Identifying key amino acid types that distinguish paralogous proteins using Shapley value based feature subset selection

Anonymous authors
Paper under double-blind review

Abstract

Paralogous proteins have a common ancestor but have diverged in functionality. Using known machine learning algorithms, we present a data-driven method to identify the key amino acid types that play a role in distinguishing a given pair of proteins that are paralogs. We use an existing Shapley value based feature subset selection algorithm, SVEA, to identify the key amino acid types adequate to distinguish pairs of paralogous proteins. We refer to these as the amino acid feature subset (AFS). For a paralog pair, say proteins P and Q, its AFS is partitioned based on protein-wise importance as $AFS(P)$ and $AFS(Q)$ using a linear classifier, SVM. To validate the significance of the AFS amino acids, we use multiple domain knowledge based methods : (a) multiple sequence alignment, and/or (b) 3D structure analysis, and/or (c) supporting evidence from biology literature. This method is computationally cheap, requires less data and can be used as an initial data-driven step for further hypothesis-driven experimental study of proteins. We demonstrate the results for 15 pairs of paralogous proteins. Code available at `https://anonymous.4open.science/r/AFS_AAC_SVM-F3D9`.

1 Introduction

Proteins form the fundamental machinery in living systems, having several vital functions such as DNA replication, catalysis, transport, environmental interaction, etc. Advancements in sequencing technologies have resulted in exponential growth of protein sequence databases (The UniProt Consortium, 2020). However, the number of experimentally verified annotations constitute a tiny fraction: only 0.57% of 250 million sequences in UniProtKB (The UniProt Consortium, 2020) have manually reviewed annotations. Experimental methods for determining biological process level functions (transcription, DNA repair, etc.) are high-throughput whereas methods for molecular function (catalysis, ligand specificity, etc.) are low-throughput and hence are not scalable. The relationship between sequence and function is subtle and has not been fully decoded yet.

Paralogs are proteins that have a common ancestor but have diverged functionally. The functional difference in two paralogous proteins is considered to arise due to evolutionary changes in the sequences (Yang et al., 2023). A typical experiment to investigate the role of an (or a group of) amino acid(s) in the function of a protein is to perform a site-directed mutagenesis experiment: replace one or more amino acids and test the effect of the sequence change (Kresge et al., 2006). In this work, we provide an algorithmic ML pipeline, consisting both feature engineering and feature subset selection, as a quick and resource-cheap test to assess the likely outcome from a site-directed mutagenesis experiment. Using this ML pipeline we investigate the following two questions:

- Does the coarse-grained information of amino acid composition carry a strong enough signal for discriminating paralogs?
- If so, which of the 20 amino acids have a potential role in the functional difference of the paralogs?

We use a diverse dataset of 15 paralog pairs. Our datasets show a range of sequence and function diversity (details in Appendix B). Longest common subsequence score ($lcss$) is a metric to quantify sequence diversity and median within-class $lcss$ is $\leqslant 0.5$ in 12 of the 15 datasets, and the median inter-class $lcss$ for the corresponding classes is less than within-class $lcss$. Functional diversity, as discerned from biology literature, also shows large diversity from subtle functional differences (e.g., trypsin/chymotrypsin) to drastic (e.g.,

lysozyme c/α-lactalbumin). Function description is fine-grained (e.g., trypsin/chymotrypsin) as well as coarse grained (e.g, GPCRs).

Our findings are that small subsets of amino acids can discern differences between pairs of paralogs. The subset sizes are between 5 to 10, the median being 8. However, an established ground truth feature subset/rankings is not available for the task that can be used for evaluating the method or comparing performances with alternative methods. For this reason, we rely on evidence from experimental biology literature that highlights the role of the AFS amino acids in the function/structure of the respective protein. We provide validations from literature, MSA (a popular computational tool to assess evolutionary conservation) and logical consistencies; for many pairs such validations are more than one.

Towards this, we view a protein as the composite of its constituent standard 20 amino acids. We use amino acid composition (AAC) features, a Shapley value (Shapley, 1953) based feature subset selection algorithm (Shapley Value based Error Apportioning, SVEA) (Tripathi et al., 2020; 2021), and a linear support vector machine (SVM) classifier (Steinwart & Christmann, 2008) as tools to identify key amino acid types that can distinguish a given a pair of proteins that are paralogs. It yields quick results based on which biologists can conduct detailed experiments which are resource-intensive (time, cost, trained manpower, etc.).

The key results from our ML pipeline experiments are:

• Using known machine learning algorithms we demonstrate a data-driven method to identify key amino acids that distinguish two paralogous proteins.

- The SVEA algorithm identifies a subset of amino acid types (referred to as AFS) adequate for distinguishing two paralogous proteins. The size of AFS ranges from 5 to 10 amino acids out of 20. (Table 1)
- For a paralog pair, say protein families P and Q, the computed AFS is partitioned into $AFS(P)$ and $AFS(Q)$ using a linear SVM, to determine the family-wise importance of AFS. (Table 1)

• Domain knowledge based validation of AFS: The significance of the amino acids in AFS was validated for 14 datasets using various methods like (a) multiple sequence alignment (MSA) and/or (b) structural analysis and/or (c) supporting evidence from literature that report structural/functional role of these amino acids.

• Logical consistencies in the pair-wise AFS of three paralogous proteins (globins, Section 3.1.7, and GPCRs, Section 3.1.8). If families P vs Q and P vs R have AFS_1 and AFS_2, then,

- we find common amino acids in $AFS_1(P)$ and $AFS_2(P)$, except for one pair.
- amino acids in $AFS_1 \cap AFS_2$ are either excluded from AFS_3, which is from Q vs R, or have much lower Shapley value in AFS_1, AFS_2, or AFS_3.

• Validation of AFS using test data (Section 3.2): The composition of amino acids is sufficient to classify several paralog pairs. A linear SVM classifies with high test scores (70-99%) using only the composition of AFS amino acids as features. (Appendix Table 5)

• AFS are top ranked features with an alternate feature ranking measure, Marginal Contribution feature importance (MCI) (Catav et al., 2021). (Appendix Table 6)

Shapley values based feature attribution methods are popular for explaining machine learning models (Rozemberczki et al., 2022). One such method is SHAP (Lundberg & Lee, 2017), which assigns attribution scores to input features based on a model's output for a given instance input. Another method is SAGE (Covert et al., 2020), which assigns feature attribution scores based on a model's loss computed at the dataset level. Unlike these methods, where feature attributions are based on a trained model, the SVEA algorithm that we use for our task assigns scores to the features based on the distribution of the data points in the feature space and their ground truth labels. The SVEA algorithm uses a function $v(S)$, which acts as a measure of inter-class linear separation between the data points in the space of the feature subset S. The scores assigned to the features are Shapley values computed using this function $v(\cdot)$. We also use an alternate feature ranking method, i.e. the Marginal Contribution Feature Importance (MCI) (Catav et al., 2021). MCI is an axiomatic approach that was proposed as an alternative to Shapley values to score and rank features. We find close agreement between the AFS computed using SVEA and the top-ranked amino acids using MCI.

Use of deep learning methods trained on large datasets is becoming commonplace in Biology; for example, prediction of molecular function via EC number or GO annotation (Bileschi et al., 2022; Sanderson et al., 2023),

identifying input sequence regions relevant to model output (Zhou et al., 2016) and learning sequence-function mapping from deep mutational scanning experiment data (Song et al., 2021). The use of large datasets for training makes this approach highly resource-intensive. The approach we present herein needs much smaller datasets and, consequently, (i) is computationally cheap and (ii) has far wider applicability since labelled data validated by wet lab experiments is limited.

2 Methodology

We discuss the main components of our methodology.

2.1 AAC features

Consider a paralogous pair of proteins, families P and Q. We first curate a set of sequences, say D_P and D_Q, from a standard protein sequence database, SwissProt (The UniProt Consortium, 2020), with n_P and n_Q number of sequences each from families P and Q respectively. For a protein sequence $\mathbf{p}^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_L^{(j)})$ of length L with $p_k^{(j)} \in \{1, 2, \cdots, 20\}$ corresponding to the standard 20 amino acids, the AAC feature $\mathbf{x}_j^{AAC} \in [0,1]^{20}$ for $\mathbf{p}^{(j)}$ is computed as follows,

$$x_{j,i}^{AAC} = \frac{1}{L} \sum_{k=1}^{L} \mathbf{1}_{\{p_k^{(j)}=i\}}, \ \forall i \in [20]$$

So $x_{j,i}^{AAC}$ is the normalised count of the standard amino acid i, $i \in \{1, 2, \cdots, 20\}$, in a protein $\mathbf{p}^{(j)}$.

2.2 Feature subset selection using SVEA

Given a set, N, of features from the protein sequences of P and Q, we try to find the features $S \subseteq N$ that contribute the most to the linear separation of P and Q sequences. With AAC features, we have $N = \{1, 2, \ldots, 20\}$ corresponding to each of the standard 20 amino acid types.

We utilise the Shapley value based feature ranking and subset selection algorithm, SVEA (Tripathi et al., 2020; 2021), to identify the most important feature subset $S \subseteq N$. Shapley value is a well known solution concept from cooperative game theory (Shapley, 1953; Narahari, 2014) for distributing the total worth of a coalition of players fairly among each of them by quantifying each players effective marginal contribution. The SVEA algorithm considers the binary classification task as a cooperative game among the features, with a function $v(S)$ as the worth of every feature subset S. $v(S)$ acts as a measure of linear separation between the classes in the feature space of S. Accounting for class-imbalance, we define $v(S)$ using a class-balanced hinge loss function $tr\_er(S)$, which is defined as,

$$\begin{aligned} tr\_er(S) &= \min_{w,\xi_j} \frac{1}{2n_P} \sum_{j=1}^{n_P} \xi_j + \frac{1}{2n_Q} \sum_{j=n_P+1}^{n_Q} \xi_j \\ \text{s.t. } & y_j \left(\sum_{i \in S} w_i x_{j,i}^{AAC} + b \right) \geqslant 1 - \xi_j, \ \forall j \in [n_P + n_Q] \\ & \xi_j \geqslant 0, \ \forall j \in [n_P + n_Q] \end{aligned}$$

and $v(S) = tr\_er(\varnothing) - tr\_er(S)$. The minimizer in the above finds a linear hyperplane with the least class-balanced hinge loss in the feature space of S. $\varnothing$ is the empty set and $tr\_er(\varnothing) = 1$, therefore, $v(S) = 1 - tr\_er(S)$. $tr\_er(S) = 0$ implies $v(S) = 1$, i.e., the two classes are completely linearly separable in the feature space of S. The maximum value of $tr\_er(S)$ possible is 1.

The Shapley value $\phi(i)$ for a feature $i \in N$ is computed as,

$$\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)).$$

Thus, $\phi(i)$ is a weighted sum of the marginal contribution of feature i to all the possible feature subsets that do not contain i. Shapley values are unique solution concepts satisfying the axioms - efficiency, symmetry and marginality (Young, 1985). The higher the $\phi(i)$, the higher the contribution of feature i to the linear separation between the classes and, consequentially, the higher the importance of feature i distinguishing the classes.

Exact Shapley value computations are known to be exponential time. Hence, they are computed using a linear time (in number of features) Monte Carlo approximation (Castro et al., 2009) in the SVEA algorithm. As the number of features is small (20), good approximations can be computed fast via larger sampling. Furthermore, we do not find significant variance in the Monte Carlo estimates across multiple runs. We also find that the amino acids with top Shapley values are consistent across these runs (see Appendix E.4 for a detailed variance analysis). More details of the SVEA algorithm are given in Appendix Section C.

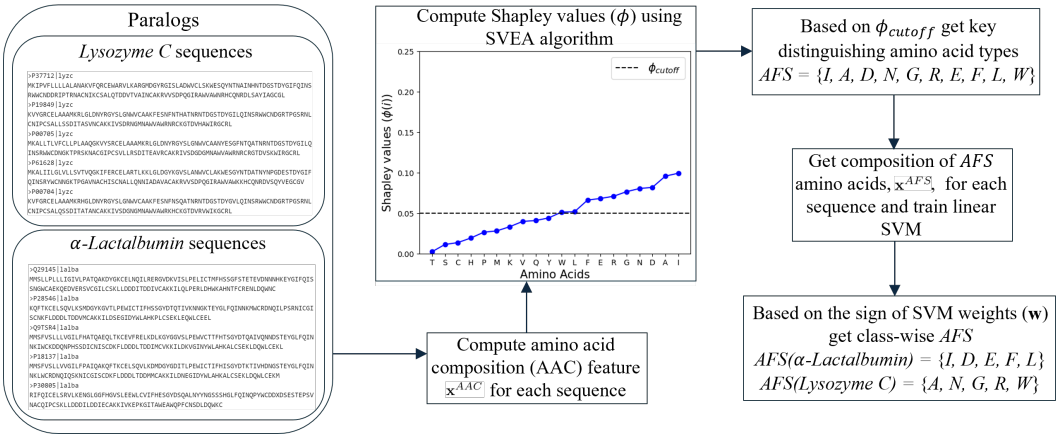


Figure 1: Flowchart summarizing the steps in our ML pipeline to compute the key amino acid types, AFS, that distinguish two paralogous proteins, using amino acid composition (AAC) features, Shapley value based SVEA algorithm for feature subset selection and class-wise feature subsets using linear SVM. Lysozyme C and α-Lactalbumin are used here as representative examples of paralog pairs. AFS identified for other paralog pairs are given in Table 1.

Data-driven cutoff for selecting AFS: The efficiency axiom of Shapley value implies, $\sum_{i=1}^{20} \phi(i) = v(N)$. If all features have equal contribution in achieving $v(N)$, then $\phi(i) = \frac{v(N)}{20}, \forall i \in N$. Consequentially, if a feature i had lesser contribution than others then $\phi(i) < \frac{v(N)}{20}$. Therefore, we set $\phi_{cutoff} = \frac{v(N)}{20}$ for selecting the key distinguishing amino acid feature subset, $AFS = \{i : \phi(i) \geqslant \phi_{cutoff}\}$. Each of the features in AFS uniquely corresponds to $d \leqslant 20$ amino acids from the standard 20.

2.3 Protein family-wise partition of AFS using SVM

We train a linear SVM, to classify P vs Q, using the composition of the amino acids in AFS as the features, i.e. using $\mathbf{x}_j^{AFS} \in [0,1]^d$, with $x_{j,i'}^{AFS} = x_{j,i}^{AAC}$ and each $i' \in \{1, 2, \cdots, d\}$ uniquely maps to a $i \in AFS$. We use these linear SVM weights $\mathbf{w} \in \mathbb{R}^d$ to divide the set AFS into disjoint sets $AFS(P)$ and $AFS(Q)$ based on the sign of the weights. Since $x_{j,i'}^{AFS} \geqslant 0 \; \forall i' \in [d]$, the sign of the linear classifier weight $w_{i'}$ indicates which class is relatively prominent in the amino acid corresponding to i'. So if the $+1$ class is P, then we divide AFS class-wise as $AFS(P) = \{i' \in [d] : w_{i'} > 0\}$ and similarly $AFS(Q) = \{i' \in [d] : w_{i'} < 0\}$. See Appendix Section D for details on SVM training.

A flowchart summarizing the steps for computing $AFS(P)$ and $AFS(Q)$ is shown in Figure 1.

2.4 Validation of AFS

Literature evidence: For 14 different paralog protein pairs, we provide supporting evidence from protein biology literature for the significance of amino acids in AFS in the functional specificity of the protein pair.

MSA analysis: We also compute multiple sequence alignment (MSA) of randomly selected sequences from D_P and D_Q and analyze the conservation of $AFS(P)$ and $AFS(Q)$ amino acids within and across the respective families (Figure 2). MSA algorithms (Edgar & Batzoglou, 2006) aim to align multiple protein sequences by inserting gaps in the sequences while optimizing an objective. The objective is usually to minimize the number of gaps inserted while maximizing an overall score that promotes the alignment of similar (based on physicochemical properties) amino acids at a given position. The alignments are often used as a tool to determine homologous relationships between proteins and identify conserved or mutated regions in them.

Structural analysis: For paralog pairs that together function as heteromers (protein complexes made up of different types of proteins), we perform structural analysis to validate the role of AFS in the heteromeric structure formed by the paralog pair (Sections 3.1.7, 3.1.3 and 3.1.4).

Using test data: We test the classifier trained in Section 2.3 on a test data. (Details on test data in Appendix Section A.1). In general, we find an imbalance in the number of sequences for the two paralogous proteins. It is known that accuracy is not a well-suited performance measure of the classifier in class imbalance settings. Therefore, we use the arithmetic mean of sensitivity and specificity (AM) to measure the performance of the classifier (Brodersen et al., 2010).

