ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE BIOTECNOLOGICHE, BIOCOMPUTAZIONALI, FARMACEUTICHE E FARMACOLOGICHE

Ciclo 38

Settore Concorsuale: 05/E1 - BIOCHIMICA GENERALE

Settore Scientifico Disciplinare: BIO/10 - BIOCHIMICA

PROTEIN FUNCTIONAL ANNOTATION WITH EMBEDDINGS AND COMPUTATIONAL APPROACH

Presentata da: Gabriele Vazzana

Coordinatore Dottorato

Maria Laura Bolognesi

Supervisore

Pier Luigi Martelli

Co-supervisore

Rita Casadio

Castrense Savojardo

Esame finale anno 2026

Abstract

The emergence of Next Generation Sequencing technologies has created an urgent need for reliable and scalable methods of protein functional annotation. Moreover, recent advancements in machine and deep learning, including protein language models (PLMs), offer unprecedented opportunities to develop methods able to generalize information collected in databases for the annotation of uncharacterized proteins. My PhD research focused on leveraging deep learning to enhance annotation strategies, with a focus on the Glutathione S-transferase (GST) superfamily, a multifunctional and hard-to-annotate enzyme group.

First, I tested the potential of protein sequence encodings based on PLMs (embeddings) in the process of functional annotation. Using an alignment algorithm designed for embeddings, I demonstrated that they capture structural information and enable accurate classification of GSTs.

Furthermore, I characterised new multifunctional traits of GSTs, providing computational evidence that canonical GSTs can bind RNA, as suggested by recent large-scale studies. By applying deep learning methods and molecular docking validation, I showed that GST–RNA interactions are theoretically possible, and proposed that this interaction occurs at the glutathione binding site.

As deep learning procedures drive modern protein structure modeling, even in low-homology scenarios, their comparison is an active research field. I analyzed protein structure models from the Alpha&ESMhFold database, which collects AlphaFold2 and ESMFold predicted models of human proteins. By mapping Pfam domains, I found that functionally relevant regions are consistently well-predicted by both methods, even when the global structures diverge.

During my period abroad in Barcelona, I analyzed evolutionary information captured by embeddings and found that, given a dataset of remote homologs, the embedding vectors representing residues aligned in a multiple structural alignment cluster together.

Overall, these studies show that embeddings and structural predictions can enhance the annotation of challenging protein families, reveal novel functional roles, and facilitate the integration of large-scale data into annotation pipelines.

Index

1. Introduction

The first chapter of this thesis provides an introduction to proteins, the central focus of my doctoral research, and the problem of protein functional annotation. Chapter 2 explores Glutathione S-transferases (GSTs), a protein superfamily notoriously hard to annotate, that has been chosen as a case study. Chapter 3 presents an overview of machine learning, with focus on the deep learning models adopted to address annotation challenges. Finally, Chapters 4, 5, and 6 detail the main research projects in which I was involved, all of which have been published in peer-reviewed journals.

1.1. The Central dogma of Molecular Biology

The flow of genetic information, from DNA to proteins, is a fundamental principle of biology, routinely referred to as the central dogma of molecular biology. This process is divided into three major steps: replication, transcription and translation. During replication the cell DNA content is duplicated so that the daughter cells receive the exact copy of the parental DNA. The transcription is the process by which protein-coding genes are transcribed into messenger RNA (mRNA) which carries the genetic information. The mRNA is then exported to the cytoplasm, where ribosomes decode its sequence to synthesize proteins through translation. Altogether, these steps ensure the accurate transmission of genetic instructions from DNA to proteins, the functional components that sustain the life of the cell (Nelson and Cox 2021).

1.2. Roles of RNA

Similarly to DNA, RNA is a nucleotide polymer, but it shows a ribose sugar instead of deoxyribose. In addition to messenger RNAs, cells produce a diverse array of non-coding

RNAs that carry out crucial regulatory and structural functions. Transfer RNAs (tRNAs) deliver amino acids to the ribosome during protein synthesis, while ribosomal RNAs (rRNAs) form the catalytic component of ribosomes. Other types, such as small nuclear RNAs (snRNAs) and microRNAs (miRNAs), play key roles in RNA splicing and post-transcriptional gene regulation, respectively (Zacco et al. 2024). Altogether, in the crowded cell environment, RNA comprises 10–20% of the cell's dry weight (approximately 20–40 mg/ml) (Berry and Pelkmans 2022).

1.3. Proteins biosynthesis and composition

Proteins are fundamental cellular components involved in virtually every aspect of life, including biosynthesis, regulation, gene expression, cell communication and more. From a chemical point of view, proteins are heteropolymers composed of 20 standard amino acid residues sequentially linked by peptide bonds. Each residue is characterized by a conserved amino group, a carboxyl group, a hydrogen atom, and a distinctive side chain (R group) bonded to a central α-carbon. The only exception is Glycine whose side chain is a hydrogen atom. The side chains confer specific physicochemical properties essential for protein folding and function. Accordingly, amino acid residues can be classified as apolar (Glycine/Gly/G, Alanine/Ala/A, Valine/Val/V, Proline/Pro/P, Leucine/Leu/L, Isoleucine/Ile/I, Methionine/Met/M), aromatic (Phenylalanine/Phe/F, Tyrosine/Tyr/Y, Tryptophan/Trp/W), polar uncharged (Serine/Ser/S, Threonine/Thr/T, Cysteine/Cys/C, Glutamine/Gln/Q, Asparagine/Asn/N), and charged (acid: Glutamate/Glu/E, Aspartate/Asp/D; basic: Lysine/Lys/K, Histidine/His/H, Arginine/Arg/R) (Nelson and Cox 2021).

1.4. Proteins hierarchical organization

The linear polypeptide sequence in which these residues are covalently joined defines the protein's primary structure. In polar environments and physiological conditions, proteins adopt specific local conformations known as secondary structure (typically recurrent patterns of α-helix or β-sheet) and an overall three-dimensional organization (tertiary structure) through folding (Nelson and Cox 2021). Protein folding is a complex and spontaneous process in which proteins adopt the conformation with the lowest Gibbs free energy ($\Delta G_{folding}$) driven by the interplay of different physicochemical properties of residues and their surrounding environment. The correct folding is essential for the final functionality of the protein; this implies a tight connection between the protein's amino acid sequence, structure, and biological role (Anfinsen 1973). Indeed, DNA mutations (i.e., errors occurring during DNA replication) can change the primary sequence of a protein and, if the affected residue is essential for proper folding or function, such a change can lead to disease. Folded proteins can be locally organized into structural units, called domains, which represent self-organized regions of the polypeptide that fold independently from the rest. A single protein can have one or more domains, and the same domain can appear in unrelated proteins. Finally, two or more polypeptide subunits, also called chains, can interact with each other with non-bonding interactions: this organization is routinely referred to as quaternary structure (Figure 1) (Nelson and Cox 2021).

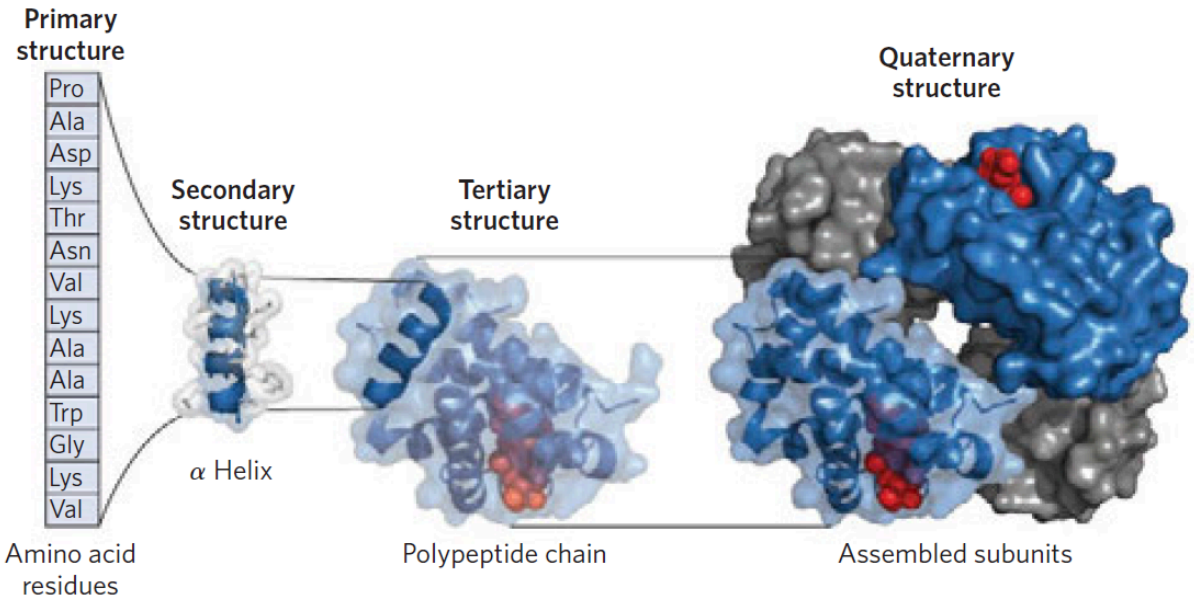


Figure 1. Hierarchical organization of protein structure (hemoglobin in this example). The primary structure resembles the sequence of residues in the polypeptide chains. The local organization of the chain is referred to as a secondary structure. The overall fold of the protein is the tertiary structure, which resembles its final three-dimensional shape. Two or more polypeptide chains can aggregate in quaternary structures.

Through the years, different experimental techniques have been designed to determine the three-dimensional structure of proteins: X-ray crystallography, Cryogenic electron microscopy (cryo-EM) and Nuclear Magnetic Resonance (NMR) (Berman 2000). However, these techniques present limitations as they are expensive and time-consuming. For this reason, extensive research efforts of the past decades have focused on the so-called "protein folding problem": developing computational methods to predict protein structures starting from the sequence (Kryshtafovych et al. 2023).

1.5. Enzymes

In living beings, crucial biochemical reactions are made possible by a special class of proteins known as enzymes. These proteins act as catalysts, accelerating reaction rates by lowering activation energy by many orders of magnitude, so that reactions proceed fast

enough to support life. An important feature of enzymes is their high specificity: it arises from a well-defined three-dimensional pocket or cleft where substrate binding and catalysis occur through an ensemble of molecular interactions and structural adjustments. The chemical reaction itself is driven by a set of catalytically active amino acid residues located within the enzyme's active site (Nelson and Cox 2021).

The Enzyme Commission (EC) number is a standardized system used to classify enzymes based on the specific reactions they catalyze (Webb 1992). Each EC number consists of four digits separated by periods, where the first digit denotes the broad class of the enzymatic reaction. There are seven primary classes: oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), ligases (EC 6), and translocases (EC 7). However, the "one enzyme, one function" paradigm, while foundational, does not encompass the full complexity of the proteome. Multifunctional (also called moonlighting or multifaceted) enzymes can perform two or more distinct, often unrelated, biochemical functions (Gupta and Uversky 2023; Bertolini et al. 2024).

1.6. Protein families and distant related homologs

Evolutionarily related proteins derive from a common ancestral protein that undergoes various selective pressures, accumulating mutations at the level of the protein sequences. These mutations are positively selected only when they provide no or minimal alterations to the protein's structure and function. Consequently, proteins with divergent sequences can adopt the same structural fold and can therefore be grouped into so-called protein families. When the structure is not available, classic approaches such as sequence alignment algorithms (like Needleman-Wunsch (Needleman and Wunsch 1970) and Smith-Waterman (Smith and Waterman 1981)) can be used to compare two protein sequences and score their similarity by means of the sequence identity. A general principle formulated in (Chothia and

Lesk 1986; Rost 1999) states that, in these cases, a sequence identity above a 30% threshold, named twilight zone, is usually sufficient to infer that two sequences share the same fold. Below this threshold, however, sequence identity alone cannot reliably indicate structural similarity. As a result, proteins with similar folds but sequence identity below the twilight zone, routinely referred to as distantly related (or remote) homologs, cannot be detected using standard alignment methods.

1.7. The problem of protein functional annotation

Protein functional annotation is the process of identifying and describing a protein's biological roles, activities, and interactions using all available data, with structure being the most informative source of information. Two of the most valuable and used databases for protein annotation are UniProt (The UniProt Consortium et al. 2023) (https://www.uniprot.org/) and the Protein Data Bank (PDB) (Berman 2000) (https://www.rcsb.org/), which collect information about protein sequences and structures, respectively. UniProt is divided into two subsections: Swiss-Prot and TrEMBL. Swiss-Prot contains proteins whose annotation has been manually revised by expert curators and include, when available, reference to the current literature and to different resources collecting information on different aspects of protein structure and function. TrEMBL contains entries with annotations carried out automatically and not revised by a curator.

After the breakthrough of Next Generation Sequencing (NGS) technologies, genomic and metagenomic projects can generate an overwhelming amount of sequence data, revealing a vast universe of putative proteins whose functions remain unknown (Sarkar 2016). This deluge of data stands in contrast to the low-throughput and time-consuming nature of experimental functional and structural characterization which remains the golden standard for the elucidation of the role of proteins in cellular processes, in the context of biological

complexity. The huge difference between the number of entries in TrEMBL (253,061,696), Swiss-Prot (573,661) and PDB experimental entries (242,066) (databases accessed on 10 September 2025) is a clear indication of knowledge gap between the different levels of complexity of the protein organization. The computational analysis of the available data has been able to identify inference rules that can be applied to fill this gap and to develop tools, often based on machine learning, that, to some extent, can provide a structural and functional annotation starting from the protein sequence.

Traditional annotation pipelines have heavily relied on the transfer of annotation upon sequence similarity or on the recognition of conserved domains and motifs that can be associated to specific functional features (Altschul et al. 1990; Finn et al. 2007; Paysan-Lafosse et al. 2023). In UniProt, one of the components of the automatic annotation pipeline is the Association-Rule-Based Annotator (ARBA): a human-readable expert rule system, trained on Swiss-Prot entries, that combines information of computationally recognized domains and taxonomic data to derive protein properties such as function, catalytic activity, pathway membership, subcellular location, protein family, protein name and more (https://www.uniprot.org/help/arba). During the manual annotation of Swiss-Prot entries, each sequence is analyzed with computational tools (including ARBA) and the results are integrated with current knowledge of related proteins; relevant results are then selected by expert curators for integration (https://www.uniprot.org/help/manual_curation).

However, this approach is scarcely effective for distantly related homologs or proteins belonging to poorly characterized families. Moreover, functional annotation is complicated by the multifaceted nature of protein roles. Therefore, it is necessary to develop new computational methods that exploit the recent impressive advances of computational technology and algorithmic development.

This thesis investigates how to address the protein functional annotation problem by leveraging advanced Artificial Intelligence techniques. My research was mainly focused on the Glutathione S-transferase (GST) superfamily: a diverse, widespread and multifunctional enzyme group that is notoriously challenging to annotate. The next chapters introduce the GST superfamily and the deep learning models adopted. Subsequently, the manuscript details two analytical workflows in which we produced novel functional annotations on the GST superfamily. Finally, a supplementary study is presented that compares protein models predicted by state-of-the-art protein structure prediction algorithms, enhancing protein domains functional annotation.

2. Glutathione S-transferase superfamily

Glutathione S-transferases (GSTs, EC 2.5.1.18) constitute a protein superfamily of multifunctional enzymes primarily involved in cellular detoxification and widely expressed across most living organisms (Mazari et al. 2023; Aloke et al. 2024). These enzymes play a crucial role in phase II detoxification: they catalyze the conjugation of the reduced glutathione (γ-glutamyl-cysteinyl-glycine, GSH. Figure 2) to toxic endogenous and exogenous electrophilic compounds.

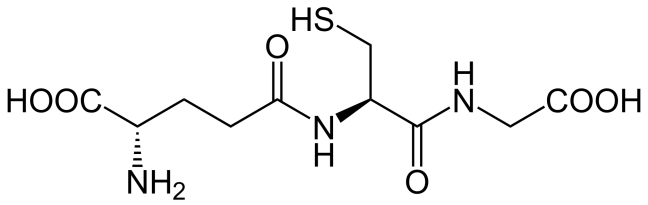
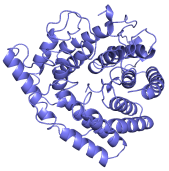
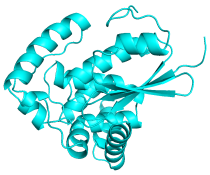
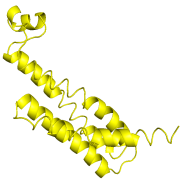


Figure 2. Structural formula of Glutathione (GSH)

GSH is present in all mammalian tissues at 1–10 mM concentrations, with the highest expression level in the liver and whose increased transcription is associated with oxidative stress. After the reaction, GSH conjugates can be easily excreted from the cell by phase III transporters (Mazari et al. 2023; Aloke et al. 2024). The superfamily is highly heterogeneous and is divided into three major groups on the basis of the subcellular localisation: cytosolic, mitochondrial and microsomal. Both cytosolic and mitochondrial enzymes are soluble proteins whereas microsomal GSTs are membrane proteins. One major feature of GSTs is that each group is characterized by the presence of remote homologs that are hard to annotate with classic alignment methods. In UniProt a total of 6 GST folds are present. As summarized in Table 1, the cytosolic group contains proteins belonging to 4 different folds, while the mitochondrial and microsomial group are characterized by one fold each.

Table 1. Overview of GST folds as reported in the UniProt database.

Group	Fold	Representative PDB
Cytosolic	**Canonical**	 8GSS
	Omega-like	 5LKD
	FosA	 1NPB
	LanC	 8D19
Mitochondrial	**Kappa**	 3RPN
Microsomal	**MAPEG**	 4AL0

Representative structures have been selected from high-quality Swiss-Prot entries.

2.1. Cytosolic GSTs

The majority of the GSTs are present in the cytosol, where 4 folds can be found: canonical, omega-like, FosA and LanC GSTs.

Canonical GSTs are the most abundant and extensively studied GSTs: these enzymes are further included into several families (named also classes) introduced on the basis of sequence similarity, structural properties and chromosomal organization in different organisms (Mazari et al. 2023; Aloke et al. 2024). Each class is identified by a distinct name (often a Greek letter) and exhibits an intra-class sequence identity above 30%; however, the inter-class sequence identity falls below the twilight zone threshold, even if the structure is largely conserved, indicating that different classes are distantly related homologs (Aloke et al. 2024).

Among canonical GST classes, some are widespread across all species (Zeta, Theta), while others are taxon-specific. For example, the Tau, Lambda, Phi and DHAR classes are plant-specific (Kumar and Trivedi 2018), Delta and Epsilon classes are insect-specific (Koirala et al. 2022) whereas Beta classes are only found in bacteria (Shehu et al. 2018). In humans a total of seven families are currently known: Mu, Sigma, Alpha, Pi, Theta, Omega, Zeta.

All canonical GSTs are small proteins (roughly 200-250 residues) with a recognizable fold: a N-terminal thioredoxin-like (βαβαββα) and an all-alpha C-terminal domain connected by a short linker (Zhuge et al. 2020). The N-terminal domain is the most conserved among canonical GSTs: it contains the GSH-binding pocket (called G-site) and the GSH-activating Cysteine, Serine or Tyrosine residues, depending on the class. The H-site is the region responsible for the interaction with the toxic compounds (routinely hydrophobic molecules); it is close to the G-site in the three-dimensional structure and mainly involves residues of the C-terminal domain. The functional unit of these enzymes mainly exists as either homo- or

heterodimeric proteins, and only subunits within the same class can form heterodimers (Mazari et al. 2023).

Omega-like GSTs, also called GSTs Xi class, were first described in *Saccharomyces cerevisiae* and assigned to the canonical Omega class (Garcerá et al. 2006). Further analysis and structural superimposition with Omega-class cytosolic GSTs, as reported in (Xun et al. 2010; Meux et al. 2011), suggest that these proteins should be considered a new GST fold. Despite assuming a GST canonical fold (a thioredoxin-like domain followed by an all-alpha domain), these proteins are complemented with additional structural features: a long N-terminal coil (roughly 80 residues), a 30 residues long loop between the strand β2 and helix α2 and finally a 20 residues long C-terminal coil. Altogether, these proteins are longer than canonical GSTs, with a total of 300 or more residues (Meux et al. 2011).

LanC GSTs were discovered after assays on mammalian LanC-like protein 1 (LANCL1), demonstrating that these proteins, mainly recognized as peptide-modifying enzymes, can catalyse the conjugation of GSH to synthetic substrates, similarly to GSTs (Huang et al. 2014). These cytosolic proteins are longer than canonical GSTs (roughly 400 residues) and show a characteristic α,α-toroid fold with a zinc ion bound at the top of the barrel by two His residues and a Cys residue (Ongpipattanakul et al. 2023).

Fosfomycin-resistance proteins (Fos-resistance) are bacterial metalloenzymes that inactivate fosfomycin through the addition of nucleophiles to the oxirane ring of the molecule (Varotsou et al. 2023). The most well characterized Fos-resistance proteins are FosA, which are manganese (Mn^{+2})- and potassium (K^{+})-dependent glutathione transferases. These bacteria-specific GSTs are short enzymes (130-140 residues) with a distinct fold organized in two βαβββ tandem domains connected by a flexible linker (Pakhomova et al. 2004; Shehu et al. 2018). The enzyme is catalytically active as homodimers (Varotsou et al. 2023).

