fine-tuned regulation of enzyme activity in response to varying cellular conditions. The study of aspartate aminotransferase (AspAT) provides an exemplary case of how proteins can retain their catalytic function while significantly altering substrate specificity through cumulative mutations. AspAT, which typically catalyzes amino group transfer between acidic amino acids such as aspartate and glutamate, was subjected to directed evolution to modify its specificity toward branched-chain amino acids like valine and isoleucine. The resulting mutant enzyme, ATB17, achieved a 2.1 x 10^6-fold increase in catalytic efficiency for valine, despite retaining its overall aminotransferase function (Oue et al., 1999). Conversely, protein pairs with high hdsc_r and low pdsc_r signify that while the functional pockets are conserved (low pdsc_r), there is considerable variation in the specific functional residues within these pockets. This pattern suggests that the core structural framework of the pocket remains intact to preserve essential biochemical functions, while the individual residues contributing to these functions exhibit differential conservation, potentially altering the enzymatic or binding properties of the proteins. The cytochrome P450 BM-3 enzyme provides a compelling example of how directed evolution can lead to novel catalytic activities while retaining substrate specificity. In particular, the P450 BM-3 variant 139-3 was engineered to shift from performing hydroxylation of fatty acids to epoxidation of alkenes, such as styrene and cyclohexene. This functional shift illustrates that, while the enzyme continues to bind the same types of substrates, its active site modifications allow it to catalyze a different reaction—epoxidation instead of hydroxylation—thereby expanding its catalytic repertoire without altering its substrate-binding properties (Farinas et al., 2004). This understanding of metric combinations—high pdsc_r with low hdsc_r and high hdsc_r with low pdsc_r—highlights the diverse mechanisms through which functional divergence can occur. It underscores the importance of employing an integrative metric approach, combining conservation-based and embedding-based measures, to capture the full spectrum of functional adaptations in duplicated proteins. By doing so, our classification framework not only preserves critical biochemical functions but also uncovers the structural and contextual adaptations that drive evolutionary innovation. Our approach underscores the importance of integrating multiple metrics to capture the complex nature of functional divergence. Conservation-based metrics ensure the retention of critical functions, while embedding-based metrics provide insights into structural and contextual adaptations that drive functional innovation. This integrative framework enhances our ability to accurately identify and interpret functional divergence, offering a robust foundation for future studies.

The sigmoid function is commonly used to transform raw scores into probabilities in the range of 0 to 1, making it ideal for probabilistic interpretations in biological data (Hastie et al., 2001). The transformation of the hpd_r scores into probabilities ranging from 0 to 1 created two distinct tails in the distribution of probabilities associated with our functionally divergent and non-divergent protein pairs. Notably, 237 values fell in the range $0.3 < prob < 0.7$, 293 were associated with $prob \geq 0.7$, and 582 with $prob \leq 0.3$ (**Fig. 48**). During our initial analysis, we calculated one-tailed-p-values to assess whether our observed scores were significantly higher than the randomized scores generated through permutation testing (Ernst, 2004). However, we consistently found that the p-values were equal to 1. This result indicates that the observed scores were consistently lower than the randomized scores. The key reason behind that is that, on average, functional residues tend to be more conserved. As a result, when randomizing the differential scores (DSc), it becomes more likely that these random associations will show greater divergence simply due to chance, rather than reflecting actual evolutionary divergence. In essence, by randomizing the differential scores, the randomized scores are most likely to be associated with functional residues purely by chances. In contrast, the real data show less divergence between functional residues because these residues are under strong selective pressure to remain conserved. This conservation reduces the observed divergence between gene pairs, leading to lower scores compared to the randomized data, where such constraints do not exist. This also suggests that functional divergence, when present, is relatively rare in the observed data, further highlighting the evolutionary importance of maintaining these functional sites. To better assess the biological relevance of the observed lower scores, we later shifted to a two-tailed p-value approach, as two-tailed tests allow for evaluation of both high and low deviations from the null distribution (Zar, 1999), in our case to evaluate whether the observed scores were significantly different (either higher or lower) than the randomized scores. The empirical p-values associated with these scores largely fell below the significance threshold of 0.05, with 840 gene pairs exhibiting significant divergence. Notably, 560 of these pairs were associated with p-values less than 1e-4 (**Fig. 48**). This demonstrates that a significant portion of the gene pairs show divergence or non-divergence that is statistically distinct from what would be expected by random change, highlighting the importance of using a two-tailed test to fully capture both sides of the distribution. The use of a two-tailed approach also provided valuable information on both positive and negative deviations. This helps understand the significance of the observed conservation and divergence pattern, especially as it allows us to differentiate between gene pairs that showed strong conservation and those with meaningful divergence. Moreover, these results suggest that specific evolutionary pressures might act differently across these gene pairs, possibly influencing their functional roles. The significant number of

gene pairs with very low p-values further emphasizes that evolutionary conservation is a predominant factor in maintaining functional integrity, while the outliers with high p-values might represent emerging adaptations or shifts in functional context.

5. Functional Enrichment and Tissue-Specific Expression Analysis

Our analysis of tissue-specific expression differences, combined with functional enrichment, provides insights into the mechanisms driving functional divergence among duplicated protein pairs. By utilizing the Human Protein Atlas (HPA) database alongside the Genotype-Tissue Expression (GTEx) dataset, we assessed absolute expression differences (exp_diff) across 50 tissues for 434 protein pairs. The resulting radial plot highlighted the liver as the most frequently identified tissue with significant expression differences, accounting for approximately 22% of the pairs. Additionally, the liver ranked third in terms of the proportion of divergent pairs, underscoring its essential role in metabolism and detoxification (Grant, 1991). The liver's prominent representation is consistent with its high concentration of enzymes critical for managing diverse biochemical reactions and responding to environmental challenges. This suggests that duplicated proteins in the liver may undergo substantial functional innovation to efficiently handle a wide array of metabolic processes. Such specialization is likely driven by the liver's central role in enzymatic and metabolic functions, necessitating adaptations that support its diverse physiological responsibilities (Rui, 2014). In contrast, the skin exhibited the highest proportion of divergent pairs, with 70% of protein pairs classified as divergent within this tissue. This indicates significant functional adaptation and specialization, likely reflecting the skin's role in barrier formation, immune responses, and interaction with the external environment (Brettmann and de Guzman Strong, 2018). The kidney followed with the second-highest proportion of divergent pairs, emphasizing its involvement in specialized metabolic processes and regulatory functions. Conversely, tissues such as the brain and skeletal muscle displayed lower proportions of divergent pairs despite being among the most populated tissues analyzed. This lower divergence reflects a higher degree of functional conservation, aligning with the critical and stable roles these tissues play in maintaining fundamental physiological functions. The brain's involvement in neural signaling and the skeletal muscle's role in contraction and movement necessitate stringent functional consistency, thereby reducing the likelihood of functional divergence among duplicated proteins in these areas. This is consistent with the observation that brain-expressed genes tend to evolve more slowly due to stronger purifying selection, whereas tissues like the liver experience faster evolution, possibly due to weaker

selective pressures (Kryuchkova-Mostacci and Robinson-Rechavi, 2015). These differential patterns of functional divergence across tissues highlight the interplay between tissue-specific demands and evolutionary pressures on duplicated proteins. Tissues with high metabolic and adaptive requirements, such as the liver and skin, facilitate extensive functional diversification to meet specialized biochemical needs. In contrast, tissues that require stringent functional consistency, like the brain, maintain higher conservation levels to preserve essential cellular processes. Overall, our findings emphasize the importance of considering tissue-specific expression profiles when investigating functional divergence. The ability of duplicated proteins to adapt and specialize in response to the unique demands of different tissues underscores the dynamic nature of the proteome in supporting complex biological systems and facilitating evolutionary innovation (Li et al., 2005).

Complementing the tissue-specific expression analysis, our functional enrichment study elucidated distinct biological functions and pathways associated with divergent and non-divergent protein pairs. Divergent protein pairs were significantly enriched in catalytic activities, including transferase, oxidoreductase, hydrolase, and lyase activities, as well as in specialized metabolic pathways such as arachidonic acid metabolism, fatty acid elongation, and steroid biosynthesis. These enrichments suggest that functional divergence in these proteins is primarily driven by the need to adapt to specialized metabolic requirements and environmental pressures, facilitating metabolic diversity and innovation. Conversely, non-divergent protein pairs showed enrichment in binding-related molecular functions and central metabolic pathways, such as nucleotide binding, DNA binding, pyruvate metabolism, and purine metabolism. While this pattern highlights the essential and conserved roles of these proteins in maintaining fundamental cellular processes and structural integrity, it is important to note that even conserved pathways can undergo significant internal innovation. Studies have shown that conserved enzymes often participate in multiple pathways, and their recruitment into new contexts is a major driver of functional divergence within these pathways (Peregrín-Alvarez et al., 2009). These findings are consistent with broader observations in metabolic network studies, which suggest that pathways associated with core metabolic functions (e.g., pyruvate metabolism, glycolysis, TCA cycle) are less prone to large-scale innovation. These pathways, enriched in our functionally conserved protein pairs, represent the backbone of cellular metabolism, ensuring essential energy production and cellular maintenance. The stability and conservation of these pathways underscore their indispensable roles in metabolism. Conversely, pathways such as arachidonic acid metabolism, steroid biosynthesis, and xenobiotic metabolism, enriched among the functionally divergent pairs, align with categories known to be more prone to internal

modularity and innovation. These pathways, which are more specialized, tend to evolve through recruitment and adaptation of existing enzymes to meet the dynamic metabolic requirements of different tissues and environmental contexts. This flexibility is critical for facilitating metabolic diversity and adaptation (Peregrín-Alvarez et al., 2009).

6. Autonomous Pipeline: Functional Divergence of AOC2 and AOC3

The implementation of our autonomous pipeline within a Google Colab environment represents a significant advancement in the accessibility and scalability of functional divergence analyses. This cloud-based, user-friendly platform enables researchers to perform useful analyses on any pair of proteins across all species, including interspecies comparisons, without the necessity for local installations or specialized computational expertise. By requiring only two UniProt identifiers as input, the pipeline accommodates a diverse range of homologous protein pairs, encompassing both enzymatic and non-enzymatic proteins. This broad applicability ensures that the pipeline can be seamlessly integrated into various biological and evolutionary studies, facilitating the exploration of functional divergence in a wide array of contexts. A key aspect of our pipeline is the calculation of the functional divergence score (hpd_r), which synthesizes multiple metrics to provide a holistic assessment of divergence. The hpd_r score integrates functional predicted hotspots, predicted residues within binding pockets, differences in the embedding space, and evolutionarily conserved differences. This multifaceted approach ensures that the score captures both sequence and structural variations that contribute to functional specialization.

In the illustrative case of the AOC2 (UniProt ID: O75106) and AOC3 (UniProt ID: Q16853) protein pair, the pipeline effectively identified significant functional divergence, predominantly driven by the pocket differential score (pdsc_r). Despite both proteins retaining conserved enzymatic activities within the semicarbazide-sensitive amine oxidase (SSAO) family, the high pdsc_r score highlighted substantial alterations in their binding pocket regions. Specifically, residues Gly463 in AOC2 and Leu469 in AOC3 were pinpointed as critical for substrate specificity, where Gly463 facilitates the accommodation of larger aromatic amines in AOC2, whereas Leu469 restricts substrate access to smaller amines in AOC3. This pocket-focused divergence underscores the pipeline's capability to detect structural adaptations that underlie functional specialization (Kaitaniemi et al., 2009). The outputs generated by the pipeline, including multiple sequence alignments, phylogenetic trees, structural visualizations, and annotated residue tables, provide an in-depth understanding of

the evolutionary and functional dynamics between protein pairs. The phylogenetic analysis using RAxML revealed that the duplication event leading to *AOC2* and *AOC3* is conserved exclusively in mammals, with species-specific duplications observed in *Bos taurus*. This finding suggests lineage-specific expansions and functional divergences within the SSAO family, aligning with evolutionary theories that posit gene duplication as a catalyst for functional innovation and adaptation (Ohno, 1970). Moreover, the integration of tissue-specific expression data from the Human Protein Atlas (HPA) adds an additional layer of functional context to our analyses. In the case of *AOC2* and *AOC3*, significant expression differences in the liver— a tissue central to metabolic and detoxification processes—highlight the physiological relevance of their functional divergence. The liver's role necessitates specialized enzymatic functions to manage diverse biochemical reactions, thereby driving the evolution of duplicated proteins to fulfill these specialized roles (Rui, 2014). The ability of the pipeline to prioritize pocket-specific changes, functional residues changes in the active site, and evolutionarily conserved differences, as evidenced by the high pdsc_r score in the AOC2/AOC3 pair, is particularly noteworthy. Binding pockets and active site residues are critical for substrate recognition, catalysis, and overall protein function, and alterations in these regions can lead to significant functional diversification (Tyzack et al., 2017). By accurately identifying such divergences, the pipeline provides valuable insights into how proteins evolve to acquire new functionalities. In summary, our autonomous pipeline offers a robust and versatile tool for the analysis of functional divergence in protein pairs across diverse species. The successful application to the AOC2 and AOC3 proteins demonstrates the pipeline's efficacy in uncovering pocket-specific functional adaptations, underscoring its potential to facilitate discoveries in protein evolution and functional genomics.

