CysDuF database: annotation and characterization of Cysteine residues in Domain of Unknown Function (DUF) proteins based on Cysteine post-translational modifications, their protein microenvironments, biochemical pathways, taxonomy, and diseases

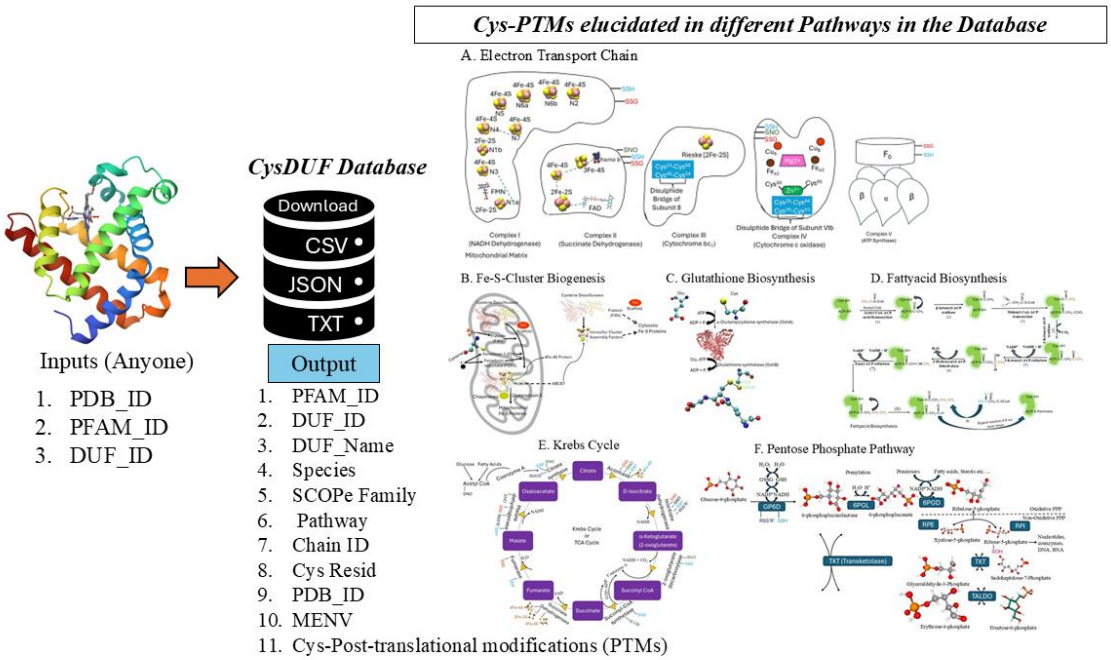
Devarakonda Himaja and Debashree Bandyopadhyay*,

Department of Biological Sciences, Birla Institute of Technology and Science, Pilani, Hyderabad Campus, Hyderabad, India 500078

***corresponding author email: banerjee.debi@hyderabad.bits-pilani.ac.in**

Abstract:

Experimental characterization and annotation of amino acids belonging to Domains of Unknown Function (DUF) proteins are expensive, and time-consuming which could be complemented by computational methods. Cysteine, being the second most reactive amino acid at the catalytic sites of enzymes, was selected for functional annotation and characterization on DUF proteins. Earlier we reported functional annotation of Cysteine on DUF proteins belonging to the COX-II family. However, holistic characterization of Cysteine functions on DUF proteins was not known, to the best of our knowledge. Here, we annotated and characterized Cysteine residues based on post-translational modifications (PTMs), biochemical pathways, diseases, taxonomy, and protein microenvironment. The information on uncharacterized DUF proteins was initially obtained from the literature and the sequence, structure, pathways, taxonomy, and disease information were retrieved from the SCOPe database using DUF IDs. Protein microenvironments (MENV) around Cysteine residues from DUF proteins were computed using protein structures (n=70342). The Cysteine PTMs were predicted using the in-house Cysteine-function prediction server, DeepCys https://deepcys.bits-hyderabad.ac.in). The accuracy of the prediction, validated against known experimental Cysteine PTMs (n=18626) was 0.79. The information was consolidated in the database (https://cysduf.bits-hyderabad.ac.in/), retrievable in downloadable formats (CSV, JSON, or TXT) using the following inputs, DUF ID, PFAM ID, or PDB ID. For the first time, we annotated Cysteine PTMs in DUF proteins belonging to seven different biochemical pathways and various species across the taxonomy, notably for the SARS-COV2 virus. The nature of MENV around Cysteine from DUF proteins was mainly buried and hydrophobic. However, in the SARS-COV2 virus, a significant number of functional Cysteine residues were exposed on the surface with hydrophilic microenvironment.



Abstract Figure: CysDuF database

Keywords: Cysteine, Post-Translational Modifications (PTMs), DUF proteins, Biochemical pathways, Diseases, Taxonomy, Protein Microenvironment

Introduction:

Cysteine has unique chemical properties due to its reactive thiol group that undergoes a wide range of redox reactions and contributes towards various biological pathways. It can act as a nucleophile (S^-) under physiological pH (pKa of cysteine thiol group is 8.1) and may serve as one of the key catalytic residues in many enzymes. Cysteine functions are broadly categorized into four groups, i) Structural cysteines, ii) metal-binding cysteines, iii) catalytic cysteines, and iv) regulatory cysteines (Marino and Gladyshev 2012). The biological functions of cysteines include redox properties, binding to co-factors, scavenging reactive oxygen species (ROS), and reactive nitrogen species (RNS), scavenging toxic heavy metal ions, etc. This variety of cysteine functions and their possible consequences on biochemical reactions make cysteine a suitable candidate for its function prediction in a given protein. With the advent of high-throughput screening, a large number of protein domains, Domains of Unknown Function (DUFs), were sequenced whose functions were uncharacterized. Experimental characterizations of amino acid functions for these DUF proteins were laborious and time-consuming. The computational approach could complement functional annotations of Cysteine amino acids on DUF proteins. A total of 4775 DUF protein families were available in the PFAM database (v 35.0) (Mistry et al. 2021), including both Domains of Unknown Function (DUFs) and Uncharacterized Protein Families (UPFs) (Mudgal et al. 2015; Mistry et al. 2021). "SUPFAM" database curated all DUF proteins and provided the external link to the SCOPe database (Pandit et al. 2004). Similarly, the "PathFams" database detected pathogen-assisted protein domains in DUF proteins (Lobb et al. 2021). The DUF proteins may belong to different biological functions, species, groups of organisms, or environmental conditions. Hence, the characterization of DUF protein function is crucial. DUF family proteins were reported to be involved in plant physiology, such as plant cell wall development, trichome development, plant stress responses, etc. (Luo et al. 2024; Lv et al. 2023). The disease-related DUF proteins were reported, such as pneumonia, neuronal diseases, viral infections, food-borne illnesses, fungal diseases, and many more (Goodacre, Gerloff, and Uetz 2014). DUF characterization was accelerated using computational techniques, such as Phylogenetic Tree, Gene Expression Analysis, GO Analysis, DALI Search Algorithm (Behrens and Spielmann 2024; Huang et al. 2019), etc. Recently, bacterial signaling proteins, from DUF families, were characterized as GGDEF and EAL domains (Tong et al. 2016). In *Oryza Sativa* (Rice), the function of the DUF568 was characterized using the phylogenetic tree, Gene expression, GO analysis, Co-expression, and protein-protein interaction (PPI) networks (Chen et al. 2023). In *Plasmodium falciparum,* DUF proteins were characterized using DALI search on Alpha Fold predictions. In *Agrobacterium tumefaciens,* DUF1127 was predicted to be involved in phosphate and carbon metabolism, using sequence similarity (Kraus et al. 2020). Similarly, DUF692 was annotated as Multicellular non-heme iron-dependent oxidative enzymes, using sequence similarity (Ayikpoe et al. 2023). Our recent study predicted post-translational modifications of Cysteine in the DUF proteins belonging to cytochrome C oxidase, subunit II-like transmembrane domains (COX II protein) (Nallapareddy et al. 2021). "Unknome" database reported experimentally annotated genes of the DUF proteins using RNA interference (RNAi) and knockdown techniques (Rocha et al. 2023). Apart from DUF sequences, only two PDB structures are available for DUF proteins. However, there are many DUF-related protein structures available in the PDB database (Burley et al. 2019). Due to the unavailability of DUF PDB structures, the structural information was extracted from the DUF-related protein PDBs, reported in the SCOPe database. The structural information was required for the computation of local protein microenvironments and subsequent characterization of biochemical pathways, taxonomic distributions, diseases, etc. The protein microenvironment around Cysteines from DUF-related proteins could be

calculated based on the structures of the globular proteins only. The protein microenvironment is known to modulate various biological activities, including molecular recognition, protein-protein interactions, alteration of amino acid pKa values, hydration and dehydration properties, etc. (Bandyopadhyay and Mehler 2008; Bhatnagar, Apostol, and Bandyopadhyay 2016a; Bhatnagar and Bandyopadhyay 2018; Najafi et al. 2025). The hypothesis in the current study is protein microenvironment will modulate the Cysteine post-translational modifications in DUF-related proteins, their biochemical pathways, and related diseases. This hypothesis was tested on four Cysteine post-translational modifications (limited due to the availability of the protein structures), namely, disulfide, metal-binding, thioether and sulfenylation (Figure 1); seven biochemical pathways, Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis, Fatty Acid Synthesis, Photosynthesis, Kreb's Cycle, and Pentose phosphate pathway; and one hundred and fifty-six diseases within four taxonomic groups, according to NCBI Taxanomy(Schoch et al. 2020).

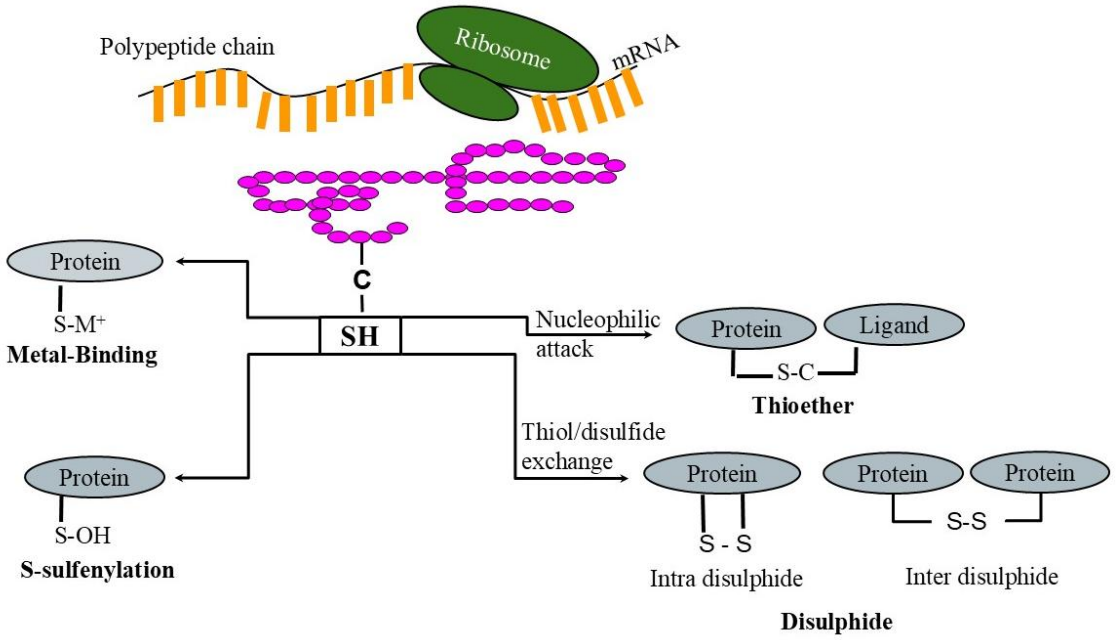


Figure 1: Schematic representation of four Cysteine post-translational modifications described in the CysDuF database. The figure was depicted using Microsoft power point 365 suite.

Methods:

1. DUF protein Dataset curation:

DUF protein dataset was curated (May 22nd, 2024) from the SUPFAM database (http://proline.biochem.iisc.ernet.in/RHD_DUFS/) using the Python library, beautifulsoup4 (version=4.12.3). The list of curated DUF proteins was filtered using two criteria. The first one was pathway names – Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis, Fatty Acid Synthesis, Photosynthesis, Kreb's Cycle, and Pentose phosphate pathway. The second criterium was catalytic Cysteine in those pathways. The filtered information was saved in CSV format that contains the following columns, Pfam Accession (ID), DUF_ID, DUF name, and SCOPe ID. The SCOPe database (Chandonia et al. 2022) was searched to extract SCOPe superfamily ID, family ID, and PDB ID, sequentially. The flow of the data curation was shown schematically (Figure 2). The PDB IDs were obtained from different experimental

sources, namely, X-ray diffraction (n=5835), NMR studies (n=233), and Electron Microscopy (n=68) (Figure S1). The structures without reported experimental methods were discarded.