Using marginal contribution feature importance (MCI): We check agreement of AFS with another feature ranking method, MCI (Catav et al., 2021). See Appendix Section E.3 for details on MCI computation.

3 Results and Discussions

3.1 Role of the amino acids identified in AFS

For 15 paralog pairs, we discuss the significance of the amino acids identified in the respective AFS (Table 1).

3.1.1 Lysozyme C and α-Lactalbumin

Literature evidence: Amino acids D and E of $AFS(\alpha\text{-Lactalbumin})$ are found in the Ca^{2+} and Zn^{2+} binding sites respectively of α-lactalbumin (Permyakov & Berliner, 2000; Permyakov, 2020). All α-lactalbumins studied so far are known to bind Ca^{2+} and Zn^{2+} whereas several (but not all) lysozymes do not bind Ca^{2+}.

MSA analysis: (Figure 2a) $AFS(\alpha\text{-Lactalbumin})$ and $AFS(\text{Lysozyme C})$ amino acids (Table 1) are significantly conserved in respective families.

3.1.2 Trypsin and Chymotrypsin

Literature evidence: Y and W get the highest Shapley value $\phi(\cdot)$ in $AFS(\text{Trypsin})$ and $AFS(\text{Chymotrypsin})$ respectively (Table 1 and Figure 7b). In experiments to convert trypsin to chymotrypsin (Hedstrom et al., 1994; Hedstrom, 2002) it has been shown that Y to W conversion in loop-3 of trypsin leads to significant increase in chymotrypsin activity. We do not find S, H and D in AFS, which are important for the function of both families and are known as the catalytic triad (Dodson & Wlodawer, 1998).

3.1.3 Tubulin-α and Tubulin-β

MSA analysis: (Appendix Figure 10) $AFS(\text{Tubulin-}\alpha)$ and $AFS(\text{Tubulin-}\beta)$ amino acids are significantly conserved in respective families.

Structural analysis of AFS: Tubulins typically exist as heterodimers, consisting of two subunits: tubulin-α and tubulin-β (Mühlethaler et al., 2021). We looked at the contact residues of a tubulin-α chain and tubulin-β chain in the 3D structure of tubulin-α/β heterodimer (PDB IDs: 3JAR, 5N5N). We see that the contact points

Table 1: AFS and its class-wise partition computed for 15 paralog pairs. The number of unique sequences from the SwissProt (The UniProt Consortium, 2020) database used for computing AFS is given inside parenthesis $(\cdot)$ for each protein family. Data collection details are in Appendix Section A.1. AFS amino acids are written in decreasing Shapley values from left to right for each paralog pair. Figures 3 and 7 show the Shapley value of the amino acids for each paralog pair. For globins and GPCRs, common acids across different AFS within a paralog triplet are colour-coded.

Paralog pair	**Amino acid feature subset**, AFS	**Class-wise** AFS **parition**
Lysozyme C (74) and α-Lactalbumin (22)	$\{I, A, D, N, G, R, E, F, L, W\}$	$AFS(\alpha\text{-Lactalbumin}) = \{I, D, E, F, L\}$ $AFS(\text{Lysozyme C}) = \{A, N, G, R, W\}$
Trypsin (66) and Chymotrypsin (17)	$\{Y, W, T, A, V, K, P\}$	$AFS(\text{Trypsin}) = \{Y, A\}$ $AFS(\text{Chymotrypsin}) = \{W, T, V, K, P\}$
Tubulin-α (117) and Tubulin-β (191)	$\{M, Q, K, N, F, I, H, A, C, Y\}$	$AFS(\text{Tubulin-}\alpha) = \{K, I, H, C, Y\}$ $AFS(\text{Tubulin-}\beta) = \{M, Q, N, F, A\}$
Histone H2A (180) and Histone H2B (177)	$\{L, G, S, M, K, N, T, Y, F\}$	$AFS(\text{Histone H2A}) = \{L, G, N\}$ $AFS(\text{Histone H2B}) = \{S, M, K, T, Y, F\}$
Interleukin-1 α (16) and Interleukin-1 β (25)	$\{C, G, T, S, V, Q, A, N, P\}$	$AFS(\text{Interleukin-1 }\alpha) = \{T, S, A, N\}$ $AFS(\text{Interleukin-1 }\beta) = \{C, G, V, Q, P\}$
Cytochrome P450 CYP3 (32) and CYP51 (32)	$\{H, F, G, K, A, P, N\}$	$AFS(\text{CYP3}) = \{F, K, P, N\}$ $AFS(\text{CYP51}) = \{H, G, A\}$
Globins		
Myoglobin (107) and Hemoglobin-α (303)	$AFS_1 = \{E, S, Y, V, K, P, I, G, C, W\}$	$AFS_1(\text{Myoglobin}) = \{E, K, I, G, W\}$ $AFS_1(\text{Hemoglobin-}\alpha) = \{S, Y, V, P, C\}$
Myoglobin (107) and Hemoglobin-β (285)	$AFS_2 =$ $\{K, V, C, E, W, N, F, M, Y, I\}$	$AFS_2(\text{Myoglobin}) = \{K, E, M, I\}$ $AFS_2(\text{Hemoglobin-}\beta) =$ $\{V, C, W, N, F, Y\}$
Hemoglobin-α (303) and Hemoglobin-β (285)	$AFS_3 = \{W, P, N, S, G\}$	$AFS_3(\text{Hemoglobin-}\alpha) = \{P, S\}$ $AFS_3(\text{Hemoglobin-}\beta) = \{W, N, G\}$
GPCRs		
Rhodopsin-like (181) and Glutamate-like (89)	$AFS_1 = \{D, Q, E, G, M, L\}$	$AFS_1(\text{Rhodopsin}) = \{M, L\}$ $AFS_1(\text{Glutamate}) = \{D, Q, E, G\}$
Secretin-like (90) and Glutamate-like (89)	$AFS_2 = \{W, H, Y, V, D\}$	$AFS_2(\text{Secretin}) = \{W, H, Y\}$ $AFS_2(\text{Glutamate}) = \{V, D\}$
Rhodopsin-like (181) and Secretin-like (90)	$AFS_3 = \{W, E, M, S, V, H, Q, A\}$	$AFS_3(\text{Rhodopsin}) = \{M, S, V, A\}$ $AFS_3(\text{Secretin}) = \{W, E, H, Q\}$
Rhodopsin-like GPCRs		
Aminergic receptors (186) and Lipid receptors (113)	$AFS_1 = \{L, P, E, W, F, M, D\}$	$AFS_1(\text{Aminergic receptors}) =$ $\{P, E, W, D\}$ $AFS_1(\text{Lipid receptors}) = \{L, F, M\}$
Aminergic receptors (186) and Peptide receptors (367)	$AFS_2 = \{L, F, E, M, K, D, V, R\}$	$AFS_2(\text{Aminergic receptors}) =$ $\{E, K, D, R\}$ $AFS_2(\text{Peptide receptors}) = \{L, F, M, V\}$
Lipid receptors (113) and Peptide receptors (367)	$AFS_3 = \{P, R, G, I, W, S, V\}$	$AFS_3(\text{Lipid receptors}) = \{R, G, S\}$ $AFS_3(\text{Peptide receptors}) = \{P, I, W, V\}$

of the tubulin-α chain in the heterodimer have more $AFS(\text{Tubulin-}\alpha)$ amino acids than $AFS(\text{Tubulin-}\beta)$. Similarly, $AFS(\text{Tubulin-}\beta)$ amino acids are more than $AFS(\text{Tubulin-}\alpha)$ at the contact point of the tubulin-β chain in the heterodimer. Thus, the amino acids identified in AFS can be considered to be significant towards the quaternary structure of tubulin-α/β heterodimer. Appendix Section E.1 has more details.

3.1.4 Histone H2A and Histone H2B

MSA analysis: (Appendix Figure 11), $AFS(\text{Histone H2A})$ and $AFS(\text{Histone H2B})$ amino acids are significantly conserved in respective families.

Structural analysis of AFS: Histones have a heterooctameric structure comprising of two H2A/H2B dimers and one H3/H4 tetramer (Dutta et al., 2001). We looked at the contact residues of an H2A chain and H2B chain in the heteroocatmer structure of histone (PDB IDs: 3KWQ, 1AOI). We find that the contact points of H2A chain in the heterooctamer have more AFS(Histone H2A) amino acids than AFS(Histone H2B). This is interesting since AFS(Histone H2A) has only three amino acids, while AFS(Histone H2B) has six amino acids. Similarly, the contact points of H2B chain in the heterooctamer have more AFS(Histone H2B) amino acids than AFS(Histone H2A). Thus, the amino acids identified in AFS can be considered to be significant towards the quaternary structure of the histone heterooctamer. See Appendix Section E.2 for more details.

3.1.5 Interleukin-1 α and Interleukin-1 β

Literature Evidence: C has the highest Shapley value and is in AFS(Interleukin-1 β). Deleting C results in loss of activity in Interleukin-1 β (Veerapandian et al., 1992). We do not find such studies for Interleukin-1 α.

MSA analysis: (Appendix Figure 12) AFS(Interleukin-1 α) and AFS(Interleukin-1 β) amino acids show significant conservation in respective families.

3.1.6 Cytochrome P450 CYP3 and CYP51

Literature evidence: H, F and G, in the respective order, have the highest Shapley value $\phi(\cdot)$ for this paralogous pair (Table 1 and Figure 7f). H and G with the highest $\phi(\cdot)$ in AFS(CYP51) have been reported (Nitahara et al., 2001; Lepesheva & Waterman, 2004; 2007; Strushkevich et al., 2010) to be important in the enzymatic activity of CYP51. Mutation of these amino acids at specific positions has been shown to result in a decrease in the activity of the enzyme (Lepesheva & Waterman, 2007; 2004). Similarly, F with the highest $\phi(\cdot)$ in AFS(CYP3) is also known to be important in the enzymatic activity of CYP3 (Qiu et al., 2008; Denisov et al., 2019; Zhang et al., 2024). A cluster of F residues in CYP3 is known to form a substrate-binding pocket with an active site (Zhang et al., 2024).

3.1.7 Globins

MSA analysis: (Figures 2b,2c and Appendix Figure 9) For the three globin paralog pairs (Table 1), we observe in the MSA, conservation of the class-wise partition of AFS in the respective families.

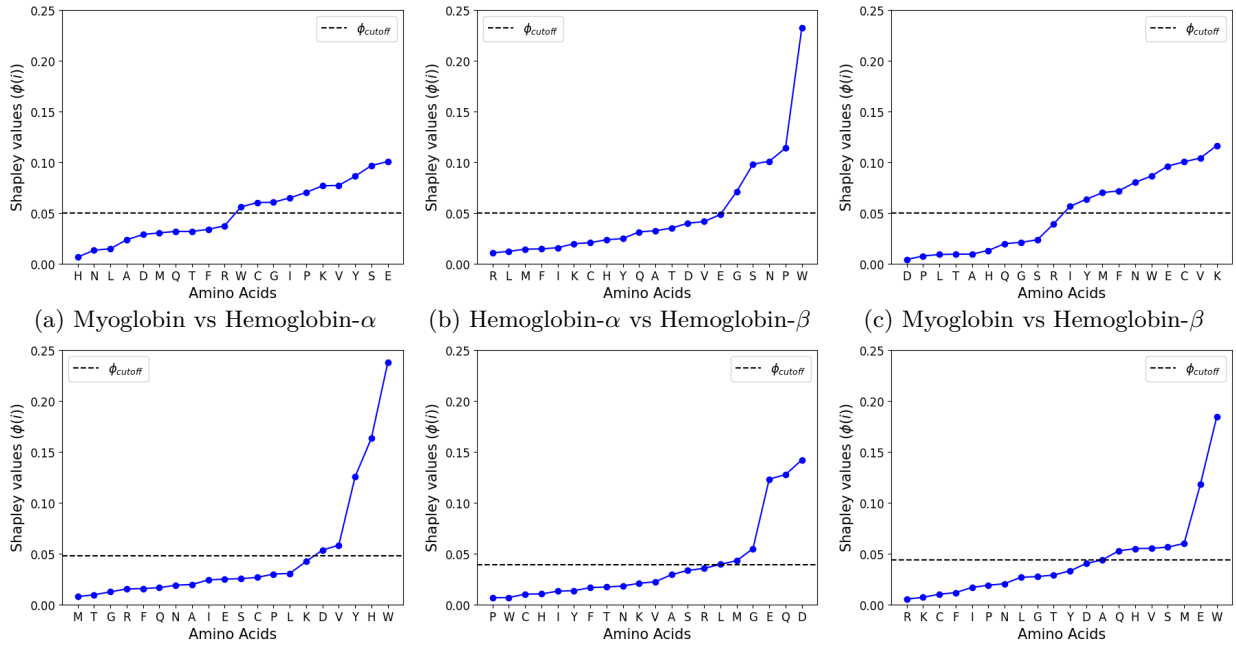
Structural analysis of AFS: Myoglobin is a monomer, while α and β chains together constitute hemoglobin, a tetramer of composition $\alpha_2\beta_2$ (Dill et al., 2017). We superimposed the 3D structures of myoglobin, hemoglobin-α and hemoglobin-β (PDB IDs: 3RGK, 1HHO) and mapped the α, β contact residues (based on (Shionyu et al., 2001)) of hemoglobin tetramer to that of myoglobin. We find that the amino acids K, E, I, which are common in AFS_1(Myoglobin) and AFS_2(Myoglobin), are less in number at the contact residues of hemoglobin tetramer and more in number at the corresponding locations in myoglobin, which is a monomer (see Figure 4).

Literature evidence: W with a significantly high Shapley value $\phi(W)$ (Figure 3b), is present in AFS_3(Hemoglobin-β). It is highly conserved at position 40 in the MSA (Figure 2c) in hemoglobin-β sequences as compared to hemoglobin-α sequences. This W at position 40 has been determined to be present in hemoglobin-β at one of its contact positions to hemoglobin-α in the tetrameric structure (Shionyu et al., 2001) and is, therefore, a structurally and functionally significant residue. C, present in AFS_1(Hemoglobin-α) and AFS_2(Hemoglobin-β), has been shown to play an important role in the tetrameric structure of hemoglobin formed by α and β hemoglobins (Kan et al., 2013).

Logical consistencies in AFS (refer to Table 1 (Globins) for AFS_1, AFS_2, AFS_3):

- $AFS_1 \cap AFS_2 = \{E, Y, V, K, I, C, W\}$. Except for W with the least Shapley value in AFS_1 (Figure 3a), the remaining are excluded from AFS_3.
- *Explanation*:V, Y, C in AFS_1(Hemoglobin-α) $\cap$ AFS_2(Hemoglobin-β) can be expected not to be *key* in AFS_3 for distinguishing α vs β hemoglobin.

Figure 2: Multiple sequence alignment of sequences from the respective families in (a), (b) and (c). Within each alignment, 15 sequences on the left are from one family, and those on the right are from the other family in each of (a), (b) and (c). The sequences are randomly selected from the train set of the families. For each aligned sequence in (a) AFS(α-Lactalbumin) amino acids are in green and AFS(Lysozyme C) are in red, in (b) the amino acids in AFS_1(Myoglobin) are in green and AFS_1(Hemoglobin-α) are in red, and in (c) the amino acids in AFS_2(Hemoglobin-α) are in green and AFS_2(Hemoglobin-β) are in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (Figures 3 and 7).



(a) Myoglobin vs Hemoglobin-α (b) Hemoglobin-α vs Hemoglobin-β (c) Myoglobin vs Hemoglobin-β

(d) Secretin-like vs Glutamate-like (e) Rhodopsin-like vs Glutamate-like (f) Rhodopsin-like vs Secretin-like

Figure 3: Shapley value ($\phi(i)$) for AAC features computed using SVEA. See Appendix Figure 7 for remaining paralogs.

• $AFS_2 \cap AFS_3 = \{W, N\}$. N is excluded from AFS_1, while W gets the least Shapley value in AFS_1 (Figure 3a).

• $AFS_3 \cap AFS_1 = \{W, P, S, G\}$. $\{P, S, G\}$ are excluded from AFS_2, while W gets the least Shapley value in AFS_1.

The Shapley value for W is very close to the cut-off in AFS_1 (Figure 3a). If it is dropped from AFS_1, then the exclusion principle illustrated above would be more prominent as in GPCRs (Section 3.1.8).