7. Case Study: AADACL2 Functional Divergence and Expression

We analyzed the functional divergence and evolutionary trajectories of the AADAC-AADACL2 protein pair, highlighting their specialized roles in human physiology and lipid metabolism. Utilizing a custom ranking system, we prioritized AADAC (P22760) and AADACL2 (Q6P093) as key candidates for further functional characterization. Our stringent selection criteria focused on significant expression differences, incomplete enzymatic annotations, and the presence of at least one characterized protein. This approach ensured that our analysis concentrated on protein pairs with potential novel functional insights, particularly emphasizing the understudied AADACL2. Specifically, among the top 30 ranked pairs, only four others exhibited a similar pattern of enzymatic classification. AADAC

(Arylacetamide deacetylase) is annotated with the Enzyme Commission (EC) number 3.1.1.3, while AADACL2 (Arylacetamide deacetylase-like 2) has an incomplete EC number 3.1.1.-. This incomplete annotation suggests potential variations in their substrate specificity, distinguishing them from other proteins in their rank.

AADAC (Arylacetamide deacetylase) plays a key role in hepatic and gastrointestinal drug metabolism by hydrolyzing amide bonds in pharmaceuticals. This activity significantly influences drug efficacy and toxicity. AADAC metabolizes drugs such as rifampicin, ketoconazole, and phenacetin, thereby either activating or deactivating these compounds (Kobayashi et al., 2012; Watanabe et al., 2009, Nagaoka et al., 2024). Beyond drug metabolism, AADAC reduces intracellular triglyceride accumulation, suggesting a protective role against non-alcoholic fatty liver disease (NAFLD). The hydrolyzed triglycerides facilitate the assembly of very low-density lipoprotein (VLDL), with apolipoprotein B (ApoB) serving as a critical structural component essential for VLDL stability (Nourbakhsh et al., 2013). The association of ApoB with AADAC within the Human Protein Atlas (HPA) cluster underscores their interconnected roles in lipid metabolism (**Fig. 53**). Structurally, AADAC shares an α/β-hydrolase fold and a catalytic triad (Ser189, Asp343, His373) essential for its hydrolytic function. Despite similarities with CES1 and CES2, AADAC exhibits distinct substrate specificities, likely due to variations in their active site architectures (Yao et al., 2018). In contrast, *AADACL2* (Arylacetamide deacetylase-like 2) is predominantly expressed in the skin, suggesting a specialized function in epidermal lipid processing. The high expression difference (exp_diff > 70) between *AADAC* and *AADACL2* highlights their functional specialization in different tissues. AADACL2 shares a conserved catalytic triad (Ser189, Asp343, His371) and residues forming the oxyanion hole (His111, Gly112, Gly113), which stabilize the transition state during hydrolysis (**Table 5**). Given its exclusive expression in the skin and its association with genes essential for skin structure and function, such as LCE1E and various keratins, AADACL2 likely plays a role in maintaining the skin barrier and lipid metabolism.

This is further supported by its clustering within the HPA cluster 13, labeled “Skin - Cornification,” contrasting with AADAC's placement in cluster 83, “Liver - Metabolism.” The structural similarities between AADACL2 and AADAC suggest that AADACL2 may also have evolved towards specialized metabolic roles, potentially possessing unidentified lipase activities that contribute to epidermal lipid processing. The tissue expression profiles from the HPA data reveal a significant divergence in the localization of AADAC and AADACL2, further emphasizing their functional specialization. *AADAC* is predominantly expressed in the liver, aligning with its roles in drug and lipid metabolism, whereas *AADACL2* exhibits almost

exclusive expression in the skin, indicating a specialized function in epidermal lipid processing (**Fig. 53**). HPA clustering reinforces this distinction: AADAC is situated within cluster 83, labeled “Liver - Metabolism,” while AADACL2 is positioned in cluster 13, labeled “Skin - Cornification.” This segregation underscores their functional divergence and reflects tissue-specific metabolic adaptations. Given these expression patterns and genetic associations, it is plausible that AADACL2 or similar proteins possess unidentified lipase activities that play a crucial role in epidermal lipid processing.

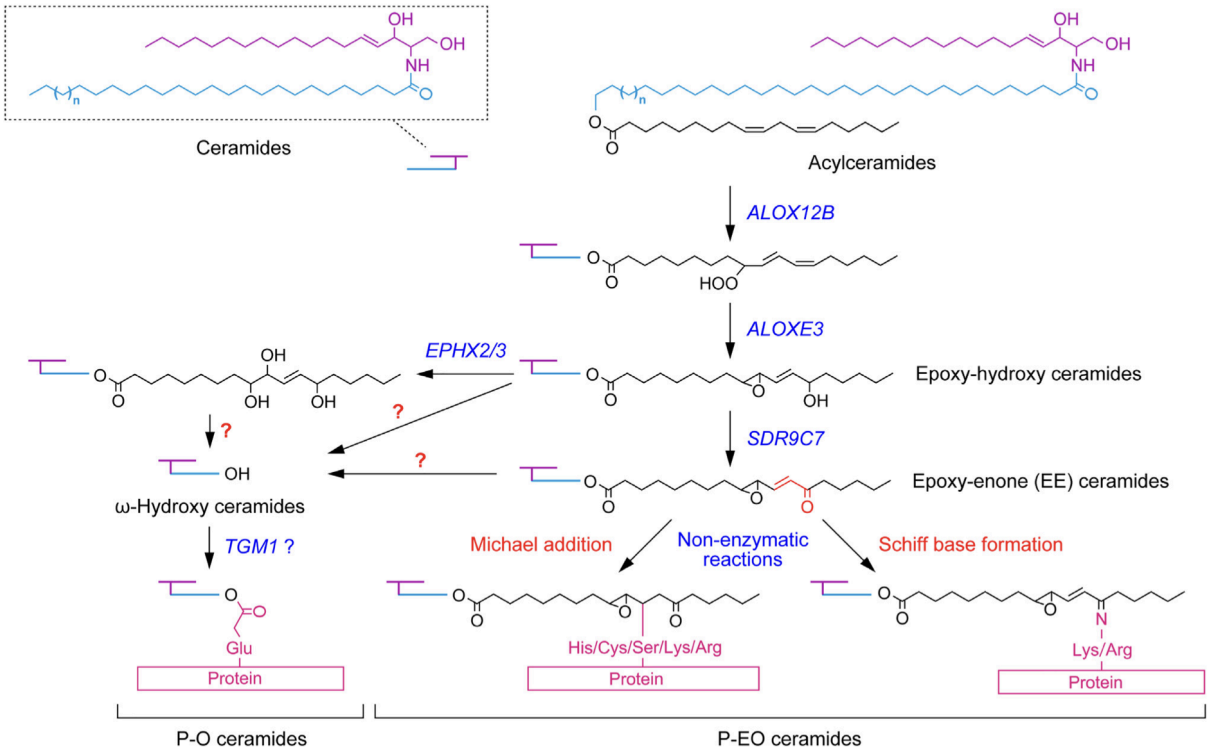


Figure 63. Two structural models and synthetic pathways of protein-bound ceramides. The diagram shows the synthesis of P-O ceramides and P-EO ceramides, including key genes involved at each step. Acyl ceramides are modified via peroxidation and epoxy-hydroxylation to produce epoxy-hydroxy ceramides. In the P-O pathway, unknown lipases cleave epoxy-hydroxy ceramides, forming ω-OH ceramides, which are then covalently attached to cornified envelope (CE) proteins. In the P-EO pathway, epoxy-hydroxy ceramides are reduced to EE ceramides, which bind to proteins through Michael addition or Schiff base formation. Reproduced from Ohno et al. (2023).

The skin serves as the body’s first line of defense, forming a multifaceted barrier against external threats while regulating water loss and thermoregulation (Romanovsky, 2014). In mammals, the formation and maintenance of this barrier rely on a complex interplay of lipids, particularly ceramides, and structural proteins within the cornified envelope (CE) of the epidermis (Wertz, 2021). Ceramides, a class of sphingolipids with long-chain fatty acids, are

essential for maintaining the skin's impermeability and structural integrity. These lipids are covalently attached to proteins, providing the scaffold necessary for the skin's protective barrier (Marekov and Steinert, 1998). Despite decades of research, the complete enzymatic pathway responsible for the processing and attachment of ceramides to the CE remains incompletely understood, with key enzymes yet to be identified (Ohno et al., 2023). Ceramides are synthesized and modified through a series of complex enzymatic reactions that result in their covalent attachment to proteins such as involucrin. This process ensures the formation of the cornified lipid envelope (CLE), a lipid-protein matrix that provides mechanical strength, prevents water loss, and acts as a shield against environmental pathogens. There are two primary pathways leading to protein-bound ceramides: (1) P-O Ceramide Pathway: Ceramides are ester-linked to proteins; (2) P-EO Ceramide Pathway: Ceramides bind through Michael addition to cysteine residues of proteins via epoxy-enone (EE ceramides). Several enzymes have been identified in the ceramide processing pathway. Notably, ALOX12B and ALOXE3 are lipoxygenases critical for the early stages of ceramide modification. These enzymes oxidize acyl ceramides to generate epoxy-hydroxy ceramides, a key intermediate in the pathway. Additionally, the enzyme SDR9C7 plays a role in converting epoxy-hydroxy ceramides into epoxy-enone (EE) ceramides. However, the precise lipase or hydrolase responsible for further processing ceramides into forms that can bind to proteins such as involucrin has remained unidentified. This step is crucial for forming the mature cornified lipid envelope (Ohno et al., 2023) (**Fig. 63**). Identifying these enzymes is vital for understanding the complete enzymatic pathway involved in skin barrier formation and maintenance.

Our evolutionary analysis revealed that *AADAC* and *AADACL2* are absent in cetaceans, a notable divergence given their presence in other vertebrate lineages. At least for *AADACL2*, this gene loss appears to be an adaptation associated with the transition to an aquatic lifestyle, where distinct metabolic requirements may have rendered these genes unnecessary. For instance, cetaceans have also lost related genes like *ALOXE3*, *TGM5*, and *filaggrin*, which play critical roles in skin barrier formation and lipid processing, respectively. In contrast, terrestrial mammals depend heavily on these genes to maintain robust skin barriers capable of resisting environmental stressors. In cetaceans, alternative mechanisms have evolved to maintain skin integrity in a marine environment, potentially through other lipid-metabolizing enzymes or structural adaptations (Espregueira Themudo et al., 2020). This raises compelling questions about the plasticity of metabolic pathways and the redundancy of lipid-processing enzymes across different taxa. Understanding such evolutionary adaptations provides deeper insights into how lipid metabolism and skin barrier

functions evolve under varying environmental pressures. Cetaceans present a fascinating case of how selective pressures can remodel skin physiology. Over millions of years, the shift from a terrestrial to an aquatic lifestyle has driven substantial anatomical and physiological changes, including the thickening of the epidermis, increased cellular turnover, and the loss of typical mammalian features such as sebaceous glands and hair. Genetic changes mirror these physiological adaptations; many genes related to skin maintenance, including those involved in keratinization, lipid metabolism, and immune response, have been lost in cetaceans. For example, the absence of *ALOXE3* in both toothed and baleen whales reflects a broader trend of gene inactivation in skin-related pathways, facilitating adaptation to a marine environment (Espregueira Themudo et al., 2020). Recent studies suggest the involvement of a yet unidentified lipase that could catalyze the hydrolysis of ceramide intermediates—an essential step in forming protein-bound ceramides crucial for skin impermeability (Ohno et al., 2023). Although the exact identity of this enzyme remains unknown, it likely plays a significant role in maintaining skin function in aquatic settings. Interestingly, further analysis identified another pair of proteins from the same family, AADACL3 (Q5VUY0) and AADACL4 (Q5VUY2), which exceeded our inclusion thresholds (probability = 0.897, p-value = 0.004, hdsc_r = 0.15, pdsc_r = 0.23, hdelta_r = 0.54). These proteins, along with AADAC and AADACL2, are included in the RAxML phylogenetic tree (**Fig. 54**), which reveals an intriguing evolutionary pattern. While AADACL3 and AADACL4 have orthologs in aquatic mammals like *Tursiops truncatus*, AADAC and AADACL2 do not. This suggests a lineage-specific retention of functionally distinct enzymes in response to differing environmental pressures, particularly between terrestrial and aquatic mammals.