2.2. Mitochondrial GSTs

Mitochondrial GSTs (also known as Kappa) are mostly found in the mitochondrial matrix and peroxisomes (Morel and Aninat 2011). Similarly to canonical GSTs, these are soluble proteins including roughly 200 residues. The subunits are organized into homodimers and contain a thioredoxin-like and all-alpha domains. However, the two domains are rearranged with respect to the canonical GST fold: the all-alpha domain is placed between the βαβ and the ββα motifs of the thioredoxin-like domain (Oakley 2011). As a result the overall structure of the two GST families is different. Given the domain sharing, a parallel evolution hypothesis has been proposed in (Ladner et al. 2004).

2.3. Microsomal GSTs

Microsomal GSTs are members of the MAPEG protein group (Membrane-Associated Proteins in Eicosanoid and Glutathione metabolism) (Morgenstern et al. 2011). Each subunit is roughly 150 residues long and consists of a 4-helix bundle that packs together to form homotrimers (Bresell et al. 2005; Sjögren et al. 2013). The MAPEG family is rich in remote homologs (Aloke et al. 2024), with sequence identities that can drop below 15%.

2.4. Catalytic reaction(s)

At the basis of the main catalytic activity of GSTs is the ability to reduce the sulfhydryl group's pK_a within reduced glutathione (GSH) from 9.0 in an aqueous environment to approximately 6.5 when GSH becomes conjugated in the enzyme's active site; once the GSH is reduced in the GST active site, it can spontaneously react by nucleophilic attack with electrophilic xenobiotics that are situated in close proximity (Aloke et al. 2024). Therefore,

GSTs catalysis occur by simultaneously activating the GSH moiety and binding hydrophobic electrophilic compounds.

Rather than the GST main activity, several GSTs enzymes can also exhibit glutathione peroxidase and isomerase activity (Aloke et al. 2024). Additionally, some GSTs, like human GSTP1, are able to facilitate S-glutathionylation reactions: this is a protein post-translational modification that attaches GSH to a solvent accessible cysteine of an acceptor protein (Mazari et al. 2023). Also, several GST isoenzymes have been found to interact with kinases involved in regulatory pathways (like stress response, cell proliferation and apoptosis), suggesting that these enzymes play relevant roles in signaling (Mazari et al. 2023).

2.5. Clinical relevance

GSTs enzymes are highly relevant for human health as they are overexpressed in several human pathologies (Lv et al. 2023), when their concentrations can rise to as much as 10% of the cell cytosolic protein content under stress conditions (Mukanganyama et al. 2011). Pi-class cytosolic GSTs are of particular interest as they are widely expressed in human tissues and their overexpression in cancer cells is highly correlated with chemoresistance (Lv et al. 2023).

3. Machine Learning for Computational Biology

Machine learning (ML) is a branch of Artificial Intelligence in which computer systems learn to perform tasks by directly learning patterns from existing data rather than relying on explicitly programmed instructions (Bishop 2016). While first machine learning programs were developed in the 1950s, the field experienced important progress only in the 2010s thanks to the advances in algorithms (like deep neural networks), hardware (like GPUs) and the beginning of the Big Data era (Baldi 2021). Building a ML model is fundamentally a data-driven procedure that tries to approximate a high-dimensional mathematical function to solve a given problem. ML processes can be divided into two main steps: training and inference. During training, the model uses the training data to adjust its internal parameters; once trained, it can be used in "inference mode" to generate predictions on new, unseen data. ML models can be trained using three distinct strategies: supervised learning, which uses labeled training data (i.e., data for which the association between the input and the output variables is known); unsupervised learning, which finds patterns in unlabeled data; and reinforcement learning, where the model learns optimal strategies to achieve a specific task (Baldi 2021).

Even though recent state-of-the-art models are based on deep neural networks (see next sections), earlier breakthroughs were driven by other influential approaches such as Hidden Markov Models (Rabiner 1989), Random Forests (Ho 1995) and Support Vector Machines (Cortes and Vapnik 1995). It is worth noting that Hidden Markov Models (HMMs) have been widely adopted in bioinformatics to model protein domains, with the emergence of databases like Pfam (Finn et al. 2007) and InterPro (Paysan-Lafosse et al. 2023).

Nowadays, machine learning is widely applied across many research domains, including computational biology. The following sections of this chapter will address three key aspects:

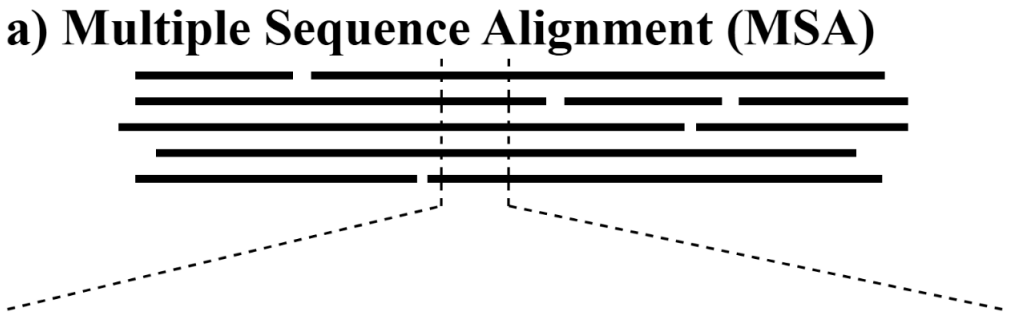
(i) classic strategies to numerically encode protein sequence data; (ii) the main ML models relevant for this thesis that expand the discussion on protein encoding; and (iii) a brief overview of recent state-of-the-art models for protein structure prediction.

3.1. Protein sequence encodings based on single sequence and Multiple Sequence Alignments

Feeding a machine learning model requires input data to be represented in numerical form. In the case of protein sequences, this translates into the challenge of numerically encoding amino acid residues. This is not a trivial task as biological sequences contain multiple layers of information like residue identity, evolutionary signals, structural and functional features. The simplest encoding scheme is the so-called one-hot encoding, in which each residue along the sequence is represented by a 20-dimensional binary vector containing a single one and nineteen zeros, univocally identifying a residue. While this uniquely identifies each amino acid, it fails to capture any deeper information. A major advance over such a simplistic approach came with the use of Multiple Sequence Alignments (MSAs). By aligning a query protein sequence against homologous sequences retrieved from large databases, an MSA is generated: it represents a matrix where homologous residue positions are arranged in the same column. From an MSA, a Sequence profile can be constructed by computing the residue frequency distribution at each MSA column (Figure 3). This provides a more informative representation that reflects the degree of conservation at each site and highlights evolutionary constraints across a protein family. The quality of MSAs is influenced by several factors. First, the choice of the alignment algorithm is crucial, as different methods try to balance speed and accuracy. Over the years several tools have been developed, including the Clustal family (Larkin et al. 2007; Sievers et al. 2011), T-Coffee (Notredame et al. 2000), MAFFT (Katoh 2002) and others. Moreover, the database used to retrieve similar sequences strongly

affects the representativity of the alignment and, consequently, the quality of the resulting profiles. Poorly represented protein families give rise to the so-called orphan protein problem, in which limited homologous sequences result in poor alignments.

The recent introduction of protein language models, based on specific deep neural network architectures, offers an alternative encoding, that will be better detailed in section 3.2.2, after a short introduction to Neural Networks



Sequence 1:	F	K	L	L	S	Q	C	L	L	V	Q	F
Sequence 2:	F	K	A	P	G	Q	T	M	L	Q	Q	S
Sequence 3:	Y	P	-	V	G	Q	-	L	L	G	L	D
Sequence 4:	F	P	P	V	V	Q	E	A	L	L	L	K
Sequence 5:	L	E	F	I	S	Q	E	E	C	-	I	Q

b) Sequence profile

Position

A:	0	0	25	0	0	0	0	20	0	0	0	0
C:	0	0	0	0	0	0	25	0	20	0	0	0
D:	0	0	0	0	0	0	0	0	0	0	0	20
E:	0	20	0	0	0	0	50	20	0	0	0	0
F:	60	0	25	0	0	0	0	0	0	0	0	20
G:	0	0	0	0	40	0	0	0	0	25	0	0
H:	0	0	0	0	0	0	0	0	0	0	0	0
K:	0	40	0	0	0	0	0	0	0	0	0	20
I:	0	0	0	20	0	0	0	0	0	0	20	0
L:	20	0	25	20	0	0	0	40	80	25	40	0
M:	0	0	0	0	0	0	0	20	0	0	0	0
N:	0	0	0	0	0	0	0	0	0	0	0	0
P:	0	40	25	20	0	0	0	0	0	0	0	0
Q:	0	0	0	0	0	100	0	0	0	25	40	20
R:	0	0	0	0	0	0	0	0	0	0	0	0
S:	0	0	0	0	40	0	0	0	0	0	0	20
T:	0	0	0	0	0	0	25	0	0	0	0	0
V:	0	0	0	40	20	0	0	0	0	25	0	0
W:	0	0	0	0	0	0	0	0	0	0	0	0
Y:	20	0	0	0	0	0	0	0	0	0	0	0

Figure 3. Representation of a sequence profile (b) computed from a Multiple Sequence Alignment (a) involving 5 protein sequences. The example shows the frequencies (in percentage) of each residue in a region of the MSA.

3.2. Neural Networks

Neural networks have emerged as the main actors of the advances in Artificial Intelligence. These models are built upon artificial neurons: computational units, originally inspired by biological neurons, that integrate signals from connected inputs and return an output through a transfer (or activation) function. Such units can be organized in a simple architecture called Perceptron, first proposed in (Rosenblatt 1958). Perceptrons, structured in two layers of input and output neurons, process the data in a feed forward manner, as data flow from the input to the output. The synaptic weights are the trainable parameters and their value is optimized during the training phase in order to reproduce as well as possible the input-output associations present in the training set. Unfortunately this model was capable of solving only linearly separable problems and was therefore abandoned soon. Artificial neurons and perceptrons can be visualized in Figure 4.

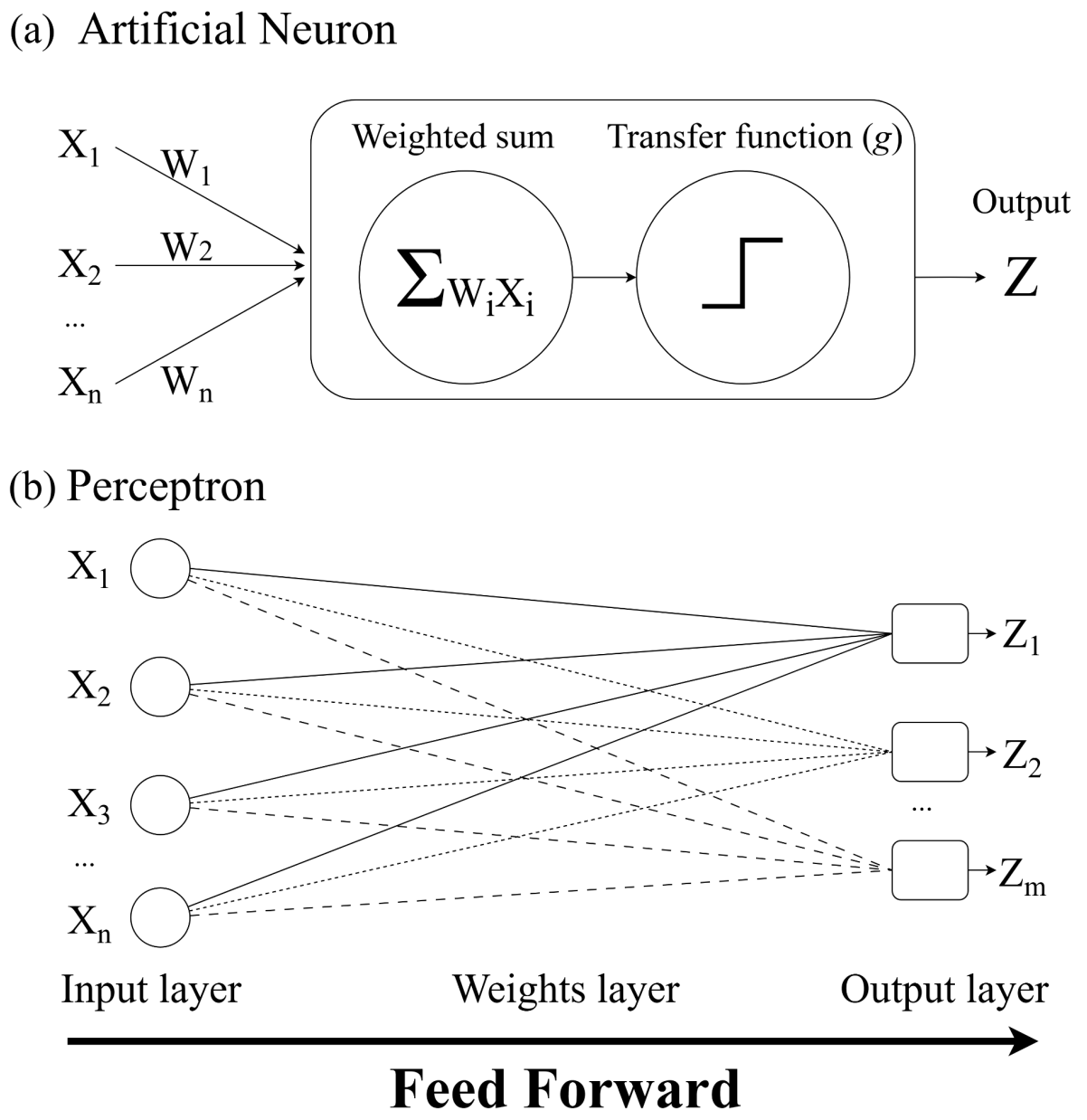


Figure 4. Artificial neurons and perceptrons. (a) A generic artificial neuron receives a set of inputs *X* and associated weights *W* (both real numbers). This unit computes a weighted sum of the inputs and the weights and applies a transfer function *g* to the result, producing the output *Z*. (b) A perceptron organized in an input and output layer of neurons, connected by a set of weights; the two layers are organized in *n* and *m* components, respectively, with *n* routinely greater than *m*. The data flows in a feed forward direction, from the input to the output.

The interest in neural networks was renewed when it was shown that the introduction of additional hidden layers between the input and output layers improve the computational ability of the networks, allowing them to address non linear problems. Consequently new

algorithms, and in particular the back-propagation, have been introduced for training multilayer networks. The efficiency of the training algorithms have limited the possibility to adopt complex (non strictly feed-forward) network architectures and to increase the number of network layers. The introduction of new computational techniques and the availability of more powerful and parallel computers have prompted the emergence of what is now known as Deep Learning, which enables the solution of far more complex problems. Over the years, the mathematical functions modeled by Deep Learning applications have become more complex, leading to improved performance but reduced interpretability. Consequently, these models are often considered "black boxes"(Linardatos et al. 2020; Li et al. 2022).

3.2.1. Transformers

Transformers are an advanced type of neural network specifically designed to handle sequential data, and were originally implemented in (Vaswani et al. 2017) to solve sequence-to-sequence tasks such as language translation. Unlike previous architectures, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), Transformers can process a full sequence in parallel rather than sequentially, thanks to the so-called self-attention mechanism. This key component enables the model to assign context-dependent weight to relationships between all positions of the input sequence, enabling it to capture both local dependencies and long-range patterns essential for understanding the global context information. Moreover, the high parallelization enables these models to take full advantage of modern hardware, and grow both in architecture and training data size.

Transformers main application is in the field of Natural language processing, where the analysed sequences are phrases whose elementary units (tokens) are the words; this framework has led to the emergence of the so-called Large Language Models (LLMs). The

scope of a Transformer is to associate two sequences (e.g., a sentence in two different languages, or the corrupted and the uncorrupted versions of a sentence). To this aim, these models use two main trainable blocks: encoder and decoder (Figure 5). After training on a huge corpora of texts, the encoder layer derives high-dimensional internal representations of the input words, routinely referred to as embeddings. Embeddings capture relationships among the input elements, which the decoder layers then exploit to construct the output sequence.

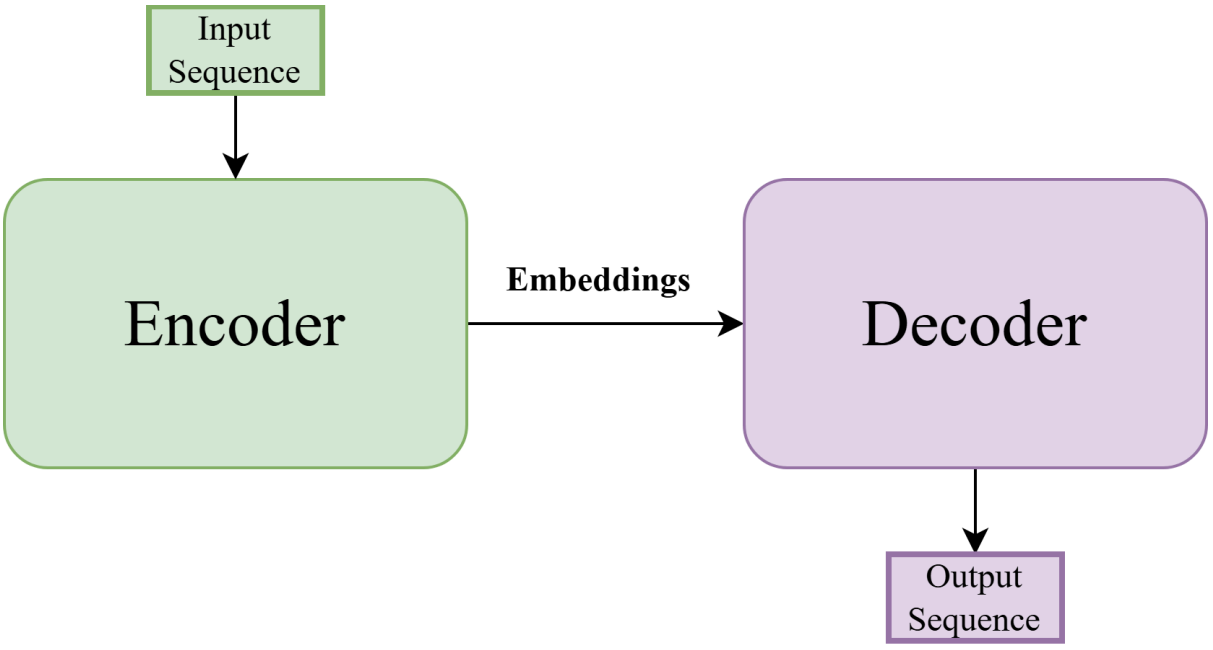


Figure 5. Simplified version of the Transformer architecture described in (Vaswani et al. 2017). For a given input sequence, encoder blocks (left) generate word embeddings that are then processed by the decoder block (right) to output a new sequence. Both blocks adopt self-attention mechanisms. The model was originally used for language translation tasks.

3.2.2. Protein Language Models

Transformers have quickly found applications in several fields including bioinformatics, with the emergence of Protein language Models (PLMs) (Elnaggar et al. 2020; Ofer et al. 2021).

Indeed, a protein sequence can be seen as a natural language sentence in which the 20 amino acid residues represent all the possible words/tokens. Similarly to LLMs, PLMs are trained on a huge amount of sequence data, making them well-suited for the big volume of modern biological databases (as discussed in Section 1.7.).

Research on PLMs has grown fast in the last few years, and different model architectures have been explored. They can be mainly classified into three types: encoder-decoder, decoder-only and encoder-only (Bepler and Berger 2021; Wang et al. 2025). Encoder–decoder architectures follow the design of the original Transformer model and are suited for text-to-text tasks. For instance, ProstT5 adopts this architecture to translate protein sequences into one-dimensional string representations that encode three-dimensional structural information (Heinzinger et al. 2024). Decoder-only architectures are commonly used for generation tasks; their training paradigm relies on the prediction of the next residue and the comparison with the real one. For example, ProtGPT is an autoregressive, decoder-only model capable of generating sequences within seconds (Ferruz et al. 2022).

During my work I focused on encoder-only models, which include the Evolutionary Scale Modeling (ESM) family of PLMs developed at Meta (Rives et al. 2021; Lin et al. 2023). These models are used to generate information-rich embeddings for each amino acid residue along the sequence. The training objective mostly adopted in these cases is called Masked Language Modeling: during training, a set of random positions along the sequence are masked and the model is tasked to predict the hidden residue by exploiting the contextual information. This training paradigm is called self-supervised because the actual label of the training data is already stored in the input (Figure 6). After training, the output block is discarded and, given a new input sequence, the resulting embeddings can be extracted.

PLMs embeddings are high-dimensional, context-aware vectorial representations of residues of a protein sequence (Weissenow and Rost 2025). The size D of these vectors depends on the

number of hidden states of the transformer layer from which the representations are extracted (typically the last one). The final encoding of the protein is therefore a matrix $e \in R^{l \times D}$, (with l being the protein length), routinely referred to as per-residue (or per-token) embedding (Figure 6). Such representations can be used as an alternative way to encode protein residues and present advantages over classic encoding strategies like MSA. First, although training a PLMs requires high computational resources, it has to be done just once; after training, embedding generation is much more cost-efficient than computing a MSA through database search. Additionally, ML models trained on PLMs embeddings are better performing than counterparts trained on MSA (Weissenow and Rost 2025).

Over the years, several PLMs have been developed, differing in their architectures and the size of their training datasets. Table 2 summarizes information about popular PLMs, highlighting the embedding dimension (*D*), number of training sequences and trainable parameters.

Table 2. Details about popular Protein Language Models (PLMs).

