International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Structural and Evolutionary Analysis of Disease-Associated Missense Variants in Phenylalanine Hydroxylase

Shaurya Chandna

Independent Researcher

Abstract: ***Phenylalanine Hydroxylase (PAH) is a critical enzyme found in the liver that catalyses the bioconversion of the dietary amino acid phenylalanine into the nonessential amino acid tyrosine (Scriver et al., 2001; Blau et al., 2010). This bioconversion is considered to be an important regulatory step in amino acid metabolism and is necessary for the maintenance of physiological levels of phenylalanine in the body. PAH belongs to the class of aromatic amino acid hydroxylases (AAAH) and acts as a monooxygenase that utilises molecular oxygen (O_2), ferrous iron (Fe^{2+}), and tetrahydrobiopterin (BH_4) as cofactors for its enzymatic activity (Fitzpatrick, 2012). Disruption of this tightly regulated enzymatic/biochemical pathway results in the accumulation of phenylalanine (Phe), a toxic substance to both the developing and fully developed nervous system (Scriver et al., 2001).***

Keywords: phenylalanine metabolism, liver enzyme function, tyrosine formation, amino acid balance, nervous system toxicity

1. Introduction

A deficiency in PAH enzymatic activity results in phenylketonuria (PKU), one of the most commonly diagnosed inborn errors of metabolism (Scriver et al., 2001). If not identified and treated early via newborn screening and dietary measures, patients with PKU will develop severe neurological complications, including, but not limited to, intellectual disability, seizures, behavioural problems, and decreased pigmentation (Blau et al., 2010). Due to the implementation of newborn screening programs and active dietary management, the once devastating neurological consequences of PKU have now been largely transformed into the much more manageable metabolic disorder that it currently is. However, the molecular basis of PKU is very complex, with some variability observed in patients suffering from PKU having the same biochemical defect, i.e. PAH deficiency.

The genetic basis for PKU is found through the identification of mutations in the PAH gene located on chromosome 12q23.2 (Scriver et al., 2001). To date, over one thousand different PAH variants, also classified as missense mutations, have been identified, with the majority of these mutations resulting in a single amino acid substitution in the polypeptide chain (ClinVar, 2024). Typically, missense mutations will result in keeping the wild-type PAH alive; nonsense mutations and frameshift mutations will result in the total absence of PAH expression. Any residual PAH activity serves to have a major impact on the severity of the disease and on the chances of having a favourable response to therapeutic intervention with BH_4.

Clinically speaking, individuals with PKU have the most variable phenotypic expression, from individuals with no PAH activity who would be classified as classic PKU to individuals who have a mild elevation of phenylalanine but do not have significant clinical manifestations of hyperphenylalaninemia (HPA). Some will have a response to BH_4 therapy, which has been shown to enhance the stability of certain missense PAH proteins or to increase residual PAH activity (Blau et al., 2010; Pey et al., 2007). A greater understanding of not only how specific missense mutations impact PAH structure and function but also how those mutations would lead to differences in biochemical activity and responses to BH_4 therapy is of utmost importance.

2. Background and Literature Review

2.1 Structure and Regulation of Phenylalanine Hydroxylase

Structurally, PAH is a homotetramer composed of three distinct domains: an N-terminal regulatory domain that regulates PAH enzymatic activity, a central catalytic domain that contains the active site for converting phenylalanine to tyrosine (hydroxylation), and a C-terminal domain that provides a scaffold for structural stability, allowing for proper functionality of the PAH molecule. If one of the regulatory, catalytic or oligomeric domains becomes disrupted, then the PAH molecule will lose functionality even though its actual active site has not been altered (Fitzpatrick, 2012).

The missense mutations located on highly conserved regions of PAH are, from a structural perspective, most likely to impair PAH functionality. Highly conserved amino acids have been selected strongly for retention due to their structural and/or functional roles in the protein or enzyme (Pupko et al., 2002). Mutations at positions within proteins that have remained unchanged throughout evolution (conserved positions) tend to be associated with a greater probability of producing a deleterious effect than mutations at positions that have not remained unchanged throughout evolution (non-conserved positions). Therefore, the relationship between the two concepts of mutation site and pathogenicity will allow for predictions to

be made about the influence of genetic variants on pathogenicity.