Our analysis revealed key differences between the binding pockets of AADAC and AADACL2, highlighting their divergent evolutionary adaptations. Specifically, AADACL2's binding pocket appears elongated, as confirmed by structural alignment and visualization with PyMOL (**Fig. 55**, bottom right panel). This elongation suggests an adaptation for accommodating larger lipid substrates, supporting the hypothesis that AADACL2 is specialized for processing more complex substrates like epoxy-hydroxy ceramides, unlike AADAC. Both AADAC and AADACL2 conserve the catalytic triad (Ser189, Asp343, His371 in AADACL2) and the oxyanion hole residues (His111, Gly112, Gly113), which are essential for their enzymatic functionality. This conservation aligns AADACL2 with related hydrolases such as CES1 and CES2. However, a mutation from tryptophan to phenylalanine at residue 115 in AADACL2, located adjacent to the oxyanion hole, indicates a fine-tuned alteration in substrate binding. This mutation could increase the binding pocket's affinity for larger lipid

molecules, enhancing AADACL2's specificity for lipid intermediates important for epidermal barrier formation.

Our docking studies provide compelling evidence that AADACL2 interacts with epoxy-hydroxy ceramides, positioning it as a likely candidate for the unidentified lipase involved in ceramide maturation. This enzymatic activity is crucial for forming the cornified lipid envelope (CLE), a key component of the epidermal barrier that ensures mechanical strength, impermeability, and protection against pathogens (Ohno et al., 2023). The structural compatibility observed, with the catalytic serine optimally positioned for substrate cleavage, supports AADACL2's role in ceramide processing. This insight fills a significant gap in our understanding of ceramide metabolism, positioning AADACL2 as a potential missing link in converting epoxy-hydroxy ceramides into ω-hydroxy ceramides, essential for CLE formation. In this investigation, AADACL2 was hypothesized to act on epoxy-hydroxy ceramides within the ceramide metabolism pathway. Based on known biochemical pathways, particularly the P-O ceramide pathway, we focused on this substrate due to its role as a key intermediate produced by the sequential actions of ALOX12B and ALOXE3 (Muñoz-Garcia et al., 2014). **Figure 64** illustrates how these intermediates are further processed by a lipase to release ω-hydroxy ceramides, critical for protein attachment in the CLE. The docking study's choice of epoxy-hydroxy ceramides provided insight into AADACL2's potential mechanism in catalyzing their cleavage and release. The 3D structure of the epoxy-hydroxy ceramide used for docking (**Figure 56**, top panel) features a sphingoid base with an amino group, hydroxyl group, and long hydrocarbon chain; an N-acyl fatty acid ranging from 30 to 36 carbons linked via an amide bond, including an omega-hydroxyl group (Akiyama, 2021). This omega-hydroxyl group is key for covalent attachment to proteins in the CLE (Nemes et al., 1999), remaining esterified with linoleic acid in the context of epoxy-hydroxy ceramides. The reactive groups—an epoxide and a hydroxyl—enhance the substrate's activity. Docking results (**Figure 56**, central panels) indicate several favorable conformations where the catalytic serine in AADACL2 is positioned within 4 angstroms of the carbonyl carbon, aligning well with expected enzymatic functionality. The highest number of such favorable conformations was found in AADACL2, supporting its hypothesized role. **Figure 56** (bottom panel) also illustrates the structure of AADACL2, with functional residues and those with high differential conservation scores (hdsc) highlighted. These findings suggest that AADACL2 could indeed process epoxy-hydroxy ceramides, providing a crucial enzymatic step in CLE formation. The docking analysis thus serves as an important basis for further experimental validation, such as enzymatic assays, to confirm AADACL2's functional role. Identifying and

characterizing this enzyme will enhance our understanding of ceramide metabolism and the molecular mechanisms underlying skin barrier formation and maintenance.

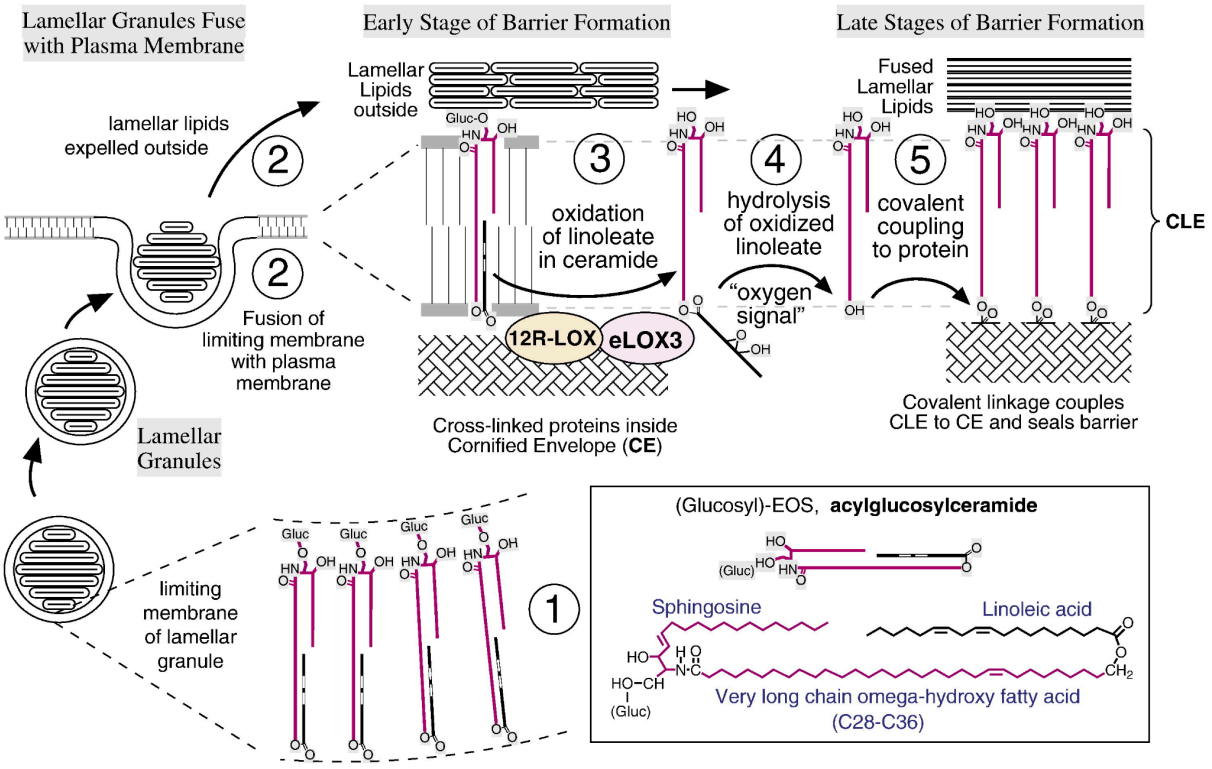


Figure 64. Ceramides, essential fatty acids, and lipoxygenases in the maturation of the epidermal water barrier. (1) Lamellar granules contain lipid disks with acyl glucosylceramide (Glc-EOS) in the limiting membrane. (2) Fusion of the lamellar granules with the plasma membrane initiates the formation of the corneocyte lipid envelope (CLE), extruding lipids extracellularly. (3) The progression towards a mature barrier involves lipoxygenase-catalyzed oxygenation of linoleate in the ceramide by 12R-LOX and eLOX3. (4) This "oxygen signal" allows esterase-catalyzed hydrolysis of the oxidized linoleate, freeing the ceramide omega-hydroxyl. (5) Transglutaminase then catalyzes the covalent attachment of the CLE to cross-linked proteins in the cornified envelope, sealing the barrier. Reproduced from Muñoz-Garcia et al. (2014).

A significant aspect of our findings lies in the association of genes related to *AADAC* and *AADACL2* with skin diseases, such as atopic dermatitis and ichthyosis. Ichthyosis encompasses a group of genetic skin disorders characterized by dry, scaly skin due to impaired lipid metabolism and barrier function. Many genes implicated in ichthyosis are involved in ceramide processing and epidermal lipid metabolism (Fischer, 2009), suggesting that AADACL2 may play a crucial role in maintaining skin barrier integrity. Although direct connections between AADACL2 and specific skin diseases have not yet been established, the identification of AADACL2 as a potential lipase involved in ceramide maturation presents

a promising avenue for medical research. By bridging this metabolic gap, AADACL2 could emerge as a novel therapeutic target for treating ichthyosis and other lipid metabolism-related skin disorders. Future studies should aim to elucidate the precise role of AADACL2 in skin pathology through in vitro and in vivo models, potentially leading to the development of targeted therapies that restore or enhance its function.

The integration of expression data from the Human Protein Atlas (HPA) with phylogenetic analyses revealed a coordinated regulation of lipid metabolism genes within tissue-specific clusters. AADAC's inclusion in the "Liver - Metabolism" cluster and AADACL2's placement in the "Skin - Cornification" cluster, alongside genes such as ApoB and various keratins, highlight a tightly regulated network of proteins essential for maintaining lipid homeostasis and barrier function. The evolutionary analysis of AADAC across several vertebrates, with notable exceptions like cetaceans, underscores their fundamental role in terrestrial mammalian physiology. This conservation, coupled with the observed functional divergence, emphasizes the importance of these enzymes in specialized metabolic processes tailored to specific tissues. Cetaceans, adapted to aquatic environments, have lost several related genes, indicating that different environmental pressures can shape the retention or loss of key metabolic enzymes. While the bioinformatic and structural predictions provide a robust framework for understanding AADACL2's potential role in ceramide metabolism, experimental validation is imperative. Future research should focus on enzymatic assays to confirm AADACL2's lipase activity and substrate specificity. Additionally, generating knockout or overexpression models in relevant cell lines or animal models could elucidate AADACL2's physiological relevance in skin lipid metabolism and barrier function. Investigating the interplay between AADACL2 and other lipid-metabolizing enzymes could unveil a network governing epidermal lipid homeostasis. Moreover, exploring the potential genetic associations between *AADACL2* variants and skin disorders in patient populations could provide valuable insights into its clinical significance. This study advances our understanding of the AADAC family by delineating the specialized functions and evolutionary adaptations of *AADAC* and *AADACL2*. The findings highlight the balance between functional specialization and evolutionary pressures that define these enzymes, offering valuable insights into both basic biology and potential clinical applications. By identifying AADACL2 as a probable key player in ceramide processing, this research not only fills a critical gap in lipid metabolism but also opens avenues for developing novel therapeutic strategies targeting skin barrier dysfunctions. The integration of bioinformatic analyses, structural modeling, and evolutionary perspectives exemplifies the power of interdisciplinary approaches in uncovering the complexities of enzyme function and evolution.

The expression of recombinant proteins in *Escherichia coli* is a widely used strategy due to its simplicity, cost-effectiveness, and well-characterized genetics (Rosano, 2014). In this study, we aimed to express the human arylacetamide deacetylase-like 2 (AADACL2) protein in *E. coli* BL21-CodonPlus DE3 cells using a modified pET-28a(+)-TEV expression vector. Despite strategic modifications to enhance solubility and expression, our results highlighted significant challenges in obtaining soluble and functional AADACL2 protein from this prokaryotic system. The *AADACL2* coding sequence (CDS) was cloned into the pET-28a(+)-TEV vector, which provides an N-terminal 6xHisTag and a TEV protease cleavage site to facilitate purification and potential downstream processing. Recognizing that eukaryotic proteins often contain signal peptides that can hinder expression in bacterial systems, we utilized SignalP 5.0 to predict and remove the N-terminal signal peptide comprising the first 18 amino acids. The decision to initiate the sequence with "FYTP" was based on the mutagenesis performed to enhance solubility by eliminating hydrophobic regions that could contribute to aggregation (Gopal and Kumar, 2013). The choice of *E. coli* BL21-CodonPlus DE3 cells was strategic due to their enhanced capacity to express eukaryotic proteins that contain rare codons. The presence of additional tRNA genes for rare codons minimizes translational pauses that can lead to misfolding or premature termination (Kaur et al., 2018). Induction with IPTG at an OD_{600} of 0.6 and harvesting at an OD_{600} of 6 aimed to maximize protein yield while maintaining cell viability.

