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

Pranav Machingal, Rakesh Busi, Nandyala Hemachandra, Petety V Balaji

***Abstract*—We view a protein as the composite of the standard 20 amino acids (ignoring their order in the protein sequence) and try to identify a set of important amino acid types whose composition is enough to distinguish two paralogous proteins. For this, we use a linear classifier with amino acid composition as features, and a Shapley value based feature subset selection algorithm. We demonstrate our method using 15 datasets of pairs of paralogous proteins. We find that the amino acid composition feature is adequate to distinguish many paralogous proteins from each other. For a pair of paralogous proteins, we are able to identify a subset of amino acids, referred to as AFS (amino acid feature subset), that are key to distinguish them, for each protein. We validate the ability of the AFS amino acids to discriminate by analyzing multiple sequence alignments of corresponding protein families and/or by providing supporting evidence from literature. We also pair-wise classify sub-families of a protein superfamily and highlight common amino acids identified in the AFS for two pairs with a common sub-family.**

***Index Terms*—Feature evaluation and selection, support vector machine, amino acid composition, proteins, globins, GPCRs, multiple sequence alignment.**

I. Introduction

Proteins form the fundamental machinery in living systems performing several important functions such as DNA replication, catalysis, transport, interaction with the environment, etc.

Advancements in sequencing technologies have resulted in exponential growth of protein sequence databases [1]. However, the number of experimentally verified annotations constitute a tiny fraction; for instance, out of around 250 million sequences in UniProtKB [1] only about 0.57 million sequences 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.

P. Machingal and N. Hemachandra are with the Department of Industrial Engineering and Operations Research, Indian Insitute of Technology Bombay, Mumbai, India (e-mail: pranav.mac@iitb.ac.in; nh@iitb.ac.in).

R. Busi and P. V. Balaji are with the Department of Biosciences and Bioengineering, Indian Insitute of Technology Bombay, Mumbai, India (e-mail:rakesh.busi@iitb.ac.in; balaji@iitb.ac.in).

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible

Recently, deep learning methods from natural language processing (NLP) [2], [3] have been adapted to create protein language models [4]–[9]. These give high-dimensional representations for protein sequences which can be used for downstream tasks like predicting the function, domains or physicochemical properties of the protein. Although these models achieve high accuracy on large datasets for coarse-grained levels of annotations, the interpretation of the high-dimensional model embeddings/vectors in terms of the biological concepts/notions is not possible. Thus, they fail to provide insights into the sequence-function relationship or lead to experimentally testable hypotheses.

Paralogs are proteins that have a common ancestor but have diverged functionally. The difference in functionality in two paralogous proteins is considered to arise due to specific evolutionary changes in the sequences [10]. The acquired mutations could lead to differences in the composition of certain amino acid types between two paralogous proteins. Figuring out these amino acid types can be useful to identify the residues that determine the functional specificity of paralogous proteins. Towards this, we view a protein as the composite of its constituent standard 20 amino acids, ignoring the ordering of the amino acids in the protein sequence. We use amino acid composition (AAC) features, Shapley value [11] based feature ranking and subset selection algorithm (Shapley Value based Error Apportioning, SVEA) [12], and support vector machine (SVM) based linear classifier [13] as tools to identify key amino acid types that can distinguish one paralogous protein family from another. The proposed method yields quick results based on which biologists can conduct detailed experiments which are resource-intensive (time, cost, trained manpower, etc.).

We perform experiments on 15 datasets (pairs of paralogous proteins). The key results from our experiments are as follows:

- On 14 out of 15 datasets, high classification scores are achieved using linear SVM and AAC features, implying many paralogous proteins can be distinguished using AAC features.
- The SVEA algorithm identifies a subset of amino acid types (referred to as AFS) that are adequate for distinguishing two paralogous proteins.
- When classifying two paralogous proteins, say families P and Q, we divide the set AFS into disjoint sets $AFS(P)$ and $AFS(Q)$ based on the sign of the linear classifier

0000–0000/00$00.00 © 2021 IEEE

coefficients. In the multiple sequence alignments (MSA) of four datasets, we find that the amino acids in $AFS(P)$ are relatively conserved at more positions in family P sequences than in family Q and similarly, $AFS(Q)$ amino acids are relatively conserved at more positions in family Q sequences than in P.
- When classifying families P vs Q and P vs R, we find common amino acids selected in $AFS(P)$ for the two separate classifications as observed in the case of globin family (Section III-B) and secretin-like GPCRs (Section III-E).
- Common amino acids in $AFS(Q)$ and $AFS(R)$ are either excluded from the AFS for Q vs R classification or have a low rank in that AFS as observed in the case of globin family (Section III-B) and aminergic, lipid, peptide receptors (Section III-E).
- The significance of the amino acids identified in the AFS was validated for nine datasets using supporting evidence from literature that report the conservation of these amino acids at structurally or functionally essential positions.

Most prior works [3], [14] focus on function prediction of protein sequences via EC (enzyme commission) numbers or GO (gene ontology) annotations. Some works like [14] utilise deep learning models and instance-level feature attribution [15] to identify regions in a given input sequence relevant to its prediction. Another method [16] uses a set of positive and unlabelled sequence data generated from deep mutational scanning experiments to determine residues important for a given functional activity. In comparison, the novelty of our work is in identifying amino acid types that distinguish paralogous proteins using only the relative frequencies of the standard 20 amino acids and ignoring their relative ordering in the protein sequences. Furthermore, unlike deep learning methods, our method (a) uses small datasets, and (b) is simple and computationally inexpensive. This can have wide applicability to the domain as the amount of labelled data validated by wet lab experiments by biologists is very limited and unlabelled data is very large.