All information was concatenated and saved in CSV format. This CSV file was utilized to develop the web server.

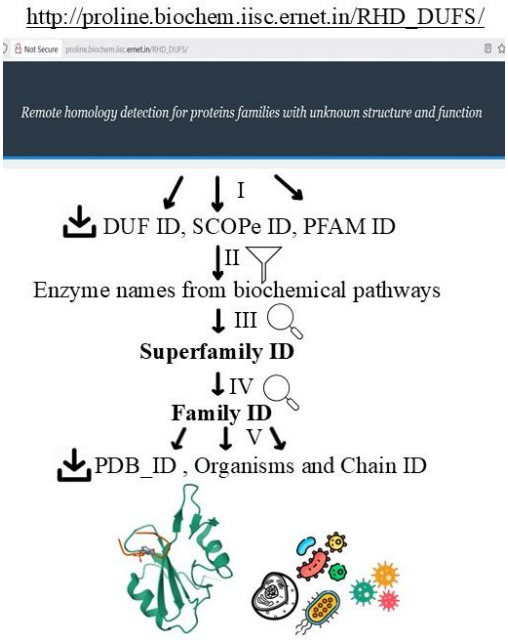


Figure 2. Steps of DUF data curation. (i) extract and download a list of PFAM ID/DUF ID/ SCOPe ID using search criteria, a) pathway names, and b) catalytic Cysteines, from (http://proline.biochem.iisc.ernet.in/RHD_DUFS) (ii) filter the downloaded list using SCOPe superfamily resulting enzyme names from 7 biochemical pathways studied here (iii) search SCOPe database with SCOP ID to extract superfamily ID (iv) search SCOPe database with superfamily ID to extract family ID (v) extract PDB ID per family ID. The figure was generated using Microsoft Power point 365 suite.

A total of 74 DUF proteins (Table 1), 6218 PDB IDs (Table S1), and 70342 Cysteine residues were reported. The maximum number of Cysteine residues belonged to the Electron Transport Chain (n=29638), followed by Glutathione metabolism (n=26656), Fe-S Cluster Biogenesis (n=24826), Fatty Acid Synthesis (n=9229), Photosynthesis (n=1145), Kreb's Cycle (n=27), and Pentose Phosphate Pathway (n=18). The biochemical pathway information was curated from the SUPFAM database[4].

There were eight cell organelles (cytoplasm, mitochondria, thylakoid membrane, periplasm, ROD Outer Segment (Eye), chloroplast, cell membrane, and nucleus) reported in the database. The cell organelle location information was curated from the PDB database.

Table 1: List of DUF IDs and biochemical pathway names, curated from the SUPFAM Database

S. N	DUF ID	Biochemical Pathways
1	DUF459	Electron Transport Chain

2	DUF460	Electron Transport Chain
3	DUF461	Electron Transport Chain
4	DUF462	Electron Transport Chain
5	DUF463	Electron Transport Chain
6	DUF464	Electron Transport Chain
7	DUF465	Electron Transport Chain
8	DUF466	Electron Transport Chain
9	DUF467	Electron Transport Chain
10	DUF468	Electron Transport Chain
11	DUF455	Electron Transport Chain, Fe-S-Cluster Biogenesis
12	DUF1863	Electron Transport Chain
13	DUF3050	Electron Transport Chain
14	DUF3291	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
15	DUF1636	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
16	DUF4405	Electron Transport Chain
17	DUF3182	Fatty Acid Synthesis and Glutathione Metabolism
18	DUF2764	Electron Transport Chain
19	DUF1175	Fatty Acid Synthesis
20	DUF521	Krebs Cycle and Fe-S-Cluster Biogenesis
21	DUF2298	Electron Transport Chain
22	DUF1015	Electron Transport Chain
23	DUF4173	Photosynthesis
24	DUF137	Electron Transport Chain
25	DUF2652	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
26	DUF1691	Electron Transport Chain
27	DUF3611	Electron Transport Chain
28	DUF899	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
29	DUF3088	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis

30	DUF1574	Electron Transport Chain
31	DUF4343	Fatty Acid Synthesis and Glutathione Metabolism
32	DUF1287	Fatty Acid Synthesis
33	DUF2214	Electron Transport Chain
34	DUF2272	Fatty Acid Synthesis
35	DUF4300	Fatty Acid Synthesis
36	DUF1624	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
37	DUF2919	Electron Transport Chain
38	DUF2231	Electron Transport Chain
39	DUF4142	Electron Transport Chain, Fe-S-Cluster Biogenesis
40	DUF2165	Electron Transport Chain
41	DUF1352	Electron Transport Chain
42	DUF3483	Electron Transport Chain
43	DUF4344	Electron Transport Chain
44	DUF4188	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
45	DUF1111	Electron Transport Chain
46	DUF2338	Pentose phosphate pathway
47	DUF2339	Pentose phosphate pathway
48	DUF2340	Pentose phosphate pathway
49	DUF2340	Electron Transport Chain
50	DUF420	Complex IV of Electron Transport Chain
51	DUF3581	Fatty Acid Biosynthesis
52	DUF4333	Complex III of Electron Transport Chain
53	DUF2387	Electron Transport Chain
54	UPF0203	Complex III of Electron Transport Chain
55	DUF1120	Complex III of Electron Transport Chain
56	DUF1298	Fatty Acid Synthesis
57	UPF0547	Electron Transport Chain

58	DUF3613	Complex III of Electron Transport Chain
59	DUF2872	Electron Transport Chain
60	DUF1451	Electron Transport Chain
61	DUF4523	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
62	DUF2414	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
63	DUF2414	Photosynthesis
64	DUF4174	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
65	DUF4350	Electron Transport Chain
66	DUF1450	Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis
67	DUF973	Photosynthesis
68	DUF1610	Electron Transport Chain
69	DUF1440	Electron Transport Chain
70	UPF0180	Electron Transport Chain
71	DUF2194	Electron Transport Chain
72	DUF2296	Electron Transport Chain
73	DUF779	Fe-S-Cluster Biogenesis
74	DUF2827	Uronic Acid Pathway

2. Computation of Cysteine protein microenvironment (MENV) embedded in the DUF proteins:

The protein microenvironments (MENV) around 70342 Cysteine thiol groups embedded in DUF proteins were computed using crystal structures. The cysteine protein microenvironment (three-dimensional spatial arrangement around Cysteine amino acid) was quantified as the summation of the hydrophobic/hydrophilic contributions (estimated by Rekker's fragmental constants) ("Rekker, R. F. The Effect of Intramolecular Hydrophobic Bonding on Partition Experiments; 1967; Vol. 86," n.d.) from the protein structure encompassed within the first contact shell (approximately 4.5 Å radius) (Bandyopadhyay and Mehler 2008) (Figure 3). The weighted summation of the Rekker's fragments constants within the first contact shell of the Cysteine amino acid was termed Hpy^A (Eq. 1)(Bandyopadhyay and Mehler 2008). Similarly, Hpy^s was expressed as the weighted summation of the Rekker's fragmental constants of solvent molecules within the first contact shell. Hpy^s was derived from Molecular Dynamics Simulations with TIP3P water models (Jorgensen et al. 1983). Summation of Hpy^A and Hpy^s, weighted by the buried fraction (ζ) was reported as total Hpy (THpy) (Eq. 3)(Bandyopadhyay and Mehler 2008). The final property descriptor, the relative hydrophobicity, rHpy, was obtained by normalizing THpy by Hpy^s. The rHpy quantity is an intrinsic property and is independent of the size of an amino acid.

Although the MENV calculation needed protein cartesian coordinates from any source, such as X-ray crystallography, NMR, SAXS, molecular modeling, etc., in this database, we selected only crystallography data. The input to the protein microenvironment, encoded in the FORTRAN language, was a three-dimensional structure and the outputs were (i) buried fraction and (ii) rHpy (Bandyopadhyay and Mehler 2008). The buried fraction was defined as the fraction of the surface of the functional group embedded within the protein (Pascual-ahuir, Silla, and Tuñon 1994); that ranges from zero to one; zero buried fraction indicates the thiol group is completely exposed to the solvent, and vice versa. The upper limit of rHpy was formulated as one indicating the Cysteine thiol group was completely immersed in the aqueous solvent. There was no lower limit of rHpy; slight variations in the lower limits were observed depending on the dataset, for example, -0.3 (Bandyopadhyay and Mehler 2008) to -0.4 (Bhatnagar, Apostol, and Bandyopadhyay 2016b). The buried fraction and rHpy together constituted protein microenvironment space around a Cysteine thiol group.

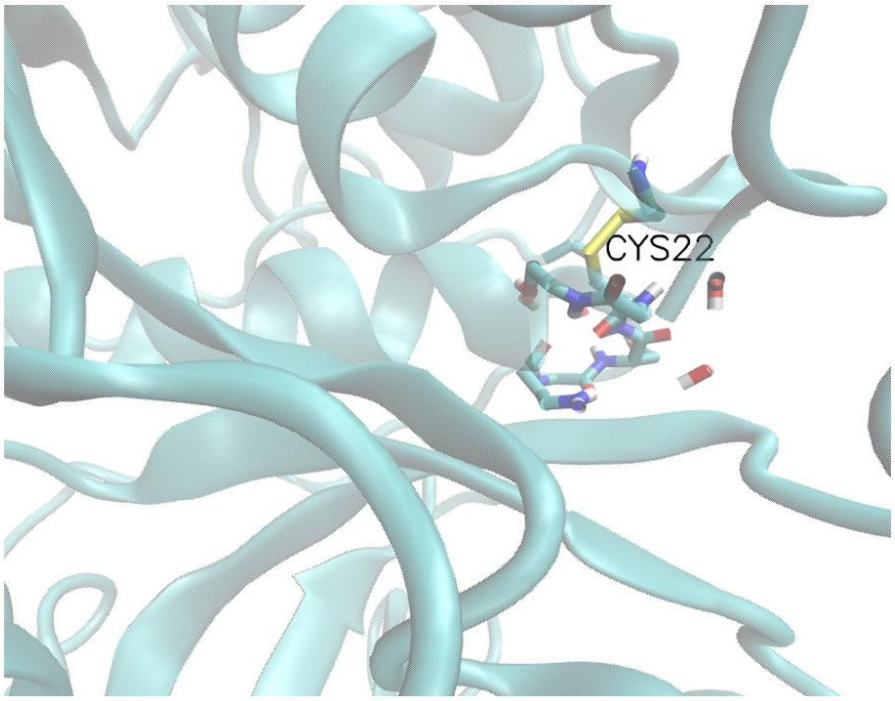


Figure 3: Depiction of Cysteine (Cys^{22}) protein microenvironment (4.5 Å radius) (from PDB ID:8PCH), in stick representation. Cysteine thiol group is depicted in yellow. The protein background is shown in cartoon representation. The figure was generated using VMD (Humphrey, Dalke, and Schulten 1996a) and Microsoft power point 365 suite.

3. Prediction of Cysteine post-translational modifications in the DUF proteins:

Cysteine post-translational modifications were predicted using the prediction server, DeepCys, based on a deep neural network and trained on protein crystal structures, developed by our group (Nallapareddy et al. 2021). Inputs to DeepCys were - the PDB ID of the DUF protein, chain ID, and the Cysteine residue

number. DeepCys, being a multiple Cysteine function prediction tool, outputs probabilities of four Cysteine post-translational modifications, namely, Disulphide, S-Sulphenylation, Thioether, and Metal-binding.

4. Clustering the protein microenvironment space around the Cysteine thiol group:

The protein microenvironment space around the Cysteine thiol group was clustered using agglomerative hierarchical clustering ("Robert C.Tryon and Daniel E. Bailey. Cluster Analysis. New York McGraw-Hill, 1970. Pp. Xvii," n.d.) implemented in a Python script and enabled with Scikit-Learn (1.1.1) and Matplotlib (3.5.3) libraries. Protein microenvironment space was divided into smaller bins of equal spacing [buried fraction = 0.1, rHpy = 0.1]. The clustering was done by using the subsampling method where only 10% subsample has been employed in the Python code. The agglomerative hierarchical clustering initially considers each bin as a single cluster. The final clusters were defined based on the proximity of a data point (buried fraction, rHpy) to its nearest cluster center. The agglomerative hierarchical clustering resulted in three clusters.