SDS-PAGE analysis revealed the presence of an induction band at approximately 47 kDa, corresponding to the expected molecular weight of AADACL2 with the HisTag and TEV site. However, the induction levels were relatively low, and the expression was not robust across all colonies tested. One of the most significant hurdles encountered was the poor solubility of AADACL2 in the *E. coli* expression system. Solubility assays demonstrated that the majority of the recombinant protein was localized in the pellet fraction after cell lysis, indicating the formation of inclusion bodies. Inclusion bodies are common when expressing eukaryotic proteins in bacteria, primarily due to the absence of eukaryotic chaperones and the inability to perform post-translational modifications necessary for proper folding (Fahnert et al., 2004). The initial attempts to solubilize the protein using sonication and resuspension in PBS were insufficient. This suggests that the protein's hydrophobic regions or disulfide bonds might be contributing to aggregation. The absence of soluble protein impeded downstream purification and functional assays, necessitating alternative strategies. To recover functional protein, we employed an in-column renaturation protocol using a urea gradient. The denaturation in 6 M urea followed by gradual renaturation was intended to refold the protein into its native conformation (**Fig. 65**).

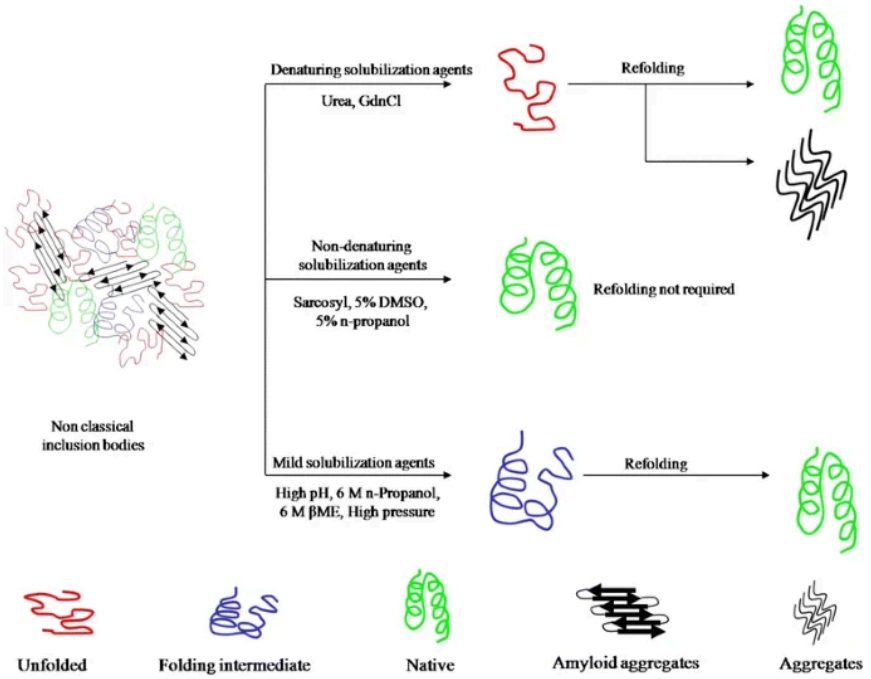


Figure 65. Model showing different solubilization methods used for the recovery of protein from inclusion bodies. Reproduced from Singh et al. (2015).

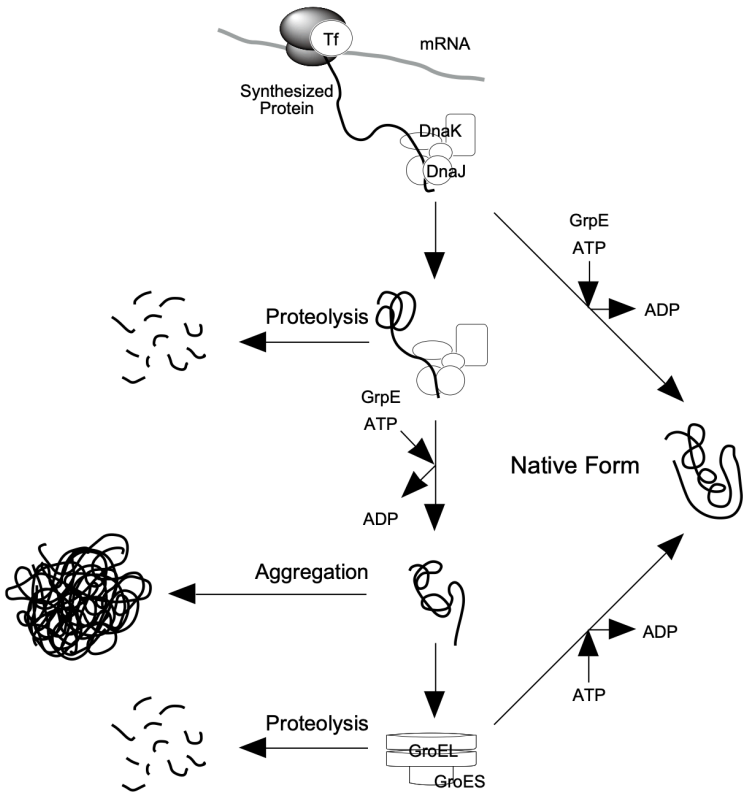


Figure 66. Model for chaperone-assisted protein folding in *E. coli*. Reproduced from Thomas et al. (1997).

While this method yielded minimal amounts of soluble protein, the lack of enzymatic activity in the PNPA assay indicated that the refolded protein was either misfolded or not present in sufficient quantities. Renaturation of proteins from inclusion bodies is inherently challenging and often requires optimization of conditions such as pH, temperature, redox environment, and the presence of folding catalysts (A. Singh et al., 2015; Yamaguchi and Miyazaki, 2014). The fact that only trace amounts of soluble protein were obtained suggests that the refolding conditions were not optimal for AADACL2 or that the protein requires specific co-factors or chaperones absent in the bacterial system. The inability to obtain soluble and active AADACL2 from *E. coli* highlights the limitations of prokaryotic expression systems for certain eukaryotic proteins. We also attempted co-expression with various chaperone systems (**Fig. 66**), including pGRO7 and pgKJE8, but these efforts did not improve solubility.

AADACL2 may require post-translational modifications, such as glycosylation or specific disulfide bond formation, which are not facilitated in bacterial hosts. Additionally, the folding machinery in *E. coli* may be inadequate for complex human proteins, leading to misfolding and aggregation. Considering these challenges, transitioning to a eukaryotic expression system, such as *Pichia pastoris*, is a logical next step. *P. pastoris* offers several advantages, including the ability to perform post-translational modifications, proper protein folding facilitated by eukaryotic chaperones, and high-yield secretion of recombinant proteins (Ahmad et al., 2014). By utilizing a eukaryotic host, we increase the likelihood of obtaining soluble, correctly folded, and functionally active AADACL2. Each step of the experimental design was carefully considered, yet the outcomes underscore the complexity of recombinant protein expression. The initial removal of the signal peptide was a strategic attempt to enhance solubility, but it may have inadvertently affected protein folding or stability. Additionally, the use of the pET-28a(+)-TEV vector and induction conditions were standard for bacterial expression, but perhaps not optimal for AADACL2. The solubility issues point to the intrinsic properties of AADACL2 that are incompatible with the bacterial cytoplasmic environment. This raises questions about the protein's structural requirements and whether alternative fusion tags might have improved solubility. Moreover, the absence of activity observed after renaturation suggests that the protein's functional conformation was not achieved. This could be due to incorrect disulfide bond formation, improper folding of active sites, or the absence of necessary cofactors.

The expression of human AADACL2 in *Pichia pastoris* GS115 cells was undertaken following unsuccessful attempts in *Escherichia coli*, where the protein predominantly formed inclusion bodies and failed to refold into a functional state. This outcome is not uncommon when expressing eukaryotic proteins in prokaryotic systems, as *E. coli* often lacks the necessary

machinery for proper folding and post-translational modifications required by human proteins (Sørensen and Mortensen, 2005). By shifting to *P. pastoris*, we aimed to exploit its eukaryotic expression system, which is capable of performing many of the post-translational modifications found in higher eukaryotes, while still offering the advantages of rapid growth and ease of genetic manipulation (Cereghino and Cregg, 2000). The pPICZ A vector was chosen for intracellular expression under the control of the methanol-inducible AOX1 promoter, allowing tight regulation of protein expression. One of the critical considerations in this study was the potential for glycosylation of AADACL2 in *P. pastoris*. Glycosylation is a common post-translational modification where carbohydrate groups are attached to proteins. In yeast like *P. pastoris*, glycosylation occurs primarily in the secretory pathway, which includes the endoplasmic reticulum (ER) and the Golgi apparatus (Helenius and Aebi, 2002). Proteins destined for secretion are translocated into the ER, where they undergo N-linked glycosylation. As they pass through the Golgi, additional carbohydrate chains can be added, often leading to hyperglycosylation (Montesino et al., 1998). By opting for intracellular expression—excluding the secretion signal—the protein remains in the cytosol rather than entering the ER. The cytosol lacks the enzymatic machinery for N-linked glycosylation. Therefore, proteins expressed intracellularly are less likely to be glycosylated, particularly N-glycosylated. This approach aims to produce proteins that are closer to their native form, minimizing glycosylation to study the protein in a state more representative of its native, unglycosylated form. It also helps preserve the protein's functional domains and activity while making the protein easier to purify due to the absence of heterogeneous glycan structures.

Despite these efforts, SDS-PAGE analysis revealed two protein bands: one at the expected ~47 kDa and another at ~70 kDa (**Fig. 61**). The higher molecular weight band is likely a hyperglycosylated form of AADACL2, suggesting that intracellular expression did not fully prevent glycosylation. The presence of hyperglycosylated AADACL2 indicates that the protein may still be subjected to glycosylation within the endoplasmic reticulum and Golgi apparatus before localization to its intracellular destination. Glycosylation can significantly impact protein solubility, folding, and activity. In some cases, glycosylation may enhance solubility and stability but may also mask functional domains or interfere with interactions (Helenius and Aebi, 2004). The methanol induction system in *P. pastoris* is known to impose metabolic stress on the cells, which can affect protein expression levels and quality. We observed that enzymatic activity peaked at 8 hours post-induction and declined thereafter (**Fig. 62**). This decline could be due to several factors: prolonged induction may lead to increased protease activity within the cells, degrading the recombinant protein; extended exposure to methanol can exhaust cellular resources and energy, reducing the efficiency of

protein synthesis (Zhang et al., 2000); overexpression may lead to misfolding and aggregation, rendering the protein inactive. Additionally, since the amount of AADACL2 in the lysates was not directly quantified, it cannot be excluded that the decrease in enzymatic activity is partly due to variations in enzyme concentration rather than intrinsic loss of function. Optimizing induction conditions, such as methanol concentration, induction time, temperature, and pH, could mitigate these issues. For instance, lower methanol concentrations or pulse feeding strategies might reduce metabolic stress (Cos et al., 2006). Efficient cell lysis is crucial for maximizing protein recovery. Yeast cells have a robust cell wall, making them more resistant to disruption than bacterial cells (Grabski, 2009; Harrison, 1991). We compared sonication, French press, and glass bead disruption methods (**Fig. 63**). Sonication was less effective, likely due to insufficient mechanical force to disrupt the yeast cell wall fully. Glass bead disruption and French press methods were more successful, with glass bead disruption slightly outperforming the French press in releasing soluble protein. These results are consistent with findings reported by Bzducha-Wrobel et al., (Bzducha-Wróbel et al., 2014). The choice of lysis method can impact not only the yield but also the integrity of the protein. Mechanical methods like French press and sonication can generate heat and shear forces, potentially denaturing proteins (Özbek and Ülgen, 2000). Maintaining samples on ice and optimizing processing times are essential to preserve protein activity (Liu et al., 2013). Despite successful expression and detection of enzymatic activity in small-scale experiments, large-scale purification of AADACL2 was unsuccessful. The protein did not bind to the nickel affinity column, and no AADACL2 was detected in the eluted fractions. Several factors could explain this outcome: hyperglycosylation may have masked the C-terminal His-tag, preventing efficient binding to the nickel resin, since glycan chains can sterically hinder the interaction between the His-tag and the immobilized metal ions; the protein may have formed aggregates that were excluded during the purification process and may not interact properly with affinity resins; the His-tag or the entire protein may have been partially degraded by proteases, a common issue in yeast expression systems, since misfolded proteins may expose hydrophobic regions that lead to aggregation or degradation (Bornhorst and Falke, 2000; Zhang et al., 2007).