In the next section, we describe our methodology in detail. In Section III we present the results of our computational experiments. Our conclusions and further discussions are provided in Section IV.

II. Methodology

Our methodology is composed of three main components which are described in detail in the following subsections.

A. AAC features

Consider two functionally divergent homologous proteins, families P and Q. We first curate a set of sequences, say P_{train} and Q_{train}, from standard protein sequence databases with n_P and n_Q number of sequences each from families P and Q respectively.

For a protein sequence $\mathbf{p} = (p_1, p_2, \ldots, p_L)$ of length L with $p_k \in \{1, 2, \cdots, 20\}$ corresponding to the standard 20 amino acids, the AAC feature $\mathbf{x}^{AAC} \in [0,1]^{20}$ for $\mathbf{p}$ is computed as follows,

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

Here feature x_i^{AAC} for a protein sequence is the normalised count of the standard amino acid i, $i \in \{1, 2, \cdots, 20\}$.

B. Feature subset selection

Given a set of N features from the protein sequences of P and Q, we try to find the features $S \subseteq N$ which contribute the most to the classification 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 [12], to identify the most important $S \subseteq N$. Shapley value is a well known solution concept from cooperative game theory [17] for distributing the total worth of a coalition of players fairly among each of them by quantifying each player's effective marginal contribution. The SVEA algorithm considers the binary classification task as a cooperative game among the features and apportions the total training error among the features using Shapley values. The Shapley value $\phi(i)$ for a feature $i \in N$ is computed as follows,

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

Where $v(S) = tr_er(\varnothing) - tr_er(S)$, $tr_er(S)$ is the minimum training hinge loss using S as features and $\varnothing$ is the empty set. The term $(v(S \cup \{i\}) - v(S))$ is the marginal contribution of feature i to the feature subset 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. We use a class-balanced hinge loss for computing $tr_er(S)$ for which $tr_er(\varnothing) = 1$. Shapley values are unique solution concepts satisfying the axioms - efficiency, symmetry and marginality [18]. With efficiency implying $\sum_{i \in N} \phi(i) = v(N) = 1 - tr_er(N)$.

The higher the $\phi(i)$, lower is the contribution of feature i to the classifier training error and consequentially higher the importance of feature i to the classification. We set $\phi_{cutoff} = \frac{1}{20}\sum_{i=1}^{20}\phi(i)$ for selecting the feature subset, $AFS = \{i : \phi(i) > \phi_{cutoff}\}$. Each of the features in AFS uniquely corresponds to $d \leqslant 20$ amino acids from the standard 20. Exact Shapley value computations are known to be exponential time, hence, the Shapley value ϕ_i for a feature i is computed using Monte Carlo based linear time (in number of features) approximation in the SVEA algorithm. As the number of features is small (20), good approximations can be computed fast via larger sampling. More details of the SVEA algorithm are given in supplementary Section S2.

C. Identifying class-wise subsets of AFS

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 $\mathbf{x}^{AFS} \in [0,1]^d$, with $x_j^{AFS} = x_i^{AAC}$ and each

$j \in \{1, 2, \cdots, d\}$ uniquely maps to a $i \in AFS$. We use this linear SVM coefficients $\mathbf{w} \in \mathbb{R}^d$ to divide the set AFS into disjoint sets $AFS(P)$ and $AFS(Q)$ based on the sign of the coefficients. Since $x_j^{AFS} \geqslant 0 \ \forall j \in [d]$, the sign of the linear classifier coefficient w_j indicates which class is relatively prominent in the amino acid corresponding to j. So if the $+1$ class is P, then we divide AFS class-wise as $AFS(P) = \{j \in [d] : w_j > 0\}$ and similarly $AFS(Q) = \{j \in [d] : w_j < 0\}$.

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 true positive and true negative rates (AM) to measure the performance of the classifier [19]. Further, we use a class-balanced version of hinge loss for training the SVM as suggested in [20] for statistical consistency with the AM score. We report the train and test scores of the trained linear SVM with AFS features on the protein family datasets considered in our computational experiments in Section III.

A flowchart summarizing the sequence of steps for computing $AFS(P)$ and $AFS(Q)$ is shown in the Figure 1. For different paralogous protein pairs we provide supporting evidence from the literature for the significance of amino acids in $AFS(P)$ and $AFS(Q)$ (in Section III). We generate a multiple sequence alignment (MSA) of a set of randomly selected sequences from P_{train} and Q_{train} and analyze the conservation of $AFS(P)$ and $AFS(Q)$ amino acids within and across the respective families (see Figure 3). When classifying families P vs Q and P vs R, we compare the amino acids in the two $AFS(P)$ sets corresponding to family P that we get from the two classifications (in Section III).

III. Identifying key distinguishing amino acid types (AFS) for paralogous proteins

We discuss the results of applying our method for identifying amino acid types that distinguish paralogous proteins using 15 datasets described in Table I. We have considered paralogous proteins wherein functional differences vary from fine-grained (e.g., trypsin/chymotrypsin) to coarse-grained (e.g., GPCR). This diversity is considered for robust evaluation of the method. All 15 datasets are taken from publicly available databases (UniProt [1] and GPCR-PEnDB [21]). 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 constitute a very large family of proteins with variations in specific functions. We chose to use an already curated dataset in this case. For each of the GPCR families considered (see Table I), the sequences are randomly split as 80%-train/20%-test. 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. However, even without removing the UniProt entries with identical sequences, no significant difference was observed in the results of the computations that may change our conclusions (see supplementary Section S3). More details regarding the method/criteria for collecting the data along with the data files (with accession numbers) are provided in the supplementary Section S1.

We train linear SVMs with AAC and AFS features respectively for all datasets using class-balanced hinge loss [20], the classification scores for which are reported in Tables II-VII. We discuss regarding the identified AFS for some of these datasets in the further subsections.