PLM	**Embedding dimension (*D*)**	**Training dataset size**	**Trainable parameters**
SecVec	32 - 1,024	33 million	93 million
ProtT5	1,024	2,1 billion	3 billion
ProtGPT2	1,280	50 million	738 million
ESM-1b	1,280	250 million	650 million
ESM2	320 - 5,120	65 million	8 million - 15 billion
ProGen	1,024 - 4,096	280 million	1.2 billion

SecVec (Heinzinger et al. 2019), ProtT5 (Elnaggar et al. 2020), ProtGPT2 (Ferruz et al. 2022), ESM-1b (Rives et al. 2021), ESM2 (Lin et al. 2023), ProGen (Madani et al. 2023).

3.2.3. Embeddings comparison

Although embeddings can represent single residues, a direct comparison between whole protein representations is not trivial due to the different dimensions of the matrices obtained from two proteins with different lengths. To address this issue, an operation called pooling can be used to compress the per-residue representations (Elnaggar et al. 2022). Pooling works by averaging the embedding matrix over the sequence length dimension, resulting in a fixed-size vector $e \in R^{1\times D}$ encoding the whole protein, called per-protein embedding (Figure 6). With this framework, single proteins can be treated as points in a high-dimensional space, and their similarity can be computed by means of distance/similarity metrics, such as cosine similarity, cosine distance, euclidean distance and others. However, despite the advantages, the averaging procedure of pooling leads to a massive loss of information (Schütze et al. 2022; Pantolini et al. 2024).

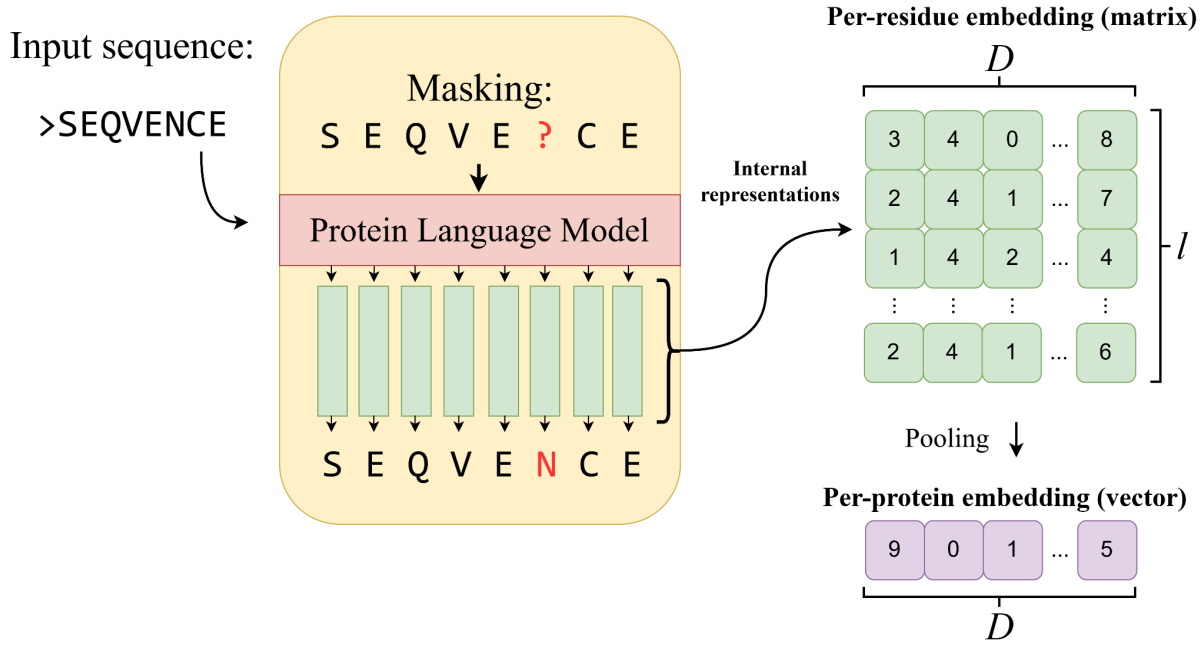


Figure 6. During training the Protein Language Model (PLM) is tasked to reconstruct a corrupted version of the protein in which residues are randomly masked. To achieve this goal the model is forced to derive meaningful and context-aware representations of each residue of the sequence (left). After training (right), these internal representations, referred to as per-residue embeddings, can be used to encode residues of a new sequence. Per-residue embeddings are represented as matrices $e \in R^{l \times D}$, where D depends on the architecture of the PLM used and l indicates the length of the input protein. Through pooling, these representations can be further compressed into compact per-protein embedding vectors $e \in R^{1 \times D}$.

A recent alternative to compare protein embedding is the Embedding-based Alignment (EBA) algorithm (Pantolini et al. 2024), that exploits the full per-residue embedding representation. The algorithm computes a pairwise distance matrix of per-residue embeddings, evaluating the Euclidean distance of all embedded residues. These values fill a matrix of dimension $l_1 \times l_2$ (where l_1 and l_2 are the lengths of the two proteins, respectively) that provides the substitution scores for a pairwise alignment based on a classic dynamic programming approach. The final score s_{align} of the algorithms is then normalized by the length of the longer or shorter sequence of the pair, obtaining two alternative normalized scores: EBA_{min} and EBA_{max} respectively (Figure 7).

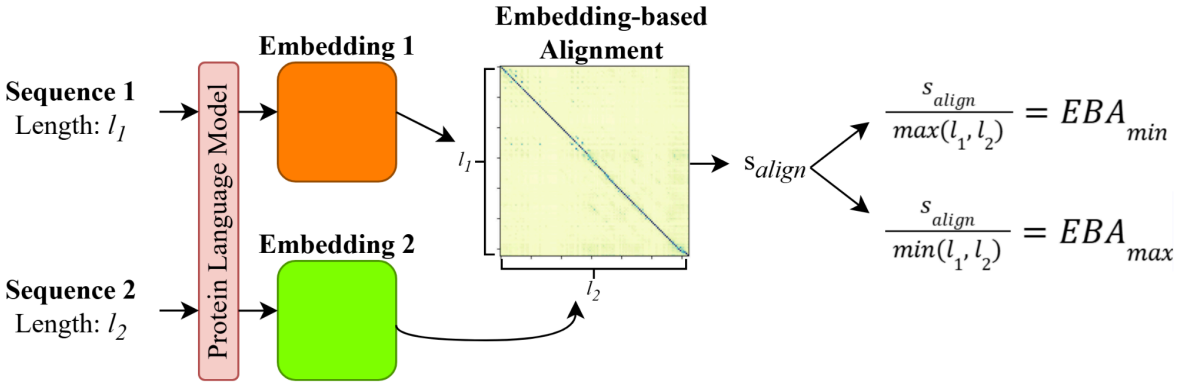


Figure 7. Procedure to compare two protein embeddings with Embedding-based Alignment (EBA). Two protein sequences of length l_1 and l_2 are embedded with a given Protein Language Model. The two embeddings can then be used as input to the EBA algorithm, which performs a dynamic programming alignment. The tool outputs a raw score (s_{align}) that can then be normalized by the length of the longest or shortest sequence of the couple, obtaining two alternative scores: EBA_{min} and EBA_{max}.

Whenever the length difference is large, the resulting higher EBA_{max} reflects the fact that the shorter sequence is entirely contained in the longest. In this cases, EBA_{min} is much lower since the longer sequence is only partially aligned. Therefore, a high EBA_{min} ensures that both sequences are globally aligned.

3.3. Protein structure prediction methods

Deep learning procedures, including transformers and protein language models, are at the basis of the modern breakthrough techniques for addressing the so called protein-folding problem, that is the implementation of a model that allows to predict the protein structure starting from its sequence, addressing the large gap between the number of known protein sequences and experimentally resolved protein structures. In recent years, major advances have been achieved by state-of-the-art deep learning models like AlphaFold2 (Jumper et al. 2021) and ESMFold (Lin et al. 2023), developed by DeepMind and Meta, respectively.

AlphaFold2 showed impressive results at the 14th Critical Assessment of Protein Structure Prediction (CASP14) (Kryshtafovych et al. 2021) and was awarded with the Nobel Prize in Chemistry in 2024. The model, given a protein sequence, integrates an attention-based module, called Evoformer, to extract evolutionary information derived from previously computed MSAs; then, the structure is predicted through a structure module. During inference, the model performs a database search to find structures from homologous sequences to use as templates. Despite the impressive performance the model presents limitations based on the issues regarding MSAs. Moreover, its prediction accuracy is highly impacted by the availability of template structures.

Similarly to AlphaFold2, ESMFold (Lin et al. 2023) adopts a structural fold but replaces the MSA block with ESM2 PLM encoding and a folding trunk; the entire process lacks any structural template search. Given the alignment-free strategy, the model is much faster and cost-efficient than AlphaFold2. For both algorithms, the quality of the predicted models can be assessed using the pLDDT (predicted Local Distance Difference Test) score, which provides a per-residue confidence measure ranging from 0 to 100 (Mariani et al. 2013).

Being based on different input encoding and procedures, the two methods can provide two different models for the same sequence and an important task is to compare them, to provide criteria for their assessment and to identify the proteins and the protein regions where the methods perform similarly or differently. To this aim, a recent database, Alpha&ESMhFold (Manfredi et al. 2024), provides a large-scale comparison between AlphaFold2 and ESMFold models of proteins derived from the human Reference Proteome. When compared to available PDB structures the two methods compute similar models and AlphaFold2 performs slightly better, while when a structural template is unavailable the two computed models can significantly diverge. I collaborated in implementing the updated version of Alpha&ESMhFolds (now under revision for publication), which now includes external

evaluations of the model quality and the possibility to functionally annotate the proteins by mapping Pfam domains (as described in chapter 6).

4. Testing the Capability of Embedding-Based Alignments on the GST Superfamily Classification: The Role of Protein Length

In (Vazzana et al. 2024) we leveraged PLMs to test the capabilities of embeddings on the GST superfamily annotation problem. In the following analysis we compared the GST class annotations obtained using Embedding-based Alignment (EBA) algorithm (Pantolini et al. 2024) with those produced by UniProt's automatic annotation framework, as defined by Association-Rule-Based Annotator (ARBA) and UniRules (MacDougall et al. 2020). ARBA rules assign protein family level annotation using InterPro (Paysan-Lafosse et al. 2023) member databases (including Pfam (Finn et al. 2007; Mistry et al. 2021)), which adopt HMM models to detect occurrences of functional domains along protein sequences.

After the selection of a Reference dataset, we performed a large-scale testing, leveraging the recent ESM2-15b PLM (Lin et al. 2023) and measuring embedding similarity with EBA (see section 3.2.3.). We found that the procedure is successful in sequence annotation, particularly when the sequence length of the proteins is conserved with respect to those included in the Reference set. With this constraint, we classified 26,180 proteins from a dataset of 64,207 unclassified GSTs extracted from UniProt.

4.1. Materials and Methods

4.1.1. Datasets generation

From UniProt (The UniProt Consortium et al. 2023) (release 2024_01), we collected GST proteins including "Glutathione S-transferase" and/or "Glutathione transferase" in the protein name, endowed with PDB structure or high-confidence AlphaFold2 model (with a per-protein

average pLDDT value ≥ 70). We retained a final Reference dataset with 284 proteins (Table 3). With a similar search we collected TrEMBL GST entries without PDB structure; when successful, ARBA rules classify these proteins as belonging to a particular GST class. After additional filtering procedures we retained a Testing set with 15,061 GSTs with class annotation and a Trial set with 64,207 GST proteins without class annotation. The three datasets are available at https://bar.biocomp.unibo.it/GST_Datasets/index.

4.1.2. Embeddings generation

Among the several available PLMs (Asgari and Mofrad 2015; Heinzinger et al. 2019; Elnaggar et al. 2020; Rives et al. 2021), we selected the Evolutionary Scale Modeling (ESM2) family of encoder-only PLMs (developed at Meta), as they represent recent, high-performing, and widely adopted models. Moreover, their embeddings have been used to train ESMFold (Lin et al. 2023). We adopted the ESM2-15b, the biggest model of the ESM2 family with 15 billion parameters, trained on 65 million sequences (Lin et al. 2023). For each protein in the three datasets, we extracted the ESM2-15b representations following the instructions and scripts available at https://github.com/facebookresearch/esm (accessed on 1 November 2022). Given an input protein of length l, ESM2-15b outputs per-residue embeddings with $D = 5,120$; therefore, the final encoding of each sequence is a matrix $e \in R^{l \times 5,120}$.

4.1.3. Embedding Based Alignment for GST annotation

To compare two per-protein embeddings we adopted the Embedding-based Alignment (EBA) method (Pantolini et al. 2024), available at https://git.scicore.unibas.ch/schwede/EBA (See section 3.2.3. for details). Following the author's suggestion (Pantolini et al. 2024) and

considering template structure conservation as an essential element of knowledge transfer (Lesk 2016) we adopted the EBA_{min} value as final score between two protein embeddings.

4.2. Results and Discussion

4.2.1. Fishing for Transfer of Annotation

The annotation transfer procedure is described in Figure 8: we computed the EBA_{min} score between each query protein embedding in the Testing/Trial dataset and all embeddings of the Reference set. The query protein inherits the class annotation of the highest-scoring reference protein, allowing a bait protein to effectively "fish" for its prey.

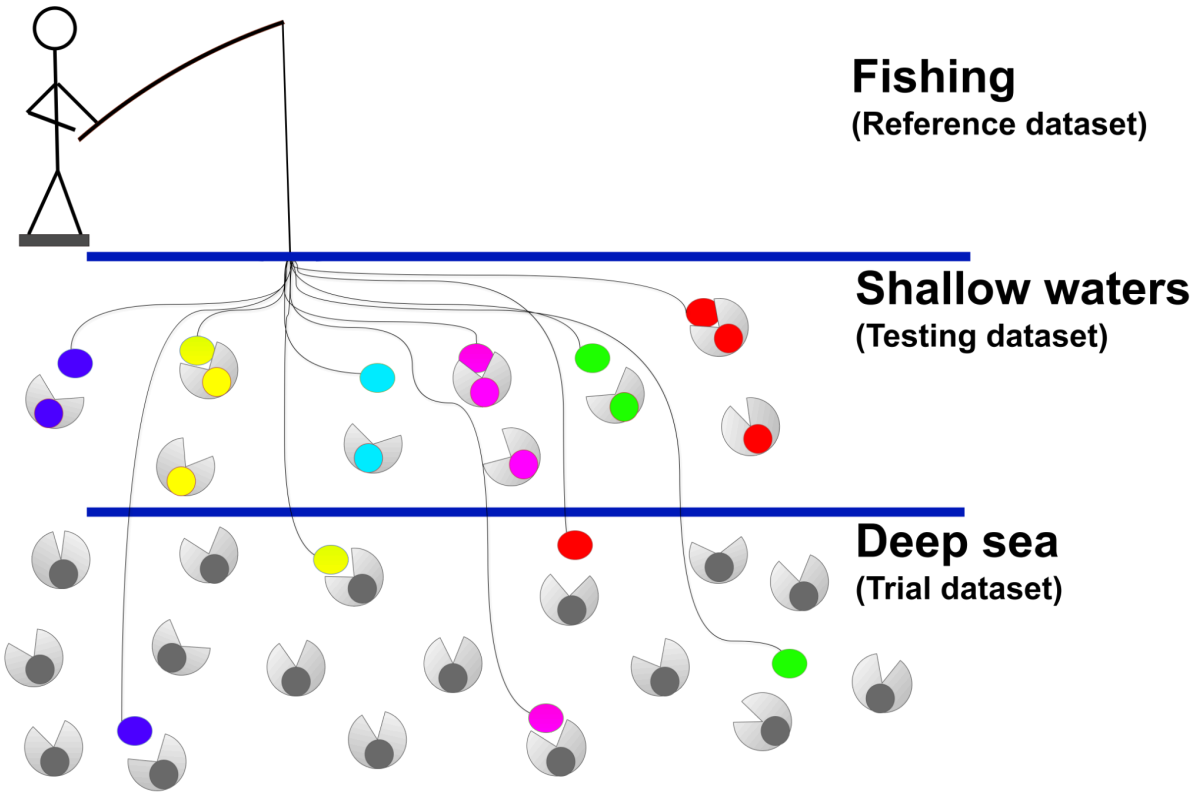
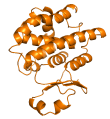


Figure 8. Embedding-based alignment for GST protein classification. "Fhishing" refers to the proteins of the Reference dataset (Table 3) that have been used as "baits" to annotate new proteins. "Shallow waters" refers to the ARBA-classified GSTs of the Testing set (Table 4), whereas "Deep Sea" refers to the ARBA-unclassified GSTs of the Trial set (Table 5). Color matching is indicative of the affinity of baits and preys.

4.2.2. The Reference Dataset

Table 3 shows the Reference dataset, which comprises a total of 284 high-quality GST proteins that have been used as seed of information for the transfer of knowledge. The set is characterized by six folds grouped into 20 GST, with the first 15 classes belonging to the canonical fold (see chapter 2 for details). The entries are listed according to their GST classes (rows) and taxonomic groups (columns, as reported in NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi)). The "Length" column shows the minimum and maximum protein lengths observed for each GST class. This range defines the "Reference Length Region" (RLR), which corresponds to the length interval ensuring fold conservation within the reference proteins of a given GST class. The "Seq. Id. Range (%)" column shows the minimum and maximum sequence identity found within a given class. The "Fold" column shows the crystal structure of one of the entries selected as representative of the fold group. From the table it is evident the taxon-specificity of some GST classes: for instance Tau, Lambda, Phi and DHAR are only found in plants (Viridiplantae) and Beta GSTs are only found in Bacteria.

Table 3. The Reference dataset with 284 sequences

Classes	Bact.	Amoeb.	Fungi	Virid.	Plat.	Nem.	Arth.	Moll.	Actin.	Amph.	Aves	Mamm.	Total class	Length	Seq. Id. Range (%)	Fold
Mu	-	-	-	-	11 (9*)	-	3 (2*)	-	-	-	1 (1)	28 (8*)	43 (20*)	211-225	22-98	
Sigma	-	-	-	-	-	9 (2)	7 (4*)	1 (1)	-	-	1	3 (2)	21 (9*)	199-249	25-94	
Alpha	-	1	-	-	-	-	-	-	-	-	2 (1)	18 (10*)	21 (11*)	222-229	29-96	
Pi	-	-	-	-	-	5 (2*)	-	-	-	2	-	12 (3)	19 (5*)	207-210	32-99	
Theta	-	-	-	-	-	-	-	-	-	-	-	12 (3)	12 (3)	240-244	40-99	
Delta-Epsilon	-	-	-	-	-	-	32 (15*)	-	-	-	-	-	32 (15*)	208-271	25-99	
Omega	-	-	-	-	-	3	2 (2*)	-	-	-	-	7 (2)	12 (4*)	240-256	23-93	
Zeta	3 (3*)	-	1 (1)	3 (1)	-	1	-	-	-	-	-	3 (2)	11 (7*)	212-221	33-95	
Rho	-	-	-	-	-	-	-	1 (1*)	1	-	-	-	2 (1*)	223-225	41	(8GSS)
DHAR	-	-	-	3 (3)	-	-	-	-	-	-	-	-	3 (3)	213-213	66-76	
Tau	-	-	-	34 (5*)	-	-	-	-	-	-	-	-	34 (5*)	217-231	30-98	
Phi	-	-	-	25 (11*)	-	-	-	-	-	-	-	-	25 (11*)	212-221	31-95	
Lambda	-	-	-	3	-	-	-	-	-	-	-	-	3	235-237	56-73	
Beta	4 (3)	-	-	-	-	-	-	-	-	-	-	-	4 (3)	201-203	36-54	
HSP26	3 (3)	-	-	-	-	-	-	-	-	-	-	-	3 (3)	202-212	22-60	
Omega-like	-	-	4 (1)	-	-	-	-	-	-	-	-	-	4 (1)	313-370	44-63	 (5LKD)
FosA	2 (2)	-	-	-	-	-	-	-	-	-	-	-	2 (2)	135-141	59	 (1NPB)
LanC	-	-	-	-	-	-	-	-	1	-	-	4 (1)	5 (1)	399-405	63-96	 (8D19)
Kappa	-	-	-	-	-	2	-	-	-	-	-	3 (2)	5 (2)	225-226	28-86	 (3RPN)
MAPEG	-	1	-	-	-	-	-	-	-	-	-	22 (5)	23 (5)	146-155	12-98	 (4AL0)
Total taxon	12 (11*)	2	5 (2)	68 (20*)	11 (9*)	20 (4*)	44 (23*)	2 (2*)	2	2	4 (2)	112 (38*)	284 (111*)			

The 284 proteins are listed by classes (rows) and taxonomic groups (columns). The first 15 classes belong to the canonical fold GSTs. The number of proteins with a PDB reference is specified inside round brackets; (*) indicates that at least one entry in the set belongs to TrEMBL. Dashed horizontal lines discriminate classes in the same sub-cellular location. The "Length" column displays the shortest and the longest protein sequence found in each class. The Seq.Id. (%) column shows the minimum and maximum sequence identity percentage found within each class (for classes with only two representatives, the sequence identity between the two is shown). Abbreviations used: Bact., Bacteria; Amoeb., Amoebozoa; Virid., Viridiplantae; Plat., Platyhelminthes; Nem., Nematoda; Arth., Arthropoda; Moll., Mollusca; Actin., Actinopterygii; Amph., Amphibia; Mamm., Mammalia.