3.1.8 G-protein coupled receptors (GPCRs)

Literature evidence: W (with highest Shapley value $\phi(\cdot)$) and H common in AFS_2(Secretin) and AFS_3(Secretin) (Table 1 and Figure 3), are well conserved at multiple positions with structural importance and functional importance in secretin-like GPCR sequences (Cary et al., 2022; Harmar, 2001). Mutating certain conserved W leads to a loss in expression of this GPCR at the cell surface, where it functions (Cary et al., 2022). H present in the intracellular loop region is also known to be important in the activation of certain secretin-like GPCRs (Harmar, 2001).

M common in AFS_1(Rhodopsin) and AFS_3(Rhodopsin) has been found to be present at important binding pockets and a position important for activation of the GPCR (Okada et al., 2001; Sakmar et al., 2002). S from AFS_3(Rhodopsin) is found at multiple major phosphorylation sites (see Okada et al. 2001 for details) in Rhodopsin.

Mutating D at two positions has been shown to affect glutamate binding of glutamate receptor GPCRs (Jingami et al., 2003). D is common in AFS_1(Glutamate) and AFS_2(Glutamate) and has highest Shapley value in AFS_1.

E and D common in AFS_1(Aminergic) and AFS_2(Aminergic) are present at binding sites of important ligands (like histamine/serotonin) of aminergic receptors (Vass et al., 2019).

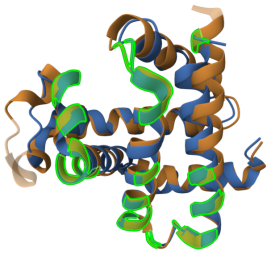
Logical consistencies in AFS of GPCRs (refer to Table 1 (GPCRs) for AFS_1, AFS_2, AFS_3):

```
3RGK.A/Myoglobin/1-40        GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETL
1HHO.A/Hemoglobin-α/1-40     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTK

3RGK.A/Myoglobin/41-80       EKFDRFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKG
1HHO.A/Hemoglobin-α/41-74    TYFPHFDL------SHGSAQVKGHGKKVADALTNAVAHVD

3RGK.A/Myoglobin/81-120      HHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
1HHO.A/Hemoglobin-α/75-114   DMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP

3RGK.A/Myoglobin/121-147     GDFGADAQGAMNKALELFRKDMASNYK
1HHO.A/Hemoglobin-α/115-141  AEFTPAVHASLDKFLASVSTVLTSKYR
```



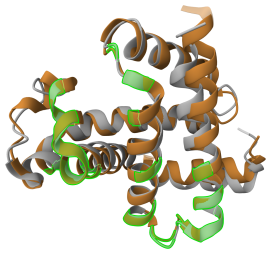
(a) alignment of myoglobin (brown) and hemoglobin-α (blue) sequences based on structure alignment (right)

```
3RGK.A/Myoglobin/1-40        GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETL
1HHO.B/Hemoglobin-β/1-38     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQ

3RGK.A/Myoglobin/41-80       EKFDRFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKG
1HHO.B/Hemoglobin-β/39-78    RFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD

3RGK.A/Myoglobin/81-120      HHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHP
1HHO.B/Hemoglobin-β/79-118   NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG

3RGK.A/Myoglobin/121-147     GDFGADAQGAMNKALELFRKDMASNYK
1HHO.B/Hemoglobin-β/119-145  KEFTPPVQAAYQKVVAGVANALAHKYH
```



(b) alignment of myoglobin (brown) and hemoglobin-β (silver) sequences based on structure alignment (right)

Figure 4: The hemologlobin-α/β tetramer contact points (as identified in Table 3 and Table 4 of Shionyu et al. (2001)) are highlighted in the hemoglobin sequences. These have also been highlighted in green in the structure alignments (right). The highlighted amino acids in the myoglobin chain correspond to the positions that align with hemologlobin-α/β tetramer contact points. The amino acids K, E, I, which are common in AFS_1(Myoglobin) and AFS_2(Myoglobin), are in **bold** in the alignment. We find that K, E, I are less in number at the contact residues of hemoglobin tetramer and more in number at the corresponding locations in myoglobin, which is a monomer. Structure alignment is done using the online tool available at `https://www.rcsb.org/alignment`, with the default parameter settings (`|algorithm: jFATCAT(rigid)||RMSD Cutoff: 3||AFP Distance Cutoff: 1600||Fragment Length: 8|`).

- $AFS_1 \cap AFS_2 = \{D\}$, is excluded from AFS_3.
- $AFS_2 \cap AFS_3 = \{W, H, V\}$, is excluded from AFS_1.
- $AFS_3 \cap AFS_1 = \{Q, E, M\}$, is excluded from AFS_2.

Logical consistencies in AFS of Rhodopsin-like GPCR subfamilies (refer to Table 1 (Rhodopsin-like GPCRs) for AFS_1, AFS_2, AFS_3):

- $AFS_1 \cap AFS_2 = \{L, E, F, M, D\}$, is excluded from AFS_3.
- $AFS_2 \cap AFS_3 = \{R, V\}$, is excluded from AFS_1.
- $AFS_3 \cap AFS_1 = \{P, W\}$ is excluded from AFS_2.

The explanations for these consistencies are similar to that in globins (Section 3.1.7).

3.2 Validation of AFS using test data

The classification scores on test data for the classifiers trained using AAC and AFS features, respectively, are reported in Appendix Table 5. Using AFS features, the test AM scores are at least 70%. For 13 of 15 paralog pairs, the scores are greater than 83%, and for 8 of 15 paralog pairs, it is greater than 90%. Details of the test data are provided in Appendix Section A.1.

3.2.1 Test scores on random feature subsets

For each paralog pair, we train classifiers using 100 random feature subsets of the same size as AFS. We compute the mean test AM scores of these classifiers with the AFS-based classifier (See Appendix Table 7). The mean test AM score drops for these random feature subsets in comparison to the AFS-based score for all 15 paralog pairs. The mean test AM further drops for 14 pairs, if the randomly selected amino acids contain only non-AFS amino acids. Details in Appendix Section E.5.

3.3 Marginal contribution feature importance (MCI) of AFS

For an AFS of size d, the top-d amino acids ranked by MCI differ with AFS only in at the most two amino acids. For 8 of 15 datasets, AFS and top-d MCI sets are the same, while only for two datasets do they differ in two amino acids. For all 15 datasets, at least the top-3 MCI amino acids are in AFS. For 11 of these datasets, at least the top-5 MCI amino acids are in AFS. (Appendix Table 6)

3.4 Relating AFS to diseases

We report some examples that link AFS amino acids to diseases. This is based on evidence from biology literature that suggests the role of these amino acids in the function/structure of the respective families. Thus, modification/mutation of these residues may be disease-causing as it could affect the structure and/or function of the protein, leading to disruption of one or more physiological processes.

Example 1: W has the highest Shapley value in AFS_2(Secretin) and AFS_3(Secretin) (Table 1). Mutating certain conserved W leads to a loss in cell-surface expression of calcitonin generelated peptide receptor (CGRPR), a secretin-like GPCR (Cary et al., 2022). A study (Ringer et al., 2017) in mice reports that disruption in CGRPR signalling accelerates Amyotrophic Lateral Sclerosis (ALS).

Example 2: (AFS of hemoglobin-β - from myoglobin vs hemoglobin-β and hemoglobin-α vs hemoglobin-β) (Table 1)

(a) W with a significantly high Shapley value $\phi(W)$ (Figure 3(b)), is present in AFS_3(Hemoglobin-β). It is also in AFS_2(Hemoglobin-β) (Table 1). W to S and W to R mutation at a certain position in hemoglobin-β has been reported to result in abnormal hemoglobins - Hemoglobin Hirose (Sasaki et al., 1978) and Hemoglobin Rothschild (Sharma et al., 1980), respectively. These have altered oxygen affinities and dissociation of hemoglobin tetramer to dimers.

(b) N is common in AFS_3(Hemoglobin-β) and AFS_2(Hemoglobin-β). Mutation/deletion of N has been associated with β-thalassemia (Thein, 2013).

(c) V has the highest Shapley value in AFS_2(Hemoglobin-β) and mutations of this residue are related to hemoglobin variants with altered structure and biochemical properties leading to varying physiological effects (Thom et al., 2013). An example is Hemoglobin Olympia (Stamatoyannopoulos et al., 1973), having V at a particular position mutated to M.

4 Conclusion

We demonstrated an ML pipeline to identify the key amino acid types, AFS, that distinguish a pair of paralogous proteins. The role of AFS in functionally distinguishing the paralog pairs was validated using various sources of domain knowledge. The robustness of this approach, as demonstrated by considering a diverse set of paralogous protein pairs, illustrates its wider applicability. Identification of AFS can be used as an initial data-driven step before doing more detailed experimental investigations, like site-directed mutagenesis (Bachman, 2013) resolving sequence-function relationship. Our experiments suggest that amino acids with high Shapley values are more likely to play a role in the functional difference between the paralogs and, hence, can be targeted for site-directed mutagenesis experiments. As the size of AFS is small (5-10 amino acids of 20), significantly fewer mutations can be tried.

As our pipeline works without using the sequence order information of the amino acids in the protein, it posits an interesting question to biologists : how amino acid composition by itself is able to distinguish paralogs given ample evidence that 3D structure and function are conserved despite sequence divergence (Lau et al., 2015)! Notably, amino acids in the AFS typically occur more than once in the sequence, but our method is silent on the specific positions where the amino acid has a functionally distinguishing role. This may be addressed by engineering features that incorporate sequence order information from the protein. However, these features can be very high-dimensional, for example, 20^k-dimensional for k-mer features. The Monte Carlo based approximation algorithm for Shapley values would require exponentially more sampling (in number of features) for good approximations.

Impact Statement

This paper presents a computationally efficient data lean ML pipeline. It can be used by biologists to decide whether they should invest valuable resources (skilled manpower, time, funds, etc.) for performing wet-lab experiments to determine amino acid(s) that are critical for functional differentiation of paralogous proteins.

References

Julia Bachman. Chapter ninteen - site-directed mutagenesis. In Jon Lorsch (ed.), *Laboratory Methods in Enzymology: DNA*, volume 529 of *Methods in Enzymology*, pp. 241–248. Academic Press, 2013. doi: 10.1016/B978-0-12-418687-3.00019-7. URL https://www.sciencedirect.com/science/article/pii/B9780124186873000197.

Khodeza Begum, Jonathon E Mohl, Fredrick Ayivor, Eder E Perez, and Ming-Ying Leung. GPCR-PEnDB: a database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors. *Database*, 2020, 11 2020. ISSN 1758-0463. doi: 10.1093/database/baaa087. URL https://doi.org/10.1093/database/baaa087.

Maxwell L Bileschi, David Belanger, Drew H Bryant, Theo Sanderson, Brandon Carter, D Sculley, Alex Bateman, Mark A DePristo, and Lucy J Colwell. Using deep learning to annotate the protein universe. *Nature biotechnology*, 40(6):932937, June 2022. ISSN 1087-0156. doi: 10.1038/s41587-021-01179-w. URL https://doi.org/10.1038/s41587-021-01179-w.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th ICPR*, pp. 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

Brian P Cary, Xin Zhang, Jianjun Cao, Rachel M Johnson, Sarah J Piper, Elliot J Gerrard, Denise Wootten, and Patrick M Sexton. New Insights into the Structure and Function of Class B1 GPCRs. *Endocrine Reviews*, 44(3):492–517, 12 2022. ISSN 0163-769X. doi: 10.1210/endrev/bnac033. URL https://doi.org/10.1210/endrev/bnac033.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. ISSN 0305-0548. doi: 10.1016/j.cor.2008.04.004. URL https://www.sciencedirect.com/science/article/pii/S0305054808000804. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance - an axiomatic approach for explaining data. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1324–1335. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/catav21a.html.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and

H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17212–17223. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf.

Ilia G. Denisov, Yelena V. Grinkova, Prithviraj Nandigrami, Mrinal Shekhar, Emad Tajkhorshid, and Stephen G. Sligar. Allosteric interactions in human cytochrome p450 cyp3a4: The role of phenylalanine 213. *Biochemistry*, 58(10):1411–1422, 2019. doi: 10.1021/acs.biochem.8b01268. URL https://doi.org/10.1021/acs.biochem.8b01268. PMID: 30785734.

K. Dill, R.L. Jernigan, and I. Bahar. *Protein Actions: Principles and Modeling*. CRC Press, 2017. ISBN 9781351815000. URL https://books.google.co.in/books?id=NHs2DwAAQBAJ.

Guy Dodson and Alexander Wlodawer. Catalytic triads and their relatives. *Trends in Biochemical Sciences*, 23(9):347–352, 1998. ISSN 0968-0004. doi: https://doi.org/10.1016/S0968-0004(98)01254-7. URL https://www.sciencedirect.com/science/article/pii/S0968000498012547.

Shuchismita Dutta, Ildikó V. Akey, Colin Dingwall, Kari L. Hartman, Tom Laue, Robert T. Nolte, James F. Head, and Christopher W. Akey. The crystal structure of nucleoplasmin-core: Implications for histone binding and nucleosome assembly. *Molecular Cell*, 8(4):841–853, 2001. ISSN 1097-2765. doi: 10.1016/S1097-2765(01)00354-9. URL https://www.sciencedirect.com/science/article/pii/S1097276501003549.

Robert C Edgar and Serafim Batzoglou. Multiple sequence alignment. *Current Opinion in Structural Biology*, 16(3):368–373, 2006. ISSN 0959-440X. doi: 10.1016/j.sbi.2006.04.004. URL https://www.sciencedirect.com/science/article/pii/S0959440X06000704. Nucleic acids/Sequences and topology.

Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts565. URL https://doi.org/10.1093/bioinformatics/bts565.

Paola Galozzi, Sara Bindoli, Andrea Doria, and Paolo Sfriso. The revisited role of interleukin-1 alpha and beta in autoimmune and inflammatory disorders and in comorbidities. *Autoimmunity Reviews*, 20(4):102785, 2021. ISSN 1568-9972. doi: 10.1016/j.autrev.2021.102785. URL https://www.sciencedirect.com/science/article/pii/S1568997221000483.

Tatiana Y. Hargrove, Kwangho Kim, Maria de Nazaré Correia Soeiro, Cristiane França da Silva, Denise da Gama Jaen Batista, Marcos Meuser Batista, Eugenia M. Yazlovitskaya, Michael R. Waterman, Gary A. Sulikowski, and Galina I. Lepesheva. Cyp51 structures and structure-based development of novel, pathogen-specific inhibitory scaffolds. *International Journal for Parasitology: Drugs and Drug Resistance*, 2:178–186, 2012. ISSN 2211-3207. doi: 10.1016/j.ijpddr.2012.06.001. URL https://www.sciencedirect.com/science/article/pii/S2211320712000206. Including Articles from Keystone Symposium on Drug Discovery for Protozoan Parasites; pp. 230270.

AJ Harmar. Family-b g-protein-coupled receptors. *Genome biology*, 2(12):REVIEWS3013, 2001. ISSN 1474-7596. doi: 10.1186/gb-2001-2-12-reviews3013. URL https://europepmc.org/articles/PMC138994.

Lizbeth Hedstrom. Serine protease mechanism and specificity. *Chemical Reviews*, 102(12):4501–4524, 2002. doi: 10.1021/cr000033x. URL https://doi.org/10.1021/cr000033x. PMID: 12475199.

Lizbeth Hedstrom, John J Perona, and William J. Rutter. Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant. *Biochemistry*, 33 29:8757–63, 1994.

Hisato Jingami, Shigetada Nakanishi, and Kosuke Morikawa. Structure of the metabotropic glutamate receptor. *Current Opinion in Neurobiology*, 13(3):271–278, 2003. ISSN 0959-4388. doi: 10.1016/S0959-4388(03)00067-9. URL https://www.sciencedirect.com/science/article/pii/S0959438803000679.

Heng-I Kan, I-Ying Chen, Muhammad Zulfajri, and Chia C. Wang. Subunit disassembly pathway of human hemoglobin revealing the site-specific role of its cysteine residues. *The Journal of Physical Chemistry B*, 117(34):9831–9839, 2013. doi: 10.1021/jp402292b. URL https://doi.org/10.1021/jp402292b. PMID: 23902424.

Nicole Kresge, Robert D. Simoni, and Robert L. Hill. The development of site-directed mutagenesis by michael smith. *Journal of Biological Chemistry*, 281(39):e31–e33, 2006. ISSN 0021-9258. doi: 10.1016/S0021-9258(19)33938-9. URL https://www.sciencedirect.com/science/article/pii/S0021925819339389.