TABLE I: 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
Globins		
Myoglobin	107	479
Hemoglobin-α	303	525
Hemoglobin-β	285	261
Trypsin-like		
Trypsin	65	45
Chymotrypsin	17	16
Tubulin		
α	117	190
β	191	347
Interleukin-1		
α	16	12
β	25	194
Histone		
H2A	180	16599
H2B	177	7599
Cytochrome P450		
CYP3	59	942
CYP5	165	679
	(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

A. *Lysozyme C and α-Lactalbumin*

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

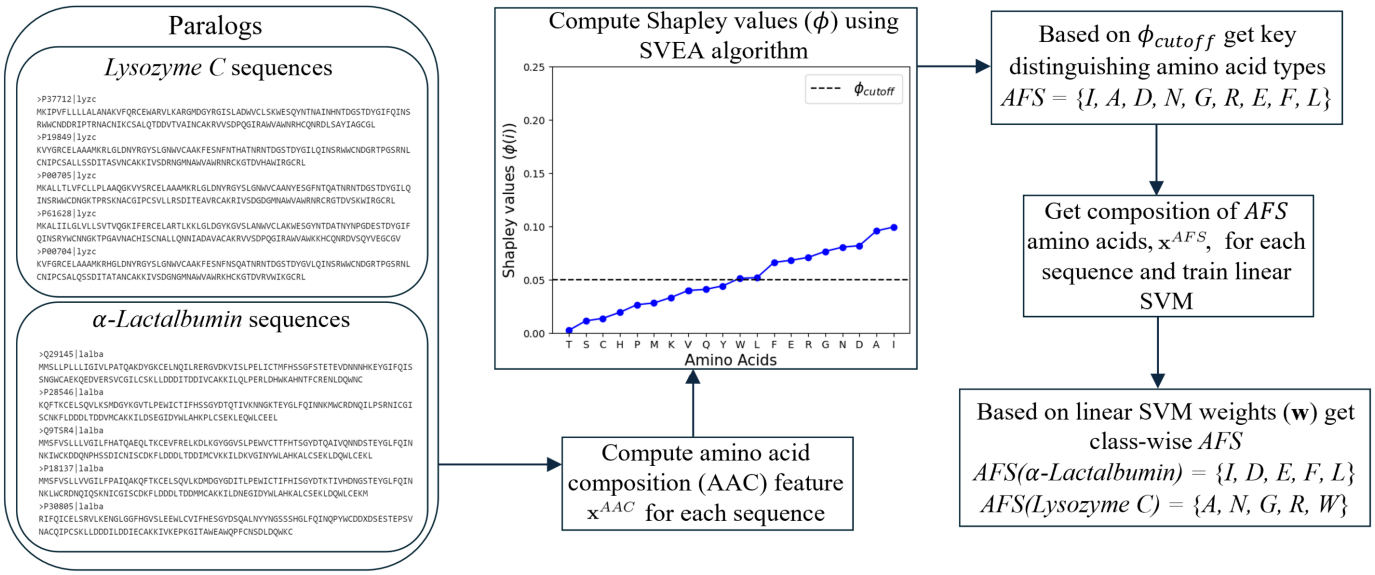


Fig. 1: Flowchart summarizing the sequence of steps used in our method to compute the key amino acid types (AFS: amino acid feature subset) that distinguish paralogous proteins, using amino acid composition (AAC) features, Shapley value based SVEA algorithm for feature subset selection and class-wise feature subsets using linear SVM. Lysozyme C and α-Lactalbumin are used here as representative examples of paralogous proteins. AFS is identified for other pairs of paralogous proteins using the same approach, and are discussed in Section III.

TABLE II: Lysozyme C vs α-Lactalbumin classification score. The train score is the mean 5-fold cross-validation score.

	AAC	AFS
Train AM.	1.0	0.993 ($\pm$0.013)
Test AM.	0.896	0.898
Train Acc.	1.0	0.99 ($\pm$0.02)
Test Acc.	0.836	0.881

to have diverged from a common ancestor millions of years ago [22].

The following are the feature subsets identified for this dataset:

- $AFS = \{I, A, D, N, G, R, E, F, L, W\}$
- $AFS(\alpha\text{-Lactalbumin}) = \{I, D, E, F, L\}$
- $AFS(\text{Lysozyme C}) = \{A, N, G, R, W\}$

We observe in the MSA of the sequences, in Figure 3a, that the identified feature sets $AFS(\alpha\text{-Lactalbumin})$ and $AFS(\text{Lysozyme C})$ show significant conservation in respective families resulting in a pattern which distinguishes the sequences in one family from the other. The amino acids D and E which are part of $AFS(\alpha\text{-Lactalbumin})$ are known to be found in the Ca^{2+} and Zn^{2+} binding sites respectively of α-lactalbumin [23], [24]. 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+}.

B. Globins

Globins are a superfamily of functionally divergent homologous protein families with a high level of fold conservation. We consider three well-known families belonging to the globins superfamily, myoglobin, hemoglobin-α and hemoglobin-β. All three are globular proteins that contain a heme group that constitutes the oxygen binding site. 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 $\alpha_2\beta_2$ [10].