With recent advances in the field of structural biology and the availability of large amounts of sequence data and structural data to the public, it has become practical to use large, publicly available databases to perform systematic analyses of missense mutations that are associated with disease. Using multiple sequence alignments across species, we can identify residues that are conserved within that gene, and we can also analyse three-dimensional protein structures to better understand how amino acid substitutions will influence the folding of the protein and its stability and interactions (Pupko et al., 2002; Thornton et al., 2007). Computational approaches have been especially helpful in addressing questions related to diseases that have many variants, like phenylketonuria (PKU), because the number of variants associated with these conditions makes it logistically and financially impossible to evaluate all variants using experimental techniques.

The recent advent of very high-accuracy models of proteins predicted by AlphaFold and the ability to use those models along with homology-derived protein structure models from the Protein Data Bank will be extremely helpful in conducting in silico analyses of genetic variants causing PKU (Jumper et al., 2021). Mapping out genetic variants to the three-dimensional structure of Phenylalanine Hydroxylase (PAH) will be important for generating biologically plausible hypotheses about the molecular mechanism of phenotypes associated with PAH mutations and the potential for therapeutic interventions (Fitzpatrick, 2012; Jumper et al., 2021).

Despite extensive clinical and genetic characterisation of PKU, the connection between individual missense mutations and biochemical defects remains poorly defined. Mutations have been classified as pathogenic simply because of an individual having a mutation and also having PKU, with no further consideration being given to the molecular mechanism of how that mutation caused PKU. Filling those gaps will require an integrated method of combining evolutionary, structural and biochemical data.

This current project utilises a computational platform combining evolutionary conservation analysis and structural mapping in order to investigate the missense mutations in Phenylalanine Hydroxylase associated with PKU. The missense variants will be analysed for patterns in their functional and phenotypic differences and compared to the conservation and three-dimensional structure of the variant. This will improve our understanding of genotype-phenotype correlations in PKU; furthermore, it will demonstrate the utility of using computational methodologies to further our understanding in biochemistry and molecular medicine.

Phenylalanine Hydroxylase (PAH) is part of the aromatic amino acid hydroxylases group, of which both Tyrosine Hydroxylase and Tryptophan Hydroxylase are members as well. All three enzymes have a distinct method of regulating activity and differ in their location, but share an overall catalytic function.

PAH is located primarily in liver cells [hepatocytes] and is responsible for producing tyrosine by hydroxylating phenylalanine, which is then converted to important biological compounds [neurotransmitters, hormones and melanin].

PAH consists of three distinct functional domains: [N]regulatory [I]catalytic [C] oligomerisation. The regulatory domain is key to the allosteric control of PAH; specifically, the domain's ability to modulate enzyme activity based on levels of intracellular phenylalanine. The binding of phenylalanine produces conformational changes in PAH, resulting in a shift in enzyme activity from low to high, allowing for the rapid metabolism of phenylalanine based on dietary intake.

The catalytic domain houses the majority of the residues that make up the active site, whereby iron is coordinated, and substrates can bind. Specifically, iron is required for the activation of oxygen during hydroxylation and is coordinated by distinct histidine and glutamate residues. Additionally, BH_4 acts as both an electron donor and stabilizing cofactor to the catalytic core. Mutations affecting iron coordination, BH_4 binding or substrate orientation often result in a significant loss of enzymatic function.

The oligomerisation domain is responsible for forming tetramers, which is necessary for the enzymatic function of PAH. If the tetramers form incorrectly due to the disruption of interactions between subunits, then the overall function of PAH will be affected even when monomeric forms retain partial function. Thus, the structural integrity of PAH supports the understanding of how mutations occurring distal to the active site can produce extreme phenotypical differences.

2.2 Genotype-Phenotype Relationships in Phenylketonuria

PKU is a complex condition with a wide range of clinical variability in patients who have it. In general, all the patients with this disease exhibit decreased or low levels of PAH enzyme activity, and they will need to be on a strict diet to limit their intake of phenylalanine. Individuals diagnosed with mild hypophenylalaninemia typically have some residual enzyme activity, so they may not require as strict a diet to control their phenylalanine levels.

The first studies of genotype and phenotype stated that individuals with nonsense and frameshift changes (mutations) would exhibit greater than 90% loss of enzyme function; therefore, their phenotypes would usually be severe. In contrast, the majority of alleles associated with PKU are missense changes with very variable clinical outcomes. Totally different mechanisms may lead to misfolded proteins that will degrade rapidly, decrease catalytic activity, or decrease the response to regulatory mechanisms.