Clinton K.Y. Lau, Louise Turner, Jakob S. Jespersen, Edward D. Lowe, Bent Petersen, Christian W. Wang, Jens E.V. Petersen, John Lusingu, Thor G. Theander, Thomas Lavstsen, and Matthew K. Higgins. Structural conservation despite huge sequence diversity allows epcr binding by the pfemp1 family implicated in severe childhood malaria. *Cell Host & Microbe*, 17(1):118–129, 2015. ISSN 1931-3128. doi: 10.1016/j.chom.2014.11.007. URL https://www.sciencedirect.com/science/article/pii/S1931312814004235.

Galina I Lepesheva and Michael R Waterman. Cyp51-the omnipotent p450. *Molecular and Cellular Endocrinology*, 215(1):165–170, 2004. ISSN 0303-7207. doi: 10.1016/j.mce.2003.11.016. URL https://www.sciencedirect.com/science/article/pii/S0303720703005148. Proceedings of the Serono Foundation for the Advancement of Medical Science Workshop on Molecular Steroidogenesis.

Galina I. Lepesheva and Michael R. Waterman. Sterol 14α-demethylase cytochrome p450 (cyp51), a p450 in all biological kingdoms. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1770(3):467–477, 2007. ISSN 0304-4165. doi: 10.1016/j.bbagen.2006.07.018. URL https://www.sciencedirect.com/science/article/pii/S0304416506002145. P450.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Andrew G. McArthur, Tove Hegelund, Rachel L. Cox, John J. Stegeman, Mette Liljenberg, Urban Olsson, Per Sundberg, and Malin C. Celander. Phylogenetic Analysis of the Cytochrome P450 3 (CYP3) Gene Family. *Journal of Molecular Evolution*, 57(2):200–211, August 2003. doi: 10.1007/s00239-003-2466-x. URL https://link.springer.com/article/10.1007/s00239-003-2466-x.

Aditya Krishna Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th ICML - Volume 28*, ICML'13, pp. III603III611. JMLR.org, 2013.

Tobias Mühlethaler, Dario Gioia, Andrea E. Prota, May E. Sharpe, Andrea Cavalli, and Michel O. Steinmetz. Comprehensive analysis of binding sites in tubulin. *Angewandte Chemie International Edition*, 60(24):13331–13342, 2021. doi: 10.1002/anie.202100273. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202100273.

Y Narahari. *Game Theory and Mechanism Design*. WORLD SCIENTIFIC / INDIAN INST OF SCIENCE, INDIA, 2014. doi: 10.1142/8902. URL https://www.worldscientific.com/doi/abs/10.1142/8902.

Yuko Nitahara, Kae Kishimoto, Yoshiyasu Yabusaki, Osamu Gotoh, Yuzo Yoshida, Tadao Horiuchi, and Yuri Aoyama. The amino acid residues affecting the activity and azole susceptibility of rat cyp51 (sterol 14-demethylase p450). *The Journal of Biochemistry*, 129(5):761–768, 2001.

Tetsuji Okada, Oliver P. Ernst, Krzysztof Palczewski, and Klaus Peter Hofmann. Activation of rhodopsin: new insights from structural and biochemical studies. *Trends in Biochemical Sciences*, 26(5):318–324, 2001. ISSN 0968-0004. doi: 10.1016/S0968-0004(01)01799-6. URL https://www.sciencedirect.com/science/article/pii/S0968000401017996.

Eugene A. Permyakov. α-actalbumin, Amazing Calcium-Binding Protein. *Biomolecules*, 10(9):1210, Aug 2020. ISSN 2218-273X. doi: 10.3390/biom10091210. URL http://dx.doi.org/10.3390/biom10091210.

Eugene A. Permyakov and Lawrence J. Berliner. α-Lactalbumin: structure and function. *FEBS Letters*, 473(3):269–274, 2000. ISSN 0014-5793. doi: 10.1016/S0014-5793(00)01546-5. URL https://www.sciencedirect.com/science/article/pii/S0014579300015465.

Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82, 2021. doi: 10.1002/pro.3943. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3943.

Pradman K. Qasba, Soma Kumar, and Dr. K. Brew. Molecular divergence of lysozymes and α-lactalbumin. *Critical Reviews in Biochemistry and Molecular Biology*, 32(4):255–306, 1997. doi: 10.3109/10409239709082574. URL https://doi.org/10.3109/10409239709082574.

Huan Qiu, Stefan Taudien, Holger Herlyn, Juergen Schmitz, Yuan Zhou, Guopei Chen, Roberta Roberto, Mariano Rocchi, Matthias Platzer, and Leszek Wojnowski. Cyp3 phylogenomics: evidence for positive selection of cyp3a4 and cyp3a7. *Pharmacogenetics and Genomics*, 18(1):5366, January 2008. ISSN 1744-6872. doi: 10.1097/fpc.0b013e3282f313f8. URL https://doi.org/10.1097/FPC.0b013e3282f313f8.

Cornelia Ringer, Sarah Tune, Mirjam A Bertoune, Hans Schwarzbach, Kazutake Tsujikawa, Eberhard Weihe, and Burkhard Schütz. Disruption of calcitonin gene-related peptide signaling accelerates muscle denervation and dampens cytotoxic neuroinflammation in sod1 mutant mice. *Cellular and molecular life sciences : CMLS*, 74(2):339358, January 2017. ISSN 1420-682X. doi: 10.1007/s00018-016-2337-4. URL https://europepmc.org/articles/PMC11107523.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5572–5579. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/778. URL https://doi.org/10.24963/ijcai.2022/778. Survey Track.

Thomas P Sakmar, Santosh T Menon, Ethan P Marin, and Elias S Awad. Rhodopsin: Insights from recent structural studies. *Annual Review of Biophysics*, 31(Volume 31, 2002):443–484, 2002. ISSN 1936-1238. doi: 10.1146/annurev.biophys.31.082901.134348. URL https://www.annualreviews.org/content/journals/10.1146/annurev.biophys.31.082901.134348.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer, deep neural networks for protein functional inference. *eLife*, 12:e80942, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.80942. URL https://doi.org/10.7554/eLife.80942.

J. Sasaki, T. Imamura, and T. Yanase. Hemoglobin hirose, a human hemoglobin variant with a substitution at the alpha1beta2 interface. subunit dissociation and the equilibria and kinetics of ligand binding. *Journal of Biological Chemistry*, 253(1):87–94, 1978. ISSN 0021-9258. doi: 10.1016/S0021-9258(17)38273-X. URL https://www.sciencedirect.com/science/article/pii/S002192581738273X.

L. S. Shapley. *17. A Value for n-Person Games*, pp. 307–318. Princeton University Press, Princeton, 1953. ISBN 9781400881970. doi: doi:10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Vijay S. Sharma, Gerald L. Newton, Helen M. Ranney, Farouq Ahmed, John W. Harris, and Elizabeth H. Danish. Hemoglobin rothschild (β37(c3)trp arg): A high/low affinity hemoglobin mutant. *Journal of Molecular Biology*, 144(3):267–280, 1980. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(80)90090-X. URL https://www.sciencedirect.com/science/article/pii/002228368090090X.

M Shionyu, K Takahashi, and M Gō. Variable subunit contact and cooperativity of hemoglobins. *J. Mol. Evol.*, 53(4-5):416–429, October 2001.

Hyebin Song, Bennett J. Bremer, Emily C. Hinds, Garvesh Raskutti, and Philip A. Romero. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. *Cell Systems*, 12(1):92–101.e8, 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2020.10.007. URL https://www.sciencedirect.com/science/article/pii/S2405471220304142.

G Stamatoyannopoulos, PE Nute, JW Adamson, AJ Bellingham, and D Funk. Hemoglobin olympia (20 valine leads to methionine): an electrophoretically silent variant associated with high oxygen affinity and erythrocytosis. *The Journal of clinical investigation*, 52(2):342349, February 1973. ISSN 0021-9738. doi: 10.1172/jci107190. URL https://europepmc.org/articles/PMC302263.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Natallia Strushkevich, Sergey A. Usanov, and Hee-Won Park. Structural basis of human cyp51 inhibition by antifungal azoles. *Journal of Molecular Biology*, 397(4):1067–1078, 2010. ISSN 0022-2836. doi: 10.1016/j.jmb.2010.01.075. URL https://www.sciencedirect.com/science/article/pii/S0022283610001324.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100. URL https://doi.org/10.1093/nar/gkaa1100.

Swee Lay Thein. The molecular basis of β-thalassemia. *Cold Spring Harbor Perspectives in Medicine*, 3(5), 2013. doi: 10.1101/cshperspect.a011700. URL http://perspectivesinmedicine.cshlp.org/content/3/5/a011700.abstract.

Christopher S. Thom, Claire F. Dickson, David A. Gell, and Mitchell J. Weiss. Hemoglobin variants: Biochemical properties and clinical correlates. *Cold Spring Harbor Perspectives in Medicine*, 3(3), 2013. doi: 10.1101/cshperspect.a011858. URL http://perspectivesinmedicine.cshlp.org/content/3/3/a011858.abstract.

Sandhya Tripathi, N Hemachandra, and Prashant Trivedi. Interpretable feature subset selection: A Shapley value based approach. In *2020 IEEE BigData*, pp. 5463–5472, 2020. doi: 10.1109/BigData50022.2020.9378102.

Sandhya Tripathi, N. Hemachandra, and Prashant Trivedi. Interpretable feature subset selection: A shapley value based approach, 2021. URL https://arxiv.org/abs/2001.03956.

Márton Vass, Sabina Podlewska, Iwan J. P. de Esch, Andrzej J. Bojarski, Rob Leurs, Albert J. Kooistra, and Chris de Graaf. Aminergic gpcrligand interactions: A chemical and structural map of receptor mutation data. *Journal of Medicinal Chemistry*, 62(8):3784–3839, 2019. doi: 10.1021/acs.jmedchem.8b00836. URL https://doi.org/10.1021/acs.jmedchem.8b00836. PMID: 30351004.

B. Veerapandian, Gary L. Gilliland, Reetta Raag, Anders L. Svensson, Yoshihiro Masui, Yoshikatsu Hirai, and Thomas L. Poulos. Functional implications of interleukin-1β based on the three-dimensional structure. *Proteins: Structure, Function, and Bioinformatics*, 12(1):10–23, 1992. doi: 10.1002/prot.340120103. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340120103.

Yixuan Yang, Tanchumin Xu, Gavin Conant, Hirohisa Kishino, Jeffrey L Thorne, and Xiang Ji. Interlocus gene conversion, natural selection, and paralog homogenization. *Molecular Biology and Evolution*, 40(9):msad198, 09 2023. ISSN 1537-1719. doi: 10.1093/molbev/msad198. URL https://doi.org/10.1093/molbev/msad198.

H. P. Young. Monotonic solutions of cooperative games. *Int. J. Game Theory*, 14(2):6572, jun 1985. ISSN 0020-7276. doi: 10.1007/BF01769885. URL https://doi.org/10.1007/BF01769885.

Yuqing Zhang, Ziying Wang, Yuchao Wang, Weikai Jin, Zheyan Zhang, Lehao Jin, Jianchang Qian, and Long Zheng. Cyp3a4 and cyp3a5: the crucial roles in clinical drug metabolism and the significant implications of genetic polymorphisms. *PeerJ*, 12:e18636, 2024.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.319. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319.

A Data collection and code

We discuss the details of the data collection procedure for the datasets used in our computational experiments.

A.1 Datasets of 15 paralog pairs

We apply our method for identifying amino acid types that distinguish paralogous proteins using the datasets described in Table 2. Only the train set is used for computing AFS, while the test set is used for computing classification scores for the linear SVM trained using the train set.

Table 2: The number of sequences in the train and test sets of the protein families considered in computational experiments.

Family	Train (Swiss-Prot)	Test (TrEMBL)
Lysozyme-like		
α-Lactalbumin	22	53
Lysozyme C	74	14
Trypsin-like		
Trypsin	66	3813
Chymotrypsin	17	281
Tubulin		
α	117	190
β	191	347
Histone		
H2A	180	16599
H2B	177	7599
Interleukin-1		
α	16	12
β	25	194
Cytochrome P450		
CYP3	32	818
CYP51	32	601
Globins		
Myoglobin	107	479
Hemoglobin-α	303	525
Hemoglobin-β	285	261
	(GPCR-PEnDB)	
	Train (80%)	Test (20%)
GPCR families		
Rhodopsin-like	181	45
↳ Lipid receptors	113	28
Peptide receptors	367	92
Aminergic receptors	186	47
Glutamate-like	89	23
Secretin-like	90	23

All datasets are taken from publicly available databases (UniProt (The UniProt Consortium, 2020) and GPCR-PEnDB (Begum et al., 2020)). Well-known pairs of paralogous proteins were curated from millions of sequences from UniProt considering the number of sequences and manually reviewed labels available for them.

For all datasets except GPCR, we use manually curated Swiss-Prot sequences for training and electronically annotated TrEMBL sequences for testing. These proteins have very specific functions. In contrast, GPCRs are a large and diverse group of transmembrane proteins that mediate cellular responses to extracellular signals. We chose to use an already curated dataset in this case. For each of the GPCR families considered (Table 2), the sequences are randomly split as 80%-train/20%-test. The use of GPCR-PEnDB data is to illustrate the effectiveness of our method with random slicing, which is inevitable when additional curated data are not available. If one or many UniProt entries in a dataset had identical sequences, then only one of them was retained, and the remaining were deleted.

The following queries were used for collecting data from UniProt (The UniProt Consortium, 2020),

- **lysozyme C**: `(protein_name:"lysozyme C") AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200]) AND (ec:3.2.1.17) AND (xref:cazy-GH22)`
- **α-lactalbumin**: `(protein_name:"alpha lactalbumin") AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200])`
- **myoglobin**: `(protein_name:"myoglobin") AND (xref:interpro-IPR002335) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **hemoglobin-α**: `(protein_name:"hemoglobin alpha") AND (xref:interpro-IPR002338) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **hemoglobin-β**: `(protein_name:"hemoglobin beta") AND (xref:interpro-IPR002337) AND (fragment:false) NOT (existence:5) NOT (existence:4)`
- **trypsin**: `(protein_name:trypsin) AND (fragment:false) AND (ec:3.4.21.4) NOT (existence:5)`
- **chymotrypsin**: `(protein_name:chymotrypsin) AND (fragment:false) AND (ec:3.4.21.1) NOT (existence:5)`
- **tubulin-α**: `(protein_name:"tubulin alpha") AND (family:"tubulin family") AND (length:[300 TO 600]) AND (fragment:false) NOT (annotation_score:1) NOT (annotation_score:2)`
- **tubulin-β**: `(protein_name:"tubulin beta") AND (family:"tubulin family") AND (length:[300 TO 600]) AND (fragment:false) NOT (annotation_score:1) NOT (annotation_score:2)`
- **interleukin-1** α `(protein_name:"interleukin-1 alpha") AND (family:il-1) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[200 TO 400]) NOT (annotation_score:1)`
- **interleukin-1** β: `(protein_name:"interleukin-1 beta") AND (family:il-1) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[200 TO 400]) NOT (annotation_score:1)`
- **Histone H2A**: `(protein_name:"histone h2a") AND (family:histone) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200])`
- **Histone H2B**: `(protein_name:"histone h2b") AND (family:histone) AND (fragment:false) NOT (existence:4) NOT (existence:5) AND (length:[* TO 200])`
- **Cytochrome P450 CYP3**: `(family:"Cytochrome P450") AND ((gene:cyp3) OR (gene:cyp3A*)) AND (fragment:false) NOT (existence:4) NOT (existence:5) NOT (annotation_score:1)`
- **Cytochrome P450 CYP51**: `(family:"Cytochrome P450") AND ((gene:cyp51) OR (gene:cyp51A*) OR (gene:cyp51B*) OR (gene:cyp51C*)) AND (fragment:false) NOT (existence:4) NOT (existence:5) NOT (annotation_score:1)`

The GPCR sequences were collected from the GPCR-PEn database (URL: `https://gpcr.utep.edu/`) (Begum et al., 2020). Sequence redundancy of the rhodopsin-like family was reduced using CD-hit (Fu et al., 2012) with 30% sequence similarity cutoff.

A.2 Code

The code to reproduce the computational experiments is available at `https://anonymous.4open.science/r/AFS_AAC_SVM-F3D9`. Protein sequences used in the computational experiments along with their UniProt IDs, are provided in the `datasets` folder as `.csv` files for each family.

B Sequence and function diversity of protein classes within a dataset

Paralogous proteins have a common ancestor but have diverged in functionality. Protein functions are an aggregate of descriptors describing protein's activity and influence at various levels. They can be at the molecular level, like binding with specific molecules and catalysing reactions, to the biological process level, like energy metabolism. In B.1, we discuss the diversity of the functions of the proteins considered in our datasets.