4.2.3. Testing and Trial Datasets

Tables 4 and 5 show the classification results of the Testing and Trial datasets, respectively. The 15,061 GSTs of the Testing dataset are endowed with a class annotation provided by ARBA rules, which leverages the presence of specific features, like InterPro signatures, within a specific taxa to infer GST class annotation. As an example, the GST Omega-class annotation can be transferred if the sequence matches two InterPro signatures (IPR004045 and IPR050983) and the organism is a Metazoa (https://www.uniprot.org/arba/ARBA00011067). To note that ARBA-annotated sequences of this dataset can be shorter or longer than sequences of the Reference dataset belonging to the same class.

The Trial dataset is endowed with 64,207 proteins that lack any class annotation, indicating that ARBA rules were unable, for these GSTs, to infer a specific class.

4.2.4. Testing the Embedding Alignment Method in Shallow waters

The main difference between ARBA and EBA is that ARBA classifies proteins after finding conserved domains and/or motifs that are typical of the GST superfamily, without any constraint on the sequence length of the protein, while EBA considers the pairwise global alignment of two embedded sequences. Adopting proteins of the Reference dataset as seed of information, we followed the knowledge that a transfer of classification is reliable when the protein fold is conserved (Lesk 2016), and this implies that the protein length is conserved. With this respect, classification results on the Testing dataset are shown in Table 4, which is divided into three column blocks according to the length of the specific ARBA-annotated entry with respect to the Reference Length Region (RLR) of the same class: within (RLR), below (<RLR) or above (>RLR) the Reference Length Region. This division identifies the

fraction of the GST proteins (within the range of length of the reference proteins, RLR) which conserve the structure. The table reports for each GST class (rows): the UniProt ARBA classifications (“Exp” columns), the EBA derived classifications (“Pred” columns), the length of the shorter and longer sequence of the block (RLR, <RLR and > RLR columns), their range of sequence identity (“Pred SI” columns) and, in the ” <RLR” and “>RLR” blocks, the number of errors (“Errors” columns). In the RLR subset we identified 76 remote homologs with respect to the Reference set distributed in four classes (Mu, Sigma, Omega and Kappa, see https://bar.biocomp.unibo.it/GST_Datasets/index.htm). In the RLR subset, only two Mu-class GSTs are misclassified as Sigma-class by EBA: UniProt IDs: A0A1I8FWQ8 and A0A1I8J3A2. However, sequence comparison of the two proteins with the Sigma and Mu-class reference proteins indicates that they share higher sequence identity with the Sigma than Mu GST reference proteins (33% and 28% sequence identity, respectively). Out of the 3D conservation range, the method is any way successful: the prediction accuracy of our method with respect to ARBA in classifying proteins is very high (99.3%).

Table 4. Compare EBA with ARBA classification in the "shallow waters" of the Testing dataset GSTs.

	ARBA	Within Reference Length Range (RLR)				Below Reference Length Range (< RLR)					Above Reference Range (>RLR)				
Class	Total	Exp	Pred		Pred SI	Exp	Pred		Pred SI	Errors	Exp	Pred		Pred SI	Errors
	(#)	(#)	(#)	RLR	(%)	(#)	(#)	<RLR	(%)	(#)	(#)	(#)	>RLR	(%)	(#)
Mu	1706	981	979	211-225	10-99	355	349	140-210	8-99	6	370	335	226-475	9-99	35
Sigma	694	592	592	199-249	20-99	66	66	109-198	21-99	-	36	36	250-499	3-99	-
Alpha	1520	734	734	222-229	17-99	495	471	113-221	13-99	24	291	289	230-487	10-99	2
Pi	609	323	323	207-210	24-99	158	158	120-206	21-99	-	128	124	211-488	20-99	4
Theta	1428	560	560	240-244	25-99	545	540	104-239	16-99	5	323	323	245-491	1-99	-
Delta-Epsilon	822	715	715	208-271	18-99	68	68	102-207	15-99	-	39	39	272-478	10-99	-
Omega	1349	556	556	240-256	11-99	617	597	101-239	6-99	20	176	176	257-474	10-99	-
Zeta	728	268	268	212-221	26-99	122	122	139-211	22-99	-	338	338	222-433	2-99	-
DHAR	10	-	-	213-213	-	7	7	107-212	6-97	-	3	3	214-465	37-99	-
Tau	1342	851	851	217-231	23-99	222	219	202-216	6-99	3	269	268	232-449	3-99	1
Phi	1711	1066	1066	212-221	25-99	177	177	149-211	20-99	-	468	468	222-491	2-99	-
HSP26	433	363	363	202-212	33-99	48	48	196-211	40-99	-	22	22	213-227	40-99	-
LanC	450	177	177	399-405	45-99	109	109	126-398	4-99	-	164	164	406-490	32-99	-
Kappa	1148	230	230	225-226	17-99	554	554	189-224	12-99	-	364	364	227-257	12-99	-
MAPEG	1111	687	687	146-155	7-99	143	143	101-145	3-99	-	281	281	157-363	6-99	-
Total	15,061	8103	8101			3686	3628			58	3272	3230			42

The Testing set includes 15,061 GST proteins classified by the ARBA rule system. From Table 3 we derived the reference length range (RLR) of GST proteins with a reference fold per each class. For the sake of fold conservation, we grouped GST proteins with a length included in the range (RLR), below the range (<RLR) and above range (>RLR). We also show the range of sequence identity per EBA-found GST class (Pred SI). EBA errors are mainly found on <RLR and >RLR regions. Exp = ARBA expected; Pred = EBA classification. Pred SI = Sequence identity among predicted and reference class (Table 3); # = Number of.

4.2.5. Fishing in the Deep Sea

We adopted EBA to classify 64,207 ARBA-unclassified GST proteins (Table 5). We show the results of each GST class (rows) and taxon (columns) only for proteins whose length is in the length range of fold conservation ("Within RLR" row). In this subset we were able to classify 26,180 proteins (41% of the total). Out of the Reference Length Region we could transfer class to another 58% GST proteins ("Below RLR" and "Above RLR" rows). The range of classes seems to increase, particularly in GST proteins from bacteria, and this can be explained by considering the new bacterial genomes recently included in TrEMBL.

Overall, we propose the EBA classification procedure as a valid annotation system, considering that sequence embeddings carry along information on structural templates, motifs and domains of the family, and are able to recover remote homologs.

Table 5. Classifying GST proteins in the “deep sea” with the Embedding-based alignment method

Class	Bacteria	Amoeb.	Fungi	Virid.	Plat.	Nematoda	Arth.	Moll.	Actin.	Amph.	Aves	Mamm.	Others	Total class
Mu	5	-	33	4	6	12	74	7	5	-	-	7	96	249
Sigma	30	-	87	14	-	480	133	82	11	3	73	130	376	1419
Alpha	21	-	13	7	-	11	1	1	2	-	1	6	49	112
Pi	-	-	37	2	-	6	4	-	-	1	-	4	33	87
Theta	13	-	-	10	-	-	1	3	2	-	-	30	5	64
Delta-Epsilon	1949	9	498	8	-	-	1642	1	5	-	-	4	74	4190
Omega	87	-	112	5	1	-	1	-	-	-	-	-	10	216
Zeta	1397	-	22	7	-	-	-	-	-	-	1	-	21	1448
Rho	524	-	22	-	-	-	-	5	197	-	-	-	9	757
DHAR	1	-	-	-	-	-	-	-	-	-	-	-	-	1
Tau	1555	-	30	2401	-	1	-	-	1	-	-	-	41	4029
Phi	3694	5	772	60	-	-	-	1	1	-	-	1	57	4591
Lambda	-	-	3	63	-	-	-	-	-	-	-	-	-	66
Beta	1569	-	4	-	-	-	1	-	-	-	-	-	15	1589
HSP26	2539	-	9	1	-	-	-	-	-	-	-	-	27	2576
Omega-like	2746	1	298	39	-	-	2	-	4	-	2	21	200	3313
FosA	306	-	-	-	-	-	-	-	-	-	-	-	2	308
LanC	-	-	-	-	-	-	-	-	-	-	-	-	1	1
Kappa	-	-	8	-	-	-	-	-	1	-	-	-	-	9
MAPEG	39	1	230	108	3	-	347	24	141	11	48	44	159	1155
Within RLR	16475	16	2178	2729	10	510	2206	124	370	15	125	247	1175	26180
Below RLR	11288	6	626	1493	66	237	443	52	426	39	159	509	677	16021
Above RLR	12917	22	3743	2260	18	119	388	45	342	15	56	148	1007	21080
Total per taxon	41075	44	6582	6851	99	883	3063	222	1157	69	345	928	2889	64207

The trial set contains 64,207 GST ARBA-unclassified proteins. GSTs classes are distributed by rows, whereas taxonomic groups are distributed by columns. The EBA method classifies 26,180 proteins whose range of lengths (within RLR) ensures structure conservation with respect to baits.

5. Can Human Canonical Glutathione S-Transferases Act as RNA-Binding Proteins?

The interaction between RNA and RNA-binding proteins (RBPs) play crucial roles in many aspects of RNA metabolism, including splicing, transport, translation, localization, stability and degradation (Curtis and Jeffery 2021). Recent RNA–interactome capture studies (RIC) (Castello et al. 2016) have revealed several proteins with previously unrecognized RNA-binding activity. With this respect, the RBP2GO database (Caudron-Herger et al. 2021) collects newly identified RBPs candidates from 13 organisms, including *Homo sapiens*. Such evidence suggests that certain proteins, already known for their important biological roles, can perform multiple functions by "moonlighting" as RBPs (Curtis and Jeffery 2021). Additionally, a novel study (Li et al. 2024) addressing the role of RBPs in mammalian spermatogenesis, has identified new candidate RBPs in mouse male germ cells (mMGCs). From this starting point (Li et al. 2024) we focused on three candidate RBPs belonging to the GST superfamily and specifically members of the canonical Mu-class GSTs. To our knowledge, there is no annotation record of interaction between GSTs and RNA. Based on this, we found two human Mu-class GSTs homologous to the three mouse proteins in the RBP2GO database. In addition, we found three more canonical GSTs belonging to the Pi, Omega and Zeta classes in the RBP2GO database, for a total of five human GSTs possibly behaving like RNA-binding proteins, given the evidence that all canonical GSTs share the same structure (Mazari et al. 2023). Overall, these studies suggest that canonical GSTs may bind RNA, although they do not provide information regarding the RNA species involved or the molecular details of the interaction. Therefore, we conducted a computational analysis of the RNA-binding sites on human GSTs by leveraging recent machine learning algorithms and docking programs (Vazzana et al. 2026). Our findings suggest that a GST–RNA interaction is

physically reliable and hints for a potential overlap between RNA-binding sites and residues responsible for binding with GSH.

5.1. Materials and Methods

5.1.1. RBP2GO Database

RBP2GO database (Caudron-Herger et al. 2021) (available at https://rbp2go.dkfz.de/ , last release August 2023) ranks proteins based on their occurrence in recent RNA proteomic studies. The database defines an RBP2GO score that ranges from 0 to 100, reflecting the probability of a protein to be an RBP. The score is calculated considering two indicators of RBP propensity: (i) the frequency of a given protein to be listed as RBP in the different datasets stored in the database and (ii) the average frequency of interactions with RBP proteins according to STRING (Szklarczyk et al. 2023) (https://string-db.org/).

5.1.2. Prediction of RNA-binding sites and Molecular Docking validation

We adopted two high-performing structure-based models to predict RNA-binding sites on the candidate RBPs GSTs: PST-PRNA (Li and Liu 2022) and GraphBind (Xia et al. 2021). Both methods adopt deep learning architectures that employ a convolutional neural network (PST-PRNA) and a hierarchical graph neural network (GraphBind). We validated the predictions obtained with Molecular Docking: this technique leverages computational methods to approximate a ligand–receptor complex with an energetically favorable geometry, aiming to (ideally) replicate an experimental binding mode (Meng et al. 2011). For our annotation task we adopted AutoDock Vina (Trott and Olson 2010; Eberhardt et al. 2021), a very popular molecular docking program that, in addition to the complex conformation, computes the binding affinity between the receptor and the ligand. During data preparation, the user can define a confined search space with a grid box on the receptor molecule to help

the program dock the ligand on a preferred region, which is useful when there is prior knowledge about putative binding sites. In our experiments, we adopted a sufficiently wide grid box to accommodate both GSH and RNA molecules.

5.1.3. Surface charge characterization

RNA molecules are polyanions due to the presence of negatively charged phosphate groups; consequently, most protein–RNA interfaces are enriched in positively charged residues (Corley et al. 2020). We characterized the electrostatic properties of protein surfaces using two tools, PDB2PQR and adaptive Poisson-Boltzmann solver (APBS) (Jurrus et al. 2018), directly from the web server (https://www.poissonboltzmann.org/). PDB2PQR prepares the structures for continuum solvation calculations and computes the protonation state of residues, whereas APBS computes the electrostatics of a protein by solving the equations of continuum electrostatics.

5.2. Results and Discussion

5.2.1. Human GSTs in RBP2GO

The five human canonical GSTs identified as candidate RBPs in the RBP2GO database are shown in Table 6. The best-ranking human GSTs are GSTP1 and GSTO1, with RBP2GO scores of 10.3 and 6.5, respectively. Even with low scores, the proteins appear in multiple RNA–proteomic studies: seven for GSTP1 (Beckmann et al. 2015; Milek et al. 2017; Mullari et al. 2017; Bao et al. 2018; Huang et al. 2018; Urdaneta et al. 2019; Backlund et al. 2020) and four for GSTO1 (Mullari et al. 2017; Queiroz et al. 2019; Urdaneta et al. 2019; Backlund et al. 2020). GSTP1 is of particular interest as it is a well-studied protein that is highly involved in several aspects of human health (Goto et al. 2009; Simic et al. 2022).

Table 6. Candidate RNA binding human GSTs in RBP2GO.

Gene Name	UniProt ID	Protein Name	[1]Class	[2]PDB	[3]RBP2GO score	[4]Pub No
GSTP1	P09211	Glutathione S-transferase P	Pi	8GSS chain A	10.3	7
GSTO1	P78417	Glutathione S-transferase omega-1	Omega	5YVN chain A	6.5	4
MAAI (GSTZ1)	O43708	Maleylacetoacetate isomerase	Zeta	1FW1 chain A	3.8	2
GSTM2	P28161	Glutathione S-transferase Mu 2	Mu	1XW5 chain A	3.1	2
GSTM3	P21266	Glutathione S-transferase Mu 3	Mu	3GTU chain B	2	1

[1]Class: refers to the class of the GST protein in UniProt.
[2]PDB: identifiers of the PDB representatives of the GST proteins.
[3]RBP2GO-score: entries are ranked according to the RBP2GO-score (described in section 5.1.1.).
[4]Pub No: number of publications in which the protein has been reported.

5.2.2. Prediction of RNA-binding sites in GST proteins

The predictions of the eight proteins (three mouse GSTs from (Li et al. 2024) and five human GSTs from Table 6) are shown in Figure 9 and Figure 10: the former shows the predicted sites on the protein sequences (displayed on a MSA obtained from multiple structural alignment computed with Foldmason (Gilchrist et al. 2024) available at https://search.foldseek.com/foldmason), whereas the latter shows the same predictions projected in the 3D structures. In Figure 9, it appears that GSH-binding residues (in red) are consistently predicted as RNA-binding sites by the two methods in the N-terminal domain of the proteins (which contains the GSH-binding site, G-site). Figure 10 shows that

RNA-binding site predictions are found close to the G-site, and that sites predicted in the C-terminal domain are mainly located in the first helix (first C-ter helix in Figure 10), which is close to the G-site in the 3D structure of the enzymes.

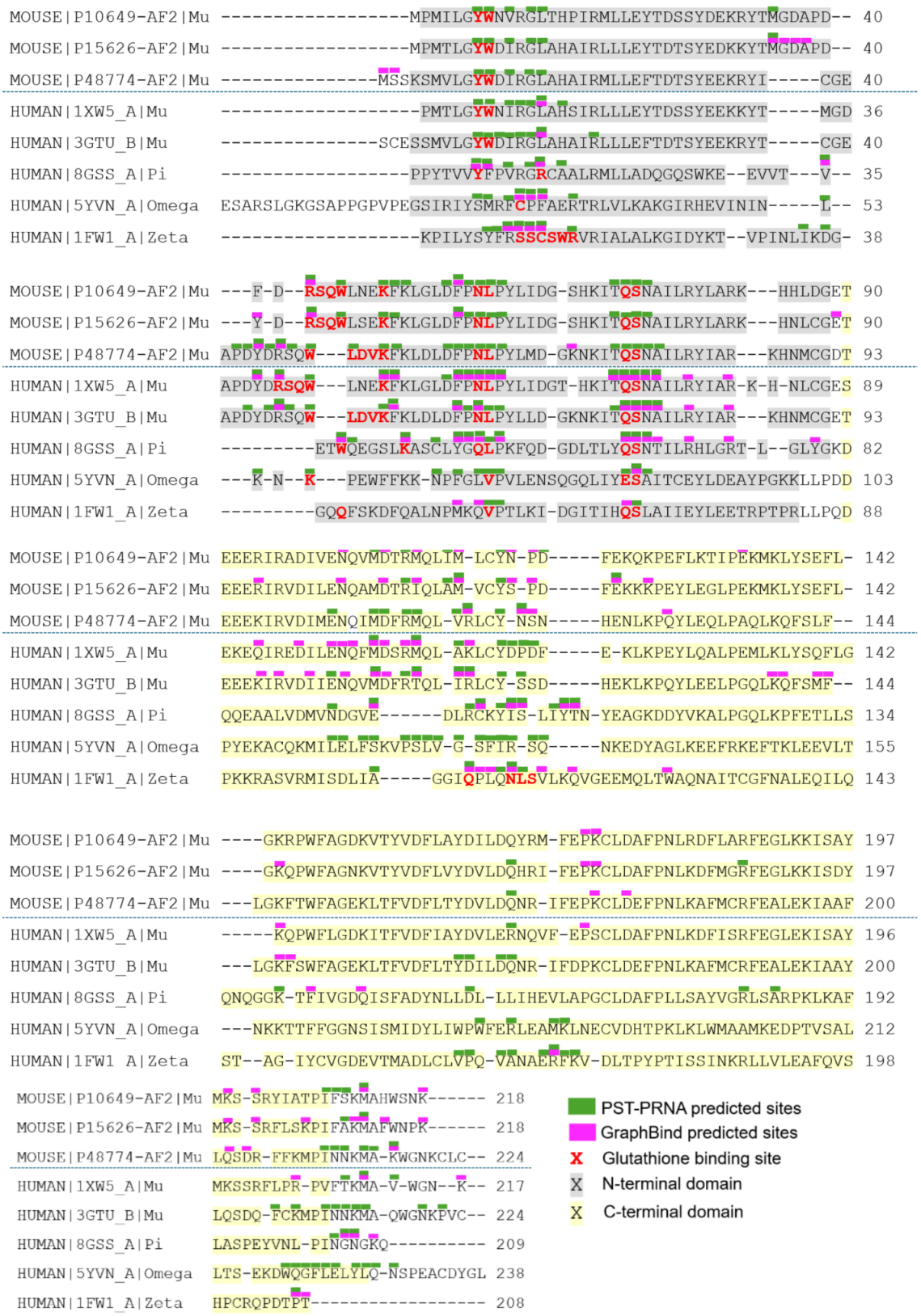


Figure 9. MSA of mouse and human GSTs after multiple structural alignment. High quality AlphaFold2 models are used for mouse GSTs. GSH-binding residues are in red. RNA-binding residues predicted with PST-PRNA and GraphBind, are colored in green and magenta, respectively. Residues belonging to the N and C-terminal domains are grey and yellow-shaded, respectively.