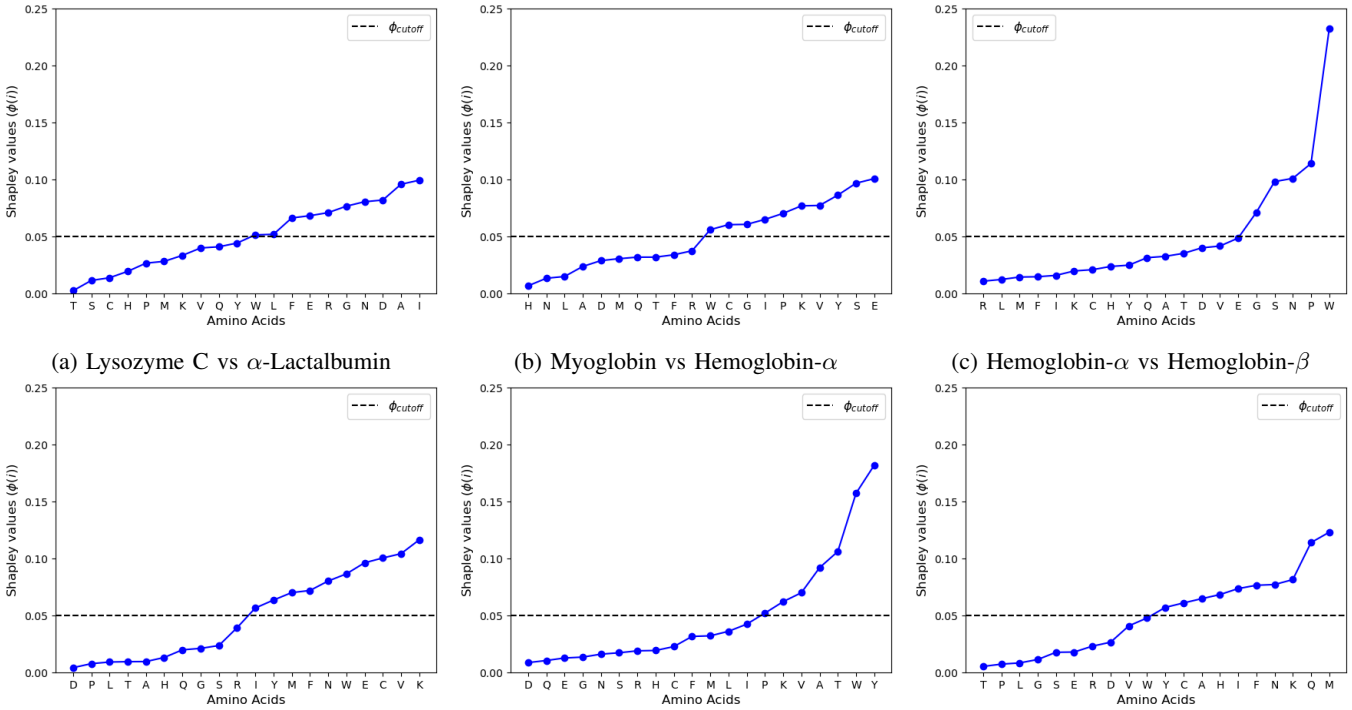
TABLE III: Globins classification scores. The train score is the mean ($\pm$1 standard deviation) 5-fold cross-validation score.

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

1) Myoglobin and Hemoglobin-α: The following are the feature subsets identified for this dataset:

- $AFS_1 = \{E, S, Y, V, K, P, I, G, C, W\}$
- $AFS_1(\text{Myoglobin}) = \{E, K, I, G, W\}$
- $AFS_1(\text{Hemoglobin-}\alpha) = \{S, Y, V, P, C\}$

We see that even with reduced feature set (AFS) a high



(a) Lysozyme C vs α-Lactalbumin (b) Myoglobin vs Hemoglobin-α (c) Hemoglobin-α vs Hemoglobin-β

(d) Myoglobin vs Hemoglobin-β (e) Trypsin vs Chymotrypsin (f) Tubulin-α vs Tubulin-β

Fig. 2: Shapley value ($\phi(i)$) for AAC features computed using Monte Carlo-based approximation algorithm, SVEA, by [12]. $AFS = \{i : \phi(i) > \phi_{cutoff}\}$.

classification score is achieved for this dataset (see Table III). Also, we can observe from the MSA in Figure 3b that the amino acids in AFS_1(myoglobin) and AFS_1(hemoglobin-α) create a pattern that distinguishes the sequences in one family from the other.

2) Myoglobin and Hemoglobin-β: The following are the feature subsets identified for this dataset:

- $AFS_2 = \{K, V, C, E, W, N, F, M, Y, I\}$
- AFS_2(Myoglobin) $= \{K, E, M, I\}$
- AFS_2(Hemoglobin-β) $= \{V, C, W, N, F, Y\}$

The MSA of the sequences of this dataset (see supplementary Figure S2) also form a distinguishing pattern based on the amino acids in AFS(Myoglobin) and AFS(Hemoglobin-β).

3) Hemoglobin-α and Hemoglobin-β: The following are the feature subsets identified for this dataset:

- $AFS_3 = \{W, P, N, S, G\}$
- AFS_3(Hemoglobin-α) $= \{P, S\}$
- AFS_3(Hemoglobin-β) $= \{W, N, G\}$

Here too we observe from the MSA in Figure 3c that the amino acids in AFS_3(Hemoglobin-α) and AFS_3(Hemoglobin-β), form a pattern which distinguishes the sequences in one family from the other. Further, amino acid W which receives a significantly high Shapley value $\phi(W)$ (see Figure 2c) is highly conserved at position 40 in the MSA in hemoglobin-β sequences as compared to hemoglobin-α sequences. This W residue at position 40 has been determined to be present in hemoglobin-β at one of its contact position to hemoglobin-α in the tetrameric structure [25] and is, therefore, a structurally and functionally significant residue.

4) Analysing AFS across globins: 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 [25]) 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 supplementary Figure S1).

The amino acids which are common in AFS_1(Hemoglobin-α) and AFS_2(Hemoglobin-β), V, Y, C, when classifying against myoglobin, are not found in AFS_3 in α vs β hemoglobin classification. Furthermore, the amino acid C has been shown to play an important role in the tetrameric structure of hemoglobin [26]. Hence, it can be expected that V, Y, C would not be *key* in AAC for distinguishing α vs β hemoglobin.