To overcome these challenges, several strategies could be employed in future studies. (1) Site-directed mutagenesis could be used to alter the Asn282 residue to prevent N-glycosylation. This approach has been successful in reducing unwanted glycosylation in other proteins (Helenius and Aebi, 2004). (2) Mammalian or insect cell expression systems might provide a more native environment for human proteins, potentially improving folding and post-translational modifications. However, these systems are more complex and costly

(Berlec and Štrukelj, 2013). (3) Modification of the expression construct, including a long linker sequence between the protein and the tag may enhance tag accessibility. (4) Including protease inhibitors during cell lysis and purification could prevent degradation of the protein and His-tag, increasing the chances of successful purification (Zhang et al., 2007). An alternative strategy to overcome the challenges associated with the expression and stability of AADACL2 is the use of ancestral sequence reconstruction (ASR). ASR involves computationally inferring the sequences of ancient proteins based on phylogenetic analysis and then synthesizing these ancestral proteins for experimental study. Remarkably, ancestral proteins have been found to exhibit enhanced stability and solubility compared to their modern counterparts. This increased stability is attributed to the adaptation of ancient organisms to extreme environmental conditions, such as higher temperatures, which necessitated more robust protein structures (Akanuma et al., 2013). For example, Risso et al. demonstrated that reconstructed Precambrian β-lactamases displayed hyperstability and increased substrate promiscuity. These ancestral enzymes were more amenable to expression and purification in heterologous systems due to their inherent stability (Risso et al., 2013). Similarly, Hobbs et al. showed that ancestral nucleoside diphosphate kinases had improved thermal stability and catalytic efficiency. By applying ASR to AADACL2, it may be possible to obtain a variant of the protein with superior folding properties, reduced aggregation tendencies, and enhanced expression levels in yeast or bacterial systems (Hobbs et al., 2022). Utilizing ancestral versions of AADACL2 could facilitate its purification and functional characterization by providing a more stable and soluble protein. This approach has the potential to bypass the issues of glycosylation and misfolding observed in modern protein expression systems. Therefore, ASR represents a promising avenue for producing functional AADACL2 and could significantly advance our understanding of its biochemical properties and physiological roles.

Bibliography

Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S.M., Roodveldt, C., Tawfik, D.S., 2005. The “evolvability” of promiscuous protein functions. Nat Genet 37, 73–76. https://doi.org/10.1038/ng1482

Ahmad, M., Hirz, M., Pichler, H., Schwab, H., 2014. Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. Appl Microbiol Biotechnol 98, 5301–5317. https://doi.org/10.1007/s00253-014-5732-5

Akanuma, S., Nakajima, Y., Yokobori, S., Kimura, M., Nemoto, N., Mase, T., Miyazono, K., Tanokura, M., Yamagishi, A., 2013. Experimental evidence for the thermophilicity of ancestral life. Proceedings of the National Academy of Sciences 110, 11067–11072. https://doi.org/10.1073/pnas.1308215110

Akiyama, M., 2021. Acylceramide is a key player in skin barrier function: insight into the molecular mechanisms of skin barrier formation and ichthyosis pathogenesis. The FEBS Journal 288, 2119–2130. https://doi.org/10.1111/febs.15497

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., 2019. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 16, 1315–1322. https://doi.org/10.1038/s41592-019-0598-1

Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., Winther, O., 2017. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33, 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37, 420–423. https://doi.org/10.1038/s41587-019-0036-z

Altenhoff, A.M., Garrayo-Ventas, J., Cosentino, S., Emms, D., Glover, N.M., Hernández-Plaza, A., Nevers, Y., Sundesha, V., Szklarczyk, D., Fernández, J.M., Codó, L., for Orthologs Consortium, the Q., Gelpi, J.L., Huerta-Cepas, J., Iwasaki, W., Kelly, S., Lecompte, O., Muffato, M., Martin, M.J., Capella-Gutierrez, S., Thomas, P.D., Sonnhammer, E., Dessimoz, C., 2020. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Research 48, W538–W545. https://doi.org/10.1093/nar/gkaa308

Altenhoff, A.M., Glover, N.M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T.M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G.H., Dessimoz, C., 2018. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Research 46, D477–D485. https://doi.org/10.1093/nar/gkx1019

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

Anfinsen, C.B., 1973. Principles that Govern the Folding of Protein Chains. Science 181, 223–230. https://doi.org/10.1126/science.181.4096.223

Arnau, J., Lauritzen, C., Petersen, G.E., Pedersen, J., 2006. Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. Protein Expression and Purification 48, 1–13. https://doi.org/10.1016/j.pep.2005.12.002

Asgari, E., Mofrad, M.R.K., 2015. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS ONE 10, e0141287. https://doi.org/10.1371/journal.pone.0141287

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29. https://doi.org/10.1038/75556

Ashrafzadeh, S., Golding, G.B., Ilie, S., Ilie, L., 2024. Scoring alignments by embedding vector similarity. Briefings in Bioinformatics 25, bbae178. https://doi.org/10.1093/bib/bbae178

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R., 2021. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 18, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Babbi, G., Baldazzi, D., Savojardo, C., Martelli, P.L., Casadio, R., 2020. Highlighting Human Enzymes Active in Different Metabolic Pathways and Diseases: The Case Study of EC 1.2.3.1 and EC 2.3.1.9. Biomedicines 8, 250. https://doi.org/10.3390/biomedicines8080250

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent Segmental Duplications in the Human Genome. Science 297, 1003–1007. https://doi.org/10.1126/science.1072047

Bansal, P., Morgat, A., Axelsen, K.B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T.B., Pozzato, M., Blatter, M.-C., Ignatchenko, A., Redaschi, N., Bridge, A., 2022. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Research 50, D693–D700. https://doi.org/10.1093/nar/gkab1016

Benton, M.J., 2014. Vertebrate Palaeontology. John Wiley & Sons.

Berlec, A., Štrukelj, B., 2013. Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. Journal of Industrial Microbiology and Biotechnology 40, 257–274. https://doi.org/10.1007/s10295-013-1235-0

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., Apweiler, R., 2009. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25, 3045–3046. https://doi.org/10.1093/bioinformatics/btp536

Bisong, E., 2019. Google Colaboratory, in: Bisong, E. (Ed.), Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress, Berkeley, CA, pp. 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7

Bornhorst, J.A., Falke, J.J., 2000. [16] Purification of proteins using polyhistidine affinity tags, in: Methods in Enzymology, Applications of Chimeric Genes and Hybrid Proteins Part A: Gene Expression and Protein Purification. Academic Press, pp. 245–254. https://doi.org/10.1016/S0076-6879(00)26058-8

Branden, C.I., Tooze, J., 1998. Introduction to Protein Structure, 2nd ed. Garland Science, New York. https://doi.org/10.1201/9781136969898

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M., 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38, 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

Brettmann, E.A., de Guzman Strong, C., 2018. Recent evolution of the human skin barrier. Experimental Dermatology 27, 859–866. https://doi.org/10.1111/exd.13689

Buljan, M., Frankish, A., Bateman, A., 2010. Quantifying the mechanisms of domain gain in animal proteins. Genome Biol 11, R74. https://doi.org/10.1186/gb-2010-11-7-r74

Bzducha-Wróbel, A., Błażejak, S., Kawarska, A., Stasiak-Różańska, L., Gientka, I., Majewska, E., 2014. Evaluation of the Efficiency of Different Disruption Methods on Yeast Cell Wall

Preparation for β-Glucan Isolation. Molecules 19, 20941–20961. https://doi.org/10.3390/molecules191220941

Callaway, E., 2020. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 588, 203–204. https://doi.org/10.1038/d41586-020-03348-4

Carrillo, H., Lipman, D., 1988. The Multiple Sequence Alignment Problem in Biology. SIAM J. Appl. Math. 48, 1073–1082. https://doi.org/10.1137/0148063

Carroll, S.B., 1995. Homeotic genes and the evolution of arthropods and chordates. Nature 376, 479–485. https://doi.org/10.1038/376479a0

Cereghino, J.L., Cregg, J.M., 2000. Heterologous protein expression in the methylotrophic yeast Pichia pastoris. FEMS Microbiology Reviews 24, 45–66. https://doi.org/10.1111/j.1574-6976.2000.tb00532.x

Chang, J.-M., Floden, E.W., Herrero, J., Gascuel, O., Di Tommaso, P., Notredame, C., 2021. Incorporating alignment uncertainty into Felsenstein's phylogenetic bootstrap to improve its reliability. Bioinformatics 37, 1506–1514. https://doi.org/10.1093/bioinformatics/btz082

Chen, L., DeVries, A.L., Cheng, C.-H.C., 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proceedings of the National Academy of Sciences 94, 3811–3816. https://doi.org/10.1073/pnas.94.8.3811

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface. 15, 20170387. https://doi.org/10.1098/rsif.2017.0387

Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. The EMBO Journal 5, 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x

Clarke, J.T., Lloyd, G.T., Friedman, M., 2016. Little evidence for enhanced phenotypic evolution in early teleosts relative to their living fossil sister group. Proceedings of the National Academy of Sciences 113, 11531–11536. https://doi.org/10.1073/pnas.1607237113

Conant, G.C., Wolfe, K.H., 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet 9, 938–950. https://doi.org/10.1038/nrg2482

Corbin, C.J., Mapes, S.M., Marcos, J., Shackleton, C.H., Morrow, D., Safe, S., Wise, T., Ford, J.J., Conley, A.J., 2004. Paralogues of Porcine Aromatase Cytochrome P450: A Novel Hydroxylase Activity Is Associated with the Survival of a Duplicated Gene. Endocrinology 145, 2157–2164. https://doi.org/10.1210/en.2003-1595

Cos, O., Ramón, R., Montesinos, J.L., Valero, F., 2006. Operational strategies, monitoring and control of heterologous protein production in the methylotrophic yeast Pichia pastoris under different promoters: A review. Microb Cell Fact 5, 17. https://doi.org/10.1186/1475-2859-5-17

Daborn, P.J., Yen, J.L., Bogwitz, M.R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T.G., ffrench-Constant, R.H., 2002. A Single P450 Allele Associated with Insecticide Resistance in Drosophila. Science 297, 2253–2256. https://doi.org/10.1126/science.1074170

Davesne, D., Friedman, M., Schmitt, A.D., Fernandez, V., Carnevale, G., Ahlberg, P.E., Sanchez, S., Benson, R.B.J., 2021. Fossilized cell structures identify an ancient origin for the teleost whole-genome duplication. Proceedings of the National Academy of Sciences 118, e2101780118. https://doi.org/10.1073/pnas.2101780118

Dehal, P., Boore, J.L., 2005. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLoS Biol 3, e314. https://doi.org/10.1371/journal.pbio.0030314

Dobson, C.M., 2003. Protein folding and misfolding. Nature 426, 884–890. https://doi.org/10.1038/nature02261

Donoghue, P., Smith, M., Sannsom, I., 2003. The origin and early evolution of chordates: molecular clocks and the fossil record, in: Donoghue, P., Smith, M. (Eds.), Telling the Evolutionary Time: Molecular Clocks and the Fossil Record. CRC Press, pp. 190–223.

Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S., 2021. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J. Chem. Inf. Model. 61, 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Eisen, J.A., 1998. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Res. 8, 163–167. https://doi.org/10.1101/gr.8.3.163

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B., 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B., 2021. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing.

Ernst, M.D., 2004. Permutation Methods: A Basis for Exact Inference. Statistical Science 19, 676–685.

Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H., Laudet, V., 2006. Neofunctionalization in Vertebrates: The Example of Retinoic Acid Receptors. PLoS Genet 2, e102. https://doi.org/10.1371/journal.pgen.0020102

Espregueira Themudo, G., Alves, L.Q., Machado, A.M., Lopes-Marques, M., da Fonseca, R.R., Fonseca, M., Ruivo, R., Castro, L.F.C., 2020. Losing Genes: The Evolutionary Remodeling of Cetacea Skin. Front. Mar. Sci. 7. https://doi.org/10.3389/fmars.2020.592375

Fahnert, B., Lilie, H., Neubauer, P., 2004. Inclusion Bodies: Formation and Utilisation, in: Physiological Stress Responses in Bioprocesses: -/-. Springer, Berlin, Heidelberg, pp. 93–142. https://doi.org/10.1007/b93995

Fang, Z., Liu, X., Peltz, G., 2023. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics 39, btac757. https://doi.org/10.1093/bioinformatics/btac757

Farinas, E.T., Alcalde, M., Arnold, F., 2004. Alkene epoxidation catalyzed by cytochrome P450 BM-3 139-3. Tetrahedron, Biocatalysts in Synthetic Organic Chemistry 60, 525–528. https://doi.org/10.1016/j.tet.2003.10.099

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17, 368–376. https://doi.org/10.1007/BF01734359

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44, D279–D285. https://doi.org/10.1093/nar/gkv1344

Fischer, J., 2009. Autosomal Recessive Congenital Ichthyosis. Journal of Investigative Dermatology 129, 1319–1321. https://doi.org/10.1038/jid.2009.57

Fischer, M.C., Foll, M., Excoffier, L., Heckel, G., 2011. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (Microtus arvalis). Molecular Ecology 20, 1450–1462. https://doi.org/10.1111/j.1365-294X.2011.05015.x

Fitch, W.M., 1970. Distinguishing Homologous from Analogous Proteins. Systematic Zoology 19, 99. https://doi.org/10.2307/2412448

Flajnik, M.F., Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat Rev Genet 11, 47–59. https://doi.org/10.1038/nrg2703

Flint, A.J., Tiganis, T., Barford, D., Tonks, N.K., 1997. Development of “substrate-trapping” mutants to identify physiological substrates of protein tyrosine phosphatases. Proc. Natl. Acad. Sci. U.S.A. 94, 1680–1685. https://doi.org/10.1073/pnas.94.5.1680

Fluss, R., Faraggi, D., Reiser, B., 2005. Estimation of the Youden Index and its Associated Cutoff Point. Biometrical Journal 47, 458–472. https://doi.org/10.1002/bimj.200410135

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., Postlethwait, J., 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151, 1531–1545. https://doi.org/10.1093/genetics/151.4.1531

Fox, J.D., Waugh, D.S., 2003. Maltose-Binding Protein as a Solubility Enhancer, in: Vaillancourt, P.E. (Ed.), E. coliGene Expression Protocols. Humana Press, Totowa, NJ, pp. 99–117. https://doi.org/10.1385/1-59259-301-1:99

Freeling, M., Thomas, B.C., 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16, 805–814. https://doi.org/10.1101/gr.3681406

Gabaldón, T., Koonin, E.V., 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet 14, 360–366. https://doi.org/10.1038/nrg3456

Gopal, G.J., Kumar, A., 2013. Strategies for the Production of Recombinant Protein in Escherichia coli. Protein J 32, 419–425. https://doi.org/10.1007/s10930-013-9502-5

Gouet, P., Courcelle, E., Stuart, D.I., M√©toz, F., 1999. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics 15, 305–308. https://doi.org/10.1093/bioinformatics/15.4.305

Grabski, A.C., 2009. Chapter 18 Advances in Preparation of Biological Extracts for Protein Purification, in: Burgess, R.R., Deutscher, M.P. (Eds.), Methods in Enzymology, Guide to Protein Purification, 2nd Edition. Academic Press, pp. 285–303. https://doi.org/10.1016/S0076-6879(09)63018-4

Grant, D.M., 1991. Detoxification pathways in the liver. J Inherit Metab Dis 14, 421–430. https://doi.org/10.1007/BF01797915

Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., Aerts, A., Altherr, M., Ashworth, L., Bajorek, E., Black, S., Branscomb, E., Caenepeel, S., Carrano, A., Caoile, C., Man Chan, Y., Christensen, M., Cleland, C.A., Copeland, A., Dalin, E., Dehal, P., Denys, M., Detter, J.C., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Georgescu, A.M., Glavina, T., Gomez, M., Gonzales, E., Groza, M., Hammon, N., Hawkins, T., Haydu, L., Ho, I., Huang, W., Israni, S., Jett, J., Kadner, K., Kimball, H., Kobayashi, A., Larionov, V., Leem, S.-H., Lopez, F., Lou, Y., Lowry, S., Malfatti, S., Martinez, D., McCready, P., Medina, C., Morgan, J., Nelson, K., Nolan, M., Ovcharenko, I., Pitluck, S., Pollard, M., Popkie, A.P., Predki, P., Quan, G., Ramirez, L., Rash, S., Retterer, J., Rodriguez, A., Rogers, S., Salamov, A., Salazar, A., She, X., Smith, D., Slezak, T., Solovyev, V., Thayer, N., Tice, H., Tsai, M., Ustaszewska, A., Vo, N., Wagner, M., Wheeler, J., Wu, K., Xie, G., Yang, J., Dubchak, I., Furey, T.S., DeJong, P., Dickson, M., Gordon, D., Eichler, E.E., Pennacchio, L.A., Richardson, P., Stubbs, L., Rokhsar, D.S., Myers, R.M., Rubin, E.M., Lucas, S.M., 2004. The DNA sequence and biology of human chromosome 19. Nature 428, 529–535. https://doi.org/10.1038/nature02399

Gu, Z., Nicolae, D., Lu, H.H.-S., Li, W.-H., 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in Genetics 18, 609–613. https://doi.org/10.1016/S0168-9525(02)02837-8

Gunasekaran, K., Ma, B., Nussinov, R., 2004. Is allostery an intrinsic property of all dynamic proteins? Proteins: Structure, Function, and Bioinformatics 57, 433–443. https://doi.org/10.1002/prot.20232

Hardison, R., 1998. Hemoglobins From Bacteria to Man: Evolution of Different Patterns of Gene Expression. Journal of Experimental Biology 201, 1099–1117. https://doi.org/10.1242/jeb.201.8.1099

Harrison, P.W., Amode, M.R., Austine-Orimoloye, O., Azov, A.G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Boddu, S., Branco Lins, P.R., Brooks, L., Ramaraju, S.B., Campbell, L.I., Martinez, M.C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N.H., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, T.E., Fatima, R., Giron, C.G., Genez, T., Grigoriadis, D., Ghattaoraya, G.S., Martinez, J.G., Gurbich, T.A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha, D., Marques-Coelho, D., Maslen, G., Merino, G.A., Mirabueno, L.P., Mushtaq, A., Hossain, S.N., Ogeh, D.N., Sakthivel, M.P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J.G., Salam, A.I.A., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F.F., Urbina-Gómez, D., Veidenberg, A., Walsh, T.A., Ware, D., Wass, E., Willhoft, N.L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G.R., Keatley, J., Loveland, J.E., Moore, B., Mudge, J.M., Naamati, G., Tate, J., Trevanion, S.J., Winterbottom, A., Frankish, A., Hunt, S.E., Cunningham, F., Dyer, S., Finn, R.D., Martin, F.J., Yates, A.D., 2024. Ensembl 2024. Nucleic Acids Research 52, D891–D899. https://doi.org/10.1093/nar/gkad1049

Harrison, S.T.L., 1991. Bacterial cell disruption: A key unit operation in the recovery of intracellular products. Biotechnology Advances 9, 217–240. https://doi.org/10.1016/0734-9750(91)90005-G

Hassanzadeh, H.R., Wang, M.D., 2016. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Presented at the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, Shenzhen, China, pp. 178–183. https://doi.org/10.1109/BIBM.2016.7822515

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations, Springer series in statistics. Springer, New York.

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., Rost, B., 2019. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20, 723. https://doi.org/10.1186/s12859-019-3220-8

Helenius, A., Aebi, and M., 2001. Intracellular Functions of N-Linked Glycans. Science 291, 2364–2369. https://doi.org/10.1126/science.291.5512.2364

Helenius, A., Aebi, M., 2004. Roles of N-Linked Glycans in the Endoplasmic Reticulum. Annual Review of Biochemistry 73, 1019–1049. https://doi.org/10.1146/annurev.biochem.73.011303.073752

Helenius, J., Aebi, M., 2002. Transmembrane movement of dolichol linked carbohydrates during N-glycoprotein biosynthesis in the endoplasmic reticulum. Semin Cell Dev Biol 13, 171–178. https://doi.org/10.1016/s1084-9521(02)00045-9

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

Hobbs, H.T., Shah, N.H., Shoemaker, S.R., Amacher, J.F., Marqusee, S., Kuriyan, J., 2022. Saturation mutagenesis of a predicted ancestral Syk-family kinase. Protein Science 31, e4411. https://doi.org/10.1002/pro.4411

Holland, P.W.H., Garcia-Fernàndez, J., Williams, N.A., Sidow, A., 1994. Gene duplications and the origins of vertebrate development. Development 1994, 125–133. https://doi.org/10.1242/dev.1994.Supplement.125

Hollingsworth, S.A., Dror, R.O., 2018. Molecular Dynamics Simulation for All. Neuron 99, 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Hughes, A.L., 1999. Phylogenies of Developmentally Important Proteins Do Not Support the Hypothesis of Two Rounds of Genome Duplication Early in Vertebrate History. J Mol Evol 48, 565–576. https://doi.org/10.1007/PL00006499

Hughes, A.L., 1997. The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society of London. Series B: Biological Sciences 256, 119–124. https://doi.org/10.1098/rspb.1994.0058

Innan, H., Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11, 97–108. https://doi.org/10.1038/nrg2689

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, 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., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Just, W., 2001. Computational Complexity of Multiple Sequence Alignment with SP-Score. Journal of Computational Biology 8, 615–623. https://doi.org/10.1089/106652701753307511

Kaessmann, H., Vinckenbosch, N., Long, M., 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10, 19–31. https://doi.org/10.1038/nrg2487

Kafri, R., Springer, M., Pilpel, Y., 2009. Genetic Redundancy: New Tricks for Old Genes. Cell 136, 389–392. https://doi.org/10.1016/j.cell.2009.01.027

Kaitaniemi, S., Elovaara, H., Grön, K., Kidron, H., Liukkonen, J., Salminen, T., Salmi, M., Jalkanen, S., Elima, K., 2009. The unique substrate specificity of human AOC2, a semicarbazide-sensitive amine oxidase. Cell. Mol. Life Sci. 66, 2743–2757. https://doi.org/10.1007/s00018-009-0076-5

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28, 27–30.

Kaur, Jashandeep, Kumar, A., Kaur, Jagdeep, 2018. Strategies for optimization of heterologous protein expression in *E. coli*: Roadblocks and reinforcements. International Journal of Biological Macromolecules 106, 803–822. https://doi.org/10.1016/j.ijbiomac.2017.08.080

Kawasaki, K., Lafont, A.-G., Sire, J.-Y., 2011. The Evolution of Milk Casein Genes from Tooth Genes before the Origin of Mammals. Molecular Biology and Evolution 28, 2053–2061. https://doi.org/10.1093/molbev/msr020

Khan, I., Maldonado, E., Vasconcelos, V., O'Brien, S.J., Johnson, W.E., Antunes, A., 2014. Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments. BMC Genomics 15, 779. https://doi.org/10.1186/1471-2164-15-779

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. Nature 217, 624–626. https://doi.org/10.1038/217624a0

Kirch, W. (Ed.), 2008. Pearson's Correlation Coefficient, in: Encyclopedia of Public Health. Springer Netherlands, Dordrecht, pp. 1090–1091. https://doi.org/10.1007/978-1-4020-5614-7_2569

Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H., 2018. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 8, 10872. https://doi.org/10.1038/s41598-018-28948-z

Kobayashi, Y., Fukami, T., Nakajima, A., Watanabe, A., Nakajima, M., Yokoi, T., 2012. Species differences in tissue distribution and enzyme activities of arylacetamide deacetylase in human, rat, and mouse. Drug Metab Dispos 40, 671–679. https://doi.org/10.1124/dmd.111.043067

Kondrashov, F.A., Kondrashov, A.S., 2006. Role of selection in fixation of gene duplications. Journal of Theoretical Biology, Special Issue in Memory of John Maynard Smith 239, 141–151. https://doi.org/10.1016/j.jtbi.2005.08.033

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of gene duplications. Genome Biol 3, research0008.1. https://doi.org/10.1186/gb-2002-3-2-research0008

Koonin, E.V., 2005. Orthologs, Paralogs, and Evolutionary Genomics1. Annual Review of Genetics 39, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Kratz, E., Dugas, J.C., Ngai, J., 2002. Odorant receptor gene regulation: implications from genomic organization. Trends in Genetics 18, 29–34. https://doi.org/10.1016/S0168-9525(01)02579-3

Kraut, J., 1988. How Do Enzymes Work? Science 242, 533–540. https://doi.org/10.1126/science.3051385

Krivák, R., Hoksza, D., 2018. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform 10, 39. https://doi.org/10.1186/s13321-018-0285-8

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M., 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Krynetski, E.Y., Schuetz, J.D., Galpin, A.J., Pui, C.H., Relling, M.V., Evans, W.E., 1995. A single point mutation leading to loss of catalytic activity in human thiopurine S-methyltransferase. Proceedings of the National Academy of Sciences 92, 949–953. https://doi.org/10.1073/pnas.92.4.949