Results:

Prediction of Cysteine post-translational modifications (PTMs) in CysDuF database:

The DUF proteins curated in the CysDuF database were related to experimentally solved structures; however, the protein functions were not annotated. Four Cysteine functions were predicted, here, using the in-house Cysteine function prediction server DeepCys, based on protein structures. By design, DeepCys can predict any one of the four Cysteine functions for a given protein, namely, disulfide, thioether, S-sulfenylation, or metal-binding. Out of 70342 cysteines in the DUF database, the majority were predicted as, thioether or metal-binding (Table 2). To note, the maximum number of Cysteine residues in this database belonged to the Electron transport chain (ETC). In Complex III of the ETC, thioether modification was reported (Daltrop et al. 2002) (Barker and Ferguson 1999). Cysteine thioether modification was also reported In the Glutathione metabolism (Townsend, Lushchak, and Cooper 2014), Fatty Acid Biosynthesis (Santiago-Tirado and Doering 2016), Kreb's Cycle (Valcarcel-Jimenez and Frezza 2023), and Pentose phosphate pathway (Marcus et al. 2003). In Complex IV of ETC, the Cysteine residues from DUF proteins were mainly predicted as two modifications, metal binding and disulfide (Nallapareddy et al. 2021). The limitation of this structure-based Cys function prediction method, DeepCys, was that it could not predict other Cysteine modifications, for example, cysteine glutathionylation, nitrosylation, or persufidation in Complex IV of ETC (Martí, Jiménez, and Sevilla 2020). We have compared our predicted results with the ground truth (experimental results) reported in the respective PDB header files.

Validation of the predicted post-translational modifications (PTMs) based on the experimental observations:

Predicted Cysteine PTMs were validated with the experimental findings reported in the respective PDB header files. There were only 18626 experimental PTMs reported for 70302 Cysteine in DUF proteins (Table 2).

Table 2: Validation of the predicted post-translational modifications of DUF Cysteines (using DeepCys) with the experimental PTMs (from PDB header files):

Cysteine PTM	Number of Experimental Cysteine PTM	Number of PTMs predicted using DeepCys	Precision	Recall	F1-score
Thioether	1853	9154	0.19	0.94	0.31
Metal-Binding	5615	2774	0.77	0.38	0.51
Disulphide	11116	5605	0.91	0.46	0.61
Glutathionylation	41	0	0	0	0
S-Sulphenylation	0	1093	0	0	0
Total	18626	18626			
Macroavg			0.37	0.35	0.28
Weighted average			0.79	0.48	0.55

Hence, the validation was restricted to 18626 Cysteines only. Four different experimental Cysteine PTMs were reported, namely, disulfide, metal-binding, thioether, and glutathionylation. Whereas, the Cysteine PTM prediction software, DeepCys, predicted disulfide, metal-binding, thioether, and sulfenylation, only. The prediction was evaluated using the confusion matrix (Figure 4). This matrix was generated from the experimental and predicted Cysteine PTM numbers (Table 2). Several evaluation metrics were used to validate the prediction performances, namely, precision, recall, F1-score, accuracy, macro average (macroavg), and weighted average (Supplementary, Eq. 1-5). The prediction performances of different Cysteine PTMs varied (Table 2). The overall accuracy of prediction was 0.79. The prediction of true positives over false positives (precision) was the best for disulfide and metal-binding. Whereas, the prediction of true positives over false negatives was the best predicted for thioether. To note, S-glutathionylation has no predictions reported and S-sulfenyaltion has no experiments reported.

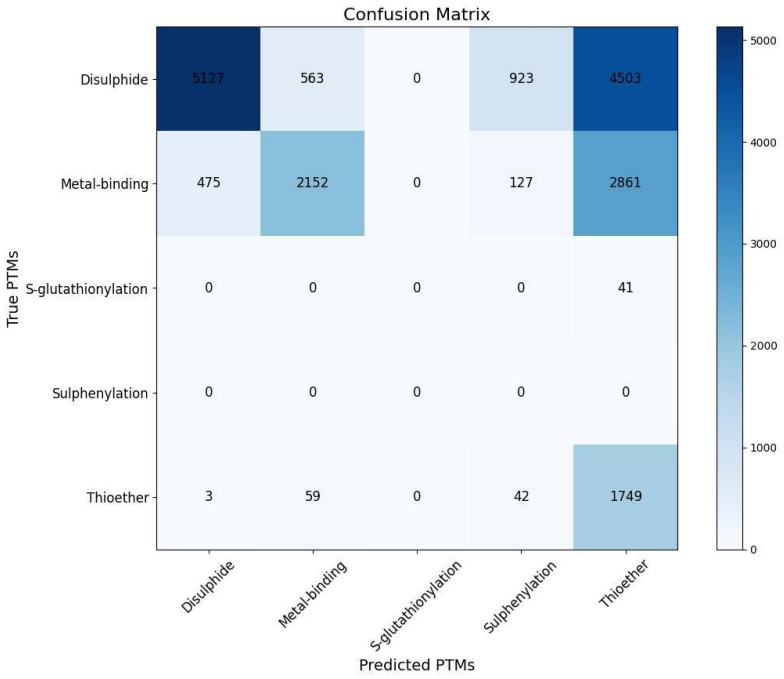


Figure 4: Confusion matrix to validate the predicted Cysteine PTMs (using DeepCys software) with the experimental (PDB header file) observations. The heatmap indicates the range of Cysteine numbers.

Diverse protein microenvironments around Cys residues in the CysDuF database:

From our earlier investigations, we observed that Cys residues were embedded in three different types of protein microenvironments, buried hydrophobic, intermediate, and exposed hydrophilic (Bhatnagar and Bandyopadhyay 2018). Here we explored two questions, i) whether diversity in the protein microenvironment existed around Cysteine in this database and ii) if it existed whether there were preferential Cysteine protein microenvironments towards different post-translational modifications, pathways, and diseases. The first question was addressed by clustering the protein microenvironment (MENV) space around all the Cysteine residues in the database. Two parameters, buried fraction (BF) and microenvironment property descriptor (rHpy), were used to cluster MENV space, using agglomerative clustering (Figure 5). The largest cluster denoted that the Cysteine MENV was deeply buried in the protein core (high average BF value of 0.98) and significantly hydrophobic (low average rHpy value of 0.08) (Table 3), hence, named as "buried-hydrophobic". To note, according to the definition of buried fraction described in the method section, BF value of one indicated that the residue was fully buried inside the protein, and BF of zero indicated full exposure of the residue to the solvent. Similarly, according to the definition, rHpy of 1 indicated that the residue microenvironment was fully governed by solvent water molecules; thus, the microenvironment was completely hydrophilic. By definition, there was no lower limit of rHpy, that denoted the hydrophobicity of the residue microenvironment. More or less, this lower limit of rHpy value was decided by the dataset, for example, -0.3, in one dataset (Bandyopadhyay and Mehler 2008) and -0.4 in another (Bhatnagar and Bandyopadhyay 2018). The second largest cluster exhibited a relatively high average buried fraction (0.81) but somehow moderate average rHpy value (0.38), indicating that the Cysteine residue despite being buried inside the protein, has a relatively hydrophilic protein microenvironment around it. This cluster appeared to be buried in nature yet hydrophilic, hence termed

as, “buried-hydrophilic”. In one of our previous studies, a similar microenvironment cluster was reported that was more exposed (average BF, 0.77) to the solvent than the “buried-hydrophilic” cluster and also more hydrophilic (0.4); hence, it was classified as an “intermediate cluster” (Bhatnagar and Bandyopadhyay 2018). The least populated cluster was “exposed-hydrophilic” where the average BF of the Cys was 0.39 and the average rHpy was 0.68.

Table 3: Statistics (average value) of Cysteine microenvironment clusters. The standard deviation (**σ**) is given within parentheses

Cluster Type	**Average Buried Fraction (σ)**	**Average rHpy (σ)**	**Average distance to centroid (Å)**	**No of Cysteines in each cluster**	**No of PDB IDs in each cluster**
Buried Hydrophobic	0.97 (0.03)	0.08 (0.12)	0.11	4517	2207
Buried Hydrophilic	0.81(0.12)	0.37(0.14)	0.15	2160	1333
Exposed Hydrophilic	0.39 (0.12)	0.67 (0.09)	0.14	366	294

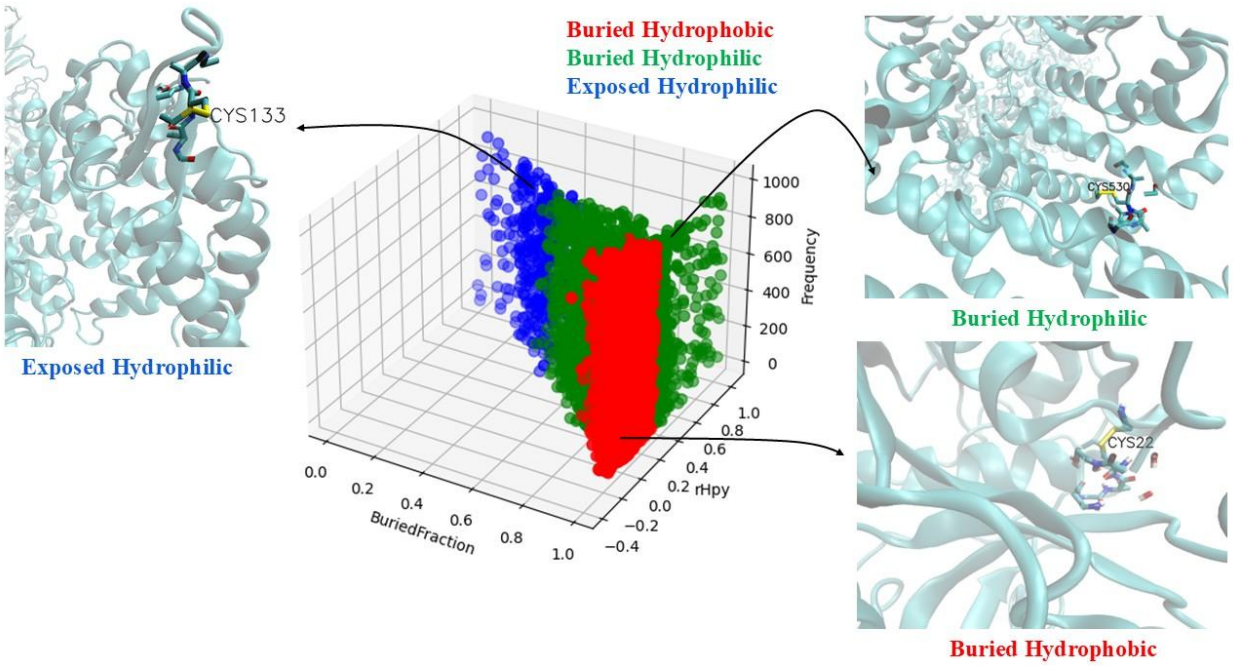


Figure 5: Distribution of Cysteine protein microenvironments, from DUF proteins, in three clusters, Buried Hydrophobic (Red), Buried Hydrophilic (Green), and Exposed Hydrophilic (Blue). The X-axis represents the

Buried Fraction; the Y-axis, rHpy; and the Z-axis, populations of Cysteine. Three insets show the relative position of the Cysteine residue in three different protein microenvironments, Buried Hydrophobic (PDB ID: 8PCH), Buried Hydrophilic (PDB ID:7XAZ), and Exposed Hydrophilic (PDB ID:7UON). The figure was generated using i) Matplotlib (Hunter 2007), ii) VMD (Humphrey, Dalke, and Schulten 1996a) and iii) Microsoft power point 365 suite

Distribution of Cysteine post-translational modifications in different microenvironments:

Here, we investigated the second question, whether the Cysteine post-translation modifications exhibited preferences towards different Cysteine protein microenvironments. To answer this question, the normalized populations of different post-translational modifications across different microenvironment clusters were compared (Table 4). The cluster population (number of Cysteines in each cluster) was normalized by the number of Cysteines, per post-translational modification. The overall trend showed that all four modifications were maximally populated in the "buried-hydrophobic" cluster, followed by "buried-hydrophilic" and "exposed-hydrophilic". This agreed with the Cysteine microenvironment distribution reported above. The Cysteine was mostly populated in the "buried-hydrophobic" cluster, matched with the hydrophobicity scale, reported elsewhere, where Cysteine exhibited the largest hydrophobic value (Bandyopadhyay and Mehler 2008). This observation indicated that the predicted Cysteine post-translational modifications, in general, followed the same trend as that of the Cysteine residue.