As paralogs have a common ancestor, high sequence similarity would suggest high evolutionary conservation in the proteins. In B.2, we discuss the extent of sequence diversity in protein classes considered in our datasets.

We see that the dataset of proteins considered in our computational experiments are diverse in their function and sequences.

B.1 Function diversity

We have considered paralogous proteins with varying functional differences. We find very subtle differences in the functions of trypsin and chymotrypsin. On the other hand, the function difference is drastic in the case of alpha-lactalbumin and lysozyme c.

Trypsin and chymotrypsin are a family of enzymes that break peptide bonds in proteins. The difference in the function of these proteins is fine-grained; trypsins cleave only the peptide bond following a basic amino acid (K and R), while chymotrypsins cleave the peptide bond following a hydrophobic amino acid (F, W, and Y) (Dodson & Wlodawer, 1998).

GPCRs constitute a large and diverse class of cell surface receptor proteins. They trigger intra-cellular pathways in response to external signals. These signals are in the form of small molecules, called ligands. Depending upon the nature of ligands and other 3D structural similarities, GPCRs are grouped into distinct classes. We consider three such classes viz., rhodopsin-like, secretin-like, and glutamate-like. Further, we consider pairwise three subfamilies of rhodopsin-like GPCRs viz., aminergic receptors, lipid receptors, and peptide receptors.

Lysozyme C and α-lactalbumin are sequence and structure homologs with mutually exclusive functions and high fold conservation. Based on phylogenetic analysis, they are considered to have diverged from a common ancestor millions of years ago (Qasba et al., 1997).

Globins are a superfamily of functionally divergent homologous protein families with a high level of fold conservation. We consider three well-known globin families viz., myoglobin, hemoglobin-α and hemoglobin-β. Myoglobin is a monomer that binds and releases oxygen as per physiological requirements. On the other hand, α and β chains together constitute hemoglobin, a tetramer of composition $\alpha_2\beta_2$ (Dill et al., 2017), that transports oxygen in red blood cells.

Tubulin-α and tubulin-β are similar to the hemoglobin-α and hemoglobin-β pair in that they both share sequence and 3D structural similarities but have subtle functional differences. One copy each of tubulin-α and tubulin-β form a functional dimer. Notably, neither two copies of tubulin-α nor two copies of tubulin-β can form a functional dimer. Tubulin-β has a catalytic activity (GTP hydrolysis) that is absent in tubulin-α. This is one of the several subtle functional differences between tubulin-α and tubulin-β.

Interleukin-1 alpha and interleukin-1 beta are both proteins involved in the immune system. They differ from each other in their occurrence within the body (on cell surface or in blood circulation), activation mechanisms, and associated signalling pathways (Galozzi et al., 2021).

Cytochrome P450 (abbreviated as CYP) is a family of proteins whose function is clearance of 'foreign' molecules (drugs; also called as xenobiotics) as well as in certain biosynthesis pathways e.g., of steroid hormones. CYP3 and CYP51 are two of the several classes of CYPs; CYP3 metabolizes lipophilic molecules (McArthur et al., 2003) whereas CYP51 is involved in steroid biosynthesis (Hargrove et al., 2012).

Hemoglobin-α/hemoglobin-β, histone H2A / histone H2B and tubulin-α/tubulin-β are paralog pairs that together function as heteromers (protein complexes made up of different protein subunits).

B.2 Sequence Diversity

The dataset of the 15 paralog pairs in our experiments comprises 21 protein families (Table 2). For these families, we compute the within-class sequence similarities (for sequences within a protein family). We also compute the inter-class sequence similarities (between sequences from two different protein families) for each paralog pair. These are shown in Appendix Figure 5. We use a longest subsequence based similarity score, *lcss*, that is defined in B.2.1. In B.2.2, we see that *lcss* significantly varies across the 21 protein families we are considering as compared to its variation between the two protein sequences of any paralog pair.

B.2.1 Longest common subsequence based similarity score (*lcss*)

We compute the longest common subsequence (*lcs*) based similarity score (*lcss*) between a pair of protein sequences. We define *lcss* between two sequences as the length of their longest common subsequence, *lcs*, divided by the length of the longest sequence from the two. For a pair of protein sequences, $\mathbf{p}^{(i)} = (p_1^{(i)}, p_2^{(i)}, \ldots, p_{L_1}^{(i)})$ of length L_1 and $\mathbf{p}^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_{L_2}^{(j)})$ of length L_2, their *lcss* is,

$$
\begin{aligned}
& lcs(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) = \max_{\mathbf{q}} k \\
\text{s.t. } & \mathbf{q} = (q_1, q_2, \ldots, q_k) \\
& (q_1 = p_{x_1}^{(i)} = p_{y_1}^{(j)}, q_2 = p_{x_2}^{(i)} = p_{y_2}^{(j)}, \ldots, q_k = p_{x_k}^{(i)} = p_{y_k}^{(j)}) \\
& x_1 < x_2 < \ldots < x_k \\
& y_1 < y_2 < \ldots < y_k
\end{aligned}
$$

lcs based similarity score, *lcss*, is defined as,

$$
lcss(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) = \frac{lcs(\mathbf{p}^{(i)}, \mathbf{p}^{(j)})}{\max(L_1, L_2)} \in [0, 1]
$$

$lcss(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) = 1$ if and only if $\mathbf{p}^{(i)} = \mathbf{p}^{(j)}$, i.e., sequences are identical. Whereas $lcss(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) = 0$ if and only if $p_x^{(i)} \neq p_y^{(j)}, \forall x, y$, i.e., there are no amino acids common to both the sequences.

B.2.2 Within-class and inter-class *lcss* for the 15 paralog pairs

Within-class *lcss*: $lcss(\mathbf{p}^{(i)}, \mathbf{p}^{(j)})$ are computed with $\mathbf{p}^{(i)}, \mathbf{p}^{(j)}$ from the same protein family. These are shown in *blue* and *magenta* in Figure 5 (with box-plots) for each of 21 protein families in the 15 paralog pairs.

- 12 of 21 protein families have median within-class *lcss* greater than 0.5. This implies less sequence diversity in this set of families from the remaining families. These are,
  - Median *lcss* $\geqslant$ 0.6 for 11 of these 12 families and $\geqslant$ 0.8 for 3 families (high level of sequence conservation).
  - For 7 out of the 15 paralog pairs, the median within-class *lcss* $>$ 0.5 for both families of a paralogous pair.
- For the remaining 9 protein families, the median within-class *lcss* is less than 0.5. This implies high sequence diversity in this set of families from the remaining families. These are,

Table 3: The median within-class *lcss* between sequences from the respective families. See boxplot in Figure 5.

Family	α-lactalbumin	lysozyme C	myoglobin	hemoglobin-α	hemoglobin-β
Median *lcss*	0.6	0.59	0.81	0.63	0.67

Family	tubulin-α	tubulin-β	cytochrome P450 CYP3
Median *lcss*	0.83	0.82	0.7

Family	interleukin-1 α	interleukin-1 β	histone H2A	histone H2B
Median *lcss*	0.72	0.66	0.65	0.68

Table 4: The median within-class *lcss* between sequences from the respective families. See boxplot in Figure 5.

Family	cytochrome P450 CYP51	rhodopsin-like receptor	aminergic receptor	lipid receptor
Median *lcss*	0.47	0.34	0.39	0.37

Family	peptide receptor	glutamate-like receptor	secretin-like receptor	trypsin	chymotrypsin
Median *lcss*	0.37	0.35	0.36	0.47	0.45

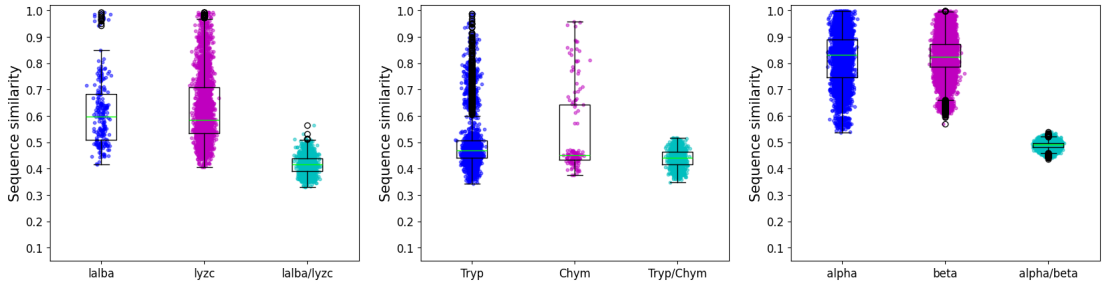
  - For 7 out of the 15 paralog pairs, the median within-class *lcss* < 0.5 for both families of a paralogous pair.
- For the paralog pair Cytochrome P450 CYP3 vs CYP51, the median sequence similarity for CYP3 is greater than 0.5, while for CYP51, it is less than 0.5.

Inter-class *lcss*: $lcss(\mathbf{p}^{(i)}, \mathbf{p}^{(j)})$ are computed with $\mathbf{p}^{(i)}, \mathbf{p}^{(j)}$ respectively from two protein families that are paralog pairs. These are shown in *cyan* in Figure 5 (with box-plots) for each of the 15 paralog pairs.

- The median inter-class *lcss* is less than 0.5 for all paralog pairs. This implies sequences of the proteins across the classes are not very similar.

Distinguishing paralog pairs based on within-class and inter-class *lcss*: If we analyse the box plots in Figure 5 - two paralog pair proteins can be considered to be distinguishable based on sequence similarity if the upper-whisker of inter-class *lcss* is lower than the lower-whiskers of the respective within-class *lcss* scores.

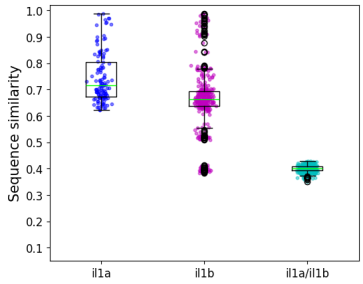
- Apart from paralog pairs, tubulin-α vs tubulin-β (Figure 5c) and interleukin-1 α vs interleukin-1 β (Figure 5d), no other paralog pair is distinguishable based on sequence similarity.
- For Trypsin vs Chymotrypsin and the 6 GPCR pairs (Figures 5b and 5j to 5o), the median inter-class *lcss* scores are close to the within-class *lcss* scores making them indistinguishable based on sequence similarity.



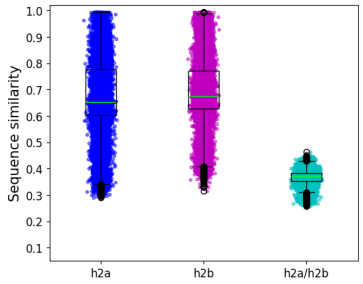
(a) Lysozyme C vs α-Lactalbumin

(b) Trypsin vs Chymotrypsin

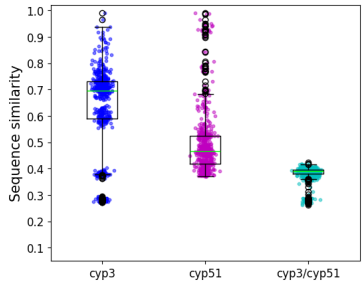
(c) Tubulin-α vs Tubulin-β



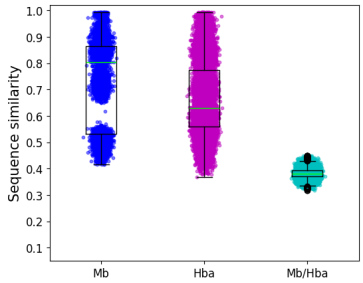
(d) Interleukin-1 α vs Interleukin-1 β



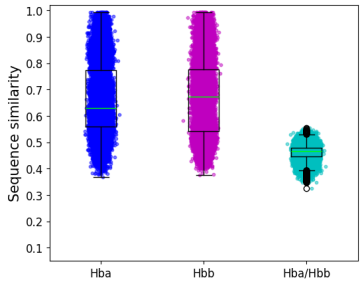
(e) Histone H2A vs Histone H2B



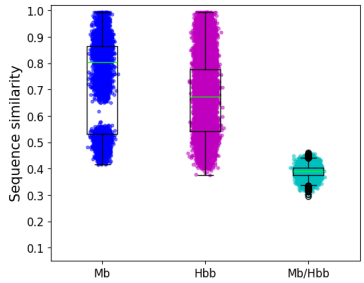
(f) Cytochrome P450 CYP3 vs CYP51



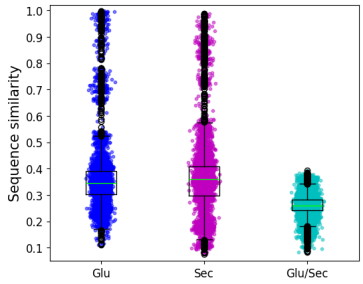
(g) Myoglobin vs Hemoglobin-α



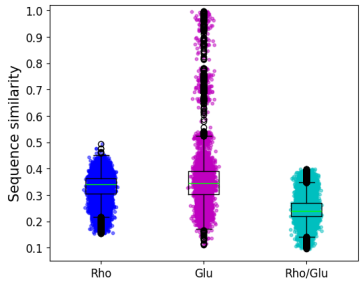
(h) Hemoglobin-α vs Hemoglobin-β



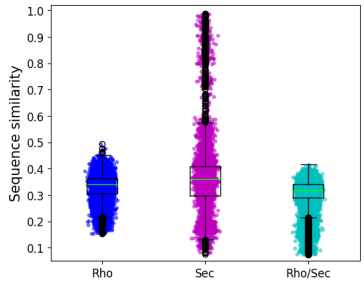
(i) Myoglobin vs Hemoglobin-β



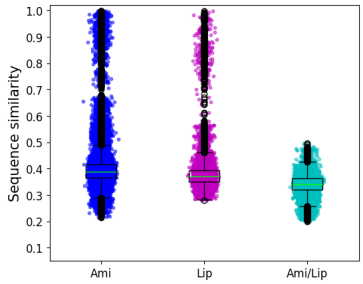
(j) Secretin-like vs Glutamate-like



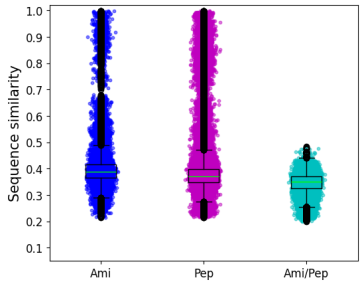
(k) Rhodopsin-like vs Glutamate-like



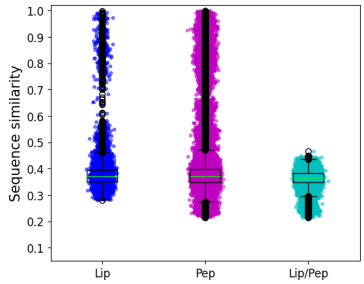
(l) Rhodopsin-like vs Secretin-like



(m) Aminergic vs Lipid receptors



(n) Aminergic vs Peptide receptors



(o) Lipid vs Peptide receptors

Figure 5: *lcss* sequence similarity scores for the 15 paralog pair datasets. In the boxplots, the lower and upper whiskers are at 1.5 IQR (inter-quantile range) values away from the first and third quartiles respectively.