Kryuchkova-Mostacci, N., Robinson-Rechavi, M., 2015. Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. PLOS ONE 10, e0131673. https://doi.org/10.1371/journal.pone.0131673

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution 35, 1547–1549. https://doi.org/10.1093/molbev/msy096

Kuwada, Y., 1911. Maiosis in the Pollen Mother Cells of Zea Mays L. (With Plate V.). 植物学雑誌 25, en163–en181. https://doi.org/10.15281/jplantres1887.25.294_163

Lallemand, T., Leduc, M., Landès, C., Rizzon, C., Lerat, E., 2020. An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice. Genes (Basel) 11, 1046. https://doi.org/10.3390/genes11091046

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A., Robles, V., 2006. Machine learning in bioinformatics. Briefings in Bioinformatics 7, 86–112. https://doi.org/10.1093/bib/bbk007

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lesk, A.M., Chothia, C., 1980. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. Journal of Molecular Biology 136, 225–270. https://doi.org/10.1016/0022-2836(80)90373-3

Li, W.-H., Gojobori, T., Nei, M., 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292, 237–239. https://doi.org/10.1038/292237a0

Li, W.H., Wu, C.I., Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2, 150–174. https://doi.org/10.1093/oxfordjournals.molbev.a040343

Li, W.-H., Yang, J., Gu, X., 2005. Expression divergence between duplicate genes. Trends in Genetics 21, 602–607. https://doi.org/10.1016/j.tig.2005.08.006

Li, Y., Wu, F.-X., Ngom, A., 2016. A review on machine learning principles for multi-view biological data integration. Brief Bioinform bbw113. https://doi.org/10.1093/bib/bbw113

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat Rev Genet 16, 321–332. https://doi.org/10.1038/nrg3920

Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., Rost, B., 2021. Protein embeddings and deep learning predict binding residues for various ligand classes. Sci Rep 11, 23916. https://doi.org/10.1038/s41598-021-03431-4

Liu, D., Zeng, X.-A., Sun, D.-W., Han, Z., 2013. Disruption and protein release by ultrasonication of yeast cells. Innovative Food Science & Emerging Technologies 18, 132–137. https://doi.org/10.1016/j.ifset.2013.02.006

Long, M., Betrán, E., Thornton, K., Wang, W., 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4, 865–875. https://doi.org/10.1038/nrg1204

Long, M., Langley, C.H., 1993. Natural Selection and the Origin of jingwei, a Chimeric Processed Functional Gene in Drosophila. Science 260, 91–95. https://doi.org/10.1126/science.7682012

Luis Villanueva-Cañas, J., Ruiz-Orera, J., Agea, M.I., Gallo, M., Andreu, D., Albà, M.M., 2017. New Genes and Functional Innovation in Mammals. Genome Biol Evol 9, 1886–1900. https://doi.org/10.1093/gbe/evx136

Lynch, M., Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290, 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Macauley-Patrick, S., Fazenda, M.L., McNeil, B., Harvey, L.M., 2005. Heterologous protein production using the Pichia pastoris expression system. Yeast 22, 249–270. https://doi.org/10.1002/yea.1208

Maddison, W.P., 1997. Gene Trees in Species Trees. Systematic Biology 46, 523–536. https://doi.org/10.1093/sysbio/46.3.523

MAGADUM, S., BANERJEE, U., MURUGAN, P., GANGAPUR, D., RAVIKESAVAN, R., 2013. Gene duplication as a major force in evolution. J Genet 92, 155–161. https://doi.org/10.1007/s12041-013-0212-8

Malatesta, M., Fornasier, E., Di Salvo, M.L., Tramonti, A., Zangelmi, E., Peracchi, A., Secchi, A., Polverini, E., Giachin, G., Battistutta, R., Contestabile, R., Percudani, R., 2024. One substrate many enzymes virtual screening uncovers missing genes of carnitine biosynthesis in human and mouse. Nat Commun 15, 3199. https://doi.org/10.1038/s41467-024-47466-3

Marekov, L.N., Steinert, P.M., 1998. Ceramides Are Bound to Structural Proteins of the Human Foreskin Epidermal Cornified Cell Envelope*. Journal of Biological Chemistry 273, 17763–17770. https://doi.org/10.1074/jbc.273.28.17763

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., Kaessmann, H., 2005. Emergence of Young Human Genes after a Burst of Retroposition in Primates. PLOS Biology 3, e357. https://doi.org/10.1371/journal.pbio.0030357

Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F., Šali, A., 2000. Comparative Protein Structure Modeling of Genes and Genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325. https://doi.org/10.1146/annurev.biophys.29.1.291

Marzluff, W.F., Gongidi, P., Woods, K.R., Jin, J., Maltais, L.J., 2002. The Human and Mouse Replication-Dependent Histone Genes. Genomics 80, 487–498. https://doi.org/10.1006/geno.2002.6850

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space.

Mintseris, J., Weng, Z., 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. U.S.A. 102, 10930–10935. https://doi.org/10.1073/pnas.0502667102

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., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H.-Y., El-Gebali, S., Fraser, M.I., Gough, J., Haft, D.R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A.P., Paysan-Lafosse, T., Pesseat, S., Potter, S.C., Qureshi, M.A., Rawlings, N.D., Redaschi, N., Richardson, L.J., Rivoire, C., Salazar, G.A., Sangrador-Vegas, A., Sigrist, C.J.A., Sillitoe, I., Sutton, G.G., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Yong, S.-Y., Finn, R.D., 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research 47, D351–D360. https://doi.org/10.1093/nar/gky1100

Montesino, R., García, R., Quintero, O., Cremata, J.A., 1998. Variation in N-Linked Oligosaccharide Structures on Heterologous Proteins Secreted by the Methylotrophic Yeast*Pichia pastoris*. Protein Expression and Purification 14, 197–207. https://doi.org/10.1006/prep.1998.0933

Muffato, M., Louis, A., Nguyen, N.T.T., Lucas, J., Berthelot, C., Roest Crollius, H., 2023. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. Nat Ecol Evol 7, 355–366. https://doi.org/10.1038/s41559-022-01956-z

Muñoz-Garcia, A., Thomas, C.P., Keeney, D.S., Zheng, Y., Brash, A.R., 2014. The importance of the lipoxygenase-hepoxilin pathway in the mammalian epidermal barrier. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, The Important Role of Lipids in the Epidermis and their Role in the Formation and Maintenance of the Cutaneous Barrier 1841, 401–408. https://doi.org/10.1016/j.bbalip.2013.08.020

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S.L., 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8, e1002764. https://doi.org/10.1371/journal.pgen.1002764

Nagaoka, M., Sakai, Y., Nakajima, M., Fukami, T., 2024. Role of carboxylesterase and arylacetamide deacetylase in drug metabolism, physiology, and pathology. Biochemical Pharmacology 223, 116128. https://doi.org/10.1016/j.bcp.2024.116128

Nathans, J., Thomas, D., Hogness, D.S., 1986. Molecular Genetics of Human Color Vision: The Genes Encoding Blue, Green, and Red Pigments. Science 232, 193–202. https://doi.org/10.1126/science.2937147

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48, 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nemes, Z., Marekov, L.N., Fésüs, L., Steinert, P.M., 1999. A novel function for transglutaminase 1: Attachment of long-chain ω-hydroxyceramides to involucrin by ester bond formation. Proceedings of the National Academy of Sciences 96, 8402–8407. https://doi.org/10.1073/pnas.96.15.8402

Ng, P.C., Henikoff, S., 2006. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu. Rev. Genom. Hum. Genet. 7, 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Nourbakhsh, M., Douglas, D.N., Pu, C.H., Lewis, J.T., Kawahara, T., Lisboa, L.F., Wei, E., Asthana, S., Quiroga, A.D., Law, L.M.J., Chen, C., Addison, W.R., Nelson, R., Houghton, M., Lehner, R., Kneteman, N.M., 2013. Arylacetamide deacetylase: A novel host factor with important roles in the lipolysis of cellular triacylglycerol stores, VLDL assembly and HCV production. Journal of Hepatology 59, 336–343. https://doi.org/10.1016/j.jhep.2013.03.022

Ohno, S., 1970. Evolution by Gene Duplication. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-86659-3

Ohno, Y., Nakamura, T., Iwasaki, T., Katsuyama, A., Ichikawa, S., Kihara, A., 2023. Determining the structure of protein-bound ceramides, essential lipids for skin barrier function. iScience 26, 108248. https://doi.org/10.1016/j.isci.2023.108248

Ohta, T., 1987. Simulating Evolution by Gene Duplication. Genetics 115, 207–213. https://doi.org/10.1093/genetics/115.1.207

Orengo, C.A., Todd, A.E., Thornton, J.M., 1999. From protein structure to function. Current Opinion in Structural Biology 9, 374–382. https://doi.org/10.1016/S0959-440X(99)80051-7

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L.L., 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research 38, D196–D203. https://doi.org/10.1093/nar/gkp931

Oue, S., Okamoto, A., Yano, T., Kagamiyama, H., 1999. Redesigning the Substrate Specificity of an Enzyme by Cumulative Effects of the Mutations of Non-active Site Residues. Journal of Biological Chemistry 274, 2344–2349. https://doi.org/10.1074/jbc.274.4.2344

Ovaere, P., Lippens, S., Vandenabeele, P., Declercq, W., 2009. The emerging roles of serine protease cascades in the epidermis. Trends in Biochemical Sciences 34, 453–463. https://doi.org/10.1016/j.tibs.2009.08.001

Özbek, B., Ülgen, K.Ö., 2000. The stability of enzymes after sonication. Process Biochemistry 35, 1037–1043. https://doi.org/10.1016/S0032-9592(00)00141-2

Papp, B., Pál, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197. https://doi.org/10.1038/nature01771

Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling — a review. Gene 238, 103–114. https://doi.org/10.1016/S0378-1119(99)00228-0

Pearlman, S.M., Serber, Z., Ferrell, J.E., 2011. A Mechanism for the Evolution of Phosphorylation Sites. Cell 147, 934–946. https://doi.org/10.1016/j.cell.2011.08.052

Peichel, C.L., Ross, J.A., Matson, C.K., Dickson, M., Grimwood, J., Schmutz, J., Myers, R.M., Mori, S., Schluter, D., Kingsley, D.M., 2004. The Master Sex-Determination Locus in Threespine Sticklebacks Is on a Nascent Y Chromosome. Current Biology 14, 1416–1424. https://doi.org/10.1016/j.cub.2004.08.030

Peregrín-Alvarez, J.M., Sanford, C., Parkinson, J., 2009. The conservation and evolutionary modularity of metabolism. Genome Biol 10, R63. https://doi.org/10.1186/gb-2009-10-6-r63

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C., Stone, A.C., 2007. Diet and the evolution of

human amylase gene copy number variation. Nat Genet 39, 1256–1260. https://doi.org/10.1038/ng2123

Pervaiz, N., Shakeel, N., Qasim, A., Zehra, R., Anwar, S., Rana, N., Xue, Y., Zhang, Z., Bao, Y., Abbasi, A.A., 2019. Evolutionary history of the human multigene families reveals widespread gene duplications throughout the history of animals. BMC Evol Biol 19, 128. https://doi.org/10.1186/s12862-019-1441-0

Phillips, A., Janies, D., Wheeler, W., 2000. Multiple Sequence Alignment in Phylogenetic Analysis. Molecular Phylogenetics and Evolution 16, 317–330. https://doi.org/10.1006/mpev.2000.0785

Powers, D.M.W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. https://doi.org/10.48550/ARXIV.2010.16061

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A., 2020. Transformer protein language models are unsupervised structure learners. https://doi.org/10.1101/2020.12.15.422761

Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A., 2021. MSA Transformer, in: Proceedings of the 38th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 8844–8856.