Table 4: Normalized Cysteine populations of different post-translation modifications across microenvironment clusters

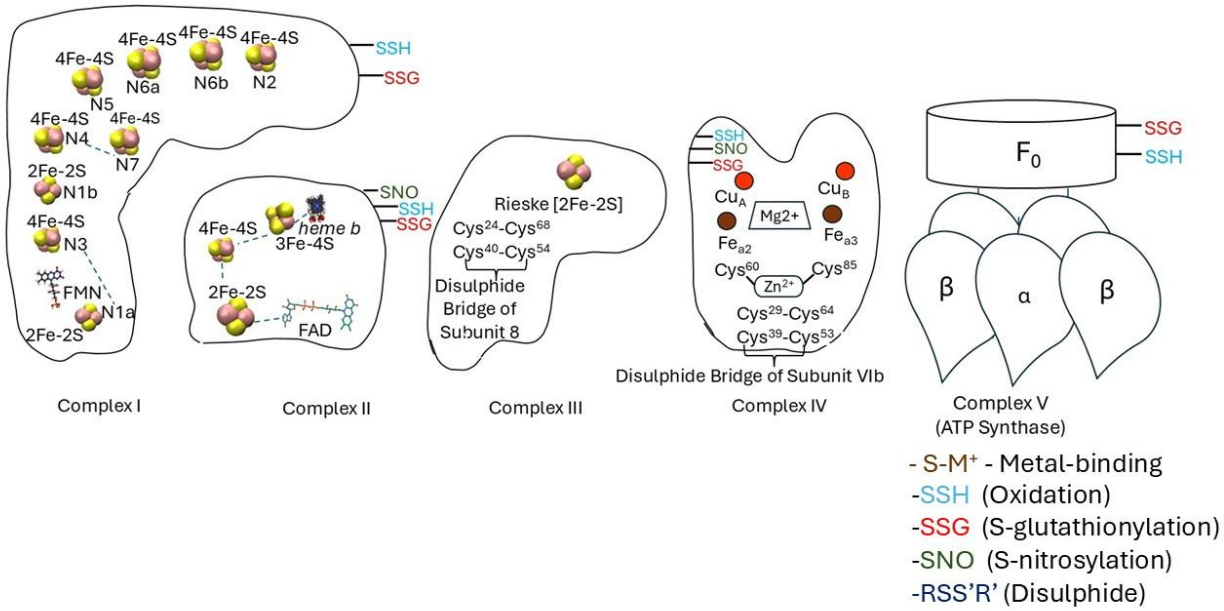
Cluster Type	Disulphide	Metal-binding	Thioether	S-Sulphenylation
Buried hydrophobic	0.62	0.66	0.63	0.62
Buried hydrophilic	0.35	0.29	0.30	0.28
Exposed hydrophilic	0.02	0.04	0.05	0.08

Preferences of Cysteine post-translational modifications and their microenvironments towards different biological pathways:

Cysteine is a dominant catalytic residue in all the biological pathways, mentioned in this database. We explored whether cysteine post-translational modifications and their microenvironments exhibited any preferences for different biological pathways. To investigate this question, the normalized populations of different Cysteine microenvironment clusters were compared across the proteins from different biological pathways (Table 5). The microenvironment cluster population (number of Cysteines in each cluster) was normalized by the number of Cysteines, per biological pathway. The Cysteine microenvironment was maximally populated in the "buried-hydrophobic" region in all the pathways, agreeing with the hydrophobic nature of the Cysteine residue. However, in the photosynthetic pathway and to some extent in Kreb's cycle, the maximum Cysteine microenvironment was populated in the "buried-hydrophilic" region. There were six cysteines from Kreb's cycle embedded in buried-hydrophilic microenvironments (Table S2), and sixty-eight from photosynthesis, also embedded in the same microenvironment (Table S3).

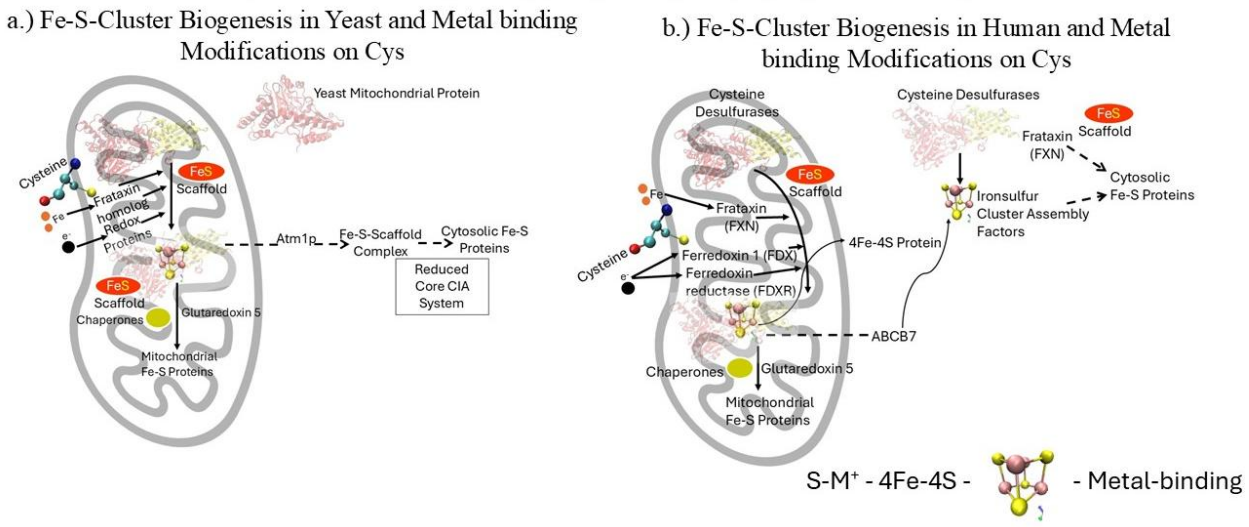
In Kreb's cycle, all six functional Cysteine residues were from the Aconitase enzyme. The predicted post-translational modifications (PTMs) were thioether, metal-binding, and sulfenylation. The reported PTMs were metal binding (as Fe-S cluster), and oxidation of the sulfhydryl group (Figure 6). Thus, the predicted and the reported PTMs are fairly similar, indicating the reliability of the database and the prediction tool (DeepCys). To note, it has been reported that the Fe-S clusters in Aconitase have a hydrophilic microenvironment created by the polar groups (Robbins and Stout 1989) that matched with our current observations – functional Cysteines from aconitase were embedded in buried -hydrophilic microenvironment. Similarly, in photosynthesis, the functional Cysteines mostly belonged to photosynthetic reaction center II proteins (like protein D1, D2, CP43, CP47, cytochrome C subunit), Cytochrome c-550, etc. A significant percentage of these Cysteines were embedded in the buried-hydrophilic microenvironments. To note, most of the photosystem II proteins were membrane proteins and not globular proteins. However, the MENV computation was designed only for globular proteins, where the surface of a protein was exposed to water molecules, in contrast to membrane proteins, exposed to the lipid bilayer. The predicted PTMs were mainly thioether and metal bindings.

a.) Electron Transport Chain(Hayashi and Stuchebrukhov 2010; Sun et al. 2005; Iwata et al. 1998; Tsukihara et al. 1996)

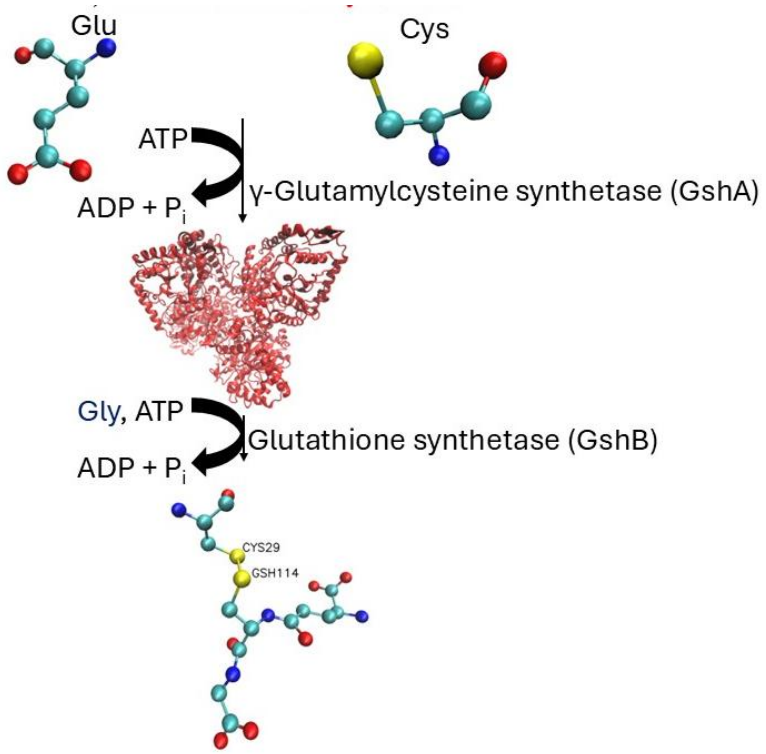


b.) Fe-S-Cluster Biogenesis(Rouault and Tong 2008)

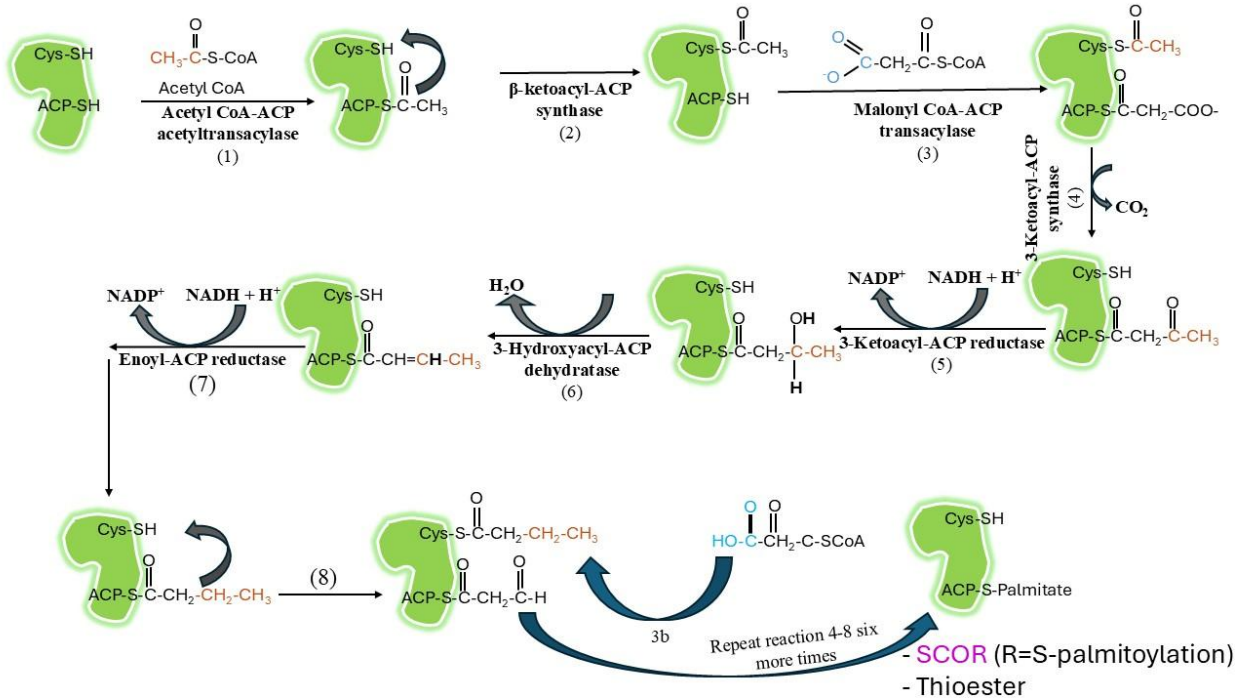
Left panel : Yeast ; Right Panel : Human