C. Trypsin and Chymotrypsin

Trypsin and chymotrypsin are a family of serine proteases [10]. They cleave peptide bonds using a spatial arrangement of three amino acids, S, H, and D, referred to as the catalytic triad. 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) [27].

The following are the feature subsets identified for this dataset:

(a) Lysozyme C (left) vs α-Lactalbumin (right) (b) Hemoglobin-α (left) vs Myoglobin (right) (c) Hemoglobin-β (left) vs Hemoglobin-α (right)

Fig. 3: Multiple sequence alignment of sequences from the respective families in (a), (b) and (c). 15 sequences on the left are from one family and 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. (a) AFS(α-Lactalbumin) amino acids are in green and AFS(Lysozyme C) are in red. (b) The amino acids in $AFS(\text{Myoglobin})_1$ are in green and $AFS(\text{Hemoglobin-}\alpha)_1$ are in red. (c) The amino acids in $AFS(\text{Hemoglobin-}\alpha)_2$ are in green and $AFS(\text{Hemoglobin-}\beta)_2$ are in red. The intensity of the color is proportional to the Shapley value $\phi(i)$ of the amino acid i (see Figure 2).

TABLE IV: Trypsin vs chymotrypsin classification scores. The train score is the mean (± 1 standard deviation) 5-fold cross-validation score.

	AAC	AFS
Train AM.	0.992 ($\pm$0.015)	0.985 ($\pm$0.019)
Test AM.	0.549	0.505
Train Acc.	0.988 ($\pm$0.024)	0.976 ($\pm$0.029)
Test Acc.	0.721	0.656

- $AFS = \{Y, W, T, A, V, K, P\}$
- $AFS(\text{Trypsin}) = \{Y, A\}$
- $AFS(\text{Chymotrypsin}) = \{W, T, V, K, P\}$

We find that despite a high 5-fold classification train score, the score on the test set is very low. This could be due to the fact the test sequences which were collected from the TrEMBL database [1] had a very low annotation score of either 1 or 2 out of 5 (mostly being 1). However, we do not find any discernible pattern in the MSA for this dataset which distinguishes one family from the other.

The amino acids Y and W get the highest $\phi(\cdot)$ score in $AFS(\text{Trypsin})$ and $AFS(\text{Chymotrypsin})$ respectively. In experiments to convert trypsin to chymotrypsin [28], [29] it has been shown that Y to W conversion in loop-3 of trypsin leads to significant increase in chymotrypsin activity. Also, we find that the catalytic triad amino acids, S, H and D, which are common to both families, are not part of the AFS

D. Tubulin-α and Tubulin-β

TABLE V: Tubulin-α vs Tubulin-β classification scores. The train score is the mean (± 1 standard deviation) 5-fold cross-validation score.

	AAC	AFS
Train AM.	0.996 ($\pm$0.009)	0.997 ($\pm$0.006)
Test AM.	0.992	0.992
Train Acc.	0.997 ($\pm$0.006)	0.994 ($\pm$0.008)
Test Acc.	0.991	0.994

Tubulin α and β heterodimers form the building blocks of microtubules, which are important components of the eukaryotic cytoskeleton [30].

The following are the feature subsets identified for this dataset:

- $AFS = \{M, Q, K, N, F, I, H, A, C, Y\}$
- $AFS(\text{Tubulin-}\alpha) = \{K, I, H, C, Y\}$
- $AFS(\text{Tubulin-}\beta) = \{M, Q, N, F, A\}$

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 find that the $AFS(\text{Tubulin-}\alpha)$ amino acids are present at more positions in contact points of tubulin-α than tubulin-β. Similarly, $AFS(\text{Tubulin-}\beta)$ amino acids are present at more positions in contact points of tubulin-β than tubulin-α. Thus, the amino acids identified in AFS can be considered to be significant towards the quaternary structure of tubulin-α/β heterodimer. Please see supplementary Section S3 for more details.

E. G-protein coupled receptors (GPCRs)

GPCRs are a large and diverse group of transmembrane proteins that mediate cellular responses to extracellular signals. A characteristic structure containing seven transmembrane helices with intermediate loops is common to the proteins in this family. We consider pair-wise classification of 3 GPCR families, rhodopsin-like, secretin-like and glutamate-like, identified in GRAFS [21] system. Further, we pair-wise classify 3 subfamilies of rhodopsin-like GPCR proteins - aminergic receptors, lipid receptors and peptide receptors.

TABLE VI: GPCRs classification scores. The train score is the mean (± 1 standard deviation) 5-fold cross-validation score.