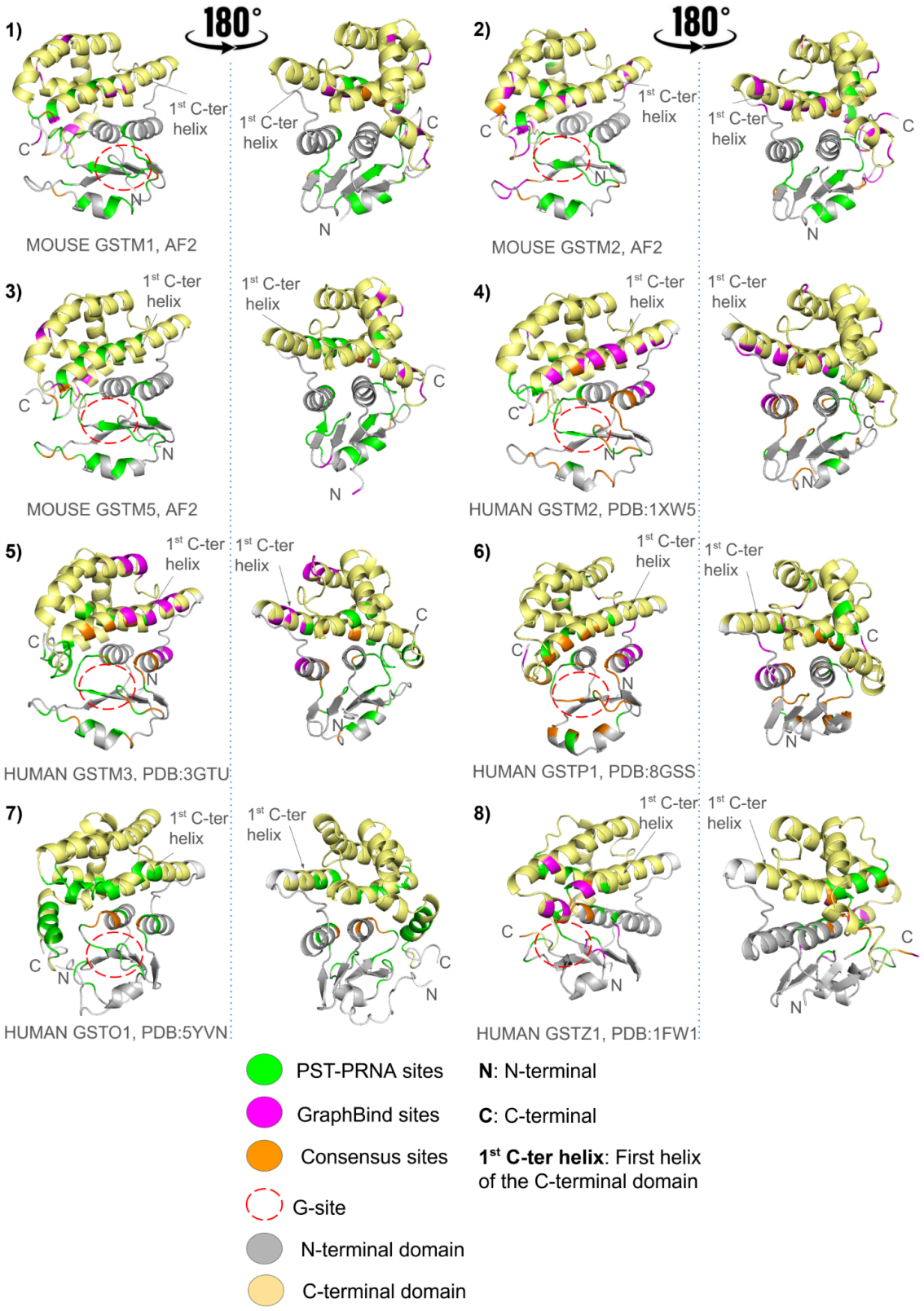
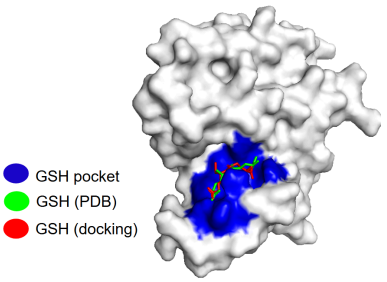
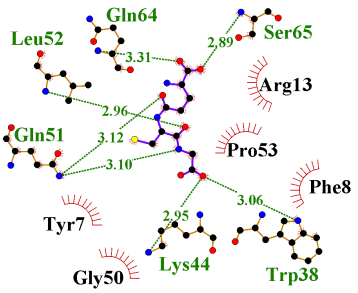
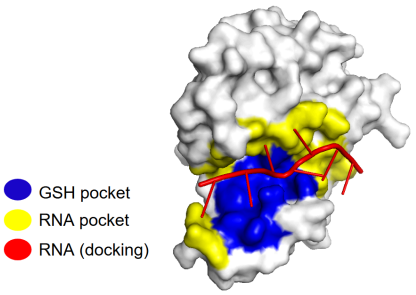
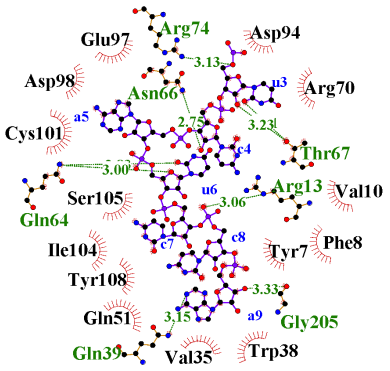
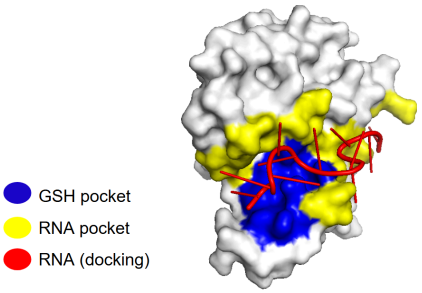
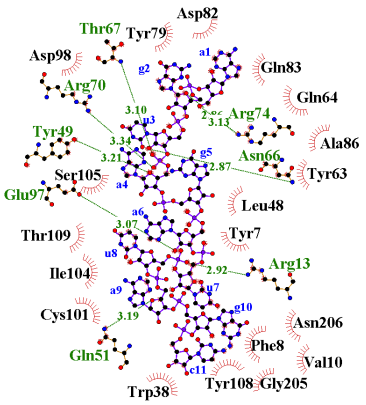
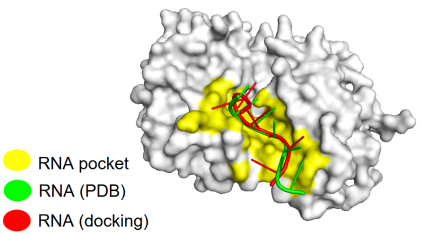
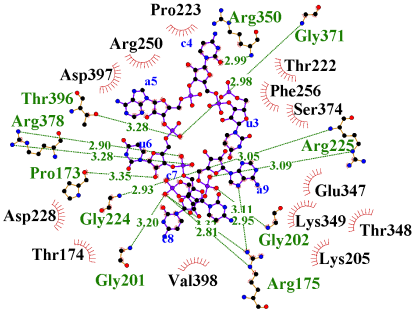


Figure 10. Predictions of Figure 9 projected on 3D models. PST-PRNA, GraphBind and consensus predicted RNA-binding residues, are colored in green, magenta and orange, respectively. The backbone is color-coded in grey (N-terminal domain) and yellow (C-terminal domain). Protein structures are shown from two perspectives: i) facing the side that displays the G-site (left), and after a 180° rotation with respect to the G-site face (right).

5.2.3. Docking validation

No clear identification of RNA sequences interacting with the GSTs has been reported in RNA proteomic studies of Table 6. To select a proper RNA substrate for the docking validation procedure we considered that short RNA molecules would better simulate protein–RNA interactions and lighten the docking procedure. We found two human miRNA molecules co-crystallized with the respective binding proteins: the 7 nucleotides primiR-18a-oligo1 in complex with DDX17 (PDB ID: 6UV4) (Ngo et al. 2019), and the 11 nucleotides pri-miRNA-18a terminal loop in complex with hnRNP A1 (PDB ID: 6DCL) (Kooshapur et al. 2018). For the docking experiment we selected GSTP1 as representative GST, being the highest-scoring human GST in the RBP2GO database (see Table 6). For sake of comparison, we performed a couple of redocking experiments (i.e., docking experimentally known binding partners): GSTP1 with GSH and the 7 nucleotides primiR-18a-oligo1 with its co-crystallized binding partner (DDX17), shown in Table 7. GST-RNA docking was performed both on the monomer and dimer of GSTP1, and the results are shown in Table 7 and Figure 11, respectively.

Table 7. Validating GST-RNA interaction with AutoDock Vina

[1]Docking partners	[2]AutoDock Best pose and Mean Affinity (kcal/mol)	[3]Literature Affinity (kcal/mol)	[4]Docking results	[5]LigPlot+
*GST Pi (8GSS) + GSH (8GSS) Redocking	-5.1 (Mean -4.7±0.2)	-5.3 ᐞ -4.6 ᵚ		
°GST Pi (8GSS) + miRNA (6UV4)	-8.5 (Mean -8.0±0.2)	Not available		
°GST Pi (8GSS) + miRNA (6DCL)	-8.6 (Mean -8.1±0.2)	Not available		
*DDX17 (6UV4) + miRNA (6UV4) Redocking	-9.6 (Mean -9.5±0.1)	-8.8 ᵀ		

Legend to Table 7.

[1]Docking partners supported by structural evidence in the PDB are tagged with (*) whereas docking partners with no structural support in the PDB are tagged with (°).

[2]Affinity of the best pose (among 20 different ones) and the mean affinity (with standard deviation) across poses computed by AutoDock Vina.

[3]Affinity computed from the experimental K_d values reported in literature.

[Δ]Affinity values computed on the dimer (Lo Bello et al. 1997).

[ω]Affinity values computed on the monomer (Fabrini et al. 2009).

[τ]Affinity values reported in (Ngo et al. 2019).

[4]Docking results: the interacting residues are derived from the 2D diagram (obtained with LigPlot+ (Laskowski and Swindells 2011)), and define the blue and yellow pockets of the GSH and RNA, respectively. Ligands of the original PDB, when available, are colored in green and docked ligands are colored in red. Interacting residues of docked structure with structural evidence show high overlap with the PDB counterpart.

[5]LigPlot+ 2D diagram of the best pose.

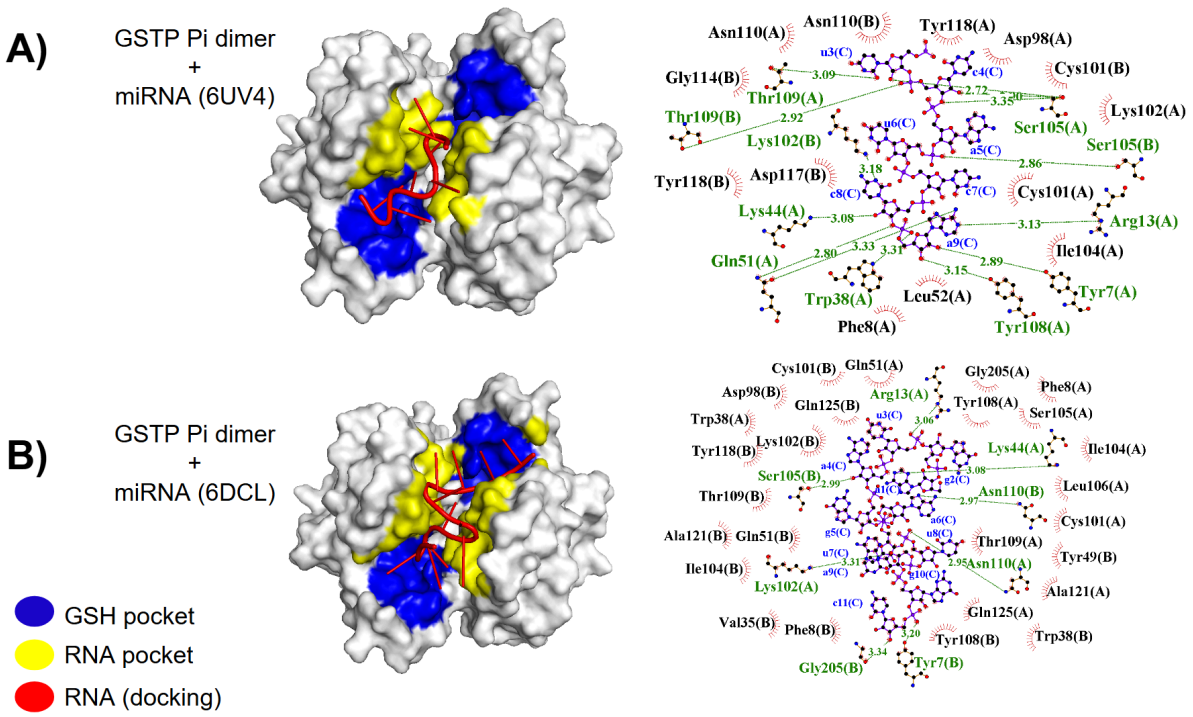


Figure 11. Docking obtained with AutoDock Vina on the GSTP1 dimer (PDB: 8GSS). LigPlot+ 2D diagrams are reported on the right side of the respective docking results. The interacting residues are derived from the corresponding LigPlot+ 2D diagram and define the blue and yellow pockets of the GSH and RNA, respectively; miRNA molecules are colored in red.

A) Best pose docking results with the 7 nucleotide-containing miRNA. Best pose: Affinity -9.3 kcal/mol. Overlapping residues with GSH-binding sites: Tyr7, Phe8, Arg13, Trp38, Lys44, Gln51, Leu52 (Chain A).

B) Best pose docking results with the 11 nucleotide-containing miRNA. Best pose: Affinity -9.6 kcal/mol. Overlapping residues with GSH-binding sites: Phe8, Arg13, Trp38, Lys44, Gln51 (Chain A) and Tyr7, Phe8, Trp38, Gln51 (Chain B).

Overall, deep learning models predictions and molecular docking are coherent, and RNA molecules successfully dock into the G-site pocket. Moreover, among the 11 residues binding GSH in the GST Pi+GSH complex, a total of six residues are also in interacting behavior in both GST Pi+miRNA docked complexes: Tyr7, Phe8, Arg13, Trp38, Gln51 and Gln64. Analysing the interaction diagram, we observed that the GST-RNA binding interactions mainly involve the backbone of the RNA molecule, either the phosphate group or ribose sugar, hinting for nonspecific interactions. Computed binding affinities were similar for both miRNA molecules and similar to the affinity of DDX17 with its miRNA (see Table 7). Favorable binding affinities have been also computed on the GST Pi dimer+miRNA (Figure 11). In Figure 12 protein surface charge reveals that the G-site of GSTs is positively charged, similarly to the RNA-binding pocket of DDX17.

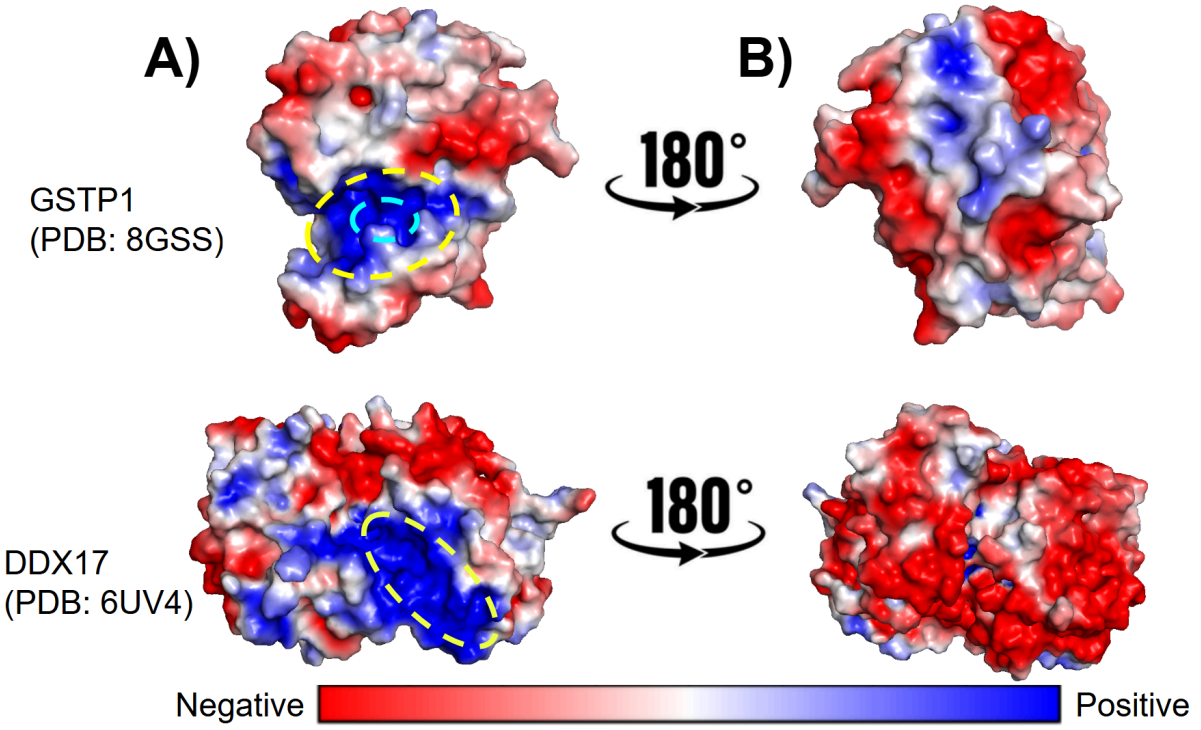


Figure 12. Structure surface charges comparison between GSTP1 (PDB ID: 8GSS chain A) and DDX17 (PDB ID: 6UV4 chain A). Electrostatic potentials are on a [-3, 3] red–white–blue colormap in units of kJ/mol. A). On the left side, cyan and yellow circles highlight the G-site pocket and RNA binding surface, respectively, of GSTP1 and DDX17. Both GSTP1 G-site pocket and DDX17 RNA-binding region show a positively charged surface. B) The right side shows the two structures rotated by 180°. Interestingly positive charges appear only on GSTP1 accessible surface, consistently with the residues predicted as RNA-binding sites by the predictors (Figure 9 and Figure 10).

These findings suggest that canonical GST-RNA interaction is possible, and that, under stress conditions, the increase in the concentration of GSH (up to millimolar concentration (Lu 2013)) and the increased expression of GSTs (Lv et al. 2023)) favor the protein role as glutathione-S transferases (Fabrini et al. 2009).

6. AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation

Alpha&ESMhFolds is a web server (Manfredi et al. 2024) collecting AlphaFold2 (Jumper et al. 2021) and ESMFold (Lin et al. 2023) models of 42,942 proteins extracted from the human Reference Proteome (UP000005640, available at UniProt (The UniProt Consortium et al. 2023)).

The original web server provides the possibility to analyse the structural superimposition of the models and the induced sequence alignment. Moreover, it presents statistics on the residue-wise quality scores computed by the two methods.

The main metric adopted for the structural comparison is the TM-score (Zhang and Skolnick 2005), a length-independent value ranging between 0 and 1, defined in Eq. 1:

$$\text{TM-score} = Max\left[\frac{1}{L_{target}}\sum_{i}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_{0\left(L_{target}\right)}}\right)}\right] \quad (1)$$

Where L_{target} is the length of target protein, L_{ali} is the number of aligned residues, d_i is the distance between the *i*th pair of aligned residues and $d_0\left(L_{target}\right)$ is a predefined parameter that normalizes the distance. A TM-score higher than 0.6 indicates a good superimposition between the compared structures.

Moreover, the database reports the model quality with the pLDDT scores associated with each residue (Mariani et al. 2013; Jumper et al. 2021). This score (ranging between 0 and

100) measures the confidence in the local structure, estimating how well the prediction would agree with an experimental structure (Mariani et al. 2013). In the database, pLDDT values are independently computed on AlphaFold2 and ESMFold; residues with pLDDT higher than 70 are routinely considered as modelled with high confidence.

The analysis of the models reveals that when an experimental structure is available in the PDB (for 2900 proteins), the models are of good quality and largely overlap; when compared with the experimental structures, AlphaFold2 models are slightly better. On the contrary, when templates are not available, the two methods compute models that can largely diverge. An interesting question is whether, despite this divergence, the resulting protein domain structures still superimpose. In (Manfredi et al. 2025) we carried a deep analysis on the structural difference between Pfam domains (Finn et al. 2007; Mistry et al. 2021) modeled by AlphaFold2 and ESMFold. We mapped available Pfam domains onto pairs of enzyme models generated with both methods and compared the corresponding regions relevant to functional annotation. Models were compared by means of TM-score and pLDDT, considering both the full models and the mapped Pfam domain regions. We find that, regardless of the global structural alignment of the pairwise models, Pfam-containing regions have a good superimposition and a pLDDT which is higher than the rest of the modeled sequence. This indicates that both methods are similarly performing in modeled regions overlapping with Pfam domains.

We recently released an updated version of Alpha&ESMhFolds, whose publication is under review, that includes the possibility to analyse and compare the structures modelled for all the Pfam domains present on the protein.

6.1. Materials and Methods

6.1.1. The Dataset

The dataset adopted for the analysis comprises 6956 human enzymes; a fraction of the total database (1314 proteins of the 6956) is endowed with a PDB structure covering at least 70% of the protein sequence. Among the remaining 5642 proteins, 3037 are included in Swiss-Prot while the remaining 2605 are listed in TrEMBL. For each enzyme, we extract the set of annotations present in Pfam (Mistry et al. 2021) downloading data available at the website (https://pfam-docs.readthedocs.io/en/latest/, version 37.0 accessed in September 2024). We collected 14,122 Pfam entries, including 9834 domains, 3391 families, 769 repeats, 93 short motifs, 19 conserved intrinsically disordered regions, and 16 coiled-coil regions, all documented also in InterPro (https://www.ebi.ac.uk/interpro/). Additionally, we ran the PfamScan tool to annotate 2578 Pfam entries with a reported active site. In total, 5684 residues were determined to be part of an active site.

6.1.2. Models comparison

For each protein, the Alpha&ESMhFolds database provides the per-residue pLDDT of the AlphaFold2 and ESMFold models (which measures the reliability of the predicted model (Mariani et al. 2013), ranging from 0 to 100), as well as the TM-score between the two models (a metric, ranging from 0 to 1, that estimates the similarity between the two 3D models (Zhang and Skolnick 2004)). We computed the local TM-score of the regions of the proteins that are covered by a Pfam entry obtained by manually extracting the corresponding segments from the predicted models. Evaluation was performed with the Foldseek program (Van Kempen et al. 2024), which computes the structural superimposition of the models and

their similarity. The code is provided at the GitHub repository accessible at https://github.com/MatteoManfredi/pfam_models.

6.2. Results and Discussion

Figure 13 shows the distribution of enzyme models as a function of the computed TM-score, color-coded depending on their structural evidence. It is evident that when a structure is present in the PDB for the protein itself (green bars) or for an homologous human protein (blue bars), the models well superimpose. On the contrary, when no homologous protein has been structurally characterized (red bars) and, in part, when a structure is available only for a non-human (often remote) homolog (yellow bars), the structures can largely diverge.