c.) Glutathione Biosynthesis

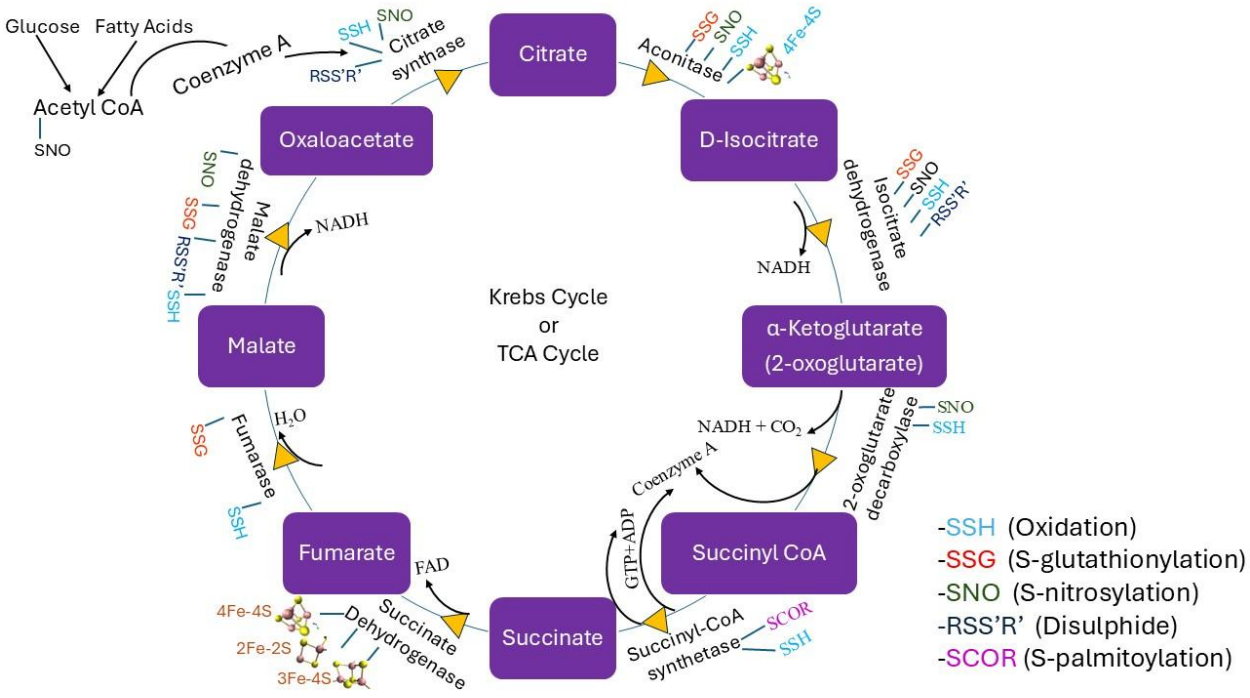


d.) Fatty acid Biosynthesis



- SCOR (R=S-palmitoylation)
- Thioester

e.) Krebs cycle(Martí, Jiménez, and Sevilla 2020)



-SSH (Oxidation)
-SSG (S-glutathionylation)
-SNO (S-nitrosylation)
-RSS'R' (Disulphide)
-SCOR (S-palmitoylation)

f.) Pentose phosphate pathway(Ge et al. 2020)

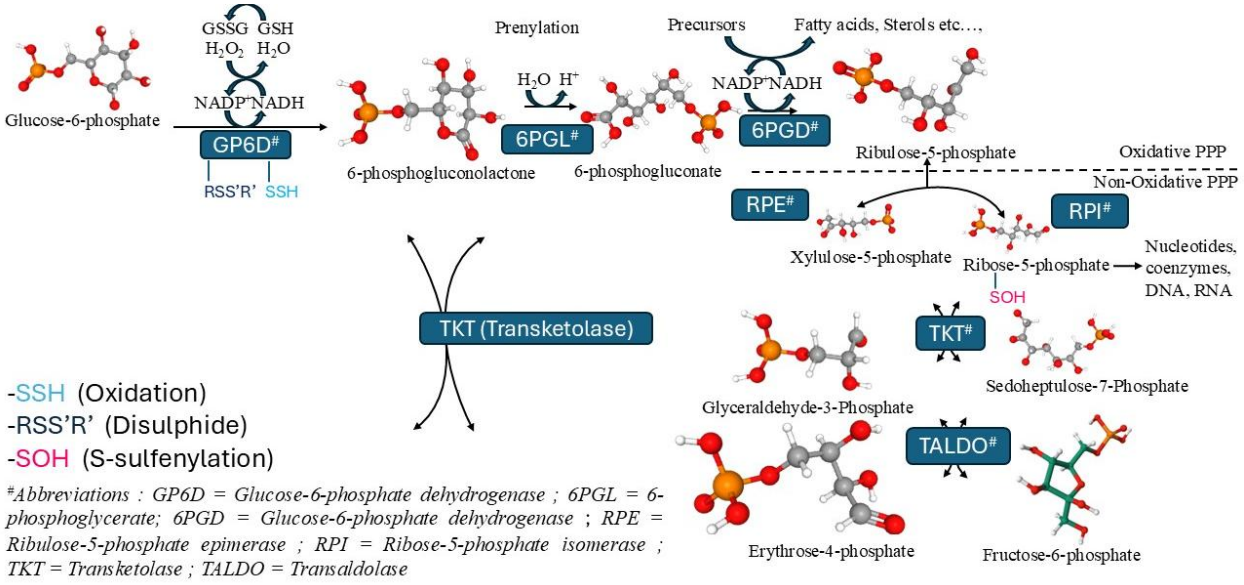


Figure 6: Schematic representations (generated using VMD (Humphrey, Dalke, and Schulten 1996b), PubChem (Kim et al. 2021), and Microsoft Power-Point 365 Suite) of Cysteine post-translational modifications (PTMs) reported in the literature in different pathways

Table 5: Normalized Cysteine populations in different biological pathways across microenvironment clusters

Cluster type	Electron Transport Chain	Glutathione Metabolism	Fe-S-Cluster Biogenesis	Fatty Acid Synthesis	Photosynthesis	Krebs Cycle	Pentose phosphate pathway
Buried hydrophobic	0.60	0.74	0.60	0.73	0.42	0.57	0.5
Buried hydrophilic	0.33	0.22	0.33	0.24	**0.54**	**0.42**	0.25
Exposed hydrophilic	0.06	0.03	0.06	0.02	0.03	0	0.25

Preferences of Cysteine post-translational modifications and their microenvironments towards different taxonomic kingdoms:

The DUF proteins were classified into four different taxonomic kingdoms, namely Bacteria, Archaebacteria, Viruses, and Eukaryotes, as per NCBI Taxonomy. (Federhen 2012). A total of 607 organisms were reported

in this database. Simple trees were constructed to represent the taxonomic variations (Figure 7 and Figure S2). The highest number of species was observed for Bacteria, pathogenic and non-pathogenic (n=342). The disease-causing bacterial species, classified according to their taxonomy were represented by a simple tree (Figure 8). The complete list of the species name and corresponding diseases were shown (Table S4). The literature report also suggested that most of the DUF proteins belonged to kingdom bacteria [3,](Goodacre, Gerloff, and Uetz 2014). The second largest kingdom in this database was Eukaryotes. The DUF proteins from Kingdom Virus (n=25), were reported for the first time. All the viruses reported were disease-causing (Table S4). Here, we explored whether cysteine post-translational modifications and their microenvironments exhibited any preferences for different taxonomic kingdoms. To investigate this question, the normalized populations of different Cysteine microenvironment clusters were compared across the proteins from different kingdoms (Table 6). The microenvironment cluster population (number of Cysteines in each cluster) was normalized by the number of Cysteines, per kingdom. Most of the Cysteine microenvironments were maximally populated in the ”buried-hydrophobic” clusters as per the hydrophobic nature of the Cysteine. However, a significant population of Cysteine microenvironment was observed in the “buried-hydrophilic” region from proteins of Archaebacteria and bacteria kingdoms. This could presumably be attributed to the extremophile nature of bacteria (n=139) out of 146 Cysteine in the same cluster. In general, the “exposed-hydrophilic” microenvironment was least populated around Cysteine residues. However, for viruses, the Cysteine microenvironment population was significant in that cluster, compared to those in other kingdoms. This observation plausibly indicated the possible exposure of the catalytic Cysteine residues on the viral protein surfaces.

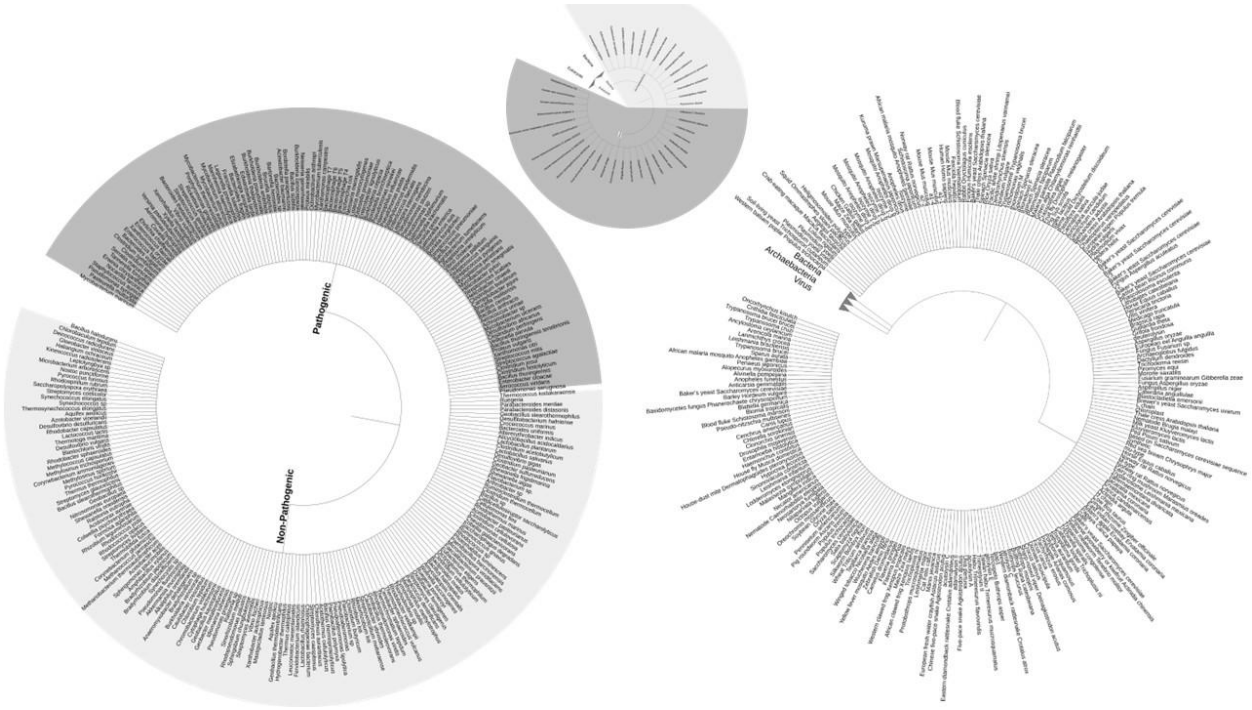


Figure 7: Simple tree representing the species in this study based on taxonomy: bacteria (left), virus and archaebacteria (middle), and eukaryotes (right). The figure was generated using Interactive Tree of Life (ITOL) version 7 (Letunic and Bork 2024)

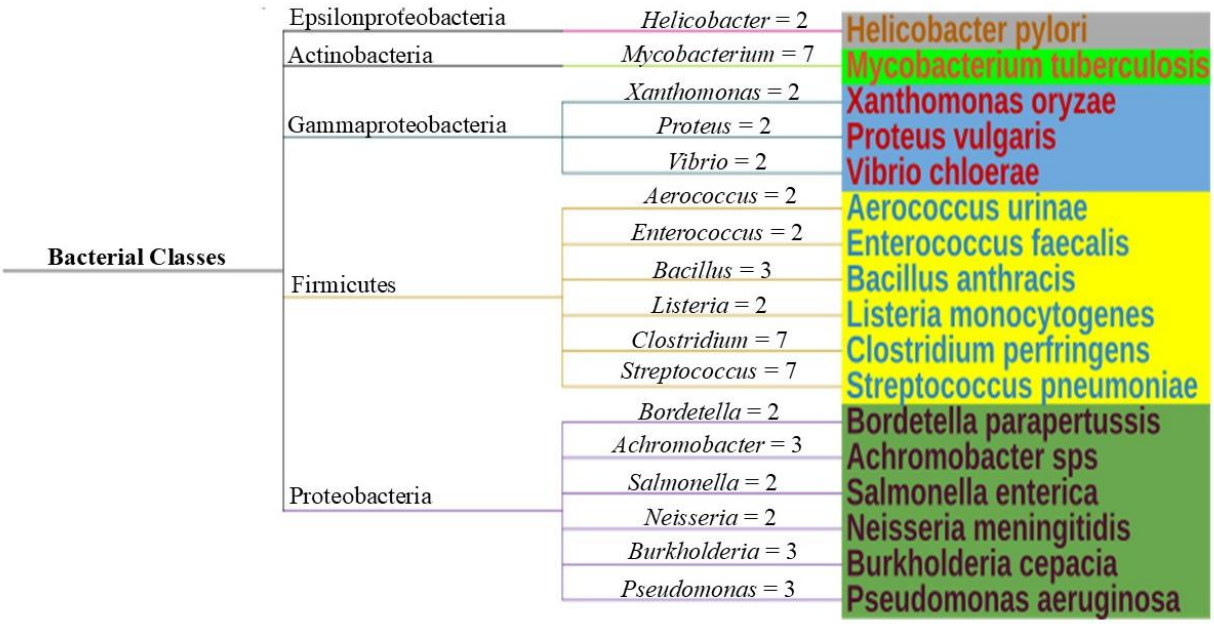


Figure 8. Simple tree for disease-causing bacteria, classified according to their taxonomy. The number of species per genera is shown on the connecting branch. One example per genera is shown for clarity. The figure was generated using Interactive Tree of Life (ITOL) version 7

Table 6: Normalized Cysteine populations in different kingdoms across microenvironment clusters. Significant numbers are reported in bold.