Ravi, V., Venkatesh, B., 2008. Rapidly evolving fish genomes and teleost diversity. Current Opinion in Genetics & Development, Genomes and evolution 18, 544–550. https://doi.org/10.1016/j.gde.2008.11.001

Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B.J., Chaisson, M., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S.C., Lama, T.M., Grutzner, F., Warren, W.C., Balakrishnan, C.N., Burt, D., George, J.M., Biegler, M.T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T., Wilkinson, M., Turner, G., Meyer, A., Kautt, A.F., Franchini, P., Detrich, H.W., Svardal, H., Wagner, M., Naylor, G.J.P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B.T., Pesout, T., Houck, M., Misuraca, A., Kingan, S.B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R.E., Putnam, N.H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R.E., Guan, D., London, S.E., Clayton, D.F., Mello, C.V., Friedrich, S.R., Lovell, P.V., Osipova, E., Al-Ajli, F.O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R.S., Makova, K.D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, Kevin, Flicek, P., Walenz, B.P., Kwak, W., Clawson, H., Diekhans, M., Nassar, L., Paten, B., Kraus, R.H.S., Crawford, A.J., Gilbert, M.T.P., Zhang, G., Venkatesh, B., Murphy, R.W., Koepfli, K.-P., Shapiro, B., Johnson, W.E., Di Palma, F., Marques-Bonet, T., Teeling, E.C., Warnow, T., Graves, J.M., Ryder, O.A., Haussler, D., O'Brien, S.J., Korlach, J., Lewin, H.A., Howe, Kerstin, Myers, E.W., Durbin, R., Phillippy, A.M., Jarvis, E.D., 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., Allen, J., Asri, M., Bzikadze, A.V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., Garcia Giron, C., Garrison, E., Gershman, A., Gerton, J.L., Grady, P.G.S., Guarracino, A., Haggerty, L., Halabian, R., Hansen, N.F., Harris, R., Hartley, G.A., Harvey, W.T., Haukness, M., Heinz, J., Hourlier, T., Hubley, R.M., Hunt, S.E., Hwang, S., Jain, M., Kesharwani, R.K., Lewis, A.P., Li, H., Logsdon, G.A., Lucas, J.K., Makalowski, W., Markovic, C., Martin, F.J., Mc Cartney, A.M., McCoy, R.C., McDaniel, J., McNulty, B.M., Medvedev, P., Mikheenko, A., Munson, K.M., Murphy, T.D., Olsen, H.E., Olson, N.D., Paulin, L.F., Porubsky, D., Potapova, T., Ryabov, F., Salzberg, S.L., Sauria, M.E.G., Sedlazeck, F.J., Shafin, K., Shepelev, V.A., Shumate, A., Storer, J.M., Surapaneni, L., Taravella Oill, A.M., Thibaud-Nissen, F., Timp, W.,

Tomaszkiewicz, M., Vollger, M.R., Walenz, B.P., Watwood, A.C., Weissensteiner, M.H., Wenger, A.M., Wilson, M.A., Zarate, S., Zhu, Y., Zook, J.M., Eichler, E.E., O'Neill, R.J., Schatz, M.C., Miga, K.H., Makova, K.D., Phillippy, A.M., 2023. The complete sequence of a human Y chromosome. Nature 621, 344–354. https://doi.org/10.1038/s41586-023-06457-y

Risso, V.A., Gavira, J.A., Mejia-Carmona, D.F., Gaucher, E.A., Sanchez-Ruiz, J.M., 2013. Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian β-Lactamases. J. Am. Chem. Soc. 135, 2899–2902. https://doi.org/10.1021/ja311630a

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U.S.A. 118, e2016239118. https://doi.org/10.1073/pnas.2016239118

Romanovsky, A.A., 2014. Skin temperature: its role in thermoregulation. Acta Physiologica 210, 498–507. https://doi.org/10.1111/apha.12231

Rosano, G.L., Ceccarelli, E.A., 2014. Recombinant protein expression in Escherichia coli: advances and challenges. Front. Microbiol. 5. https://doi.org/10.3389/fmicb.2014.00172

Rui, L., 2014. Energy Metabolism in the Liver. Compr Physiol 4, 177–197. https://doi.org/10.1002/cphy.c130024

Schrödinger, LLC, 2015c. The PyMOL Molecular Graphics System, Version 1.8.

Schwede, T., 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31, 3381–3385. https://doi.org/10.1093/nar/gkg520

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning Important Features Through Propagating Activation Differences, in: Proceedings of the 34th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 3145–3153.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7, 539. https://doi.org/10.1038/msb.2011.75

Singh, A., Upadhyay, V., Upadhyay, A.K., Singh, S.M., Panda, A.K., 2015. Protein recovery from inclusion bodies of Escherichia coli using mild solubilization process. Microbial Cell Factories 14, 41. https://doi.org/10.1186/s12934-015-0222-8

Singh, P.P., Arora, J., Isambert, H., 2015. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. PLOS Computational Biology 11, e1004394. https://doi.org/10.1371/journal.pcbi.1004394

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Sørensen, H.P., Mortensen, K.K., 2005. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. Microb Cell Fact 4, 1. https://doi.org/10.1186/1475-2859-4-1

Spence, M.A., Kaczmarski, J.A., Saunders, J.W., Jackson, C.J., 2021. Ancestral sequence reconstruction for protein engineers. Current Opinion in Structural Biology, Engineering and Design ● Membranes 69, 131–141. https://doi.org/10.1016/j.sbi.2021.04.001

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stärk, H., Dallago, C., Heinzinger, M., Rost, B., 2021. Light attention predicts protein location from the language of life. Bioinformatics Advances 1, vbab035. https://doi.org/10.1093/bioadv/vbab035

Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T.-B.G., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., Gupta, R., Paul Bennett, E., Mandel, U., Brunak, S., Wandall, H.H., Levery, S.B., Clausen, H., 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32, 1478–1488. https://doi.org/10.1038/emboj.2013.79

Strasser, B., Mlitz, V., Hermann, M., Rice, R.H., Eigenheer, R.A., Alibardi, L., Tschachler, E., Eckhart, L., 2014. Evolutionary origin and diversification of epidermal barrier proteins in amniotes. Mol Biol Evol 31, 3194–3205. https://doi.org/10.1093/molbev/msu251

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., Inouye, M., 1966. Frameshift Mutations and the Genetic Code. Cold Spring Harb Symp Quant Biol 31, 77–84. https://doi.org/10.1101/SQB.1966.031.01.014

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., the UniProt Consortium, 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932. https://doi.org/10.1093/bioinformatics/btu739

Taylor, T.D., Noguchi, H., Totoki, Y., Toyoda, A., Kuroki, Y., Dewar, K., Lloyd, C., Itoh, T., Takeda, T., Kim, D.-W., She, X., Barlow, K.F., Bloom, T., Bruford, E., Chang, J.L., Cuomo, C.A., Eichler, E., FitzGerald, M.G., Jaffe, D.B., LaButti, K., Nicol, R., Park, H.-S., Seaman, C., Sougnez, C., Yang, X., Zimmer, A.R., Zody, M.C., Birren, B.W., Nusbaum, C., Fujiyama, A., Hattori, M., Rogers, J., Lander, E.S., Sakaki, Y., 2006. Human chromosome 11 DNA sequence and analysis including novel gene identification. Nature 440, 497–500. https://doi.org/10.1038/nature04632

Teichmann, S.A., Babu, M.M., 2004. Gene regulatory network growth by duplication. Nat Genet 36, 492–496. https://doi.org/10.1038/ng1340

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L.J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G., Raj, S., Raposo, P., Rice, D.L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A.J., Aimo, L., Argoud-Puy, G., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Batista Neto, T.M., Blatter, M.-C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., De Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Sundaram, S., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Zhang, J., 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51, D523–D531. https://doi.org/10.1093/nar/gkac1052

Thomas, J.G., Ayling, A., Baneyx, F., 1997. Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from E. coli. Appl Biochem Biotechnol 66, 197–238. https://doi.org/10.1007/BF02785589

Pauling, L., Zuckerkandl, E., Henriksen, T., Lövstad, R., 1963. Chemical Paleogenetics. Molecular "Restoration Studies" of Extinct Forms of Life. Acta Chem. Scand. 17 supl., 9–16. https://doi.org/10.3891/acta.chem.scand.17s-0009

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., Tawfik, D.S., 2007. The Stability Effects of Protein Mutations Appear to be Universally Distributed. Journal of Molecular Biology 369, 1318–1332. https://doi.org/10.1016/j.jmb.2007.03.069

Tóth-Petróczy, Á., Tawfik, D.S., 2011. Slow protein evolutionary rates are dictated by surface–core association. Proceedings of the National Academy of Sciences 108, 11151–11156. https://doi.org/10.1073/pnas.1015994108

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, 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., Hassabis, D., 2021. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596. https://doi.org/10.1038/s41586-021-03828-1

Tyzack, J.D., Furnham, N., Sillitoe, I., Orengo, C.M., Thornton, J.M., 2017. Understanding enzyme function evolution from a computational perspective. Current Opinion in Structural Biology, Protein–nucleic acid interactions • Catalysis and regulation 47, 131–139. https://doi.org/10.1016/j.sbi.2017.08.003

Uchida, Y., Holleran, W.M., 2008. Omega-*O*-acylceramide, a lipid essential for mammalian survival. Journal of Dermatological Science 51, 77–87. https://doi.org/10.1016/j.jdermsci.2008.01.002

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., Pontén, F., 2015. Tissue-based map of the human proteome. Science 347, 1260419. https://doi.org/10.1126/science.1260419

Van De Peer, Y., Maere, S., Meyer, A., 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet 10, 725–732. https://doi.org/10.1038/nrg2600

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E., 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19, 327–335. https://doi.org/10.1101/gr.073585.107

Warden, S.M., Richardson, C., O'Donnell, J., Stapleton, D., Kemp, B.E., Witters, L.A., 2001. Post-translational modifications of the β-1 subunit of AMP-activated protein kinase affect enzyme activity and cellular localization. Biochemical Journal 354, 275–283. https://doi.org/10.1042/bj3540275

Watanabe, A., Fukami, T., Nakajima, M., Takamiya, M., Aoki, Y., Yokoi, T., 2009. Human Arylacetamide Deacetylase Is a Principal Enzyme in Flutamide Hydrolysis. Drug Metab Dispos 37, 1513–1520. https://doi.org/10.1124/dmd.109.026567

Webb, B., Sali, A., 2016. Comparative Protein Structure Modeling Using MODELLER. CP in Bioinformatics 54. https://doi.org/10.1002/cpbi.3

Wertz, P.W., 2021. Lipid Metabolic Events Underlying the Formation of the Corneocyte Lipid Envelope. Skin Pharmacol Physiol 34, 38–50. https://doi.org/10.1159/000513261

Wolfe, K., 2000. Robustness—it's not where you think it is. Nat Genet 25, 3–4. https://doi.org/10.1038/75560

Yamaguchi, H., Miyazaki, M., 2014. Refolding Techniques for Recovering Biologically Active Recombinant Proteins from Inclusion Bodies. Biomolecules 4, 235–251. https://doi.org/10.3390/biom4010235

Yang, K.K., Wu, Z., Arnold, F.H., 2019. Machine-learning-guided directed evolution for protein engineering. Nat Methods 16, 687–694. https://doi.org/10.1038/s41592-019-0496-6

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., Nielsen, R., 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution 17, 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236

Yao, J., Chen, X., Zheng, F., Zhan, C.-G., 2018. Catalytic reaction mechanism for drug metabolism in human carboxylesterase-1: Cocaine hydrolysis pathway. Mol Pharm 15, 3871–3880. https://doi.org/10.1021/acs.molpharmaceut.8b00354

Yokoyama, S., 2002. Molecular evolution of color vision in vertebrates. Gene 300, 69–78. https://doi.org/10.1016/S0378-1119(02)00845-4

Yoshikuni, Y., Ferrin, T.E., Keasling, J.D., 2006. Designed divergent evolution of enzyme function. Nature 440, 1078–1082. https://doi.org/10.1038/nature04607

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yu, H., Ma, S., Li, Y., Dalby, P.A., 2022. Hot spots-making directed evolution easier. Biotechnology Advances 56, 107926. https://doi.org/10.1016/j.biotechadv.2022.107926

Zakon, H.H., Lu, Y., Zwickl, D.J., Hillis, D.M., 2006. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: Convergent molecular evolution. Proc. Natl. Acad. Sci. U.S.A. 103, 3675–3680. https://doi.org/10.1073/pnas.0600160103

Zallot, R., Harrison, K.J., Kolaczkowski, B., de Crécy-Lagard, V., 2016. Functional annotations of paralogs: a blessing and a curse. Life 6, 39.

Zar, J.H., 1999. Biostatistical Analysis. Prentice Hall.

Zhang, J., 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J., Zhang, Y., Rosenberg, H.F., 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30, 411–415. https://doi.org/10.1038/ng852

Zhang, W., Inan, M., Meagher, M.M., 2000. Fermentation strategies for recombinant protein expression in the methylotrophic yeastPichia pastoris. Biotechnol. Bioprocess Eng. 5, 275–287. https://doi.org/10.1007/BF02942184

Zhang, Y., Liu, R., Wu, X., 2007. The proteolytic systems and heterologous proteins degradation in the methylotrophic yeastPichia pastoris. Ann. Microbiol. 57, 553–560. https://doi.org/10.1007/BF03175354