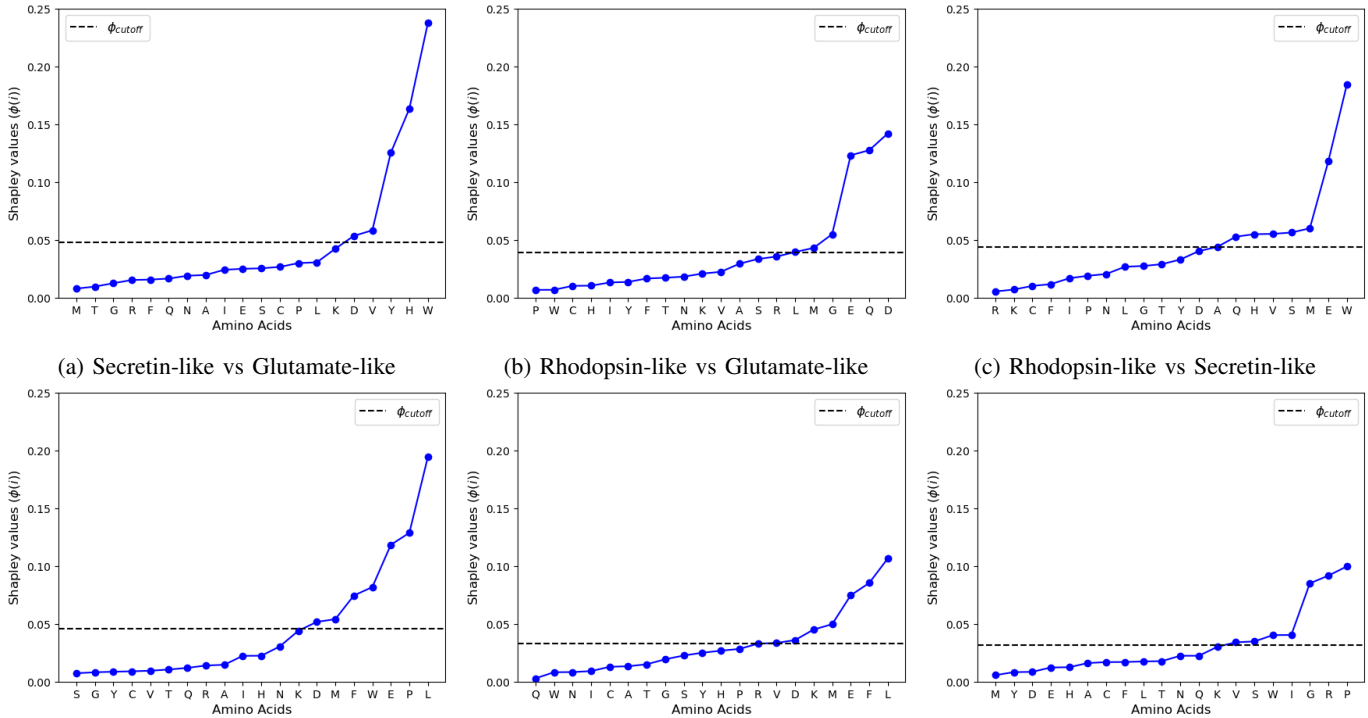
Dataset		AAC	AFS
Secretin-like vs Glutamate-like	**Train AM.**	0.933 ($\pm$0.042)	0.95 ($\pm$0.032)
	Test AM.	0.888	0.845
	Train Acc.	0.933 ($\pm$0.042)	0.95 ($\pm$0.032)
	Test Acc.	0.889	0.844
Rhodopsin-like vs Glutamate-like	**Train AM.**	0.884 ($\pm$0.042)	0.85 ($\pm$0.045)
	Test AM.	0.967	0.934
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

1) Rhodopsin-like and Glutamate-like: The following are the feature subsets identified for this dataset:

- $AFS_1 = \{D, Q, E, G, M, L\}$
- $AFS_1(\text{Rhodopsin}) = \{M, L\}$
- $AFS_1(\text{Glutamate}) = \{D, Q, E, G\}$

2) Secretin-like and Glutamate-like: The following are the feature subsets identified for this dataset:

- $AFS_2 = \{W, H, Y, V, D\}$



(a) Secretin-like vs Glutamate-like (b) Rhodopsin-like vs Glutamate-like (c) Rhodopsin-like vs Secretin-like

(d) Aminergic receptors vs Lipid receptors (e) Aminergic receptors vs Peptide receptors (f) Lipid receptors vs Peptide receptors

Fig. 4: Shapley value ($\phi(i)$) for AAC features computed using Monte Carlo based approximation algorithm, SVEA, by [12] for GPCR datasets. $AFS = \{i : \phi(i) > \phi_{cutoff}\}$.

- AFS_2(Secretin) $= \{W, H, Y\}$
- AFS_2(Glutamate) $= \{V, D\}$

3) Rhodopsin-like and Secretin-like: The following are the feature subsets identified for this dataset:

- $AFS_3 = \{W, E, M, S, V, H, Q, A\}$
- AFS_3(Rhodopsin) $= \{M, S, V, A\}$
- AFS_3(Secretin) $= \{W, E, H, Q\}$

4) Analysing AFS of secretin-like GPCRs: W which receives the highest $\phi(\cdot)$ score in AFS_2(Secretin) and AFS_3(Secretin), has been reported by [31] to be well conserved at multiple positions with structural and functional importance in secretin-like GPCR sequences.

5) Aminergic receptors and Lipid receptors: The following are the feature subsets identified for this dataset:

- $AFS_1 = \{L, P, E, W, F, M, D\}$
- AFS_1(Aminergic receptors) $= \{P, E, W, D\}$
- AFS_1(Lipid receptors) $= \{L, F, M\}$

6) Aminergic receptors and Peptide receptors: The following are the feature subsets identified for this dataset:

- $AFS_2 = \{L, F, E, M, K, D, V, R\}$
- AFS_2(Aminergic receptors) $= \{E, K, D, R\}$
- AFS_2(Peptide receptors) $= \{L, F, M, V\}$