Domain Kingdom	Eukaryotes	Archaebacteria	Viruses	Bacteria
Buried hydrophobic	0.66	0.53	0.55	0.53
Buried hydrophilic	0.29	**0.40**	0.31	**0.38**
Exposed hydrophilic	0.03	0.06	**0.13**	0.08

Preferences of Cysteine post-translational modifications and their microenvironments towards different diseases:

There were twenty diseases reported in CysDuF database caused by 156 different species (Figure 9). Most of those were bacterial species (n=101). The full list of pathogens and the diseases caused by those are reported (Table S4).

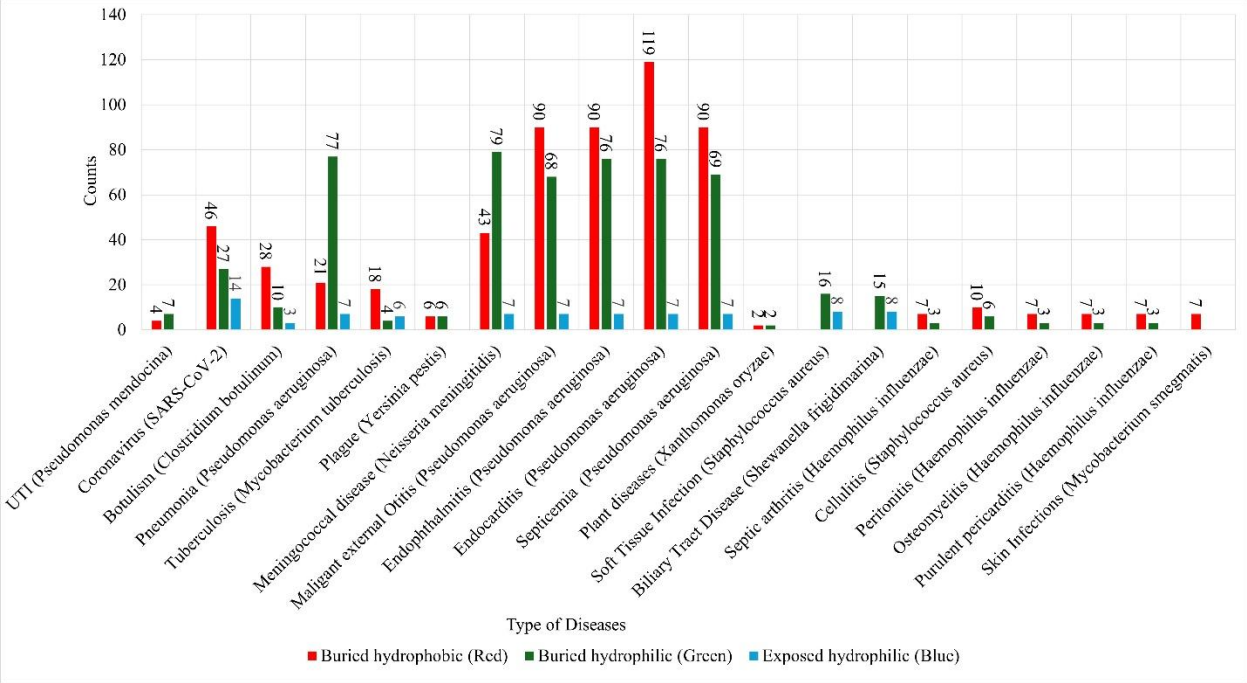


Figure 9: Counts of functional Cysteines across twenty different diseases, categorized according to protein microenvironments. The figure is generated using Microsoft Excel 365 suite

One hundred and forty-two Cysteine residues were present in the DUF proteins belonging to disease-causing bacterial species. Those 142 Cysteine residues were classified, into thirteen bacterial infections, categorized based on anatomy (organs) (Figure 10).

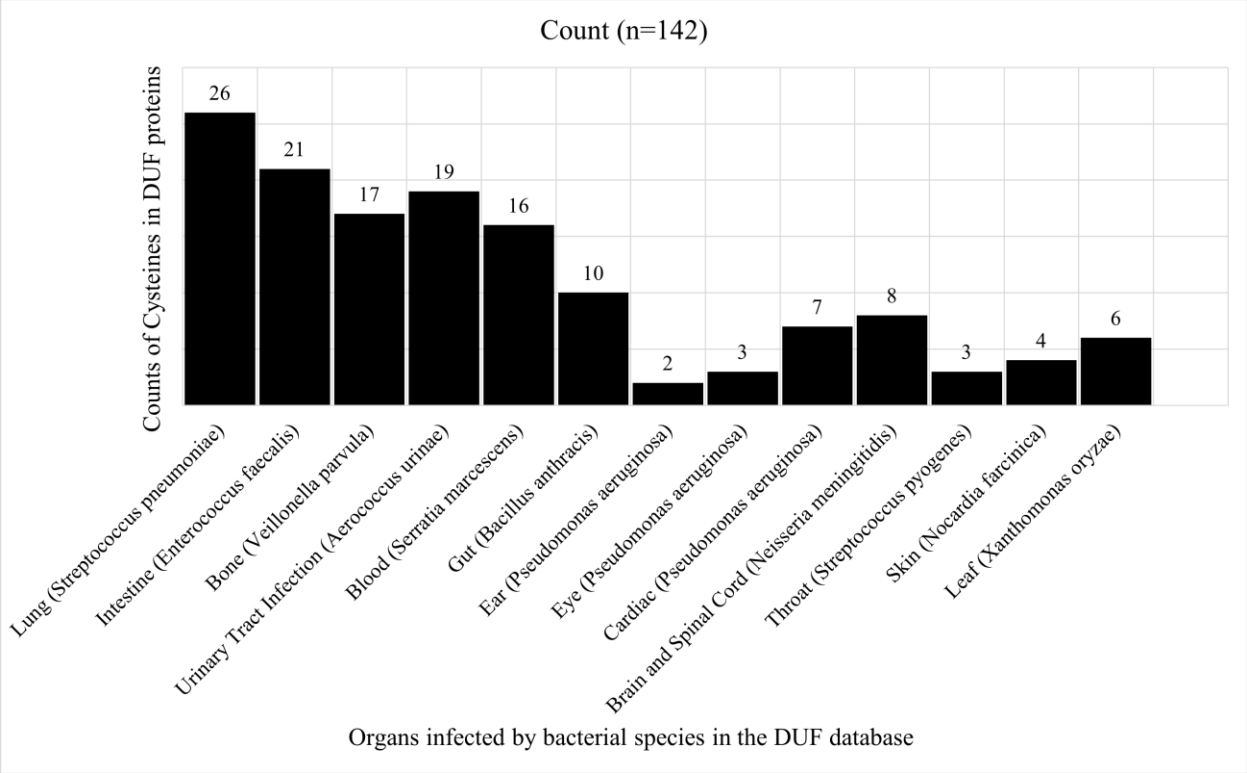


Figure 10: Disease-causing bacteria infecting different organs, categories based on anatomy. Counts of Cysteine residues present in DUF proteins per disease category are shown. The figure was generated using Microsoft Excel 365 suite.

DUF proteins involved in viral diseases (n=10) were classified as Animal inherited diseases specifically infecting human (Table S4). The DUF proteins related to SARS-COV-2 virus causing lung diseases were reported for the first time, in this database. A few fungal diseases (n=8) associated with DUF proteins were reported those mainly invade plants. The parasitic (worm) infections (n=14), were caused by Liver Fluke (n=5), Hookworm (n=2), and parasitic worm (n=7) (Table S4). The protozoan diseases (n=15) reported in this DUF database were mostly animal-inherited (n=13). Two human protozoan diseases were reported causing Gastric, by *Entamoeba histolytica* (n=1) and Sexually Transmitted Diseases/Urinary Tract Infections, caused by *Trichomonas vaginalis* (n=1). There were eight plant diseases (n=8) reported here caused by bacteria and fungi.

We explored whether cysteine post-translational modifications and their microenvironments exhibited any preferences toward diseases. To investigate this question, the normalized populations of different Cysteine microenvironment clusters were compared across the proteins from different diseases (Table S5). The microenvironment cluster population (number of Cysteines in each cluster) was normalized by the number of Cysteines, per disease. Most of the Cysteine microenvironments were maximally populated in ”buried hydrophobic” cluster, as per the hydrophobic nature of the Cysteine. Some outliers were observed, where functional cysteines from disease-causing viruses and bacteria (namely, Coronavirus, *Clostridium botulinum*, *Mycobacterium tuberculosis, Shewanella frigidimarina)* were embedded in the exposed-hydrophilic microenvironment (Figure 9). The maximum population of Cysteine microenvironments in the “buried-hydrophilic” cluster was observed in bacteria causing pneumonia, soft tissue, and biliary tract infection. In the previous section, we reported that functional Cysteines from Virus

and Bacteria kingdoms were populated in the “exposed or buried hydrophilic” microenvironment (Table 6).

The observation of solvent-exposure of catalytic Cysteine from viruses in DUF proteins was supported by the crystallographic observations: an example, Cys111, catalytic residue from MERS Corona Virus (DUF ID: DUF1175) was exposed on the protein surface and underwent disulfide bond formation with β-mercaptoethanol in the crystal structure (PDB ID: 4R3D); (Ali Dahhas et al. 2023). This Cysteine111 in our database was identified in the exposed-hydrophilic microenvironment, with the predicted S-sulfenylation modification (an oxidized state of the thiol group). The same Cysteine residue was reported to undergo ROS-induced oxidative stress leading to thiol-disulfide disbalance and further oxidation of cysteine, such as sulfenylation (Yang 2022). In the DUF protein (DUF: DUF455) from *Mycobacterium tuberculosis* (tuberculosis causing-bacteria), Cys70 formed a zwitter ionic catalytic triad with His110 and Asp127 and the thiolate acted as a nucleophile, thus the Cysteine required hydrophilic microenvironment, concurring with our observation (PDB:4BGF) (Abuhammad et al. 2013). The presence of thioether bonds in the “exposed hydrophilic” microenvironment, around Cysteines from DUF proteins (DUF: DUF4333) in *Shewanella frigidimarina* causing Soft tissue infection and Biliary Tract diseases were reported in the literature (Bamford, V et al 1999) (PDB:1QO8), (PDB:1QJB).

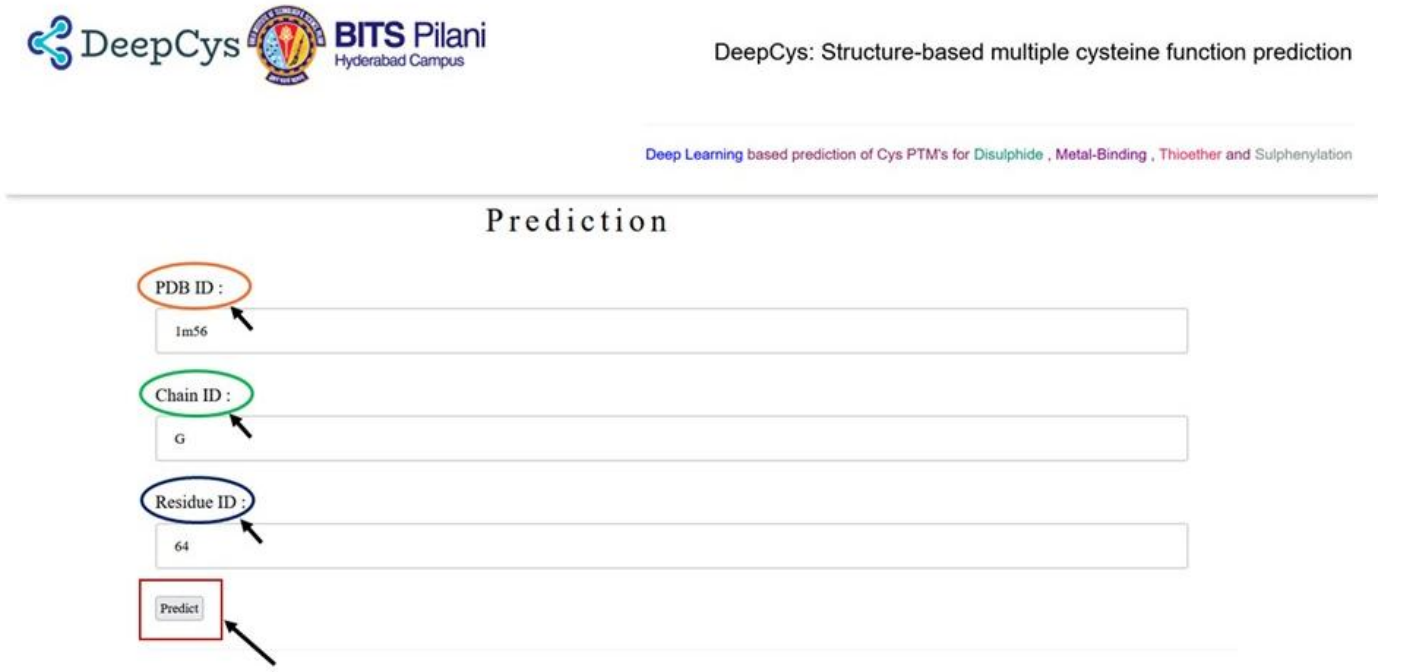
Web Application:

a) DeepCys Web Application:

A user-friendly web application DeepCys (https:/deepcys.bits-hyderabad.ac.in) was built using the Flask web framework. The input, output, and work flow of the web application are shown (Figure 11a). The web application is deployed using the NGINX and http reverse proxy server. The structure-based prediction tool can be accessed by clicking the prediction button on the navigation bar. The web application has a form that requests three inputs corresponding to a cysteine namely, (a) PDB ID of the protein, (B) Chain ID, and (C) Residue of the Cys. Based on these inputs additional parameters were internally computed to predict four probability values and the most probable Cysteine modifications.