A small group of individuals with PKU have a positive response to BH_4 supplementation. The supplementation of these

individuals has involved the use of pharmacological doses of BH_4 that either stabilise mutant PAH proteins, improve the efficiency of protein folding, and/or improve the overall performance of the PAH enzyme. The structural analyses of these mutations that respond to BH_4 supplementation indicate that they will either be located outside of the catalytic core or in a position that will indirectly affect the PAH protein's ability to bind BH_4 or to maintain structural integrity. The importance of the structural context in predicting the response to BH_4 supplementation is highlighted.

Although there is a large amount of clinical data on the PAH variants, a large number of them have insufficient molecular data associated with them. Variants that are thought to be causative and to have a similar predicted impact on enzyme function may have highly variable phenotypic outcomes based on the location, structure, and/or interaction networks of the variants with respect to how they influence protein dynamics. Therefore, the predictive power of genotype/phenotype data for PAH variants will be limited to only using sequence data.

2.3 Evolutionary Conservation and Functional Significance

Evolutionary conservation is a widely used method for assessing the functional significance of individual amino acids. Conserved residues across species are under more selective pressure than non-conserved residues, implying a more important function with respect to protein structure and/or protein function. Mutations that occur at conserved residues are potentially more damaging to the protein than those at non-conserved residues.

Highly conserved residues in PAH are primarily in the catalytic domain near the active site and within cofactor binding sites (Fitzpatrick, 2012). Conservation analysis has been used to help prioritise pathogenic variants and to differentiate between pathogenic and benign variations. However, conservation is not the only way to evaluate whether a residue affects the function of a protein; some conserved residues provide an indirect effect on function by stabilising the protein structurally rather than by a direct catalytic action.

When considering the implications of evolutionary conservation, it is important to do so in combination with structural and biochemical information. Residues that appear conserved in sequence alignment may have very different structural contexts; the functional consequence(s) related to the mutation will vary greatly. Therefore, the value of conservation analysis is enhanced when it is used in conjunction with three-dimensional structural data.

2.4 Structural Studies of PAH and Disease-Associated Variants

The high-resolution crystal structures of PAH have significantly advanced the understanding of the architecture of the enzyme and its regulatory mechanisms. The structural data provide the spatial relationship between the functional domains and conformational changes associated with the activation of the enzyme and the inhibition of the enzyme's activity. Structural studies have provided evidence for the molecular basis for the binding of BH_4 (tetrahydrobiopterin) to PAH and the stabilising effects of BH_4 on the enzyme.

The mapping of disease-associated mutations to the structures of PAH provides evidence of a non-random clustering of pathogenic variants in functionally important regions of PAH. Multiple severe phenylketonuria (PKU) variants tend to cluster within regions immediately surrounding the active site and at the oligomer interface, within which changes of any extent would severely destabilise the overall complex (Pey et al., 2007; Erlandsen et al., 2004). However, the majority of variant associations for milder phenotypes fall on solvent-exposed or dynamic regions, where structural variation is tolerated, and function can be preserved despite deviating from the optimal structure.

The recent evolution of protein structure prediction techniques has provided a wealth of structural data for PAH and its variants. The accuracy of predictions made using AlphaFold has provided detailed models of PAH; many areas of the protein can now be analysed at the residue level, even where data is not available from experimental studies. These structural models have facilitated the screening of large variant datasets and generated hypotheses for future experimental work, thereby providing the opportunity to query areas of the protein where prior experimental data were sparse.

Computational methods for the interpretation of genetic variation have become an invaluable tool in the investigation of rare diseases, particularly those with significant levels of allelic diversity, such as PKU. By employing in silico approaches, researchers can systematically evaluate how a variant affects PAH structure, stability, and function. Computation methodology, including identity conservation analysis, structural modelling, and molecular view, is complementary when considering pathogenicity.

Recognising that computational-based predictions should not replace experimental-based validation, researchers can employ computational approaches to prioritise molecular variants for further study and improve the correlation between genotype and phenotype. Integrative frameworks which combine evolutionary, structural and clinical information are particularly useful when attempting to understand complex metabolic diseases and will inform future development of therapeutic strategies.

2.5 Rationale for the Present Study

Despite decades of investigation into PKU, many remain uncovered in regard to the mechanism of action of the PVH missense variants. Previous studies have either focused on one missense variant or utilised one method of analysis, giving them limited reproducibility. Integrative methodology is needed to link the functional impact of variants to evolutionary conservation and structural context.

The present study proposes to fill that void by conducting a combined evolutionary and structural analysis of a cohort of pathogenic PAH missense variants. The ultimate goal of this work is to elucidate the patterns by which PAH is rendered dysfunctional through missense mutations across various functional domains and ranges of phenotypic severity in order to generate a better understanding of the relationship between genotype and phenotype in PKU.