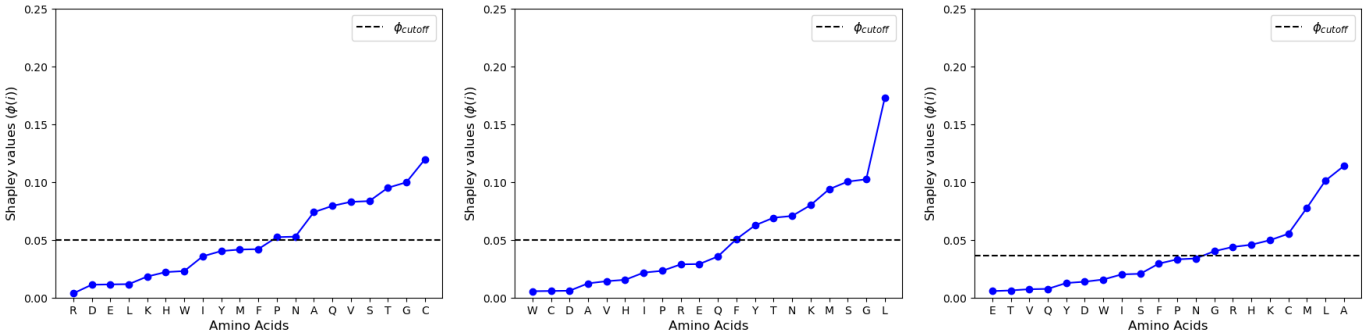
7) Lipid receptors and Peptide receptors: The following are the feature subsets identified for this dataset:

- $AFS_3 = \{P, R, G, I, W, S, V\}$
- AFS_3(Lipid receptors) $= \{R, G, S\}$
- AFS_3(Peptide receptors) $= \{P, I, W, V\}$

8) Analysing AFS of lipid, peptide receptor GPCRs: We do not find L, F in the AFS_3, which could be because L, F are part of AFS_1(Lipid receptors) and AFS_2(Peptide receptors) when these families are classified against Aminergic receptors. Hence, it can be expected that L, F would not be *key* in AAC to distinguish lipid vs peptide receptors.

TABLE VII: Classification scores for 3 datasets. The train score is the mean (± 1 standard deviation) 5-fold cross-validation score.

Dataset		AAC	AFS
Interleukin-1 α vs Interleukin-1 β	**Train AM.**	0.98 ($\pm$0.04)	0.98 ($\pm$0.04)
	Test AM.	0.979	0.985
	Train Acc.	0.975 ($\pm$0.05)	0.975 ($\pm$0.05)
	Test Acc.	0.961	0.971
Histone H2A vs Histone H2B	**Train AM.**	0.983 ($\pm$0.016)	0.983 ($\pm$0.01)
	Test AM.	0.91	0.934
	Train Acc.	0.983 ($\pm$0.016)	0.983 ($\pm$0.01)
	Test Acc.	0.889	0.922
Cytochrome P450 CYP3 vs Cytochrome P450 CYP5	**Train AM.**	0.875 ($\pm$0.042)	0.877 ($\pm$0.01)
	Test AM.	0.842	0.835
	Train Acc.	0.879 ($\pm$0.036)	0.866 ($\pm$0.001)
	Test Acc.	0.831	0.822



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

(b) Histone H2A vs Histone H2B

(c) Cytochrome P450 CYP3 vs Cytochrome P450 CYP5

Fig. 5: Shapley value ($\phi(i)$) for AAC features computed using Monte Carlo based approximation algorithm, SVEA, by [12] for GPCR datasets. $AFS = \{i : \phi(i) > \phi_{cutoff}\}$.

F. Interleukin α and Interleukin β

The following are the feature subsets identified for this dataset:

- $AFS = \{C, G, T, S, V, Q, A, N, P\}$
- AFS(Interleukin-1 α) $= \{T, S, A, N\}$
- AFS(Interleukin-1 β) $= \{C, G, V, Q, P\}$

G. Histone H2A and Histone H2B

The following are the feature subsets identified for this dataset:

- $AFS = \{L, G, S, M, K, N, T, Y, F\}$
- AFS(Histone H2A) $= \{L, G, N\}$
- AFS(Histone H2B) $= \{S, M, K, T, Y, F\}$

H. Cytochrome P450 CYP3 and Cytochrome P450 CYP5

The following are the feature subsets identified for this dataset:

- $AFS = \{A, L, M, C, K, H, R, G\}$
- AFS(Cytochrome P450 CYP3) $= \{L, M, C, K, R\}$
- AFS(Cytochrome P450 CYP5) $= \{A, H, G\}$

IV. Discussion

Paralogs are protein sequences that have a common ancestor and different functionality. We first demonstrate that AAC is a good feature set for many paralog sequences as SVM using these features achieves high test arithmetic mean (of true positive rate and true negative rate) scores; these datasets also have a significant class imbalance. We further observed in 14 out of 15 datasets in our experiment that the reduced feature set AFS achieves high classification test scores comparable to that with full 20 AAC features. This demonstrates the robustness of the feature subset selection algorithm, SVEA, by [12], which we used to identify AFS. The amino acids identified in AFS seem to contain the set of amino acids whose composition differs in the two paralogous proteins. We argue that the low test score for trypsin vs chymotrypsin classification can be due to the high amount of noisy labels in the test set curated from TrEMBL database [1]. Thus, we demonstrate that the SVEA algorithm is an effective tool to identify AFS, which is the set of key amino acid types that distinguish paralogous protein families. We also show that using a linear classifier (viz. linear SVM), we can further divide the amino acids in AFS to class-wise subsets to predict which amino acid types in AFS are important to which protein of the paralog pair.

As we use AAC features that do not contain position information of the amino acids, we relied on MSA or 3D structures to find conservation patterns of the identified feature subsets in the respective families. Engineering features that incorporate sequence order information from the protein (for e.g., k-mers), can be used to additionally identify the positions of the distinguishing amino acids in the protein sequences. However, these features can be very high-dimensional, for example, 20^k-dimensional for k-mer features. Shapley value computation from high-dimensional features can be expensive. This is because the Monte Carlo based approximation algorithm would require exponentially more sampling (in the number of features) for good approximations of Shapley value. In the future we plan to look into this aspect of efficient feature subset selection for high-dimensional features.