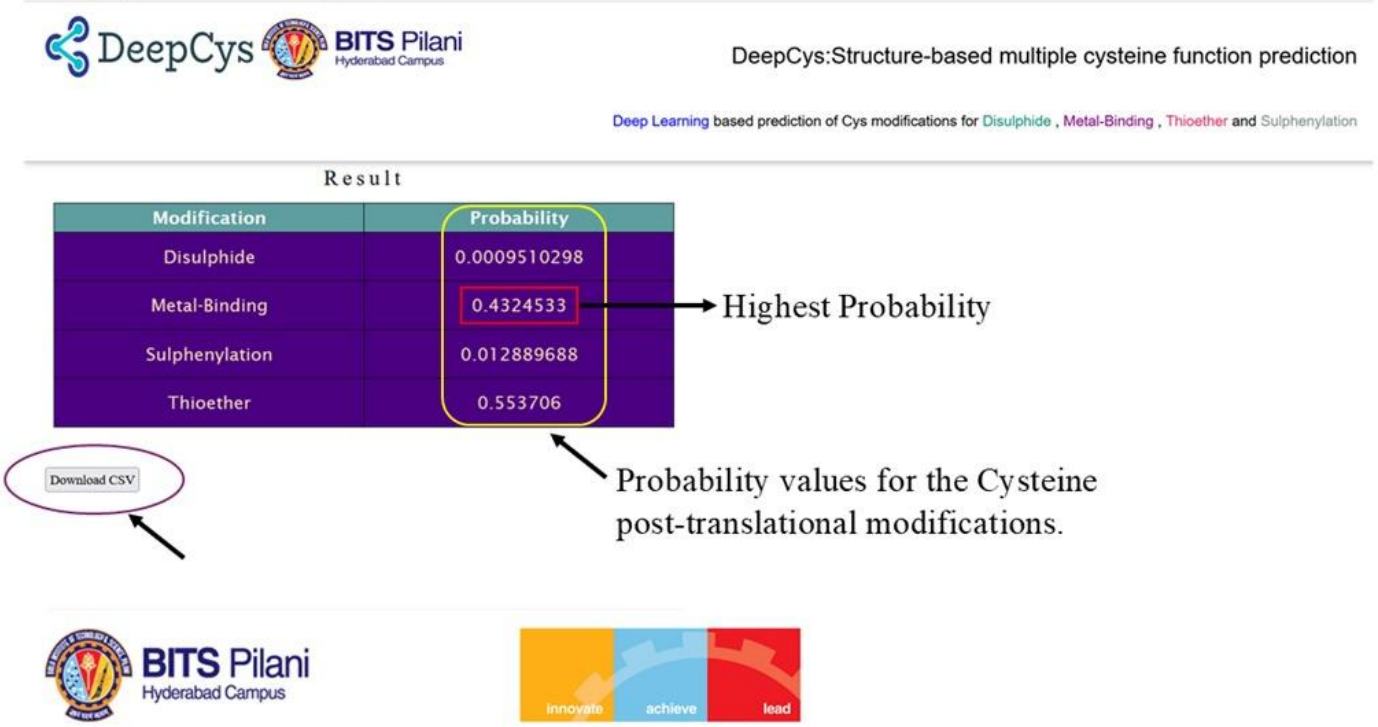
a.) DeepCys -Structure-based prediction tool.

i.) Input for the DeepCys WebServer :-



433

ii.) Output for the DeepCys WebServer :-

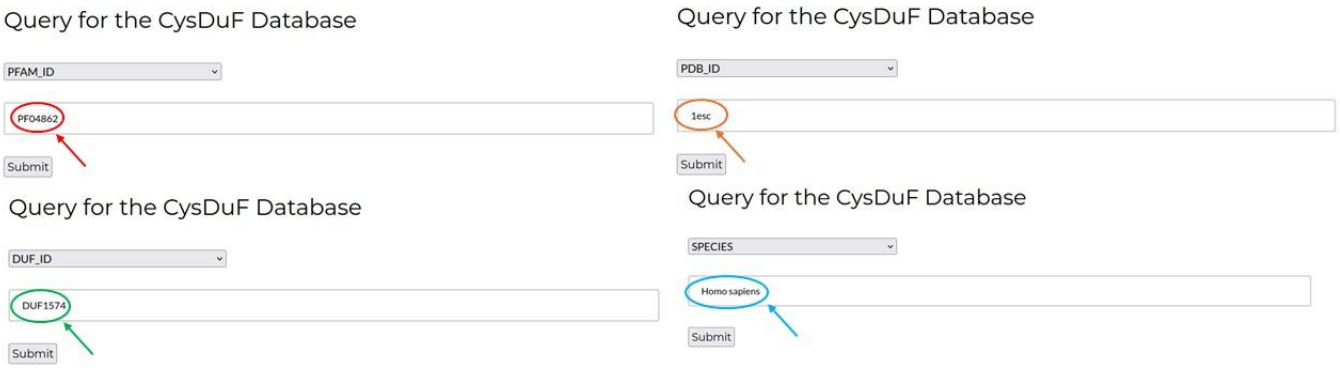


434

435

b.) CysDuF Database.

i.) Input for the CysDuF database webserver :-



ii.) Output for the CysDuF webserver :-

Results

PFAM_ID	DUF_ID	DUF_Name	Name of the DUF	Species	SCOPe Superfamily	SCOPe Family	Pathway	ChainID	PDB_ID	DeepCys_Results	Cys Residue Number	BuriedFraction	rHpy

Figure 11: Web Application for a) DeepCys – Structure-Based Prediction Tool and b) CysDUF Database. The web application screenshots were processed using Microsoft power point 365 suite.

b) DUF Database Web Application:

A user-friendly web application DUF Database (https://cysduf.bits-hyderabad.ac.in/) was built using the Flask web framework. The flowchart for input, output, and the internal storage of information used in this web application is shown (Figure 11b). The web application is deployed using the NGINX and HTTP reverse proxy server. The DUF database application has a form that requests any one of three inputs - PDB ID, DUF ID, or PFAM ID. The results are downloadable in multiple formats, CSV, Text or JSON.

Conclusions:

With the advent of high-throughput structure prediction methods, a large number of protein structures, including DUF proteins were experimentally solved which required functional characterization. The rigor, expense, and time required for experimental characterization, could be reduced by computational approaches. Aim of this study was to characterize and annotate the functions of catalytic Cysteine in DUF proteins, using computational methods. Annotation and characterization of functional Cysteine in DUF proteins were performed on seven biochemical processes, namely, Electron Transport Chain, Glutathione Metabolism, Fe-S-Cluster Biogenesis, Fatty Acid Synthesis, Photosynthesis, Kreb's Cycle, and Pentose phosphate pathway. Cysteine post-translation modifications were predicted using DeepCys software, and the results were validated with the experimental findings reported in the PDB header files. Structure-based protein microenvironment computation was done using software, developed earlier. The sequence, structure, microenvironment, disease, biochemical pathways related to the DUF proteins, and their functional Cysteines were consolidated in a database, CysDUF. This database was the first of its kind that stores and retrieves Cysteine functional annotations in DUF proteins and elucidated on seven different pathways. The detailed elucidation of Cysteine protein microenvironments in all the DUF proteins revealed

that, in general, Cysteine residues were embedded in buried hydrophobic microenvironments. However, in certain viral proteins, functional Cysteine residues were embedded in the exposed and hydrophilic microenvironments. This secondary database would serve as a reference guide to the functional Cysteines of DUF proteins and related information. There is a scope for improvement in the Cysteine function prediction, as the current method predicts only four Cysteine post-translational modifications, this was due to the limited availability of PDB crystal structure data while training the Deep Neural Network. The prediction method could be complemented using the sequence-based method, albeit, less accurate compared to the structure-based method, where sufficient data is available for a larger number of Cysteine post-translational modifications to train Deep Neural Network models. Prediction of a larger number of Cysteine modifications would add further significance to the database.

Acknowledgement:
HD acknowledges the financial support from the Indian Council of Medical Research (ICMR)- Senior Research Fellow (SRF), File No: BMI/11(99)/2022; DB acknowledges the financial support from the Department of Science and Technology (DST), Science and Engineering Research Board (SERB), India, File No: EMR/2017/002953

References:

Abuhammad, Areej, Edward D. Lowe, Michael A. McDonough, Patrick D. Shaw Stewart, Stefan A. Kolek, Edith Sim, and Elspeth F. Garman. 2013. “Structure of Arylamine *N* -Acetyltransferase from *Mycobacterium Tuberculosis* Determined by Cross-Seeding with the Homologous Protein from *M. Marinum* : Triumph over Adversity.” *Acta Crystallographica Section D Biological Crystallography* 69 (8): 1433–46. https://doi.org/10.1107/S0907444913015126.

Ali Dahhas, Mohammed, Hamad M. Alkahtani, Ajamaluddin Malik, Abdulrahman A Almehizia, Ahmed H. Bakheit, Siddique Akber Ansar, Abdullah S. AlAbdulkarim, Lamees S.Alrasheed, and Mohammad A. Alsenaidy. 2023. “Screening and Identification of Potential MERS-CoV Papain-like Protease (PLpro) Inhibitors; Steady-State Kinetic and Molecular Dynamic Studies.” *Saudi Pharmaceutical Journal* 31 (2): 228–44. https://doi.org/10.1016/j.jsps.2022.12.007.

Ayikpoe, Richard S., Lingyang Zhu, Jeff Y. Chen, Chi P. Ting, and Wilfred A. Van Der Donk. 2023. “Macrocyclization and Backbone Rearrangement During RiPP Biosynthesis by a SAM-Dependent Domain-of-Unknown-Function 692.” *ACS Central Science* 9 (5): 1008–18. https://doi.org/10.1021/acscentsci.3c00160.

Bandyopadhyay, Debashree, and Ernest L. Mehler. 2008. “Quantitative Expression of Protein Heterogeneity: Response of Amino Acid Side Chains to Their Local Environment.” *Proteins: Structure, Function, and Bioinformatics* 72 (2): 646–59. https://doi.org/10.1002/prot.21958.

Barker, Paul D, and Stuart J Ferguson. 1999. “Still a Puzzle: Why Is Haem Covalently Attached in c-Type Cytochromes?” *Structure* 7 (12): R281–90. https://doi.org/10.1016/S0969-2126(00)88334-3.

Behrens, Hannah Michaela, and Tobias Spielmann. 2024. “Identification of Domains in Plasmodium Falciparum Proteins of Unknown Function Using DALI Search on AlphaFold Predictions.” *Scientific Reports* 14 (1): 10527. https://doi.org/10.1038/s41598-024-60058-x.

Bhatnagar, Akshay, Marcin I. Apostol, and Debashree Bandyopadhyay. 2016a. “Amino Acid Function Relates to Its Embedded Protein Microenvironment: A Study on Disulfide-bridged Cystine.” *Proteins: Structure, Function, and Bioinformatics* 84 (11): 1576–89. https://doi.org/10.1002/prot.25101.