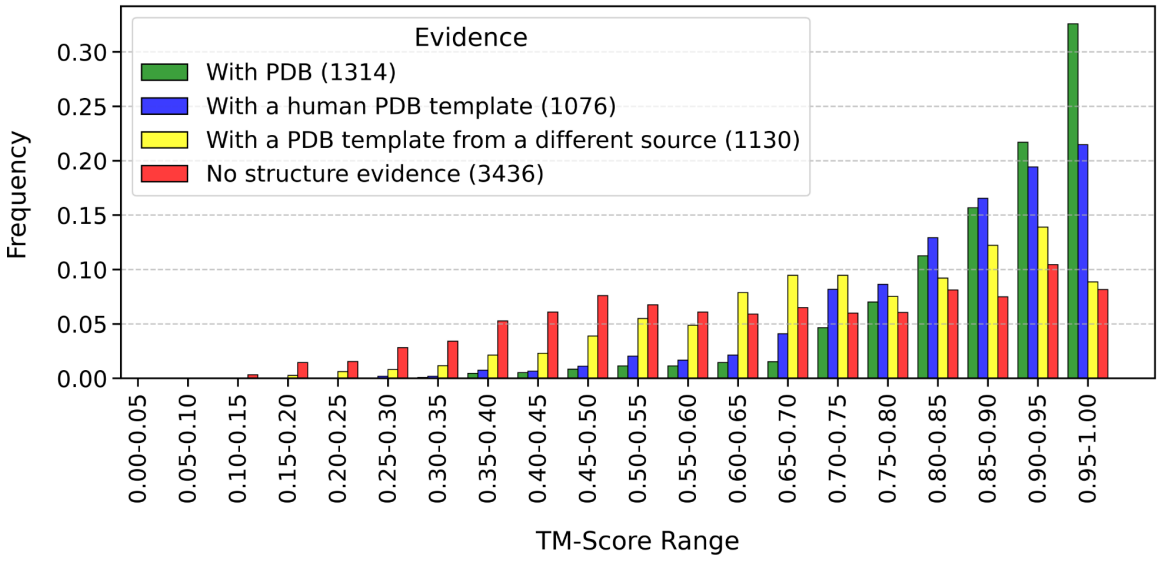


Figure 13. Distribution of pairwise AlphaFold and ESMFold models of human enzymes dataset as a function of their superimposition as evaluated with the TM-score.

Summing up, human enzyme model pairs can be grouped into three categories: i) models with an underlying PDB structure of the enzyme, ii) models without PDB in the background which superimpose (TM-score ≥ 0.6), or iii) which do not superimpose (TM-score <0.6).

We therefore mapped Pfam domains on the models. In the human enzyme set, 1517 unique Pfam domains are shared by 5204 enzyme proteins for a total of 9834 occurrences. Included

in domains are those containing an active site (249) which map into 2459 human enzymes. Table 8 shows the enzyme distributions according to the different Pfam types, as described on the Pfam website (see section 6.1.1.). The Domain type is particularly interesting since it contains information on the functional annotation and, when known, also on the active sites, which are conserved in the families.

Table 8. The human enzyme set is distributed on the 6 types of Pfam entries.

Pfam Type[1]	**# Entries in Pfam database**	**# Unique Pfam in the HES**[2]	**# Enzymes with Pfam**	**# Pfam Occurrences**	**Range of Pfam lengths**[3]
Domains	9,147	1,517	5,204	9,834	16 - 713
Domains with annotated active site	773	249	2,459	2,577	34 - 655
Families	11,536	683	2,738	3,391	11 - 1,444
Repeats	859	78	324	769	14 - 517
Short Motifs	122	21	77	93	15 - 61
Intrinsically disordered regions	122	9	19	19	60 - 165
Coiled-coil regions	193	8	14	16	35 - 331

[1]Pfam classifies its entries into 6 types. Domains containing an active site are reported as annotated by the PfamScan tool.
[2]HES = Human Enzyme Set comprising 6,956 enzymes.
[3]The minimum and maximum length of the Pfam types included in the dataset, (number of residues).
= Number of

After mapping the Pfam database on the pairwise models, we compare the TM-scores of the global models with those evaluated for the Pfam domain regions. Similarly, we compare the local quality of the prediction of the global models with that restricted to the Pfam domain regions by computing the average pLDDT score (rescaled to a 0-1 range). Results are shown in Figure 14. TM-scores of the pairwise models are higher when the models have an underlying PDB structure (1047 enzymes); when structural information is lacking, 2766

enzymes have global models with a pairwise TM-score ≥ 0.6, and 1391 enzymes have global models with TM-score < 0.6. Considering the Pfam domains, TM-scores remain high when evaluated on the region of the mapped domain, irrespective of the superimposition of global enzyme pairwise models. The same holds for the average predicted local evaluation of the quality of the models (pLDDT), which is slightly higher in AlphaFold2 than in ESMFold.

When we restrict the analysis to the enzymes with an active site, derived by mapping Pfam domains containing an active site, again we can observe (Figure 15) that the superimposition of the Pfam regions (TM-score) in the pairwise models increases rather irrespectively of the superimposition and the quality of the global enzyme models. This is particularly evident when the global models diverge (TM-score < 0.6, Figure 16). Moreover, the mapping performed with PfamScan allows the annotation of 807 proteins with 858 active sites from 117 Pfam domains. These annotations are lacking in the associated files at UniProt.

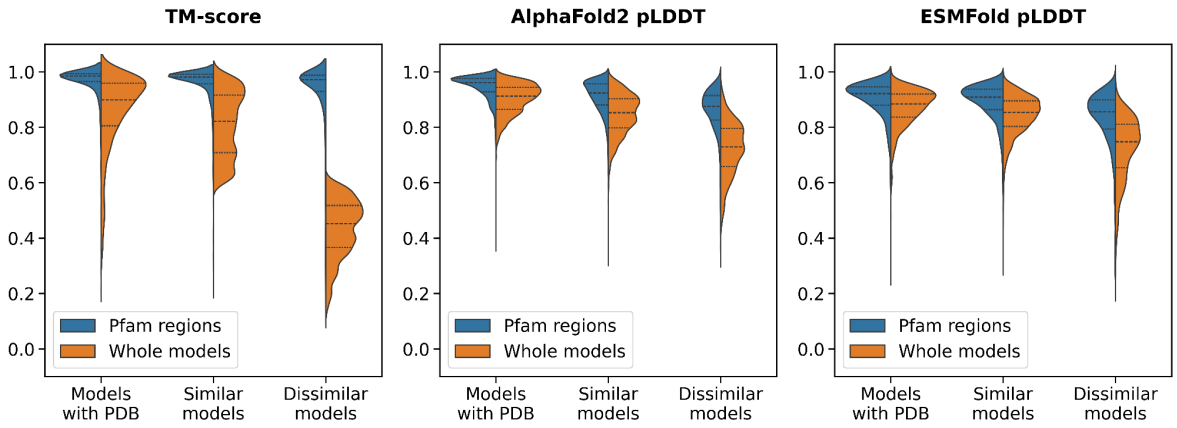


Figure 14. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains. The titles of the plots report the measure of the related y-axis. The x-axis reports the observations for three subsets: (i) enzymes with a known PDB structure, (ii) with similar models (TM-score ≥ 0.6), and (iii) with dissimilar models (TM-score < 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

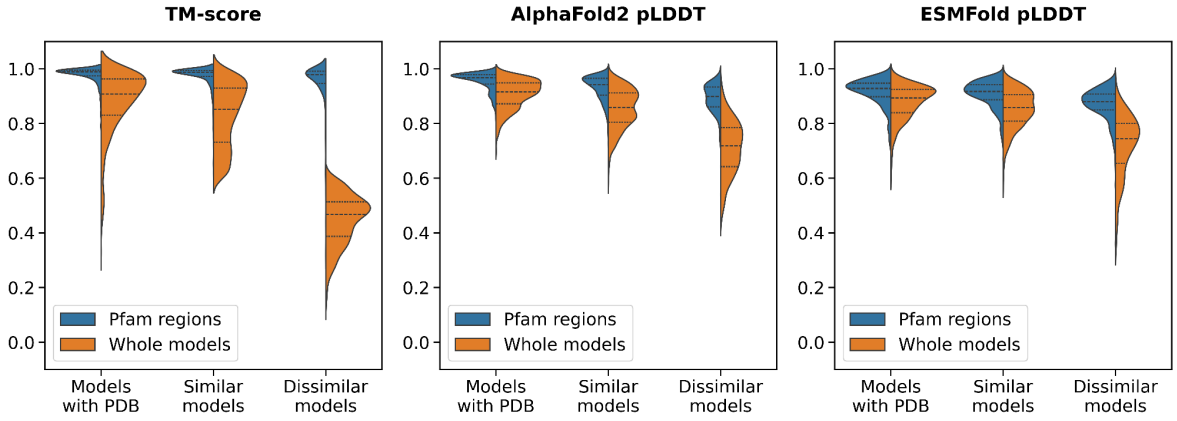


Figure 15. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains with an active site. The titles of the plots report the measure of the related y-axis. The x-axis reports the observations for three subsets: (i) enzymes with a known PDB structure, (ii) with similar models (TM-score ≥ 0.6), and (iii) with dissimilar models (TM-score < 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

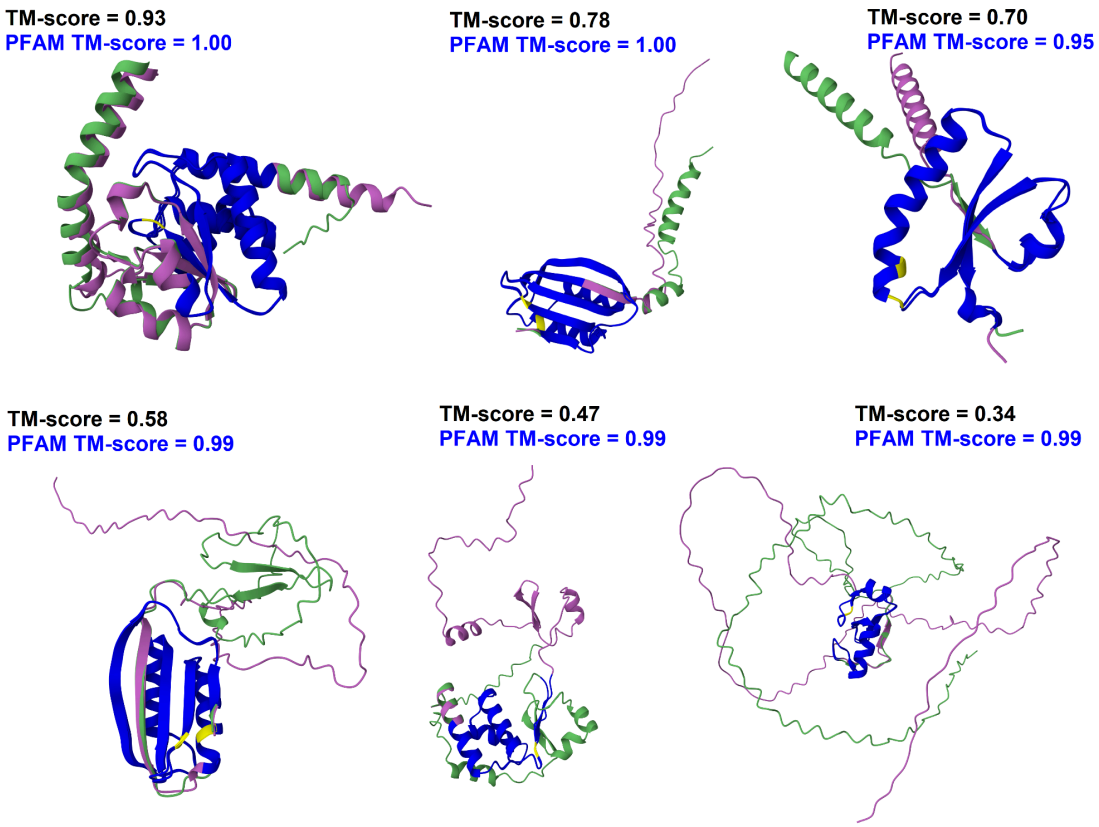


Figure 16. Examples of enzymes at different levels of models TM-scores showing that the superimposition remains good in the region covered by a Pfam domain. In all images, the green model is obtained with ESMFold and the purple one with AlphaFold2. Regions covered by a Pfam domain and the active sites are highlighted in blue and yellow, respectively. The TM-score of the global models and the regions covered by the Pfam domain are highlighted in black and blue, respectively.

Overall, these results suggest that both AlphaFold2 and ESMFold methods are equally good in grasping the information carried out by Pfam models. The analysis described in this work have been included in an updated version of the Alpha&ESMhFolds webserver (whose manuscript is currently under review), enabling users to analyze and compare the predicted structures of all Pfam domains within a given protein.

7. Conclusions

During my PhD, I adopted recent advances in machine learning for computational biology to tackle the problem of protein functional annotation. The glutathione S-transferase (GST) superfamily served as an ideal case study: these enzymes are essential for detoxification, yet their diversity across the tree of life makes their annotation challenging.

In the first part of this thesis, I showed that embedding-based alignment can reliably classify GSTs, and can be used to annotate proteins that are currently poorly annotated in UniProt.

Beyond their detoxification role, GSTs are capable of performing several other functions; therefore, despite being known for their relevance in human diseases, their real impact remains underestimated. As a proof of concept, we found evidence from large-scale RNA proteomic studies suggesting that canonical GSTs may also bind RNA. Adopting deep learning methods and molecular docking we showed that such an interaction is theoretically possible.

Additionally, we employed models from the Alpha&ESMhFolds database to evaluate and compare the ability of AlphaFold2 and ESMFold to resolve protein domains. We found that, regardless of the overall quality of the predicted structure, functional domains are consistently well resolved by both methods, making them reliable tools for transfer of knowledge.

Finally, during my internship abroad, I explored how Protein Language Model embeddings capture evolutionary information, providing insights into the interpretability of these representations. We found that, even in protein families enriched with distant homologs such as GSTs, the embedding vectors corresponding to residues aligned in a Multiple Sequence Alignment tend to cluster together. Furthermore, when comparing different Protein Language Models, this clustering effect became more evident with increasing model size and number of

trainable parameters, indicating that larger models capture evolutionary information more effectively than smaller ones. This preliminary result opens the possibility of developing a workflow that leverages Protein Language Model embeddings to accelerate Multiple Sequence Alignment generation. The code developed for this project was integrated into a Nextflow pipeline. Nextflow is a scientific workflow system for bioinformatics and biological data analysis that facilitates the creation of scalable and reproducible workflows.

Overall, these studies highlight the potential of deep learning to interpret biological complexity. They demonstrate how computational tools can expand our ability to annotate proteins, discover new functional roles and guide future experimental investigations, opening promising avenues for biotechnological research.

Acknowledgements

This journey could not have been possible without the personal and professional help I received during these three years.

First and foremost, I would like to thank Prof. Rita Casadio, Prof. Pier Luigi Martelli and Prof. Castrense Savojardo for guiding me through these years, and for all the teachings and opportunities they gave me.

Thanks to Prof. Cedric Notredame for hosting me in Barcelona and making me feel at home, and for all the support he gave me.

Thanks to all my colleagues and friends, both in Bologna and Barcelona, for their patience and for tolerating me: Giovanni Madeo, Maria Giulia Prado, Elisa Bertolini, Giulia Babbi, Alessio Vignoli, Cristina Araiz, Julia Mir, Luisa Santus, Mathys Grapotte and Suzanne Jin.

I must reserve special thanks to Matteo Manfredi, my close lab mate, for his invaluable friendship and support.

Thanks to my family for making all this possible: Alice Li Greci, Sabatino Vazzana, Massimo Vazzana, Nunziata Li Greci and Anna Gallotta.

Lastly, thanks to William Carnabuci, my beloved friend.

References

Aloke C, Onisuru OO, Achilonu I (2024) Glutathione S-transferase: A versatile and dynamic enzyme. Biochemical and Biophysical Research Communications 734:150774. https://doi.org/10.1016/j.bbrc.2024.150774

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

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

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

Backlund M, Stein F, Rettel M, et al (2020) Plasticity of nuclear and cytoplasmic stress responses of RNA-binding proteins. Nucleic Acids Research 48:4725–4740. https://doi.org/10.1093/nar/gkaa256

Baldi P (2021) Deep Learning in Science, 1st edn. Cambridge University Press

Bao X, Guo X, Yin M, et al (2018) Capturing the interactome of newly transcribed RNA. Nat Methods 15:213–220. https://doi.org/10.1038/nmeth.4595

Beckmann BM, Horos R, Fischer B, et al (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat Commun 6:10127. https://doi.org/10.1038/ncomms10127

Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. Cell Systems 12:654-669.e3. https://doi.org/10.1016/j.cels.2021.05.017

Berman HM (2000) The Protein Data Bank. Nucleic Acids Research 28:235–242. https://doi.org/10.1093/nar/28.1.235

Berry S, Pelkmans L (2022) Mechanisms of cellular mRNA transcript homeostasis. Trends in Cell Biology 32:655–668. https://doi.org/10.1016/j.tcb.2022.05.003

Bertolini E, Babbi G, Savojardo C, et al (2024) MultifacetedProtDB: a database of human proteins with multiple functions. Nucleic Acids Research 52:D494–D501. https://doi.org/10.1093/nar/gkad783

Bishop CM (2016) Pattern Recognition and Machine Learning, Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009). Springer New York, New York, NY

Bresell A, Weinander R, Lundqvist G, et al (2005) Bioinformatic and enzymatic characterization of the MAPEG superfamily. The FEBS Journal 272:1688–1703. https://doi.org/10.1111/j.1742-4658.2005.04596.x

Castello A, Horos R, Strein C, et al (2016) Comprehensive Identification of RNA-Binding Proteins by RNA Interactome Capture. In: Dassi E (ed) Post-Transcriptional Gene Regulation. Springer New York, New York, NY, pp 131–139

Caudron-Herger M, Jansen RE, Wassmer E, Diederichs S (2021) RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. Nucleic Acids Research 49:D425–D436. https://doi.org/10.1093/nar/gkaa1040

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

Corley M, Burns MC, Yeo GW (2020) How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. Molecular Cell 78:9–29. https://doi.org/10.1016/j.molcel.2020.03.011

Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Learning 20:273–297. https://doi.org/10.1023/A:1022627411411

Curtis NJ, Jeffery CJ (2021) The expanding world of metabolic enzymes moonlighting as RNA binding proteins. Biochemical Society Transactions 49:1099–1108. https://doi.org/10.1042/BST20200664

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

Elnaggar A, Heinzinger M, Dallago C, et al (2020) ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing

Elnaggar A, Heinzinger M, Dallago C, et al (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell 44:7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Fabrini R, De Luca A, Stella L, et al (2009) Monomer−Dimer Equilibrium in Glutathione Transferases: A Critical Re-Examination. Biochemistry 48:10473–10482. https://doi.org/10.1021/bi901238t

Ferruz N, Schmidt S, Höcker B (2022) ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13:4348. https://doi.org/10.1038/s41467-022-32007-7

Finn RD, Tate J, Mistry J, et al (2007) The Pfam protein families database. Nucleic Acids Research 36:D281–D288. https://doi.org/10.1093/nar/gkm960

Garcerá A, Barreto L, Piedrafita L, et al (2006) *Saccharomyces cerevisiae* cells have three Omega class glutathione S-transferases acting as 1-Cys thiol transferases. Biochemical Journal 398:187–196. https://doi.org/10.1042/BJ20060034

Gilchrist CLM, Mirdita M, Steinegger M (2024) Multiple Protein Structure Alignment at Scale with FoldMason

Goto S, Kawakatsu M, Izumi S, et al (2009) Glutathione S-transferase π localizes in mitochondria and protects against oxidative stress. Free Radical Biology and Medicine 46:1392–1403. https://doi.org/10.1016/j.freeradbiomed.2009.02.025

Gupta MN, Uversky VN (2023) Moonlighting enzymes: when cellular context defines specificity. Cell Mol Life Sci 80:130. https://doi.org/10.1007/s00018-023-04781-0

Heinzinger M, Elnaggar A, Wang Y, et al (2019) Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20:723. https://doi.org/10.1186/s12859-019-3220-8

Heinzinger M, Weissenow K, Sanchez JG, et al (2024) Bilingual language model for protein sequence and structure. NAR Genomics and Bioinformatics 6:lqae150. https://doi.org/10.1093/nargab/lqae150

Ho (1995) Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press, Montreal, Que., Canada, pp 278–282

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang C, Chen M, Pang D, et al (2014) Developmental and Activity-Dependent Expression of LanCL1 Confers Antioxidant Activity Required for Neuronal Survival. Developmental Cell 30:479–487. https://doi.org/10.1016/j.devcel.2014.06.011

Huang R, Han M, Meng L, Chen X (2018) Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. Proc Natl Acad Sci USA 115:. https://doi.org/10.1073/pnas.1718406115

Jumper J, Evans R, Pritzel A, et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

Jurrus E, Engel D, Star K, et al (2018) Improvements to the APBS biomolecular solvation software suite. Protein Science 27:112–128. https://doi.org/10.1002/pro.3280

Katoh K (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Koirala S, Moural T, Zhu F (2022) Functional and Structural Diversity of Insect Glutathione S-transferases in Xenobiotic Adaptation. Int J Biol Sci 18:5713–5723. https://doi.org/10.7150/ijbs.77141

Kooshapur H, Choudhury NR, Simon B, et al (2018) Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. Nat Commun 9:2479. https://doi.org/10.1038/s41467-018-04871-9