3. Methods

This study examined how some of the different missense mutations associated with disease in the Phenylalanine Hydroxylase (PAH) gene are likely to affect molecular mechanisms. To do this, an in silico analytical framework was used, which includes evaluating bioinformatics methods for evolutionary conservation; mapping to 3D protein structure; performing qualitative functional analysis; and generating original hypotheses regarding the pathogenicity of each variant. All bioinformatics tools that were used for this study were based solely on publicly available data, providing for complete replicability. The study involved no form of laboratory testing or collection of patient data.

3.1 Variant Identification and Selection Criteria

The pathogenic missense mutations of the PAH gene that were used for this study were sourced from the ClinVar database, which collects clinically-related genetic mutations and their associated phenotypes (ClinVar, 2024).

Variants were selected based on the following:
1) Mutations classified as having 'pathogenic' or 'likely pathogenic' status
2) Mutations associated with the phenotypes of 'phenylketonuria' or 'hyperphenylalaninemia'
3) Mutations that produced a single amino acid change

To achieve an adequate range of clinical severity for each missense variant, mutations viewed as having a 'classic PKU phenotype' and a more 'tolerable' or 'BH_4 responsive phenotype' were selected. The final selection of missense variants adequately represented the range and locations of known mutations throughout the entire PAH gene, inclusive of the regulatory, catalytic and oligomerisation domains. This process ultimately yielded a selection of ten missense pathological variants possessing biological relevance, while allowing for expansive analytical depth.

3.2 Retrieval of Protein Sequences and Domain Annotation

The canonical protein sequence for the human PAH protein was obtained from the UniProt database (UniProt Consortium, 2024). The protein sequence domains of the regulatory, catalytic and oligomerisation regions were identified based on the UniProt description of each listed protein. These descriptions were used to classify and interpret each of the above-listed variations. Orthologous PAH sequences were retrieved from several vertebrate species with the aid of the UniProt and the National Centre for Biotechnology Information protein databases to aid in evolutionary studies. The selected species covered a broad phylogenetic distance but maintained enough sequence similarity so that reliable alignments could be made.

3.3 Multiple Sequence Alignment and Conservation Assessment

PAH orthologous multiple sequence alignments were made with the help of Clustal Omega, which is a gradual alignment methodology that has been optimised for accuracy in comparing protein sequences (Sievers et al., 2011). The default settings for Clustal Omega were used to limit bias in the algorithms used for the comparisons. The quality of the alignments was evaluated through visual analysis, ensuring that there were compatible alignments of the conserved motifs and functional regions.

The conservation of the amino acid residues was evaluated qualitatively based on their degree of amino acid identity among the various species at their respective variant sites. Residues that were conserved among the majority of species were identified as very conserved, while those residues displaying limited variability were identified as moderately conserved. Variants that exhibited low conservation were also documented, but will be viewed with caution, as a low degree of conservation could reflect an adaptation to different species or a structural modification.

3.4 Structural Data Acquisition

The three-dimensional structure of PAH was obtained from the Protein Data Bank (Berman et al., 2000). To obtain the structure of the PAH protein of interest, the search focused primarily on structures of the catalytic domain and tetrameric assembly that have been experimentally determined using X-ray crystallography. In areas of the PAH protein where there are no structures of sufficient quality and resolution to construct an alignment, the three-dimensional protein structure of PAH was constructed using predicted protein structures that originated from AlphaFold.

The criteria used for selecting the predicted protein structures included the confidence scores. This study did not perform structural modifications or energy minimizations, as it was intended to qualitatively analyse (not quantitatively) molecular simulations.

3.5 Structural Mapping and Visualisation of Variants

Missense variants were mapped to PAH structures using PyMOL to measure the distance of each missense variant position to active sites, iron-binding residues, BH_4-binding domains, and inter-subunit interfaces (Schrödinger, LLC, 2023).

By classifying the variants functionally, it was possible to infer the probable biochemical effect(s) of the variants/variants by

evaluating evolutionarily conserved sequences and with respect to their structural context. Variants were qualitatively classified into functional categories based on how the mutations would be likely to affect (1) Catalytic Activity, (2) Stability/Folding of Protein, (3) Binding to Cofactor (BH_4), and (4) Oligomerisation and Quaternary Structure.