C The SVEA algorithm for AFS

Algorithm 1 ϕ_i Monte-carlo approximation algorithm as suggested in (Tripathi et al., 2020; 2021)

Input: Feature set $N = \{1, 2, \ldots, 20\}$, Number of sample permutations $samPerm$, Datasets (D_P, D_Q), Set of coalitions $Sam\_co\_set = [()]$
Initialise: $v(()) = 0$, $\hat{\phi}_i := 0 \forall i \in N$
Append N to Sam_co_set.
for $s = 1, 2, \ldots, samPerm$ **do**
  Take $\pi \in PermSet(N)$ with probability $\frac{1}{20!}$.
  for $i = 1, 2, \ldots, 20$ **do**
    Compute $Pred^i(\pi) = \{\pi(1), \pi(2), \ldots \pi(k-1) | i = \pi(k)\}$
    if $Pred^i(\pi)$ not in Sam_co_set **then**
      Compute $v(Pred^i(\pi)) = 1 - tr\_er(Pred^i(\pi))$.
      Append $Pred^i(\pi)$ to Sam_co_set.
    end if
    if $Pred^i(\pi) \cup i$ not in Sam_co_set **then**
      Compute $v(Pred^i(\pi) \cup \{i\}) = 1 - tr\_er(Pred^i(\pi) \cup \{i\})$.
      Append $Pred^i(\pi) \cup \{i\}$ to Sam_co_set.
    end if
    $\hat{\phi}_i = \hat{\phi}_i + v(Pred^i(\pi) \cup \{i\}) - v(Pred^i(\pi)))$
  end for
end for
$\hat{\phi}_i = \frac{\hat{\phi}_i}{samPerm}, \forall i \in N$

D SVM training for AFS partition

We provide details for the linear SVM classifier discussed in Section 2.3. We use 5-fold cross-validation to tune the SVM regularisation hyperparameter C from $\{0.1, 1, 10, 100, 1000\}$ that gives the best average classification score for the 5 folds. C is inversely proportional to the strength of regularisation. In general, we find that there is an imbalance in the number of sequences that we find for the two paralogous proteins, i.e. say $n_P >> n_Q$. It is known that accuracy is not a well-suited performance measure of the classifier in class imbalance settings. Therefore, we use the arithmetic mean of sensitivity and specificity (AM) to measure the performance of the classifier (Brodersen et al., 2010). Further, we use a class-balanced version of hinge loss for training the SVM as suggested in (Menon et al., 2013) for statistical consistency with the AM score. Appendix Table 5 reports the train and test scores of the trained linear SVM with AAC and AFS features, respectively, on the protein family datasets (See Appendix Table 2) considered in our computational experiments.

E More details for computational experiments

We present here details on some of the computational experiments that have been referenced in the main paper.

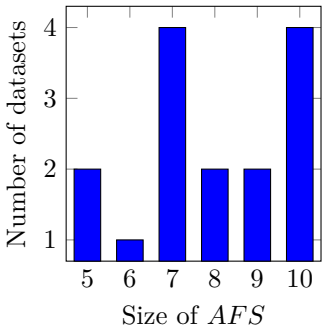
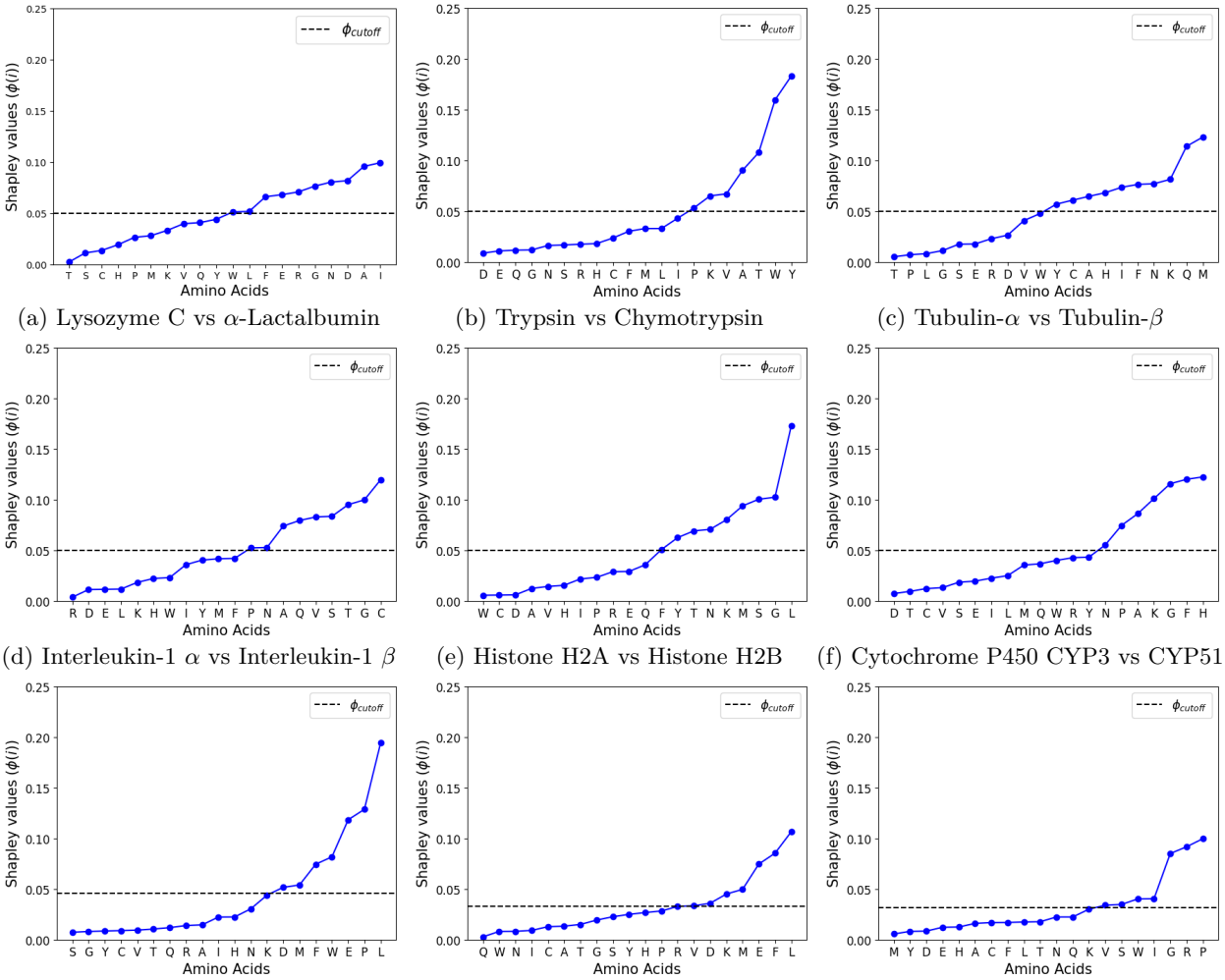


Figure 6: The sizes of the AFS for the 15 datasets.



(a) Lysozyme C vs α-Lactalbumin (b) Trypsin vs Chymotrypsin (c) Tubulin-α vs Tubulin-β

(d) Interleukin-1 α vs Interleukin-1 β (e) Histone H2A vs Histone H2B (f) Cytochrome P450 CYP3 vs CYP51

(g) Aminergic vs Lipid receptors (h) Aminergic vs Peptide receptors (i) Lipid vs Peptide receptors

Figure 7: (Continued from Figure 3) Shapley value ($\phi(i)$) for AAC features computed using SVEA.

Table 5: Classification scores for different pairs of paralogous proteins using the train/test datasets described in Table 2, using AAC and *AFS* features. The *AFS* amino acids computed for each pair are given in Table 1. The train score is the mean (±1 standard deviation) 5-fold cross-validation score. **AM** is the arithmetic mean of specificity and sensitivity. **Acc** is the accuracy.

(a) Lysozyme C vs α-Lactalbumin

	AAC	**AFS**
Train AM	1.0	0.993 (±0.013)
Test AM	0.896	0.898
Train Acc	1.0	0.99 (±0.02)
Test Acc	0.836	0.881

(b) Trypsin vs chymotrypsin

	AAC	**AFS**
Train AM	0.992 (±0.015)	0.977 (±0.031)
Test AM	0.873	0.835
Train Acc	0.988 (±0.024)	0.965 (±0.047)
Test Acc	0.844	0.756

(c) Tubulin-α vs Tubulin-β

	AAC	**AFS**
Train AM	0.996 (±0.009)	0.997 (±0.006)
Test AM	0.992	0.992
Train Acc	0.997 (±0.006)	0.994 (±0.008)
Test Acc	0.991	0.994

(d) Histone H2A vs Histone H2B

	AAC	**AFS**
Train AM	0.983 (±0.016)	0.983 (±0.01)
Test AM	0.91	0.934
Train Acc	0.983 (±0.016)	0.983 (±0.01)
Test Acc	0.889	0.922

(e) Globins

Dataset		**AAC**	**AFS**
Myoglobin vs Hemoglobin-α	**Train AM**	0.998 (±0.003)	0.994 (±0.009)
	Test AM	0.968	0.97
	Train Acc	0.998 (±0.005)	0.995 (±0.006)
	Test Acc	0.969	0.971
Myoglobin vs Hemoglobin-β	**Train AM**	1.0 (±0.0)	1.0 (±0.0)
	Test AM	0.957	0.936
	Train Acc	1.0 (±0.0)	1.0 (±0.0)
	Test Acc	0.949	0.919
Hemoglobin-α vs Hemoglobin-β	**Train AM**	0.983 (±0.008)	0.976 (±0.007)
	Test AM	0.961	0.935
	Train Acc	0.983 (±0.008)	0.976 (±0.006)
	Test Acc	0.966	0.947

(f) GPCRs

Dataset		**AAC**	**AFS**
Secretin-like vs Glutamate-like	**Train AM**	0.933 (±0.042)	0.95 (±0.032)
	Test AM	0.888	0.845
	Train Acc	0.933 (±0.042)	0.95 (±0.032)
	Test Acc	0.889	0.844
Rhodopsin-like vs Glutamate-like	**Train AM**	0.884 (±0.042)	0.85 (±0.045)
	Test AM	0.967	0.934
	Train Acc	0.867 (±0.038)	0.837 (±0.032)
	Test Acc	0.956	0.926
Rhodopsin-like vs Secretin-like	**Train AM**	0.917 (±0.051)	0.878 (±0.065)
	Test AM	0.934	0.846
	Train Acc	0.908 (±0.06)	0.863 (±0.073)
	Test Acc	0.941	0.853
Aminergic vs Lipid receptors	**Train AM**	0.949 (±0.014)	0.943 (±0.005)
	Test AM	0.922	0.843
	Train Acc	0.943 (±0.017)	0.94 (±0.008)
	Test Acc	0.92	0.84
Aminergic vs Peptide receptors	**Train AM**	0.835 (±0.06)	0.818 (±0.053)
	Test AM	0.844	0.79
	Train Acc	0.83 (±0.06)	0.819 (±0.051)
	Test Acc	0.827	0.784
Lipid vs Peptide receptors	**Train AM**	0.829 (±0.022)	0.76 (±0.035)
	Test AM	0.845	0.709
	Train Acc	0.838 (±0.018)	0.75 (±0.032)
	Test Acc	0.858	0.725

(g) Interleukin-1 α vs Interleukin-1 β

	AAC	**AFS**
Train AM	0.98 (±0.04)	0.98 (±0.04)
Test AM	0.979	0.985
Train Acc	0.975 (±0.05)	0.975 (±0.05)
Test Acc	0.961	0.971

(h) Cytochrome P450 CYP3 vs Cytochrome P450 CYP51

	AAC	**AFS**
Train AM	0.967 (±0.041)	0.933 (±0.062)
Test AM	0.902	0.92
Train Acc	0.969 (±0.038)	0.936 (±0.062)
Test Acc	0.894	0.908

E.1 Tubulin

The inter-chain contact residues from the tubulin-α/β heterodimer were identified using ChimeraX 1.4 (Pettersen et al., 2021). The *Contacts* tool available in *Tools* → *Structure Analysis* was used with settings as shown in Figure 8. For PDB ID:3JAR we count the residues of chain-A (tubulin-α) and chain-B (tubulin-β) which are in contact with the residues of other tubulin chains. Similarly, for PDB ID:5N5N we count the residues of chain-G (tubulin-α) and chain-B (tubulin-β) which are in contact with the residues of other tubulin chains. The code for counting the AFS residues at the identified contact points of the respective chains is available at `https://anonymous.4open.science/r/AFS_AAC_SVM-F3D9`.

E.2 Histone

The inter-chain contact residues of histone H2A and H2B were identified from its heterooctameric structure comprising of two H2A/H2B dimers and one H3/H4 tetramer, using ChimeraX 1.4. The *Contacts* tool available in *Tools* → *Structure Analysis* was used with settings as shown in Figure 8. For PDB ID: 1AOI and 3KWQ, we count the residues of an H2A and an H2B chain, which are in contact with other histone chains in the heterooctameric structure. The code for counting the AFS residues at the identified contact points of the respective chains is available at `https://anonymous.4open.science/r/AFS_AAC_SVM-F3D9`.

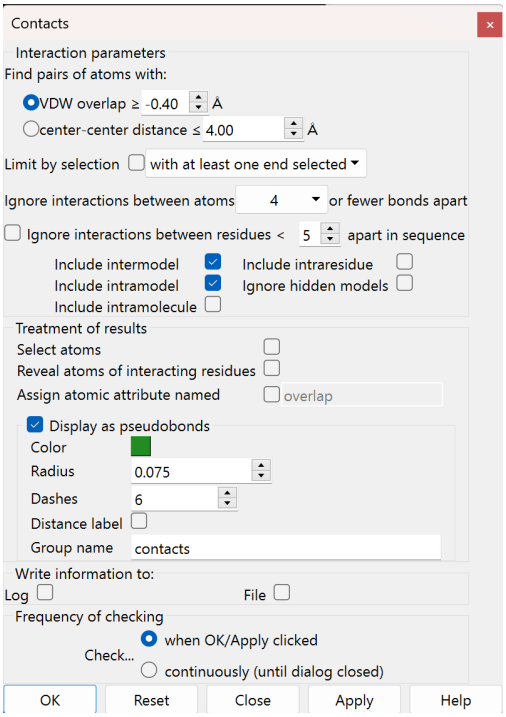


Figure 8: ChimeraX 1.4 settings for identifying inter-chain contact points from the tubulin-α/β heterodimer and from the histone heterooctamer

Figure 9: Multiple sequence alignment of hemoglobin-β and myoglobin sequences. 15 sequences on the left are from hemoglobin-β and on the right are from myoglobin. The sequences are randomly selected from the train set of the protein families. AFS(Myoglobin) amino acids are in green and AFS(Hemoglobin-β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 3c)

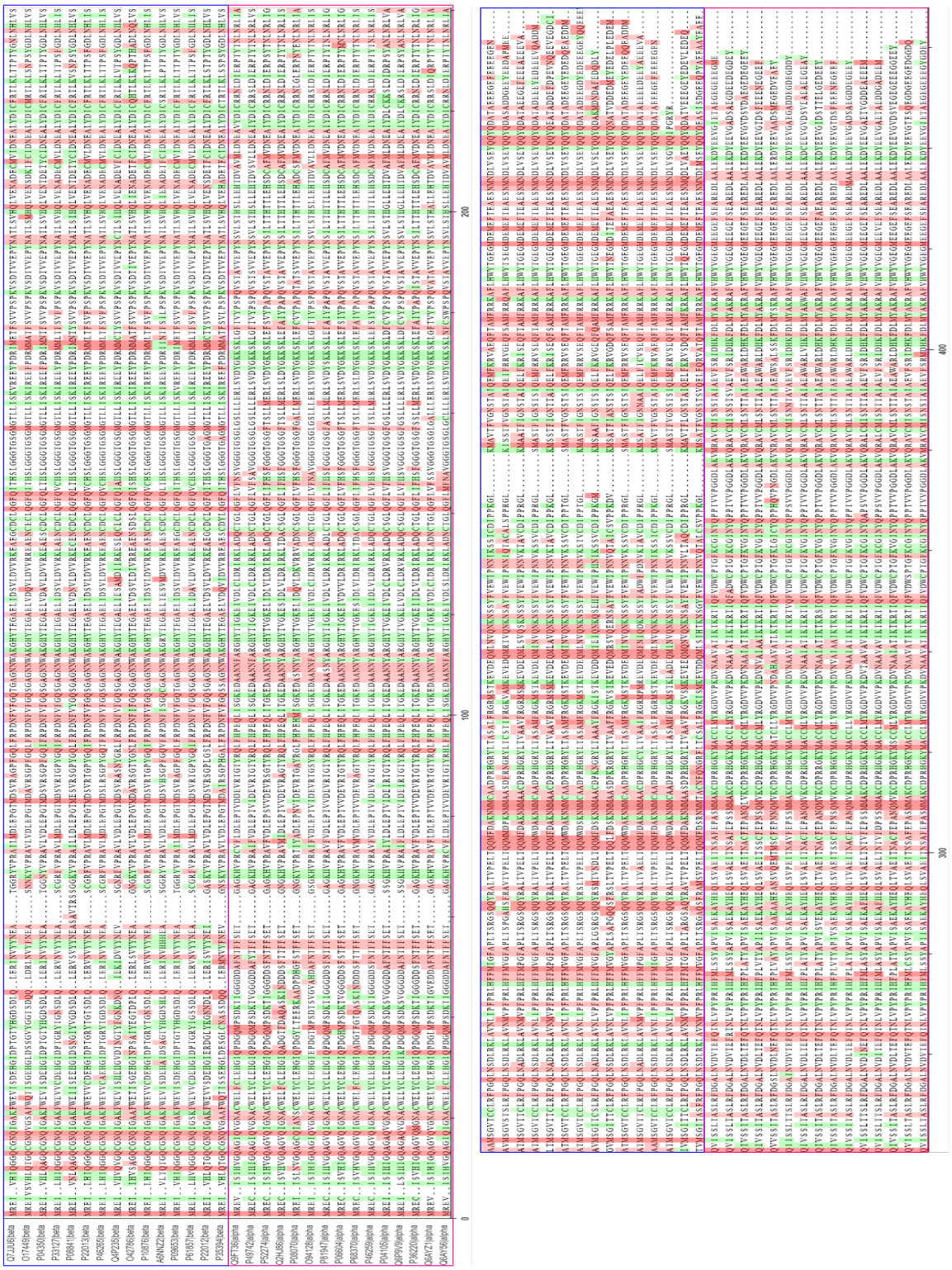


Figure 10: Multiple sequence alignment of tubulin-α and tubulin-β sequences. 15 sequences on the left are from tubulin-β and on the right are from tubulin-α. The sequences are randomly selected from the train set of the protein families. AFS(Tubulin-α) amino acids are in green and AFS(Tubulin-β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 7c)