Kryshtafovych A, Schwede T, Topf M, et al (2023) Critical assessment of methods of protein structure prediction (CASP)—Round XV. Proteins 91:1539–1549. https://doi.org/10.1002/prot.26617

Kryshtafovych A, Schwede T, Topf M, et al (2021) Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Proteins 89:1607–1617. https://doi.org/10.1002/prot.26237

Kumar S, Trivedi PK (2018) Glutathione S-Transferases: Role in Combating Abiotic Stresses Including Arsenic Detoxification in Plants. Front Plant Sci 9:751. https://doi.org/10.3389/fpls.2018.00751

Ladner JE, Parsons JF, Rife CL, et al (2004) Parallel Evolutionary Pathways for Glutathione Transferases: Structure and Mechanism of the Mitochondrial Class Kappa Enzyme rGSTK1-1. Biochemistry 43:352–361. https://doi.org/10.1021/bi035832z

Larkin MA, Blackshields G, Brown NP, et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Laskowski RA, Swindells MB (2011) LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. J Chem Inf Model 51:2778–2786. https://doi.org/10.1021/ci200227u

Lesk AM (2016) Introduction to protein science: architecture, function, and genomics, 3rd ed. Oxford university press, Oxford

Li P, Liu Z-P (2022) PST-PRNA: prediction of RNA-binding sites using protein surface topography and deep learning. Bioinformatics 38:2162–2168. https://doi.org/10.1093/bioinformatics/btac078

Li X, Xiong H, Li X, et al (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 64:3197–3234. https://doi.org/10.1007/s10115-022-01756-8

Li Y, Wang Y, Tan Y-Q, et al (2024) The landscape of RNA binding proteins in mammalian spermatogenesis. Science 386:eadj8172. https://doi.org/10.1126/science.adj8172

Lin Z, Akin H, Rao R, et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379:1123–1130. https://doi.org/10.1126/science.ade2574

Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23:18. https://doi.org/10.3390/e23010018

Lo Bello M, Oakley AJ, Battistoni A, et al (1997) Multifunctional Role of Tyr 108 in the Catalytic Mechanism of Human Glutathione Transferase P1-1. Crystallographic and Kinetic Studies on the Y108F Mutant Enzyme. Biochemistry 36:6207–6217. https://doi.org/10.1021/bi962813z

Lu SC (2013) Glutathione synthesis. Biochimica et Biophysica Acta (BBA) - General Subjects 1830:3143–3153. https://doi.org/10.1016/j.bbagen.2012.09.008

Lv N, Huang C, Huang H, et al (2023) Overexpression of Glutathione S-Transferases in Human Diseases: Drug Targets and Therapeutic Implications. Antioxidants 12:1970. https://doi.org/10.3390/antiox12111970

MacDougall A, Volynkin V, Saidi R, et al (2020) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics 36:4643–4648. https://doi.org/10.1093/bioinformatics/btaa485

Madani A, Krause B, Greene ER, et al (2023) Large language models generate functional protein sequences across diverse families. Nat Biotechnol 41:1099–1106. https://doi.org/10.1038/s41587-022-01618-2

Manfredi M, Savojardo C, Iardukhin G, et al (2024) Alpha&ESMhFolds: A Web Server for Comparing AlphaFold2 and ESMFold Models of the Human Reference Proteome. Journal of Molecular Biology 436:168593. https://doi.org/10.1016/j.jmb.2024.168593

Manfredi M, Vazzana G, Savojardo C, et al (2025) AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation. Computational and Structural Biotechnology Journal 27:461–466. https://doi.org/10.1016/j.csbj.2025.01.008

Mariani V, Biasini M, Barbato A, Schwede T (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29:2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Mazari AMA, Zhang L, Ye Z-W, et al (2023) The Multifaceted Role of Glutathione S-Transferases in Health and Disease. Biomolecules 13:688. https://doi.org/10.3390/biom13040688

Meng X-Y, Zhang H-X, Mezei M, Cui M (2011) Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. CAD 7:146–157. https://doi.org/10.2174/157340911795677602

Meux E, Prosper P, Ngadin A, et al (2011) Glutathione Transferases of Phanerochaete chrysosporium. Journal of Biological Chemistry 286:9162–9173. https://doi.org/10.1074/jbc.M110.194548

Milek M, Imami K, Mukherjee N, et al (2017) DDX54 regulates transcriptome dynamics during DNA damage response. Genome Res 27:1344–1359. https://doi.org/10.1101/gr.218438.116

Mistry J, Chuguransky S, Williams L, et al (2021) Pfam: The protein families database in 2021. Nucleic Acids Research 49:D412–D419. https://doi.org/10.1093/nar/gkaa913

Morel F, Aninat C (2011) The glutathione transferase kappa family. Drug Metabolism Reviews 43:281–291. https://doi.org/10.3109/03602532.2011.556122

Morgenstern R, Zhang J, Johansson K (2011) Microsomal glutathione transferase 1: mechanism and functional roles. Drug Metabolism Reviews 43:300–306. https://doi.org/10.3109/03602532.2011.558511

Mukanganyama S, Bezabih M, Robert M, et al (2011) The evaluation of novel natural products as inhibitors of human glutathione transferase P1-1. Journal of Enzyme Inhibition and Medicinal Chemistry 26:460–467. https://doi.org/10.3109/14756366.2010.526769

Mullari M, Lyon D, Jensen LJ, Nielsen ML (2017) Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification. J Proteome Res 16:2762–2772. https://doi.org/10.1021/acs.jproteome.7b00042

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

Nelson DL, Cox MM (2021) Lehninger principles of biochemistry, Eighth edition. Macmillan International Higher Education, New York

Ngo TD, Partin AC, Nam Y (2019) RNA Specificity and Autoregulation of DDX17, a Modulator of MicroRNA Biogenesis. Cell Reports 29:4024-4035.e5. https://doi.org/10.1016/j.celrep.2019.11.059

Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. Journal of Molecular Biology 302:205–217. https://doi.org/10.1006/jmbi.2000.4042

Oakley A (2011) Glutathione transferases: a structural perspective. Drug Metabolism Reviews 43:138–151. https://doi.org/10.3109/03602532.2011.558093

Ofer D, Brandes N, Linial M (2021) The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal 19:1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022

Ongpipattanakul C, Liu S, Luo Y, et al (2023) The mechanism of thia-Michael addition catalyzed by LanC enzymes. Proc Natl Acad Sci USA 120:e2217523120. https://doi.org/10.1073/pnas.2217523120

Pakhomova S, Rife CL, Armstrong RN, Newcomer ME (2004) Structure of fosfomycin resistance protein FosA from transposon *Tn2921*. Protein Science 13:1260–1265. https://doi.org/10.1110/ps.03585004

Pantolini L, Studer G, Pereira J, et al (2024) Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. Bioinformatics 40:btad786. https://doi.org/10.1093/bioinformatics/btad786

Paysan-Lafosse T, Blum M, Chuguransky S, et al (2023) InterPro in 2022. Nucleic Acids Research 51:D418–D427. https://doi.org/10.1093/nar/gkac993

Queiroz RML, Smith T, Villanueva E, et al (2019) Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). Nat Biotechnol 37:169–178. https://doi.org/10.1038/s41587-018-0001-2

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286. https://doi.org/10.1109/5.18626

Rives A, Meier J, Sercu T, et al (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118:e2016239118. https://doi.org/10.1073/pnas.2016239118

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65:386–408. https://doi.org/10.1037/h0042519

Rost B (1999) Twilight zone of protein sequence alignments. Protein Engineering, Design and Selection 12:85–94. https://doi.org/10.1093/protein/12.2.85

Sarkar S (2016) The Big Data Deluge in Biology: Challenges and Solutions. Global J Technol Optim 01: https://doi.org/10.4172/2229-8711.S1e103

Schütze K, Heinzinger M, Steinegger M, Rost B (2022) Nearest neighbor search on embeddings rapidly identifies distant protein relations. Front Bioinform 2:1033775. https://doi.org/10.3389/fbinf.2022.1033775

Shehu D, Abdullahi N, Alias Z (2018) Cytosolic Glutathione S-transferase in Bacteria:A Review. Pol J Environ Stud 28:515–528. https://doi.org/10.15244/pjoes/85200

Sievers F, Wilm A, Dineen D, et al (2011) Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539. https://doi.org/10.1038/msb.2011.75

Simic P, Pljesa I, Nejkovic L, et al (2022) Glutathione Transferase P1: Potential Therapeutic Target in Ovarian Cancer. Medicina 58:1660. https://doi.org/10.3390/medicina58111660

Sjögren T, Nord J, Ek M, et al (2013) Crystal structure of microsomal prostaglandin E_2 synthase provides insight into diversity in the MAPEG superfamily. Proc Natl Acad Sci USA 110:3806–3811. https://doi.org/10.1073/pnas.1218504110

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

Szklarczyk D, Kirsch R, Koutrouli M, et al (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research 51:D638–D646. https://doi.org/10.1093/nar/gkac1000

The UniProt Consortium, Bateman A, Martin M-J, et al (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51:D523–D531. https://doi.org/10.1093/nar/gkac1052

Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461. https://doi.org/10.1002/jcc.21334

Urdaneta EC, Vieira-Vieira CH, Hick T, et al (2019) Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. Nat Commun 10:990. https://doi.org/10.1038/s41467-019-08942-3

Van Kempen M, Kim SS, Tumescheit C, et al (2024) Fast and accurate protein structure search with Foldseek. Nat Biotechnol 42:243–246. https://doi.org/10.1038/s41587-023-01773-0

Varotsou C, Ataya F, Papageorgiou AC, Labrou NE (2023) Structural Studies of Klebsiella pneumoniae Fosfomycin-Resistance Protein and Its Application for the Development of an Optical Biosensor for Fosfomycin Determination. IJMS 25:85. https://doi.org/10.3390/ijms25010085

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention Is All You Need

Vazzana G, Martelli PL, Casadio R (2026) Can Human Canonical Glutathione S-Transferases Act as RNA-Binding Proteins? J Comput Biophys Chem 25:627–640. https://doi.org/10.1142/S2737416525500607

Vazzana G, Savojardo C, Martelli PL, Casadio R (2024) Testing the Capability of Embedding-Based Alignments on the GST Superfamily Classification: The Role of Protein Length. Molecules 29:4616. https://doi.org/10.3390/molecules29194616

Wang L, Li X, Zhang H, et al (2025) A Comprehensive Review of Protein Language Models

Webb EC (1992) Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, San Diego

Weissenow K, Rost B (2025) Are protein language models the new universal key? Current Opinion in Structural Biology 91:102997. https://doi.org/10.1016/j.sbi.2025.102997

Xia Y, Xia C-Q, Pan X, Shen H-B (2021) GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Research 49:e51–e51. https://doi.org/10.1093/nar/gkab044

Xun L, Belchik SM, Xun R, et al (2010) S-Glutathionyl-(chloro)hydroquinone reductases: a novel class of glutathione transferases. Biochemical Journal 428:419–427. https://doi.org/10.1042/BJ20091863

Zacco E, Broglia L, Kurihara M, et al (2024) RNA: The Unsuspected Conductor in the Orchestra of Macromolecular Crowding. Chem Rev 124:4734–4777. https://doi.org/10.1021/acs.chemrev.3c00575

Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33:2302–2309. https://doi.org/10.1093/nar/gki524

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710. https://doi.org/10.1002/prot.20264

Zhuge X-L, Xu H, Xiu Z-J, Yang H-L (2020) Biochemical Functions of Glutathione S-Transferase Family of Salix babylonica. Front Plant Sci 11:364. https://doi.org/10.3389/fpls.2020.00364

Appendix: Publications

molecules

MDPI

Article

Testing the Capability of Embedding-Based Alignments on the GST Superfamily Classification: The Role of Protein Length

Gabriele Vazzana, Castrense Savojardo, Pier Luigi Martelli * and Rita Casadio *

Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, 40126 Bologna, Italy; gabriele.vazzana2@unibo.it (G.V.); castrense.savojardo2@unibo.it (C.S.)
* Correspondence: pierluigi.martelli@unibo.it (P.L.M.); rita.casadio@unibo.it (R.C.)

Abstract: In order to shed light on the usage of protein language model-based alignment procedures, we attempted the classification of Glutathione S-transferases (GST; EC 2.5.1.18) and compared our results with the ARBA/UNI rule-based annotation in UniProt. GST is a protein superfamily involved in cellular detoxification from harmful xenobiotics and endobiotics, widely distributed in prokaryotes and eukaryotes. What is particularly interesting is that the superfamily is characterized by different classes, comprising proteins from different taxa that can act in different cell locations (cytosolic, mitochondrial and microsomal compartments) with different folds and different levels of sequence identity with remote homologs. For this reason, GST functional annotation in a specific class is problematic: unless a structure is released, the protein can be classified only on the basis of sequence similarity, which excludes the annotation of remote homologs. Here, we adopt an embedding-based alignment to classify 15,061 GST proteins automatically annotated by the UniProt-ARBA/UNI rules. Embedding is based on the Meta ESM2-15b protein language. The embedding-based alignment reaches more than a 99% rate of perfect matching with the UniProt automatic procedure. Data analysis indicates that 46% of the UniProt automatically classified proteins do not conserve the typical length of canonical GSTs, whose structure is known. Therefore, 46% of the classified proteins do not conserve the template/s structure required for their family classification. Our approach finds that 41% of 64,207 GST UniProt proteins not yet assigned to any class can be classified consistently with the structural template length.

Keywords: Glutathione S-transferases; protein language models; protein classification; functional annotation; embedding-based alignment

Citation: Vazzana, G.; Savojardo, C.; Martelli, P.L.; Casadio, R. Testing the Capability of Embedding-Based Alignments on the GST Superfamily Classification: The Role of Protein Length. *Molecules* **2024**, *29*, 4616. https://doi.org/10.3390/molecules29194616

Academic Editor: Takeshi Kikuchi

Received: 6 September 2024
Revised: 19 September 2024
Accepted: 20 September 2024
Published: 29 September 2024

1. Introduction

After the success of Large Language Models (LLMs) for natural language processing tasks, transformer-based deep-learning architectures [1] have taken hold in the field of computational biology, with the consequent emergence of a counterpart adapted to protein sequences, known as protein Language Models (pLMs) [2,3]. Several pLMs have been implemented in the past few years, mainly differing in relation to the number of sequences included in the training set (of particular relevance are the models developed by Rost's lab [3,4] and the more recent ESM family of models developed by the MetaAI group [5,6]). Recently, pLMs have emerged as a new and powerful mapping procedure which allows the representation of a protein sequence considering the knowledge that the protein can derive from its family and/or superfamily, in the multifaceted protein universe [7]. This procedure, referred to as "embedding", is "context-aware" [7] and it is often adopted to generate input to train downstream predictive tools with machine and/or deep learning approaches, replacing the classic method based on the time-consuming generation of Multiple Sequence Alignments (MSAs). The embedding procedures have been increasing the performance of relevant predictive tasks, including protein secondary structure [3], protein–protein interaction [8,9] and three-dimensional (3D) protein structure prediction [6]. Different

embedding-based methods made it possible to quantify sequence similarity [3,10,11], to cluster proteins into families [12], to generate evolutionary landscapes [13,14] and to search for structure–structure similarities [15,16], just to mention some of the applications.

Summing up, we may conclude that the embedding procedures succeed in carrying along information derived from the protein family/superfamily, including sequence profile and template structure conservation. It is still debatable whether embedding is sufficient to recognise remote homologs and perform functional annotation, as recently discussed [17,18].

Now, in light of these advancements, a key question remains: to which extent can we stress the embedding procedure for sequence alignments? In order to test this, here we tackle the annotation problem of the GST superfamily, with a new method for determining sequence embedding distances, outperforming previous ones in remote homology detection (EBA, embedding-based alignment, [18] and references therein). We choose the Glutathione S-Transferases superfamily (GST, EC 2.5.1.18, [19]) for its functional and structural characteristics. According to the literature, the superfamily includes three major groups, cytosolic, mitochondrial and microsomal, with at least 20 documented families (or classes), active in the different cell compartments. Although these enzymes function in the same cellular compartment, their structure remains conserved despite low sequence identity across classes and different organisms. A total of 75% of the classes share the same functional fold and are active in the cytosol together with the other three structural classes; two other folds are active in mitochondria and in microsomes, respectively [20–26]. The complex relation between sequence and structure makes the annotation process difficult (see Supplementary S1 [27–37] for an extended description of the classes).

In the following, we test the capabilities of embedding-based alignment in the task of assigning sequences to the different GST classes as done in UniProt with an automatic procedure defined by the ARBA/Uni rules [38]. After the selection of a reference set, we undertake large-scale testing, adopting the recent MetaAI ESM2-15b pLM and measuring sequence distance with EBA [18]. We find that the procedure is successful in sequence annotation, particularly when the sequence length of the proteins is conserved with respect to those included in the reference set. With this constraint, we classify another 26,180 proteins from 64,207 unclassified GSTs in UniProt, enriching the number of proteins in the different classes, and generating a set of sequences for future experimental investigations.

2. Results and Discussion

2.1. Fishing for Transfer of Annotation

Our procedure is described in Figure 1. Basically, we generate a reference set of the GST protein superfamily which acts as a representative set of the functional and structural properties of the proteins in the superfamily. The set is carefully selected and contains proteins with a reference PDB structure and/or a high-quality AlphaFold2 model, along with an experimental validation of the function. Then, each protein of the reference set is embedded with the selected protein language model and becomes a bait. The encoding procedure allows for carrying information on the structure and on the conserved sequence motifs of the family [2–5]. The embedded bait is then aligned with the EBA alignment procedure [18], with a testing set from UniProt, filtered with the UniProt/ARBA rules, and annotated in a specific GST class. The different reference classes are color coded in Figure 1, and fishing in shallow waters is successful when a prey with the same color is captured (in this case, the assigned annotation obtained with embedding alignments matches with the one already present in the testing set). Finally, we enter with the procedure into the deep sea, to search for new proteins to add to a specific family. After validation of the procedure in the previous step, we now classify proteins without any verification.

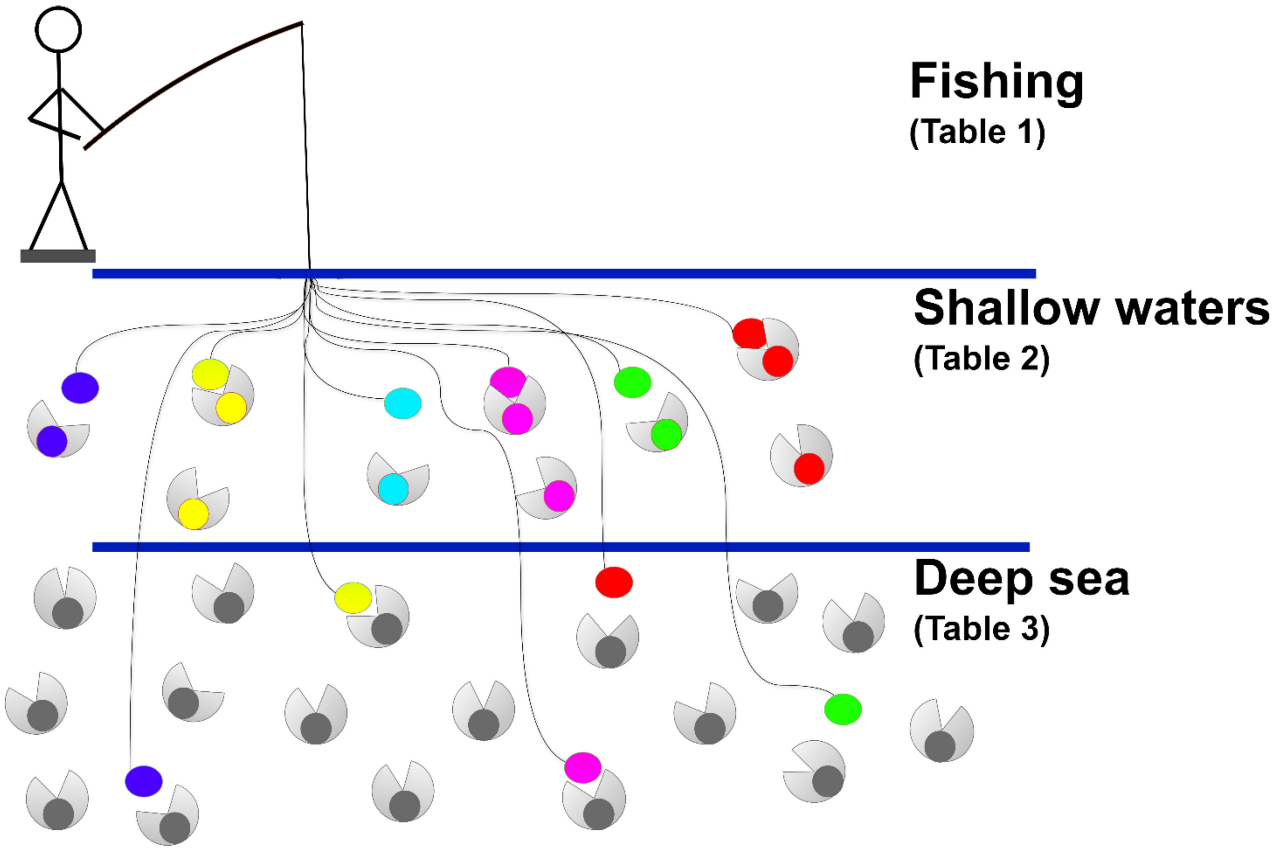
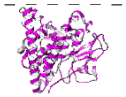
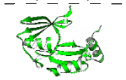
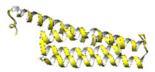


Figure 1. Embedding-based alignment for GST protein classification. A reference protein set of well-curated proteins (described in Table 1) is adopted as baits for "fishing" ARBA GST classified proteins (testing set, in shallow waters, Table 2) and ARBA GST unclassified proteins (in deep sea, Table 3). For details, see text and Section 3. Color matching is indicative of the affinity of baits and preys.