The classifications were determined based on the underlying principles that support the relationship between functional and structural proteins and the previously described genotype/phenotype correlations, although there was a lack of objective numerical scoring or predictive algorithms of pathogenicity, as the objective was not to provide a predictive ranking but rather to provide a mechanistic understanding.

3.6 Data Presentation and Interpretation

To present the results of the study, descriptive and tabular illustrations are provided to present the relationship between variant location, evolutionary conservation, and predicted effects on function. There were no specific restrictions on structural observations because there was no direct evidence to support or refute causal relationships for any one mutation; therefore, the overall focus was on identifying general trends for the range of variants, rather than for any individual variant.

3.7 Limitations of the Computational Approach

Several limitations exist for computational analyses, including the fact that evolutionary conservation typically only speculates about functional relevance, as it does not factor into context-specific effects or compensatory mutations. Structural models cannot represent the dynamic behaviours of proteins (i.e., they are typically presented as static/fixed conformations), nor can they represent the dynamic nature of proteins at the regulatory stage of folding.

The genotype/phenotype relationships associated with PKU are influenced by both genetic modifier genes and other extrinsic factors (i.e., environmental effects and dietary control) and will provide influence on the functional interpretations; therefore, the functional interpretations provided herein should be viewed as potential hypotheses that will require additional experimental confirmations through biochemical assays or cellular models.

4. Results

4.1 Overview of Selected PAH Variants

A selection of 10 missense mutations related to disease was picked to analyse in the PAH due to the fact that they contained clinical descriptions and were related to Phenotype. There is a clear variation in severity for each mutation as they all range from classical phenylketonuria (PKU) to mild PKU or those that will respond to tetrahydrobiopterin (BH_4). These mutations can be placed in each of the 3 functional domains of the PAH (the regulatory, catalytic, and oligomerisation domains); the location of the mutation will allow comparison of which mutation in each mutational area is impacting the expected biochemical function.

For example, severe mutations are usually located in the catalytic and oligomerisation domains, while mild mutations are more often located in the regulatory or peripheral domains of the enzyme, as defined by structural components.

4.2 Conservation Among Variant Residues

Multiple sequence alignment of the PAH from many species of vertebrate animals indicated differing degrees of conservation among the mutants being analysed. The amino acids associated with classical PKU disease showed a very high level of conservation and were typically invariant; this indicates that these amino acids are stronglevolutionarilyry conserved and play a critical role in either the structural characterisation of the PAH or in its catalytic character.

For example, the amino acids of milder phenotypes and BH_4-responsive aminases were demonstrated to exhibit moderate levels of evolutionary conservation. There was very little variability at these replacement sites; therefore, these mutations should retain residual enzymatic function since their sites exhibit some evolutionary constraint and have allowed for a moderate level of replacement during evolution. Only a few rare variations were observed at positions in which low conservation was recorded. Rare amino acid variations are preferentially located in exposed areas of the protein, or in parts of the protein that are flexible. Because these variations are not highly impactful upon PAH function relative to other types of variants, rare pathogenic PAH variants in patients are expected to have a milder phenotype than classic PKU patients.

4.3 Domain-Specific Pathogenic Variant Associations

Integration of structural localisation and conservation clearly indicated a domain-specific association with pathogenic variants; most of the variants located in the catalytic domain were associated with severe phenotypes, particularly those that were located close to the active site or iron-coordination residues. It was presumed that these variants would prevent proper substrate binding, activation of substrate, or electron transfer from the substrate to oxygen, thus leading to near-total elimination of PAH activity.

The oligomerisation domain also contained many likely pathogenic variants. Structural mapping indicated that many mutations associated with classic PKU occur at the interface of the four monomers that comprise the tetrameric PAH protein and would therefore influence the tetrameric structure. The mutations at these interfaces would likely result in a destabilisation of the tetrameric structure, thereby resulting in a decrease in overall PAH activity, even though each monomer may retain some degree of activity.

The regulatory domain, however, did not contain pathogenic variants that were associated with severe phenotypes. The structural analysis suggests that variants in the regulatory

domain are unlikely to have a direct effect on the catalytic mechanism, but would likely subsequently affect the flexibility of the protein or the ability of the protein to exhibit allosteric properties. These data would support a relatively mild reduction in PAH activity due to regulatory upstream pathways and would also be consistent with the potential response of these patients to BH_4 supplementation.