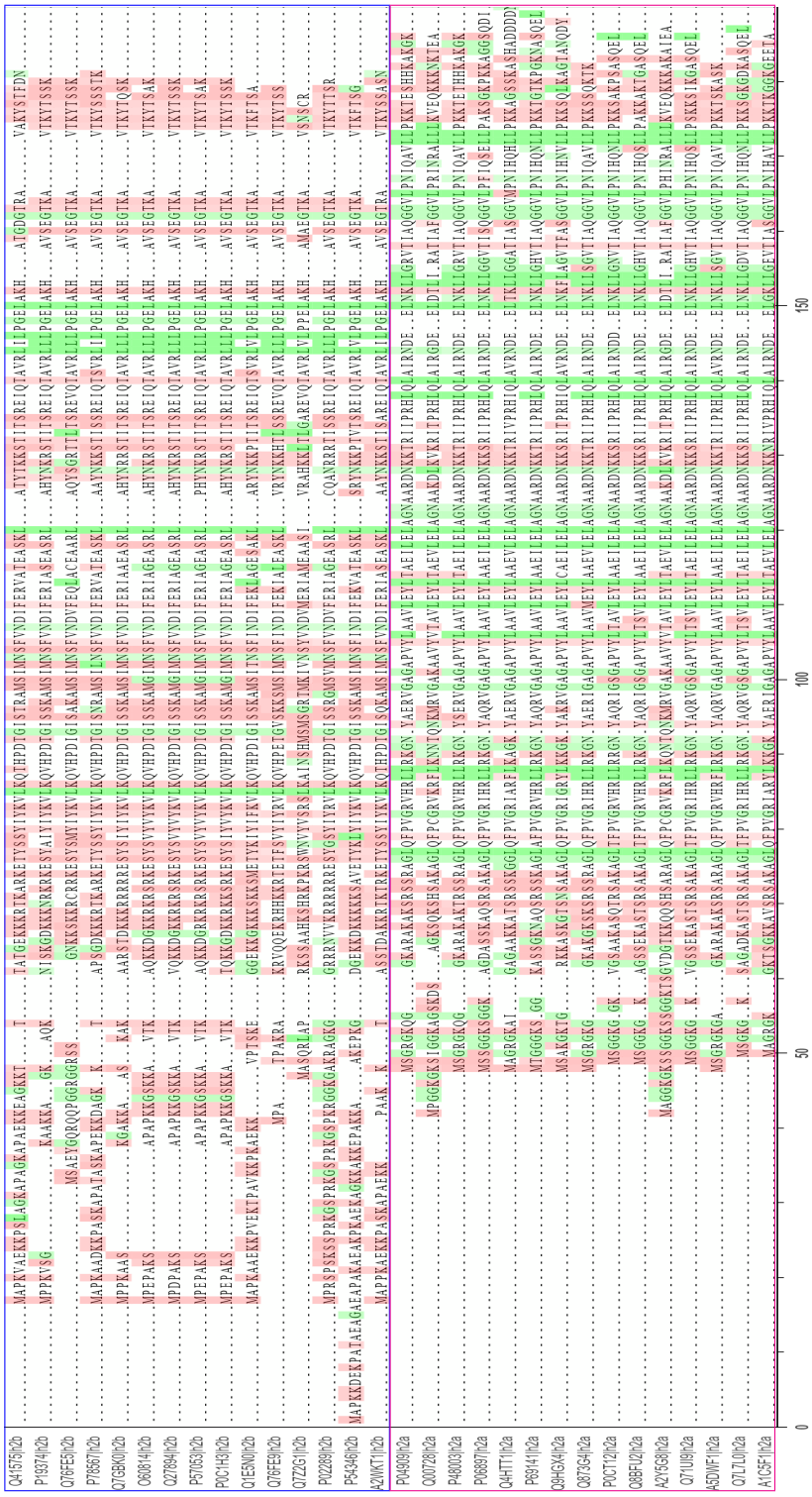


Figure 11: Multiple sequence alignment of histone H2A and histone H2B sequences. 15 sequences on the left are from histone H2B and on the right are from histone H2B. The sequences are randomly selected from the train set of the protein families. AFS(Histone H2A) amino acids are in green and AFS(Histone H2B) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 7e)

Figure 12: Multiple sequence alignment of interleukin-1 α and interleukin-1 β sequences. 15 sequences on the left are from interleukin-1 β and on the right are from interleukin-1 α. The sequences are randomly selected from the train set of the protein families. AFS(Interleukin-1 α) amino acids are in green and AFS(Interleukin-1 β) in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (See Figure 7d)

E.3 Marginal contribution feature importance (MCI) (Catav et al., 2021) for AFS

We validate the top-selected features, AFS, in our experiments using the MCI feature rankings (Catav et al., 2021). The MCI method, by design, was aimed at addressing the issue of correlated features. Catav et al. (2021), show that Shapley value-based feature rankings are altered when correlated features are introduced. Therefore, they propose the MCI method using a different set of axioms to circumvent the issue of correlated features. Thus, we validate the Shapley value-based feature rankings with those by MCI. We find strong agreement in the two rankings (see Table 6), suggesting a lack of correlated features for the datasets considered in our experiments. We use Shapley value instead of MCI in our main pipeline for computing AFS, as the efficiency axiom of the Shapley value allows us to draw a fair cutoff line for selecting the important features. Since MCI doesnt satisfy the efficiency axiom, the same cutoff line may not be applicable to it.

For a feature i, its MCI score is defined as,

$$MCI(i) = \max_{S \subseteq N \setminus \{i\}} v(S \cup \{i\}) - v(S),$$

Here, $v(\cdot)$ is the same as that defined in Section 2.2. We compare the amino acids with the top-d (d = size of AFS) MCI scores to the AFS in Table 6. MCI is computed using the same approximation scheme as in Appendix Section C Algorithm 1 with appropriate modifications.

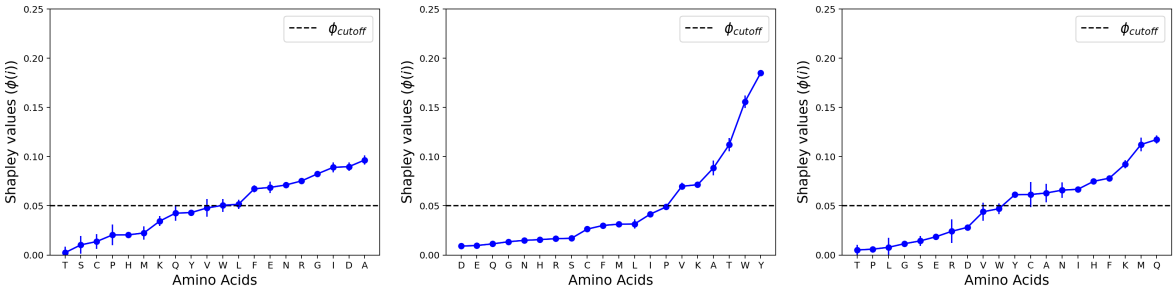
Table 6: AFS comparison with the amino acids having the top-d MCI (Catav et al., 2021) scores. Here, d is the size of AFS for the respective dataset. The amino acids that differ in the two sets are in **bold and underlined**, with their counts mentioned in the rightmost column. For 8 of 15 datasets, AFS and top-d MCI sets are the same, while only for two datasets do they differ in two amino acids. For all 15 datasets, at least the top-3 MCI amino acids are in AFS. For 11 of these datasets, at least the top-5 MCI amino acids are in AFS.

Paralog pair	top-d MCI amino acids (rank-1 → rank-d)	AFS	Difference count
Lysozyme C (74) and α-Lactalbumin (22)	$\{I, A, D, G, R, F, N, E, W, L\}$	$\{I, A, D, N, G, R, E, F, L, W\}$	0
Trypsin (66) and Chymotrypsin (17)	$\{Y, W, T, A, K, V, \underline{\boldsymbol{I}}\}$	$\{Y, W, T, A, V, K, \underline{\boldsymbol{P}}\}$	1
Tubulin-α (117) and Tubulin-β (191)	$\{Q, M, K, H, F, I, N, A, Y, C\}$	$\{M, Q, K, N, F, I, H, A, C, Y\}$	0
Histone H2A (180) and Histone H2B (177)	$\{L, G, K, S, M, T, N, F, Y\}$	$\{L, G, S, M, K, N, T, Y, F\}$	0
Interleukin-1 α (16) and Interleukin-1 β (25)	$\{G, C, T, V, Q, S, A, \underline{\boldsymbol{I}}, P\}$	$\{C, G, T, S, V, Q, A, \underline{\boldsymbol{N}}, P\}$	1
Cytochrome P450 CYP3 (32) and CYP51 (32)	$\{H, F, G, K, A, P, N\}$	$\{H, F, G, K, A, P, N\}$	0
Globins			
Myoglobin (107) and Hemoglobin-α (303)	$\{V, Y, E, K, S, G, W, I, C, P\}$	$\{E, S, Y, V, K, P, I, G, C, W\}$	0
Myoglobin (107) and Hemoglobin-β (285)	$\{V, K, E, C, W, N, F, Y, M, I\}$	$\{K, V, C, E, W, N, F, M, Y, I\}$	0
Hemoglobin-α (303) and Hemoglobin-β (285)	$\{W, S, N, P, \underline{\boldsymbol{V}}\}$	$\{W, P, N, S, \underline{\boldsymbol{G}}\}$	1
GPCRs			
Rhodopsin-like (181) and Glutamate-like (89)	$\{D, E, Q, G, L, \underline{\boldsymbol{I}}\}$	$\{D, Q, E, G, \underline{\boldsymbol{M}}, L\}$	1
Secretin-like (90) and Glutamate-like (89)	$\{W, H, Y, V, D\}$	$\{W, H, Y, V, D\}$	0
Rhodopsin-like (181) and Secretin-like (90)	$\{W, E, H, Q, S, M, V, A\}$	$\{W, E, M, S, V, H, Q, A\}$	0
Rhodopsin-like GPCRs			
Aminergic receptors (186) and Lipid receptors (113)	$\{L, E, P, \underline{\boldsymbol{K}}, F, D, \underline{\boldsymbol{I}}\}$	$\{L, P, E, \underline{\boldsymbol{W}}, F, \underline{\boldsymbol{M}}, D\}$	2
Aminergic receptors (186) and Peptide receptors (367)	$\{L, E, K, F, M, \underline{\boldsymbol{H}}, R, D\}$	$\{L, F, E, M, K, D, \underline{\boldsymbol{V}}, R\}$	1
Lipid receptors (113) and Peptide receptors (367)	$\{R, G, P, \underline{\boldsymbol{K}}, I, V, \underline{\boldsymbol{T}}\}$	$\{P, R, G, I, \underline{\boldsymbol{W}}, \underline{\boldsymbol{S}}, V\}$	2

E.4 Variance in Shapley value estimates and AFS

We computed the Shapley value estimates across 10 Monte Carlo runs with different random seeds. We do not find significant variation in these estmiates across the different runs as seen in Figure 13.

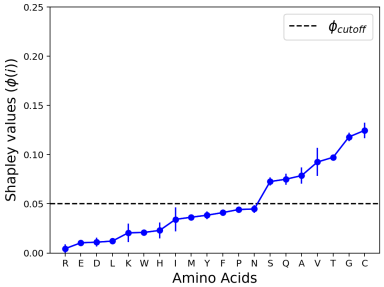
We also looked at the AFS identified using the Shapley value in each run. We do not find substantial differences in the AFS. As a stronger check of consistency, for each paralog pair, we computed the intersection of the AFSs across the 10 runs, see Table 7. We find that for each paralog pair, this intersection of AFSs recovers at least 66.6% (two-thirds) of the amino acids from the (single-run) AFS reported in the paper. For 14 pairs, greater than 75% is recovered. For 4 pairs, 100% is recovered. Furthermore, the amino acids with the top-3 Shapley values reported in the paper are present in the intersection of AFSs for all paralog pairs. For 13 pairs, the top-5 amino acids are in the intersection.



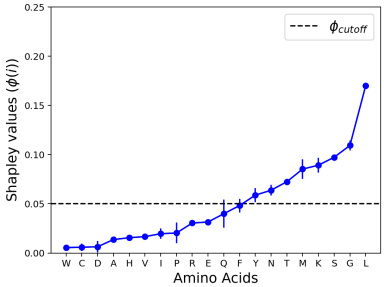
(a) Lysozyme C vs α-Lactalbumin

(b) Trypsin vs Chymotrypsin

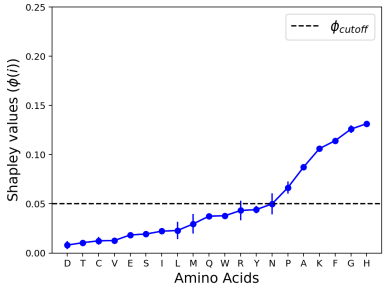
(c) Tubulin-α vs Tubulin-β



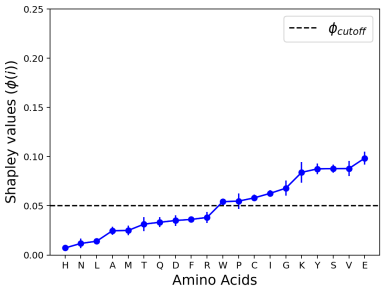
(d) Interleukin-1 α vs Interleukin-1 β



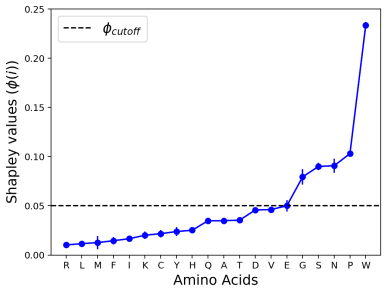
(e) Histone H2A vs Histone H2B



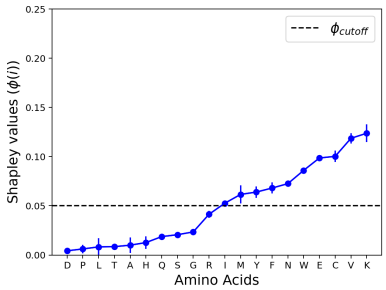
(f) Cytochrome P450 CYP3 vs CYP51



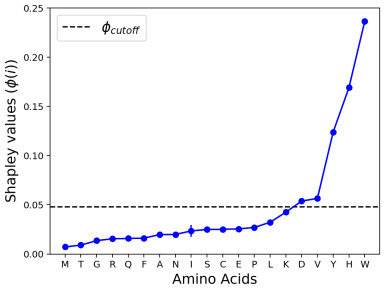
(g) Myoglobin vs Hemoglobin-α



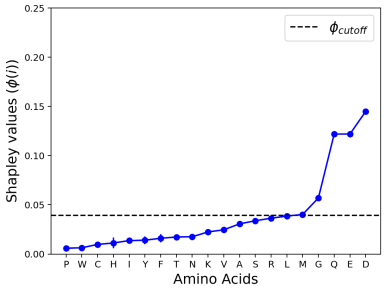
(h) Hemoglobin-α vs Hemoglobin-β



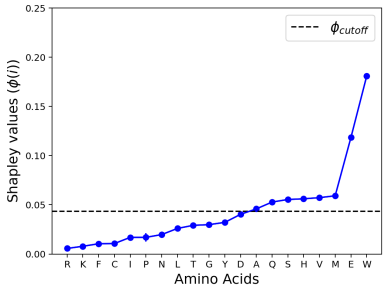
(i) Myoglobin vs Hemoglobin-β



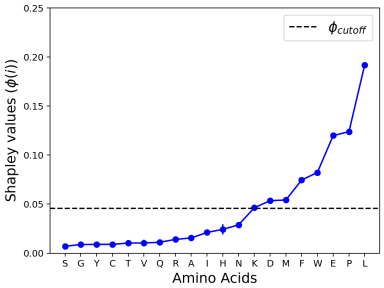
(j) Secretin-like vs Glutamate-like



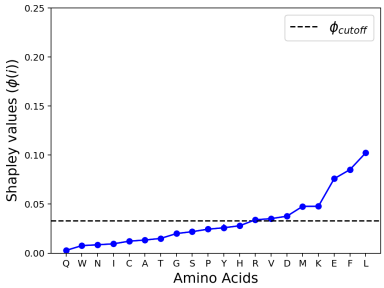
(k) Rhodopsin-like vs Glutamate-like



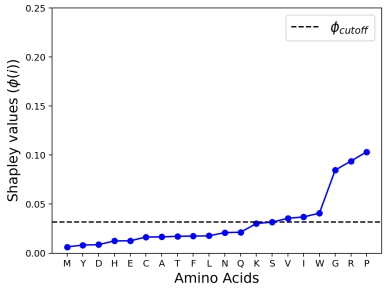
(l) Rhodopsin-like vs Secretin-like



(m) Aminergic vs Lipid receptors



(n) Aminergic vs Peptide receptors



(o) Lipid vs Peptide receptors

Figure 13: Mean ($\pm$ standard deviation) of Shapley value estimates across 10 Monte Carlo runs with different random seeds. The cut-off ϕ_{cutoff} here is computed based on the mean Shapley values.