References

[1] The UniProt Consortium, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 11 2020. [Online]. Available: https://doi.org/10.1093/nar/gkaa1100

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] M. L. Bileschi, D. Belanger, D. H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M. A. DePristo, and L. J. Colwell, "Using deep learning to annotate the protein universe," *Nature biotechnology*, vol. 40, no. 6, p. 932—937, June 2022. [Online]. Available: https://doi.org/10.1038/s41587-021-01179-w

[4] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell Systems*, vol. 12, no. 6, pp. 654–669.e3, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405471221002039

[5] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[6] D. Ofer, N. Brandes, and M. Linial, "The language of proteins: NLP, machine learning & protein sequences," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750–1758, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037021000945

[7] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, and Y. S. Song, "Evaluating protein transfer learning with tape," in *Advances in NeuIPS*, 2019.

[8] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "BERTology Meets Biology: Interpreting Attention in Protein Language Models," 2020. [Online]. Available: https://arxiv.org/abs/2006.15222

[9] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *PNAS*, vol. 118, no. 15, 2021. [Online]. Available: https://www.pnas.org/content/118/15/e2016239118

[10] K. Dill, R. Jernigan, and I. Bahar, *Protein Actions: Principles and Modeling*. CRC Press, 2017. [Online]. Available: https://books.google.co.in/books?id=NHs2DwAAQBAJ

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

[12] S. Tripathi, N. Hemachandra, and P. Trivedi, "Interpretable feature subset selection: A Shapley value based approach," in *2020 IEEE BigData*, 2020, pp. 5463–5472.

[13] I. Steinwart and A. Christmann, *Support Vector Machines*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[14] T. Sanderson, M. L. Bileschi, D. Belanger, and L. J. Colwell, "Proteinfer, deep neural networks for protein functional inference," *eLife*, vol. 12, p. e80942, feb 2023. [Online]. Available: https://doi.org/10.7554/eLife.80942

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of IEEE CVPR*, June 2016.

[16] H. Song, B. J. Bremer, E. C. Hinds, G. Raskutti, and P. A. Romero, "Inferring protein sequence-function relationships with large-scale positive-unlabeled learning," *Cell Systems*, vol. 12, no. 1, pp. 92–101.e8, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405471220304142

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

[18] H. P. Young, "Monotonic solutions of cooperative games," *Int. J. Game Theory*, vol. 14, no. 2, p. 65–72, jun 1985. [Online]. Available: https://doi.org/10.1007/BF01769885

[19] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th ICPR*, 2010, pp. 3121–3124.

[20] A. K. Menon, H. Narasimhan, S. Agarwal, and S. Chawla, "On the statistical consistency of algorithms for binary classification under class imbalance," in *Proceedings of the 30th ICML - Volume 28*, ser. ICML'13. JMLR.org, 2013, p. III–603–III–611.

[21] K. Begum, J. E. Mohl, F. Ayivor, E. E. Perez, and M.-Y. Leung, "GPCR-PEnDB: a database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors," *Database*, vol. 2020, 11 2020. [Online]. Available: https://doi.org/10.1093/database/baaa087

[22] P. K. Qasba, S. Kumar, and D. K. Brew, "Molecular divergence of lysozymes and α-lactalbumin," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 32, no. 4, pp. 255–306, 1997. [Online]. Available: https://doi.org/10.3109/10409239709082574

[23] E. A. Permyakov and L. J. Berliner, "α-Lactalbumin: structure and function," *FEBS Letters*, vol. 473, no. 3, pp. 269–274, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0014579300015465

[24] E. A. Permyakov, "α-actalbumin, Amazing Calcium-Binding Protein," *Biomolecules*, vol. 10, no. 9, p. 1210, Aug 2020. [Online]. Available: http://dx.doi.org/10.3390/biom10091210

[25] M. Shionyu, K. Takahashi, and M. Gō, "Variable subunit contact and cooperativity of hemoglobins," *J. Mol. Evol.*, vol. 53, no. 4-5, pp. 416–429, Oct. 2001.

[26] H.-I. Kan, I.-Y. Chen, M. Zulfajri, and C. C. Wang, "Subunit disassembly pathway of human hemoglobin revealing the site-specific role of its cysteine residues," *The Journal of Physical Chemistry B*, vol. 117, no. 34, pp. 9831–9839, 2013, pMID: 23902424. [Online]. Available: https://doi.org/10.1021/jp402292b

[27] W. Ma, C. Tang, and L. Lai, "Specificity of trypsin and chymotrypsin: Loop-motion-controlled dynamic correlation as a determinant," *Biophysical Journal*, vol. 89, no. 2, pp. 1183–1193, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006349505727661

[28] L. Hedstrom, J. J. Perona, and W. J. Rutter, "Converting trypsin to chymotrypsin: residue 172 is a substrate specificity determinant." *Biochemistry*, vol. 33 29, pp. 8757–63, 1994.

[29] L. Hedstrom, "Serine protease mechanism and specificity," *Chemical Reviews*, vol. 102, no. 12, pp. 4501–4524, 2002, pMID: 12475199. [Online]. Available: https://doi.org/10.1021/cr000033x

[30] T. Mühlethaler, D. Gioia, A. E. Prota, M. E. Sharpe, A. Cavalli, and M. O. Steinmetz, "Comprehensive analysis of binding sites in tubulin," *Angewandte Chemie International Edition*, vol. 60, no. 24, pp. 13 331–13 342, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202100273

[31] B. P. Cary, X. Zhang, J. Cao, R. M. Johnson, S. J. Piper, E. J. Gerrard, D. Wootten, and P. M. Sexton, "New Insights into the Structure and Function of Class B1 GPCRs," *Endocrine Reviews*, vol. 44, no. 3, pp. 492–517, 12 2022. [Online]. Available: https://doi.org/10.1210/endrev/bnac033