4.4 Structural Localisation of Variants

Three-dimensional structural mapping of pathogenic variants showed that variant positions tend to cluster in a non-random pattern within regions of the PAH protein that are functionally significant. Most of the pathogenic variants associated with classic PKU will be found in the immediate vicinity of the catalytic core or in the vicinity of the substrate binding channel or iron-coordination site. Therefore, pathogenic variants in these areas are likely to destabilise the active site architecture or reduce or eliminate essential catalytic contacts.

Most milder variants have been found distally from the core of the catalytic domain and in the periphery of the regulatory domain, where they may potentially affect local folding of the protein and/or regulatory properties without adversely impacting the overall catalytic properties of the catalytic core. Structural visualisation suggests that while substitutions at the peripheral localisation locations would likely result in decreased efficiency of the enzyme, the enzyme retains some level of function.

Those variants that have been associated with BH_4-responsiveness tended to be in proximity to components of the PAH protein that physically associate with BH_4, or in close association with a variety of chemically-stabilising regions (Blau et al., 2010; Pey et al., 2007). Therefore, the finding that these variants are associated with BH_4 interactions within the PAH protein is consistent with the premise that BH_4 supplementation could benefit patients by stabilising the PAH protein or compensating for decreased catalytic efficiency of mutant PAH proteins.

4.5 Integrated Functional Interpretation

A classification scheme was developed to estimate the anticipated biochemical consequences of variants by integrating conservation and structural location of the variant. Variants that are in highly-conserved positions within either the catalytic domain or the oligomerisation domain are predicted to have severe effects on PAH function, and consequently would be consistent with the phenotype seen in classic PKU patients and hence contain highly-conserved structure and provide integral components in the catalytic and/or structural function of PAH and that regulatory mechanisms could not compensate for.

Variants in moderately-conserved positions or peripheral regions are predicted to result in partial impairment of enzyme function; patients would exhibit milder PKU phenotypes. Adequate residual enzyme activity could be achieved in these patients to preclude the development of extreme hyperphenylalaninemia, with appropriate dietary control and/or supplementation with BH_4.

The data provide clear evidence that the relationship between residue conservation, structural context, and clinical severity prediction is well-defined, and that an integrative approach yields a more comprehensive understanding of the probable functional effects of missense variants in PAH. A consideration of both sequence and structural-level data will equally add to the value of variant analysis.

5. Discussion

The goal of this research was to determine how disease-related missense mutations disrupt the function of Phenylalanine Hydroxylase (PAH) by integrating structural data with evolutionary constraints. By situating each amino acid substitution in its evolutionary context and 3-dimensional protein structure, rather than analysing variations in isolation, a more coherent interpretation has been established about how genetic variation contributes to the phenotypic spectrum of phenylketonuria (PKU) (Thornton et al., 2007).

When the variations were examined together, a strong connection between evolutionary conservation and the severity of the disease was noted. Classic PKU-associated variant substitutions were predominantly located at amino acid residues that were highly conserved across all vertebrate species, indicating significant selection pressure for conservation at each of these positions. Such high levels of conservation suggest that the amino acids at these positions are critical for maintaining either the enzyme's structural integrity or its functional enzymatic activity (efficiency). Therefore, because substitutions occurring at highly conserved residues are generally intolerated, the molecular basis for the profound reduction in PAH activity in classic PKU is explained.

The addition of structural maps allowed for the elucidation of an aspect of spatial relationships that was previously unexamined. It was noted that many variants associated with severe phenotype classifications clustered within proximity to the catalytic core and also at the interfaces of the subunits involved in the formation of PAH oligomers. Variants located close to the active site are very likely to disrupt the correct positioning of substrates, coordination of iron, and activation of molecular oxygen, all of which are necessary for the hydroxylation of the substrate. Likewise, substitutes at oligomerisation sites are likely to prevent the formation of stable tetramers, thereby resulting in decreased efficiency of the enzyme as a whole, even if each monomer has a partial catalytic capability. The data presented above emphasise that PAH function is dependent on both the presence of an active site and the proper higher-order assembly of the enzyme.

However, many of the variants associated with mild PKU phenotypes, or with a response to tetrahydrobiopterin (BH_4), exhibited a different distribution. Substitutions occur mostly at residues with moderate conservation on more peripheral parts

(i.e., the edge) of the protein, particularly in the regulatory domain or near the boundaries of the catalytically active domain, with these substitutions typically having little or no negative effect on catalysis. Rather, substitutions often have a greater effect on conformational flexibility, control of regulatory parameters, and local stability. In these cases, some residual PAH activity may be maintained, and some degree of phenylalanine metabolism may remain under physiologic or therapeutic conditions.