Table 7: Comparison of AFS computed in 1 run with the intersection of AFS computed over 10 runs. The amino acids that differ in the two sets are in **bold and underlined**. The percentage of overlap in the two sets is given in the rightmost column.

Paralog pair	AFS in 1 run	Intersection of AFS over 10 runs	overlap ratio (%)
Lysozyme C and α-Lactalbumin	$\{I, A, D, N, G, R, E, F, \underline{\boldsymbol{L}}, \underline{\boldsymbol{W}}\}$	$\{I, A, D, N, G, R, E, F\}$	80
Trypsin and Chymotrypsin	$\{Y, W, T, A, V, K, \underline{\boldsymbol{P}}\}$	$\{Y, W, T, A, V, K\}$	86
Tubulin-α and Tubulin-β	$\{M, Q, K, \underline{\boldsymbol{N}}, F, I, H, A, C, \underline{\boldsymbol{Y}}\}$	$\{M, Q, K, F, I, H, A, C\}$	80
Histone H2A and Histone H2B	$\{L, G, S, M, K, N, T, \underline{\boldsymbol{Y}}, \underline{\boldsymbol{F}}\}$	$\{L, G, S, M, K, N, T\}$	78
Interleukin-1 α and Interleukin-1 β	$\{C, G, T, S, V, Q, A, \underline{\boldsymbol{N}}, \underline{\boldsymbol{P}}\}$	$\{C, G, T, S, V, Q, A\}$	78
Cytochrome P450 CYP3 and CYP51	$\{H, F, G, K, A, P, \underline{\boldsymbol{N}}\}$	$\{H, F, G, K, A, P\}$	86
Globins			
Myoglobin and Hemoglobin-α	$\{E, S, Y, V, K, \underline{\boldsymbol{P}}, I, G, C, \underline{\boldsymbol{W}}\}$	$\{E, S, Y, V, K, I, G, C\}$	80
Myoglobin and Hemoglobin-β	$\{K, V, C, E, W, N, F, M, Y, \underline{\boldsymbol{I}}\}$	$\{K, V, C, E, W, N, F, M, Y\}$	90
Hemoglobin-α and Hemoglobin-β	$\{W, P, N, S, G\}$	$\{W, P, N, S, G\}$	100
GPCRs			
Rhodopsin-like and Glutamate-like	$\{D, Q, E, G, \underline{\boldsymbol{M}}, \underline{\boldsymbol{L}}\}$	$\{D, Q, E, G\}$	67
Secretin-like and Glutamate-like	$\{W, H, Y, V, D\}$	$\{W, H, Y, V, D\}$	100
Rhodopsin-like and Secretin-like	$\{W, E, M, S, V, H, Q, A\}$	$\{W, E, M, S, V, H, Q, A\}$	100
Rhodopsin-like GPCRs			
Aminergic receptors and Lipid receptors	$\{L, P, E, W, F, M, D\}$	$\{L, P, E, W, F, M, D\}$	100
Aminergic receptors and Peptide receptors	$\{L, F, E, M, K, \underline{\boldsymbol{D}}, V, \underline{\boldsymbol{R}}\}$	$\{L, F, E, M, K, V\}$	75
Lipid receptors and Peptide receptors	$\{P, R, G, I, W, \underline{\boldsymbol{S}}, V\}$	$\{P, R, G, I, W, V\}$	86

E.5 AFS comparison with random feature subsets

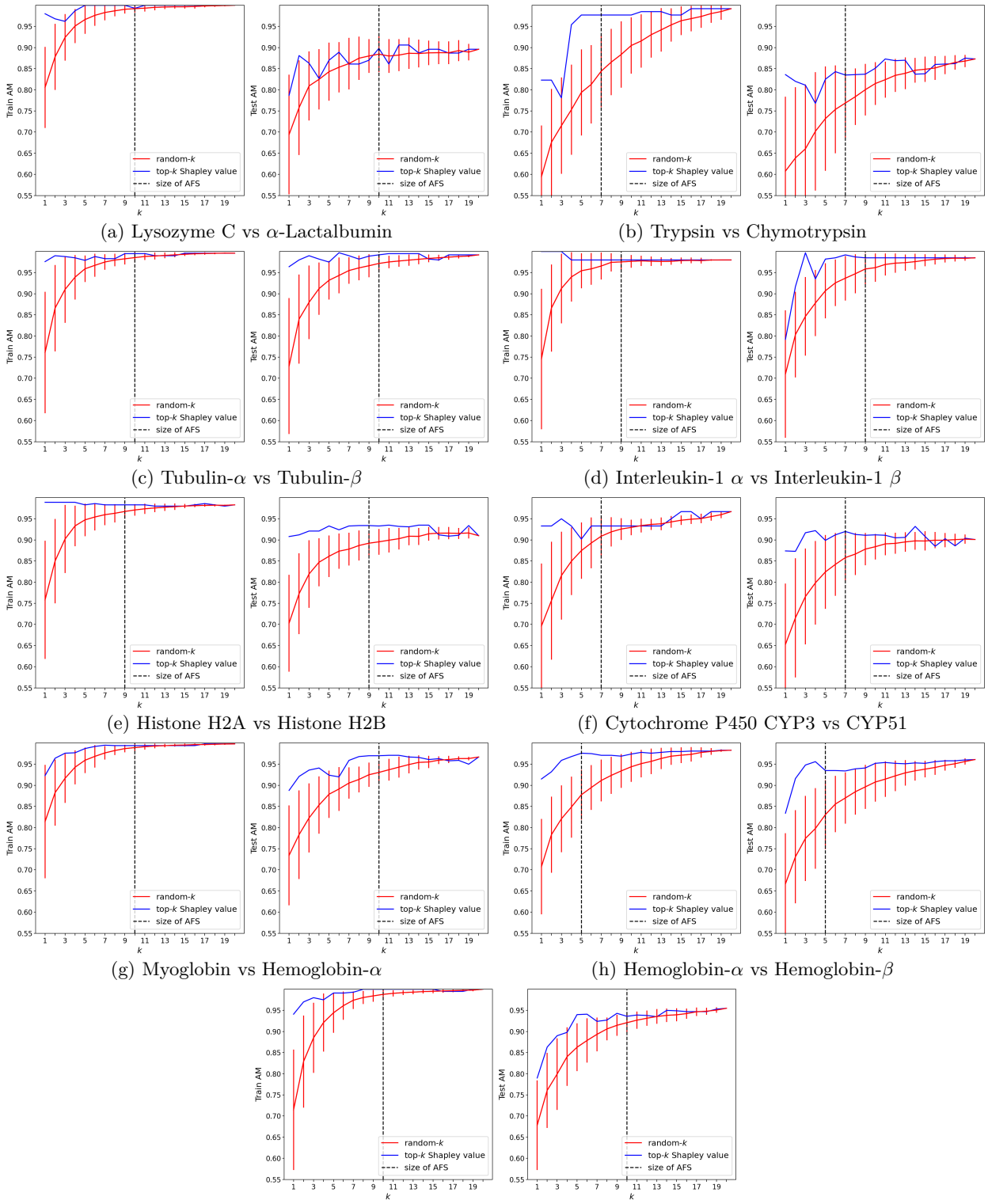
We sample 100 random feature subsets of the same size as the AFS and train classifiers for each paralog pair. We compute the mean test classification scores across these subsets, see Table 8. The mean test AM score drops for these randomly selected amino acid subsets for all 15 paralog pairs. The mean test AM further drops if randomly selected amino acids contain only non-AFS amino acids for 14 pairs.

Reason for increase in score for α-lactalbumin/lysozyme-C pair using non-AFS subset: There is a marginal increase in test score (0.8%) with non-AFS random subset for α-lactalbumin/lysozyme-C. We suspect this to be due to the noisy nature of the dataset as the test set consists of unreviewed TrEMBL UniProt entries. For example: we find that one of the sequences in the test set (UniProt ID: T1PBP9) is identified as *C-type lysozyme/alpha-lactalbumin family* in UniProt. Another sequence (UniProt ID: Q5M8G0) is identified in UniProt as *Novel protein similar to lysozyme C*. Furthermore, the class imbalance in the test data is the reverse of that in the train data.

Table 8: The classification test AM scores using AFS and random amino acid subset features for each paralog pair. The size of the random subsets here is the same size as the respective AFS. The '**Random Subset**' column shows the mean (± standard deviation) of the test scores for 100 randomly sampled feature subsets. The '**Non-AFS random subset**' is similar, but the random features are sampled from the respective non-AFS amino acids.

Paralog Pairs	**AFS**	**Random subset**	**Non-AFS random subset**
α-Lactalbumin / Lysozyme C	0.898	0.881 (±0.044)	0.906 (±0.0)
Myoglobin / Hemoglobin-α	0.970	0.927 (±0.032)	0.824 (±0.0)
Myoglobin / Hemoglobin-β	0.936	0.918 (±0.022)	0.823 (±0.0)
Hemoglobin-α / Hemoglobin-β	0.935	0.836 (±0.075)	0.754 (±0.058)
Trypsin / Chymotrypsin	0.835	0.773 (±0.075)	0.724 (±0.086)
Tubulin-α / Tubulin-β	0.992	0.973 (±0.017)	0.869 (±0.0)
Rhodopsin-like / Glutamate-like	0.934	0.834 (±0.077)	0.743 (±0.052)
Glutamate-like / Secretin-like	0.845	0.705 (±0.083)	0.670 (±0.065)
Rhodopsin-like / Secretin-like	0.846	0.797 (±0.072)	0.689 (±0.046)
Aminergic receptors / Lipid receptors	0.843	0.782 (±0.076)	0.637 (±0.038)
Aminergic receptors / Peptide receptors	0.790	0.764 (±0.045)	0.660 (±0.025)
Lipid receptors / Peptide receptors	0.709	0.653 (±0.073)	0.566 (±0.061)
Interleukin-1 α / Interleukin-1 β	0.985	0.954 (±0.038)	0.818 (±0.030)
Histone H2A / Histone H2B	0.934	0.899 (±0.032)	0.791 (±0.028)
Cytochrome P450 CYP3 / CYP51	0.920	0.865 (±0.043)	0.723 (±0.061)

Figure 14 shows how the train and test AM scores vary for different feature subset sizes for each paralog pair. We see that feature subsets based on the top-k Shapley values have in general higher train/test AM scores in comparison to random-k feature subsets.



(a) Lysozyme C vs α-Lactalbumin

(b) Trypsin vs Chymotrypsin

(c) Tubulin-α vs Tubulin-β

(d) Interleukin-1 α vs Interleukin-1 β

(e) Histone H2A vs Histone H2B

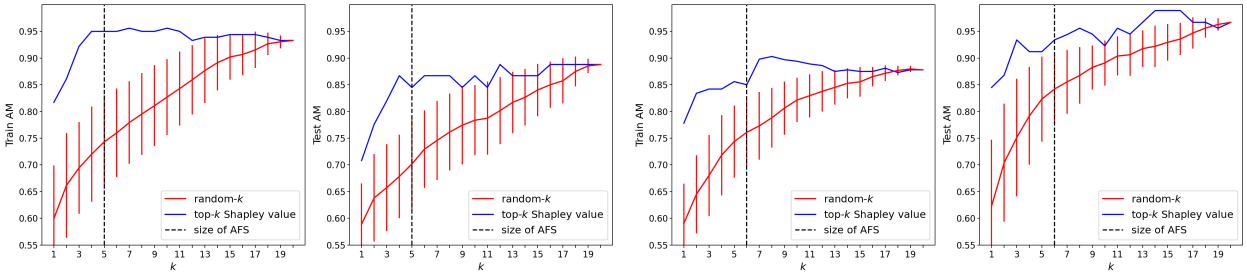
(f) Cytochrome P450 CYP3 vs CYP51

(g) Myoglobin vs Hemoglobin-α

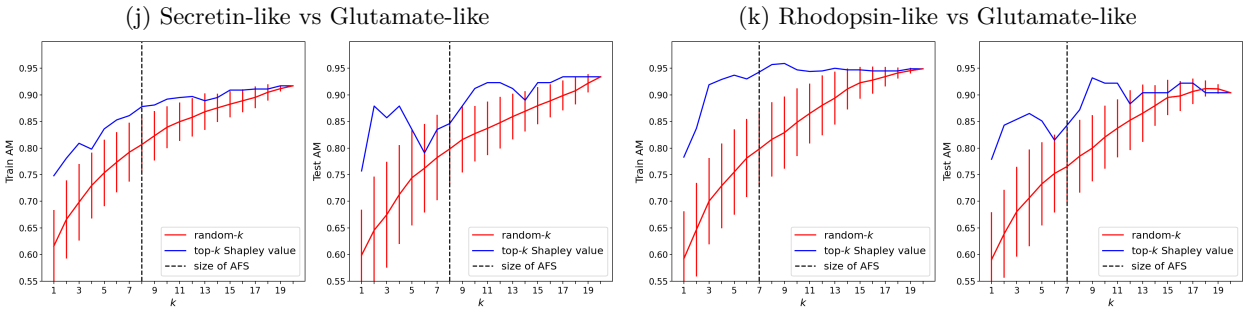
(h) Hemoglobin-α vs Hemoglobin-β

(i) Myoglobin vs Hemoglobin-β

Figure 14: Classification scores (train and test AM) for feature subsets of different sizes. The blue line shows the score for top-k features based on the Shapley value. The red line shows the mean ($\pm$ standard deviation) of the scores for 100 randomly selected subsets of size k. The train AM score here is the mean 5-fold cross-validation score on the training dataset. Figure is continued on page 37.



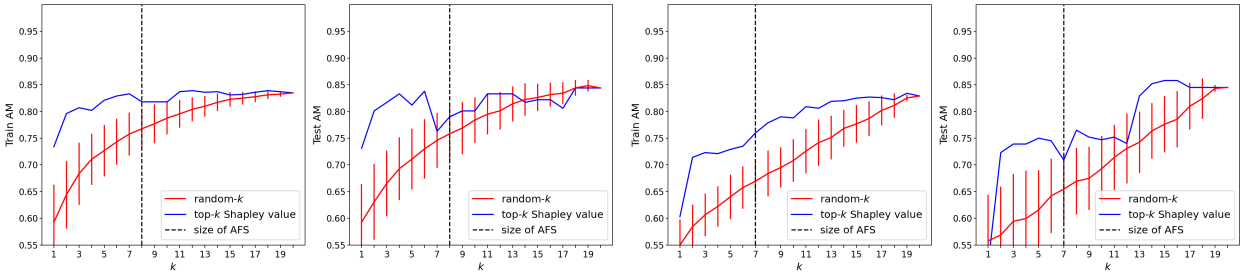
(j) Secretin-like vs Glutamate-like



(k) Rhodopsin-like vs Glutamate-like

(l) Rhodopsin-like vs Secretin-like

(m) Aminergic vs Lipid receptors



(n) Aminergic vs Peptide receptors

(o) Lipid vs Peptide receptors

Figure 14: continued from page 36.