———. 2016b. “Amino Acid Function Relates to Its Embedded Protein Microenvironment: A Study on Disulfide-bridged Cystine.” *Proteins: Structure, Function, and Bioinformatics* 84 (11): 1576–89. https://doi.org/10.1002/prot.25101.
Bhatnagar, Akshay, and Debashree Bandyopadhyay. 2018. “Characterization of Cysteine Thiol Modifications Based on Protein Microenvironments and Local Secondary Structures.” *Proteins: Structure, Function, and Bioinformatics* 86 (2): 192–209. https://doi.org/10.1002/prot.25424.
Burley, Stephen K, Helen M Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, et al. 2019. “RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy.” *Nucleic Acids Research* 47 (D1): D464–74. https://doi.org/10.1093/nar/gky1004.
Chandonia, John-Marc, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, and Steven E Brenner. 2022. “SCOPe: Improvements to the Structural Classification of Proteins – Extended Database to Facilitate Variant Interpretation and Machine Learning.” *Nucleic Acids Research* 50 (D1): D553–59. https://doi.org/10.1093/nar/gkab1054.
Chen, Kai, Yilin Wang, Xiaoyan Nong, Yichi Zhang, Tang Tang, Yun Chen, Qikun Shen, Changjie Yan, and Bing Lü. 2023. “Characterization and in Silico Analysis of the Domain Unknown Function DUF568-Containing Gene Family in Rice (Oryza Sativa L.).” *BMC Genomics* 24 (1): 544. https://doi.org/10.1186/s12864-023-09654-1.
Daltrop, Oliver, James W. A. Allen, Anthony C. Willis, and Stuart J. Ferguson. 2002. “*In Vitro* Formation of a *c* -Type Cytochrome.” *Proceedings of the National Academy of Sciences* 99 (12): 7872–76. https://doi.org/10.1073/pnas.132259099.
Federhen, S. 2012. “The NCBI Taxonomy Database.” *Nucleic Acids Research* 40 (D1): D136–43. https://doi.org/10.1093/nar/gkr1178.
Ge, Tongxin, Jiawen Yang, Shihui Zhou, Yuchen Wang, Yakui Li, and Xuemei Tong. 2020. “The Role of the Pentose Phosphate Pathway in Diabetes and Cancer.” *Frontiers in Endocrinology* 11 (June):365. https://doi.org/10.3389/fendo.2020.00365.
Goodacre, Norman F., Dietlind L. Gerloff, and Peter Uetz. 2014. “Protein Domains of Unknown Function Are Essential in Bacteria.” Edited by Claire M. Fraser. *mBio* 5 (1): e00744-13. https://doi.org/10.1128/mBio.00744-13.
Hayashi, Tomoyuki, and Alexei A. Stuchebrukhov. 2010. “Electron Tunneling in Respiratory Complex I.” *Proceedings of the National Academy of Sciences* 107 (45): 19157–62. https://doi.org/10.1073/pnas.1009181107.
Huang, Wei, Song Hong, Guirong Tang, Yuzhen Lu, and Chengshu Wang. 2019. “Unveiling the Function and Regulation Control of the DUF3129 Family Proteins in Fungal Infection of Hosts.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 374 (1767): 20180321. https://doi.org/10.1098/rstb.2018.0321.
Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996a. “VMD: Visual Molecular Dynamics.” *Journal of Molecular Graphics* 14 (1): 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
———. 1996b. “VMD: Visual Molecular Dynamics.” *Journal of Molecular Graphics* 14 (1): 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.
Iwata, So, Joong W. Lee, Kengo Okada, John Kyongwon Lee, Momi Iwata, Bjarne Rasmussen, Thomas A. Link, S. Ramaswamy, and Bing K. Jap. 1998. “Complete Structure of the 11-Subunit Bovine Mitochondrial Cytochrome Bc_1 Complex.” *Science* 281 (5373): 64–71. https://doi.org/10.1126/science.281.5373.64.

Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. 1983. “Comparison of Simple Potential Functions for Simulating Liquid Water.” *The Journal of Chemical Physics* 79 (2): 926–35. https://doi.org/10.1063/1.445869.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, et al. 2021. “PubChem in 2021: New Data Content and Improved Web Interfaces.” *Nucleic Acids Research* 49 (D1): D1388–95. https://doi.org/10.1093/nar/gkaa971.

Kraus, Alexander, Mareen Weskamp, Jennifer Zierles, Miriam Balzer, Ramona Busch, Jessica Eisfeld, Jan Lambertz, Marc M. Nowaczyk, and Franz Narberhaus. 2020. “Arginine-Rich Small Proteins with a Domain of Unknown Function, DUF1127, Play a Role in Phosphate and Carbon Metabolism of Agrobacterium Tumefaciens.” Edited by Anke Becker. *Journal of Bacteriology* 202 (22). https://doi.org/10.1128/JB.00309-20.

Letunic, Ivica, and Peer Bork. 2024. “Interactive Tree of Life (iTOL) v6: Recent Updates to the Phylogenetic Tree Display and Annotation Tool.” *Nucleic Acids Research* 52 (W1): W78–82. https://doi.org/10.1093/nar/gkae268.

Lobb, Briallen, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. 2021. “PathFams: Statistical Detection of Pathogen-Associated Protein Domains.” *BMC Genomics* 22 (1): 663. https://doi.org/10.1186/s12864-021-07982-8.

Luo, Chengke, Maryam Akhtar, Weifang Min, Xiaorong Bai, Tianli Ma, and Caixia Liu. 2024. “Domain of Unknown Function (DUF) Proteins in Plants: Function and Perspective.” *Protoplasma* 261 (3): 397–410. https://doi.org/10.1007/s00709-023-01917-8.

Lv, Peiyun, Jinlu Wan, Chunting Zhang, Aiman Hina, G M Al Amin, Naheeda Begum, and Tuanjie Zhao. 2023. “Unraveling the Diverse Roles of Neglected Genes Containing Domains of Unknown Function (DUFs): Progress and Perspective.” *International Journal of Molecular Sciences* 24 (4): 4187. https://doi.org/10.3390/ijms24044187.

Marcus, Yehouda, Hagit Altman-Gueta, Aliza Finkler, and Michael Gurevitz. 2003. “Dual Role of Cysteine 172 in Redox Regulation of Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase Activity and Degradation.” *Journal of Bacteriology* 185 (5): 1509–17. https://doi.org/10.1128/JB.185.5.1509-1517.2003.

Marino, Stefano M, and Vadim N Gladyshev. 2012. “Analysis and Functional Prediction of Reactive Cysteine” 287 (7): 4419–25. https://doi.org/10.1074/jbc.R111.275578.

Martí, María Carmen, Ana Jiménez, and Francisca Sevilla. 2020. “Thioredoxin Network in Plant Mitochondria: Cysteine S-Posttranslational Modifications and Stress Conditions.” *Frontiers in Plant Science* 11 (September):571288. https://doi.org/10.3389/fpls.2020.571288.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Mudgal, Richa, Sankaran Sandhya, Nagasuma Chandra, and Narayanaswamy Srinivasan. 2015. “De-DUFing the DUFs: Deciphering Distant Evolutionary Relationships of Domains of Unknown Function Using Sensitive Homology Detection Methods.” *Biology Direct* 10 (1): 38. https://doi.org/10.1186/s13062-015-0069-2.

Najafi, Saeed, Samuel Lobo, M. Scott Shell, and Joan-Emma Shea. 2025. “Context Dependency of Hydrophobicity in Intrinsically Disordered Proteins: Insights from a New Dewetting Free Energy-Based Hydrophobicity Scale.” *The Journal of Physical Chemistry B* 129 (7): 1904–15. https://doi.org/10.1021/acs.jpcb.4c06399.

Nallapareddy, Vamsi, Shubham Bogam, Himaja Devarakonda, Shubham Paliwal, and Debashree Bandyopadhyay. 2021. “DeepCys : Structure-based Multiple Cysteine Function Prediction Method Trained on Deep Neural Network: Case Study on Domains of Unknown Functions Belonging to

598 COX2 Domains." *Proteins: Structure, Function, and Bioinformatics* 89 (7): 745–61.
599 https://doi.org/10.1002/prot.26056.
600 Pandit, Shashi B, Rana Bhadra, Vs Gowri, S Balaji, B Anand, and N Srinivasan. 2004. "SUPFAM: A Database
601 of Sequence Superfamilies of Protein Domains." *BMC Bioinformatics* 5 (1): 28.
602 https://doi.org/10.1186/1471-2105-5-28.
603 Pascual-ahuir, J. L., E. Silla, and I. Tuñon. 1994. "GEPOL: An Improved Description of Molecular Surfaces.
604 III. A New Algorithm for the Computation of a Solvent-excluding Surface." *Journal of*
605 *Computational Chemistry* 15 (10): 1127–38. https://doi.org/10.1002/jcc.540151009.
606 "Rekker, R. F. The Effect of Intramolecular Hydrophobic Bonding on Partition Experiments; 1967; Vol. 86."
607 n.d.
608 Robbins, A. H., and C. D. Stout. 1989. "The Structure of Aconitase." *Proteins: Structure, Function, and*
609 *Bioinformatics* 5 (4): 289–312. https://doi.org/10.1002/prot.340050406.
610 "Robert C.Tryon and Daniel E. Bailey. Cluster Analysis. New York McGraw-Hill, 1970. Pp. Xvii." n.d.
611 Rocha, João J., Satish Arcot Jayaram, Tim J. Stevens, Nadine Muschalik, Rajen D. Shah, Sahar Emran,
612 Cristina Robles, Matthew Freeman, and Sean Munro. 2023. "Functional Unknomics: Systematic
613 Screening of Conserved Genes of Unknown Function." Edited by Ian Dunham. *PLOS Biology* 21
614 (8): e3002222. https://doi.org/10.1371/journal.pbio.3002222.
615 Rouault, Tracey A., and Wing Hang Tong. 2008. "Iron–Sulfur Cluster Biogenesis and Human Disease."
616 *Trends in Genetics* 24 (8): 398–407. https://doi.org/10.1016/j.tig.2008.05.008.
617 Santiago-Tirado, Felipe H., and Tamara L. Doering. 2016. "All about That Fat: Lipid Modification of
618 Proteins in Cryptococcus Neoformans." *Journal of Microbiology* 54 (3): 212–22.
619 https://doi.org/10.1007/s12275-016-5626-6.
620 Schoch, Conrad L, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda
621 Khovanskaya, Detlef Leipe, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation,
622 Resources and Tools." *Database* 2020 (January):baaa062.
623 https://doi.org/10.1093/database/baaa062.
624 Sun, Fei, Xia Huo, Yujia Zhai, Aojin Wang, Jianxing Xu, Dan Su, Mark Bartlam, and Zihe Rao. 2005. "Crystal
625 Structure of Mitochondrial Respiratory Membrane Protein Complex II." *Cell* 121 (7): 1043–57.
626 https://doi.org/10.1016/j.cell.2005.05.025.
627 Tong, Sen-Miao, Ying Chen, Sheng-Hua Ying, and Ming-Guang Feng. 2016. "Three DUF1996 Proteins
628 Localize in Vacuoles and Function in Fungal Responses to Multiple Stresses and Metal Ions."
629 *Scientific Reports* 6 (1): 20566. https://doi.org/10.1038/srep20566.
630 Townsend, Danyelle M., Volodymyr I. Lushchak, and Arthur J.L. Cooper. 2014. "A Comparison of
631 Reversible Versus Irreversible Protein Glutathionylation." In *Advances in Cancer Research*,
632 122:177–98. Elsevier. https://doi.org/10.1016/B978-0-12-420117-0.00005-0.
633 Tsukihara, Tomitake, Hiroshi Aoyama, Eiki Yamashita, Takashi Tomizaki, Hiroshi Yamaguchi, Kyoko
634 Shinzawa-Itoh, Ryosuke Nakashima, Rieko Yaono, and Shinya Yoshikawa. 1996. "The Whole
635 Structure of the 13-Subunit Oxidized Cytochrome c Oxidase at 2.8 Å." *Science* 272 (5265): 1136–
636 44. https://doi.org/10.1126/science.272.5265.1136.
637 Valcarcel-Jimenez, Lorea, and Christian Frezza. 2023. "Fumarate Hydratase (FH) and Cancer: A Paradigm
638 of Oncometabolism." *British Journal of Cancer* 129 (10): 1546–57.
639 https://doi.org/10.1038/s41416-023-02412-w.
640 Yang, Moua. 2022. "Redox Stress in COVID-19: Implications for Hematologic Disorders." *Best Practice &*
641 *Research Clinical Haematology* 35 (3): 101373. https://doi.org/10.1016/j.beha.2022.101373.
642
643