By examining the structural position of the BH_4-responsive variant(s), insights into the mechanism for BH_4 therapeutic response may be gathered. In addition to acting as a cofactor in the catalysis of PAH, BH_4 plays a role in stabilising PAH via assisting proper protein folding and increasing residual activity in some mutant proteins. Therefore, the closer the substitution is to the surface of BH_4 or within a flexible structural element, the greater the positive effects will be by increasing BH_4 availability. In this way, structural context can assist in demonstrating how genotype data regarding potential therapeutic response can be obtained from data related to therapeutic response, as it may not otherwise be demonstrable solely based on genotype.

This research demonstrates that while evolutionary conserved residue information is helpful, it cannot independently predict phenotype, as many mutations occurring at these highly conserved (i.e., not a typographical error) residues are much more likely to lead to severe disease compared to mutations at less conserved residues; however, the nature of the structural changes introduced by substitution of a particular residue will vary based on the functional implications associated with that substitution. Additionally, although many evolutionarily conserved residues may indirectly contribute to functional ability by providing stability to the tertiary or quaternary structure(s) of the protein or proteins (i.e., they provide support), they do not directly participate in functional activity (i.e., catalysing reactions) of the protein or proteins. Similarly, substituting an evolutionarily conserved residue could create an extremely dysfunctional effect if disrupting the protein's ("disturbing") critical networks of interactions and/or regulatory dynamics, whereas substituting a less frequently (i.e., less evolutionarily conserved) residue could also yield an extremely dysfunctional effect (i.e., by creating an equally dysfunctional effect on another.

This paper emphasises the requirement for analysis methods combining 3-D geometry with constraints at the sequence level. The results presented in this article correspond to previous genotype-phenotype analysis (i.e. PAH function influences phenotype for PKU patients), although as of this time, there is not yet a clear explanation available for all the variants described here. By combining several types of data mentioned above, we have produced an integrated research framework that serves to create a pathway for quality of life assessment for severely versus mildly affected PKU patients. The integration framework will continue to expand as new variants are identified and will provide valuable assistance in interpreting other rare metabolic disorders that currently present significant clinical challenges.

From a more general perspective, the research demonstrates a continued trend toward computational analysis of both biochemical and molecular biology data. The rapid development of genetic sequencing alone has outpaced the experimental characterisation of the effects caused by specific variants and highlighted the necessity to develop robust in silico methodology for prioritising and interpreting variants. Even though the results from computational modelling cannot provide the same level of experimental validation as traditional laboratory techniques can provide, the results produced can still provide valuable biological hypotheses for developing future experimental research plans in the fields of both biochemistry and cellular biology.

Whereas this study's analyses should have their limitations noted regarding all analyses being conducted using in silico methods (versus validated by experimental observations), using molecular dynamics to study protein flexibility and compute alternative metrics of variations based on conservation methods; this study serves as a starting place for establishing an experimental validation methodology of the predicted affect of selected PKU variant proteins based on the above-mentioned limitations of the framework used.

(Expansion of the study resulting from future research will include in vitro enzyme assays on validation of computationally predicted variant effects and performance among protein variants. Ongoing research aimed at enlarging the number of variants analysed and also including new findings related to clinical outcomes from future research will ultimately provide further understanding concerning the GxP relationship between rare Mendelian disorders and affected proteins.)

This study, demonstrates that the application of both evolutionary/structural and evolutionary analyses may support the idea that finding evolutionary relationships can provide useful and relevant interpretative methods for determining molecular pathophysiology associated with missense mutations that could lead to a more complete understanding of how proteins do or do not perform as part of their normal functioning in metabolic diseases such as PKU and reinforces the use of computational methodologies as valuable research resources for undergraduate researchers.

6. Conclusion and Future Directions

This study applied an integrated evolutionary and structural framework to examine disease-associated missense variants in Phenylalanine Hydroxylase (PAH), with the aim of clarifying molecular mechanisms underlying phenylketonuria (PKU). By combining sequence conservation analysis with three-dimensional structural mapping, the work provides a biologically grounded approach for interpreting how specific amino acid substitutions influence PAH function and contribute to the observed phenotypic spectrum.

The findings demonstrate that variants associated with classic PKU are predominantly located at highly conserved residues within structurally and functionally critical regions of the enzyme, particularly the catalytic and oligomerisation domains. Substitutions at these sites are predicted to disrupt essential catalytic interactions or destabilise higher-order assembly, resulting in severe loss of enzymatic activity. In contrast, variants associated with milder phenotypes or Tetrahydrobiopterin (BH_4) responsiveness tends to occur at moderately conserved or structurally peripheral positions, where partial enzymatic function may be preserved. These patterns support a model in which both evolutionary constraint and structural context jointly shape disease severity.

Importantly, this work illustrates that neither sequence conservation nor structural analysis alone is sufficient to fully explain genotype- phenotype relationships in PKU. Instead, meaningful interpretation arises from integrating these perspectives, allowing variants to be evaluated in terms of both functional importance and spatial context within the protein.

This integrative approach helps explain why some mutations distant from the active site exert profound functional effects, while others closer to catalytic regions retain partial activity.

Beyond PAH and PKU, the framework presented here has broader relevance for the interpretation of missense variants in metabolic enzymes and other proteins with complex structural regulation. As genetic sequencing continues to identify large numbers of rare variants, computational approaches such as those used in this study will play an increasingly important role in prioritising variants for experimental validation and guiding hypothesis-driven research (Jumper et al., 2021).

Several avenues for future work emerge from this analysis. Incorporating molecular dynamics simulations could provide insight into how specific mutations alter protein flexibility and regulatory transitions. Experimental validation through in vitro enzymatic assays or cellular expression systems would be essential to confirm predicted functional effects. Expanding the analysis to larger variant datasets and integrating clinical outcome data could further refine genotype–phenotype correlations and improve predictive accuracy.

In summary, this study demonstrates the value of combined evolutionary and structural analyses for understanding disease-associated missense mutations. By situating genetic variation within a rigorous biochemical framework, it contributes to a more nuanced understanding of PAH dysfunction in PKU and highlights the potential of computational methods as powerful tools in undergraduate biochemistry research.

References

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

[2] Blau, N., van Spronsen, F. J., & Levy, H. L. (2010). Phenylketonuria. *The Lancet, 376*(9750), 1417–1427. https://doi.org/10.1016/S0140-6736(10)60961-0

[3] ClinVar. (2024). *ClinVar: Public archive of interpretations of clinically relevant variants*. National Centre for Biotechnology Information. https://www.ncbi.nlm.nih.gov/clinvar/

[4] Erlandsen, H., Pey, A. L., Gamez, A., Perez, B., Desviat, L. R., Aguado, C., Koch, R., Surendran, S., Tyring, S., Matalon, R., Scriver, C. R., & Martinez, A. (2004). Correction of kinetic and stability defects by tetrahydrobiopterin in phenylketonuria-associated phenylalanine hydroxylase mutants. *Proceedings of the National Academy of Sciences of the United States of America, 101*(48), 16903–16908. https://doi.org/10.1073/pnas.0407346101

[5] Fitzpatrick, P. F. (2012). Allosteric regulation of phenylalanine hydroxylase. *Archives of Biochemistry and Biophysics, 519*(2), 194–201. https://doi.org/10.1016/j.abb.2011.10.019

[6] 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[7] Pey, A. L., Perez, B., Desviat, L. R., Martinez, M. A., Aguado, C., Erlandsen, H., Gamez, A., Stevens, R. C., Thórólfsson, M., Ugarte, M., & Martinez, A. (2007). Mechanisms underlying responsiveness to tetrahydrobiopterin in phenylketonuria. *Human Mutation, 28*(9), 831–845. https://doi.org/10.1002/humu.20537

[8] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., & Ben-Tal, N. (2002). Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics, 18*(S1), S71–S77. https://doi.org/10.1093/bioinformatics/18.suppl_1.S71

[9] Schrödinger, LLC. (2023). *The PyMOL Molecular Graphics System* (Version 2.5). https://pymol.org/

[10] Scriver, C. R., Kaufman, S., Eisensmith, R. C., & Woo, S. L. C. (2001). The hyperphenylalaninemias. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), *The metabolic and molecular bases of inherited disease* (8th ed., pp. 1667–1724). McGraw-Hill.

[11] 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. *Molecular Systems Biology, 7*, 539. https://doi.org/10.1038/msb.2011.75

[12] Thornton, J. W., Need, E., & Crews, D. (2007). Resurrecting the ancestral steroid receptor: Ancient origin of estrogen signalling. *Science, 301*(5640), 1714–1717. https://doi.org/10.1126/science.1086185

[13] UniProt Consortium. (2024). UniProt: The universal protein knowledgebase. *Nucleic Acids Research, 52*(D1), D551–D558. https://doi.org/10.1093/nar/gkad982