2.2. The Reference Set

Our reference set (Ref set) is detailed in Table 1. It contains 284 well-annotated proteins from SwissProt, with documented experimental evidence. When listing per taxon, the presence of specific classes in particular taxon (e.g., Phi, Tau, Lambda only in Viridiplantae, Beta and HSP 26 only in bacteria) is evident. A more detailed description of the GST superfamily is available in Supplementary S1. It appears that, for the time being, four different folds are adopted by GST proteins functionally active in the cytosol; however, the most populated one is conserved for 15 classes, active in the cytosol and collecting proteins that are or are not distantly related homologs (see rightmost columns in Table 1 where the length variability together with the sequence identity range are reported). Considering a 30% sequence identity, the threshold between homologs and distantly related ones, six classes indeed contain distantly related homologs, where the conserved structure in the PDB testifies to the inclusion in the class (family). Interestingly, in Table 1, 245 proteins share the same fold and are distributed in 15 classes. Other folds are present in the cytosol of fungi (omega-like), bacteria (FosA) and mammals (LanC). The length of these proteins is different from the previous ones. Finally, kappa and MAPEG have been reported in mammals and are active in the mitochondria and in the microsomes, respectively. Both classes include remote homologs and folds different from the cytosolic ones. The proteins in the different classes share less than 30% sequence identity (Table S1).

Table 1. The reference dataset (REFset) with 284 sequences.

Classes	Bact.	Amoeb.	Fungi	Virid.	Plat.	Nem.	Arth.	Moll.	Actin.	Amph.	Aves	Mamm.	Total Class	Length	Seq. Id. Range (%)	Structure
Mu	-	-	-	-	11 (9 *)	-	3 (2 *)	-	-	-	1 (1)	28 (8 *)	43 (20 *)	211–225	22–98	
Sigma	-	-	-	-	-	9 (2)	7 (4 *)	1 (1)	-	-	1	3 (2)	21 (9 *)	199–249	25–94	
Alpha	-	1	-	-	-	-	-	-	-	-	2 (1)	18 (10 *)	21 (11 *)	222–229	29–96	
Pi	-	-	-	-	-	5 (2 *)	-	-	-	2	-	12 (3)	19 (5 *)	207–210	32–99	
Theta	-	-	-	-	-	-	-	-	-	-	-	12 (3)	12 (3)	240–244	40–99	
Delta-Epsilon	-	-	-	-	-	-	32 (15 *)	-	-	-	-	-	32 (15 *)	208–271	25–99	
Omega	-	-	-	-	-	3	2 (2 *)	-	-	-	-	7 (2)	12 (4 *)	240–256	23–93	
Zeta	3 (3 *)	-	1 (1)	3 (1)	-	1	-	-	-	-	-	3 (2)	11 (7 *)	212–221	33–95	
Rho	-	-	-	-	-	-	-	1 (1 *)	1	-	-	-	2 (1 *)	223–225	41	
DHAR	-	-	-	3 (3)	-	-	-	-	-	-	-	-	3 (3)	213–213	66–76	
Tau	-	-	-	34 (5 *)	-	-	-	-	-	-	-	-	34 (5 *)	217–231	30–98	
Phi	-	-	-	25 (11 *)	-	-	-	-	-	-	-	-	25 (11 *)	212–221	31–95	(8GSS)
Lambda	-	-	-	3	-	-	-	-	-	-	-	-	3	235–237	56–73	
Beta	4 (3)	-	-	-	-	-	-	-	-	-	-	-	4 (3)	201–203	36–54	
HSP26	3 (3)	-	-	-	-	-	-	-	-	-	-	-	3 (3)	202–212	22–60	
Omega-like	-	-	4 (1)	-	-	-	-	-	-	-	-	-	4 (1)	313–370	44–63	 (5LKD)
FosA	2 (2)	-	-	-	-	-	-	-	-	-	-	-	2 (2)	135–141	59	 (1NPB)
LanC	-	-	-	-	-	-	-	-	1	-	-	4 (1)	5 (1)	399–405	63–96	 (8D19)
Kappa	-	-	-	-	-	2	-	-	-	-	-	3 (2)	5 (2)	225–226	28–86	 (3RPN)
MAPEG	-	1	-	-	-	-	-	-	-	-	-	22 (5)	23 (5)	146–155	12–98	 (4AL0)
Total Taxon	12 (11 *)	2	5 (2)	68 (20 *)	11 (9 *)	20 (4 *)	44 (23 *)	2 (2 *)	2	2	4 (2)	112 (38 *)	284 (111 *)			

Legend to Table 1. The 284 proteins of the reference set are listed according to their classes (rows) and taxonomic groups (columns). The GST superfamily includes four folds in the cytosol and two folds in mitochondria and microsomes, respectively. Cytosolic GSTs comprise 15 classes with the same fold. Three other folds are cytosolic, and two are found in mitochondria (Kappa) and in microsomes (MAPEG), respectively, for a grand total of 20 classes. Taxa are listed, according to the classification adopted in NCBI at phylum (Mollusca,

Arthropoda, Nematoda and Platyhelminthes), superclass (Actinopterygii) or class (Mammalia, Aves and Amphibia) for metazoan, at kingdom level for Viridiplantae and Fungi and at superkingdom level for Bacteria. Amoebozoa are also included. The number of proteins with a PDB reference is specified inside round brackets; (*) indicates that at least one entry in the set belongs to TrEMBL. Entries without a PDB reference are endowed with high-quality AlphaFold2 models (see Materials). Dashed horizontal lines discriminate classes in the same sub-cellular location. The "Length" column displays the shortest and the longest protein sequence found in each class. The Seq.Id. (%) column shows the minimum and maximum sequence identity percentage found within each class (for classes with only two representatives, the sequence identity between the two is shown). Abbreviations used: Bact., Bacteria; Am., Amoebozoa; Fu., Fungi; Vir., Viridiplantae; Plat., Platyhelminthes; Nem., Nematoda; Arth., Arthropoda; Moll., Mollusca; Act., Actinopterygii; Amph., Amphibia; Mamm., Mammalia., DHAR, dehydroascorbate reductase; HSP26, Heat Shock protein 26 kDa; FosA, Fosfomycin resistance; and LanC, LanC-like. A close inspection of the literature on the arthropoda proteins indicates that they belong to the delta or epsilon classes, closely related in both sequence and structure [39,40], and we included these proteins in one single delta-epsilon class (as also suggested by the Conserved Domain Database (CDD), https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. (accessed on 14 June 2024)). As to the Rho class, we added it after the reclassification of previous theta into rho GST proteins, found in marine organisms [41]. All of the structures shown are from PDB current release (https://www.rcsb.org/ (accessed on 1 January 2024)). From structural alignment we find that the root mean square deviation (RMSD) within class is less than 3 Å.

Table 2. Testing the embedding-based alignment (EBA) towards the ARBA GST classification.

	ARBA*	Within Reference Length Range (RLR)*				Below Reference Length Range (< RLR)*					Above Reference Range (>RLR)*				
Class	**Total**	**°Exp**	**°Pred**		**°Pred SI***	**°Exp**	**°Pred**		**°Pred SI***	**Errors**	**°Exp**	**°Pred**		**°Pred SI***	**Errors**
	(#)	**(#)**	**(#)**	**RLR**	**(%)**	**(#)**	**(#)**	**<RLR**	**(%)**	**(#)**	**(#)**	**(#)**	**>RLR**	**(%)**	**(#)**
Mu	1706	981	979	211–225	10–99	355	349	140–210	8–99	6	370	335	226–475	9–99	35
Sigma	694	592	592	199–249	20–99	66	66	109–198	21–99	-	36	36	250–499	3–99	-
Alpha	1520	734	734	222–229	17–99	495	471	113–221	13–99	24	291	289	230–487	10–99	2
Pi	609	323	323	207–210	24–99	158	158	120–206	21–99	-	128	124	211–488	20–99	4
Theta	1428	560	560	240–244	25–99	545	540	104–239	16–99	5	323	323	245–491	1–99	-
Delta-Epsilon	822	715	715	208–271	18–99	68	68	102–207	15–99	-	39	39	272–478	10–99	-
Omega	1349	556	556	240–256	11–99	617	597	101–239	6–99	20	176	176	257–474	10–99	-
Zeta	728	268	268	212–221	26–99	122	122	139–211	22–99	-	338	338	222–433	2–99	-
DHAR	10	-	-	213–213	-	7	7	107–212	6–97	-	3	3	214–465	37–99	-
Tau	1342	851	851	217–231	23–99	222	219	202–216	6–99	3	269	268	232–449	3–99	1
Phi	1711	1066	1066	212–221	25–99	177	177	149–211	20–99	-	468	468	222–491	2–99	-
HSP26	433	363	363	202–212	33–99	48	48	196–211	40–99	-	22	22	213–227	40–99	-
LanC	450	177	177	399–405	45–99	109	109	126–398	4–99	-	164	164	406–490	32–99	-
Kappa	1148	230	230	225–226	17–99	554	554	189–224	12–99	-	364	364	227–257	12–99	-
MAPEG	1111	687	687	146–155	7–99	143	143	101–145	3–99	-	281	281	157–363	6–99	-
Total	15,061	8103	8101			3686	3628			58	3272	3230			42

Legend to Table 2. * The testing set includes 15,061 GST proteins classified by the ARBA rule system [38]. * From Table 1, we derived the reference length range (RLR) of GST proteins with a reference fold per each class present in the set (see Section 3, Table 1). For the sake of fold conservation, we clustered GST proteins as proteins with a length included in the range (RLR), below the range (<RLR) and above range (>RLR). We show also the range of sequence identity per EBA-found GST class (Pred SI). EBA errors are particularly on GST proteins with lower or higher length than those in the range of fold conservation. See text for details and discussion. Only two mu proteins in the reference length range are misclassified by EBA: (UniProt IDs: A0A1I8FWQ8, A0A1I8J3A2) are classified as sigma by the embedding procedure. A sequence comparison of the two proteins with the sigma and mu GST proteins indicates that they share higher sequence identity with the sigma than with mu GST reference proteins (33% and 28% sequence identity, respectively). Moreover, the sigma classification is supported by the presence of CDD sigma domains in the annotation. °Exp = ARBA expected; Pred = EBA classification. Pred SI = Sequence identity among predicted and reference class (Table 1); # = Number of. In the testing dataset, among the mu-class GSTs, a set of 29 similar sequences (sequence identity >40%) above the reference length range are misclassified.

InterPro annotations for this group of proteins reveal the presence of canonical GST domains together with an extra Elongation Factor 1B domain, suggesting that the canonical fold is not conserved. Most of the remaining errors are found in the "below reference length range" region of the alpha and omega classes. Among these, 30 are due to the normalization procedure of the method, as the similarity alignment score (s_align) is higher for the correct class. Indeed, when a test protein is shorter than the representative entries of the class in the reference dataset, the EBA_{min} score for the correct classification is penalized with respect to shorter sequences of a different class, possibly resulting in misclassifications. In the case of omega errors they are always classified as tau, with the former class showing longer sequences in the reference dataset with respect to the latter. Interestingly, these two classes show deep structural similarities, with the tau class lacking an N-terminal extension typical of omega cGSTs [42]. The remaining misclassifications are either normalization-derived or driven by sequence similarity with the bait proteins.

Table 3. Classifying GST proteins in the "deep sea" with the embedding-based alignment method.

Class	Bacteria	Amoeb.	Fungi	Virid.	Plat.	Nematoda	Arth.	Moll.	Actin.	Amph.	Aves	Mamm.	Others	Total Class
Mu	5	-	33	4	6	12	74	7	5	-	-	7	96	249
Sigma	30	-	87	14	-	480	133	82	11	3	73	130	376	1419
Alpha	21	-	13	7	-	11	1	1	2	-	1	6	49	112
Pi	-	-	37	2	-	6	4	-	-	1	-	4	33	87
Theta	13	-	-	10	-	-	1	3	2	-	-	30	5	64
Delta-Epsilon	1949	9	498	8	-	-	1642	1	5	-	-	4	74	4190
Omega	87	-	112	5	1	-	1	-	-	-	-	-	10	216
Zeta	1397	-	22	7	-	-	-	-	-	-	1	-	21	1448
Rho	524	-	22	-	-	-	-	5	197	-	-	-	9	757
DHAR	1	-	-	-	-	-	-	-	-	-	-	-	-	1
Tau	1555	-	30	2401	-	1	-	-	1	-	-	-	41	4029
Phi	3694	5	772	60	-	-	-	1	1	-	-	1	57	4591
Lambda	-	-	3	63	-	-	-	-	-	-	-	-	-	66
Beta	1569	-	4	-	-	-	1	-	-	-	-	-	15	1589
HSP26	2539	-	9	1	-	-	-	-	-	-	-	-	27	2576
Omega-like	2746	1	298	39	-	-	2	-	4	-	2	21	200	3313
FosA	306	-	-	-	-	-	-	-	-	-	-	-	2	308
LanC	-	-	-	-	-	-	-	-	-	-	-	-	1	1
Kappa	-	-	8	-	-	-	-	-	1	-	-	-	-	9
MAPEG	39	1	230	108	3	-	347	24	141	11	48	44	159	1155
Within RLR	16,475	16	2178	2729	10	510	2206	124	370	15	125	247	1175	26,180
Below RLR	11,288	6	626	1493	66	237	443	52	426	39	159	509	677	16,021
Above RLR	12,917	22	3743	2260	18	119	388	45	342	15	56	148	1007	21,080
Total per Taxon	41,075	44	6582	6851	99	883	3063	222	1157	69	345	928	2889	64,207

Legend to Table 3. The trial set contains 64,207 GST ARBA unclassified proteins. The EBA method classifies 26,180 proteins whose range of lengths (within reference length range (RLR)) ensures structure conservation with respect to baits, about 41% of the total. A total of 58% is classified below and above the range (below and above RLR), and another 1% is not classified. The spreading of the classes in different taxa from those in Table 1 is discussed in the text. The trial dataset contains more bacteria genera (700) than the reference and testing dataset (200). Proteins classified in the reference length region belong to plant symbionts (Rhizobium, Sinorhizobium and Rhizobiales), plant pathogens (Acidovorax), photosynthetic bacteria (Synechoccus and Nostoc) and soil bacteria (Acinetobacter, Azospirillum, Myxococcus, Streptomyces, Sphingomonas and Variovorax). The classification led to the enrichment of new GST classes for bacteria not found in the reference and testing datasets. As an example, a couple of "new" alpha-class bacterial Myxococcus proteins (UniProt IDs: F8CAS1, Q1D6B3) share 34% sequence identity with alpha-class proteins in the reference dataset. More functional and structural studies are necessary for data validation.

2.3. Testing the Embedding Alignment Method

After embedding the reference proteins, we tested the EBA procedure (see Section 3) to classify the protein of the testing set, already classified by the UniRule/ARBA automatic annotation system of UniProt. Results are shown in Table 2. The main difference between ARBA and EBA is that ARBA classifies after finding conserved domains and/or motifs that are typical of the GST superfamily, without any constraint on the sequence length of the protein, while EBA considers the pairwise global alignment of any two protein embedded sequences. One should also consider that 15 different classes of the canonical cytosolic GSTs share the same fold, and therefore the same InterPro domains. This makes their classification difficult, considering that within classes remote homologs are also present. In this respect, Table 1 lists our baits along with the length range associated with the different folds, and their range of sequence identity. Since all of the baits are complete proteins, we follow the knowledge that a transfer of classification is reliable when the protein fold is conserved [43], and this implies that the protein length is conserved. Accordingly, in Table 2, we divide EBA-classified GST proteins into three groups: within, below and above the length range of fold conservation. This division identifies the fraction of the GST proteins (within the range of length of the of the baits) which conserve the structure. In this subset, the number of remote homologs with respect to the reference set is 76 (five in mu, three in sigma, thirteen in omega and fifty-five in kappa, see https://bar.biocomp.unibo.it/GST_Datasets/index.htm (accessed on 19 September 2024). Out of the 3D conservation range, the method is any way successful (Table 2). Overall, the prediction accuracy of our method with respect to ARBA in classifying proteins is very high (99.3%). Interestingly enough, it appears that the differences in classification between our method and ARBA rules in Table 2 are mainly confined in regions below and above the range of fold conservation (Table 2), when the protein length is lower or higher than those of the reference set. A closer inspection indicates that in these regions, the predicted protein can be included into the prey (lower length) or can include it (higher length). In other words, in these regions, the embedding-based alignment captures domains and motifs like the ARBA rules. Proteins in the below regions contain fewer motifs/domains than the prey while in the above regions they include extra motifs/domains. This is consistent with the notion that structure is not conserved; therefore, the definition of remote homologs fails. In these regions, errors are mainly due to the fact that baits from other classes share a higher identity with the GST protein at hand. Results are also detailed by taxon (Table S2).

2.4. Fishing in the Deep Sea

We tried EBA to classify 64,207 ARBA-unclassified GST proteins (Table 3). We show only results obtained on GST proteins whose length is in the length range of fold conservation and classify 41% of the total. Out of the safety range we can transfer class to another 58% GST proteins. The range of classes seems to increase, particularly in GST proteins from bacteria, and this can be explained by considering the new bacterial genomes recently included in TrEMBL. However, more functional and structural studies are necessary for data validation. Setting a length interval for structure conservation highlights more reliable predictions.

3. Materials and Methods

3.1. Dataset Generation

In order to address our task, we downloaded three different datasets from UniProt (release 2024_01, https://www.uniprot.org/ (accessed on 24 January 2024)).

3.1.1. The Reference Dataset

We collected all of the GST proteins including "Glutathione S-transferase" and/or "Glutathione transferase" in the protein name, endowed with a PDB structure or a high-quality AlphaFold2 model [44], and excluding fragments. For each of the proteins linked to a PDB structure, we selected a representative based on resolution, sequence coverage

(higher than 70%) and, when possible, in complex with the glutathione substrate. We checked that proteins lacking PDB structures are endowed with high-quality AlphaFold2 models (with a per protein average pLDDT (predicted Local Distance Difference Test) value ≥ 70 [44]), whose root mean square value (RMSD) to the backbone of the 3D representatives of the class is $\leq$1.5 Å. We retained 284 proteins (Table 1), characterized by six structural types, and grouped into 20 GST classes. We also grouped GST reference proteins in relation to their taxa (https://www.ncbi.nlm.nih.gov/taxonomy (accessed on 15 June 2024), [45]). The reference dataset is available at https://bar.biocomp.unibo.it/GST_Datasets/index.htm (accessed on 19 September 2024).

3.1.2. The Testing and Trial Datasets

With a similar search we collected all of the TrEMBL (https://www.uniprot.org/ (accessed on 24 January 2024) sequences named "Glutathione S-transferase" and/or "Glutathione Transferase". Routinely, the protein name is automatically assigned, together with the Enzyme Commission (EC) number, when the sequence entry annotation satisfies either UniRule ID UR000000494 or ARBA ID ARBA00012452, respectively (https://www.uniprot.org/help/arba (accessed on 24 January 2024), [38,46]). The rules are routinely based on the automatic recognition of GST-specific motifs and/or domains in the sequence. In some cases, ARBA rules, satisfying class-specific InterPro [47] signatures, assign a specific class to the protein (ARBA rules are present for 14 of the 20 classes, https://www.uniprot.org/arba (accessed on 19 September 2024) (see above)). After filtering out all of the entries with a "Caution" statement in the "Function" field of the protein file and proteins with sequence identity values higher than 95% to the reference set, we retained a testing set with 15,061 GST proteins annotated with an assigned class and a trial set with 64,207 GST proteins without classification.

3.2. Embedding Procedure

3.2.1. Embedding Generation

Among the pLMs currently available, the MetaAI ESM2 encoding set has been used to train ESMFold [6], an advanced protein tertiary structure prediction method ([48], and references therein). We adopt the most recent ESM2 pLM: ESM2-15b trained on 65 million proteins [6]. For each protein in the three datasets, we extracted the ESM2-15b representations following the instructions and scripts available at https://github.com/facebookresearch/esm (accessed on 1 November 2022). Given an input protein sequence of length l, the pLM outputs meaningful distributed vector representations for each amino acid residue of the protein at hand. The size D of the vectors depends on the number of hidden states of the transformer layers from which the representations are extracted (routinely the last one). ESM2-15b outputs vectors with $D = 5120$. The final encoding of the protein is therefore a matrix $e \in R^{l \times D}$, (with l as the protein length), routinely referred to as per-residue protein embedding.

3.2.2. Embedding-Based Alignment

We compared per-residue GST protein embeddings exploiting the embedding-based alignment (EBA) method [18]. The algorithm (available at https://git.scicore.unibas.ch/schwede/EBA (accessed on 1 January 2023)) computes a pairwise distance matrix of per-residue embeddings, evaluating the Euclidean distance of all embedded residues. These values fill a matrix of dimension $l_1 \times l_2$ (where l_1 and l_2 are the lengths of the two proteins, respectively), which provides the substitution scores for the pairwise alignment based on a classic dynamic programming approach. The tool includes also an optimizing intermediate step, called "signal enhancement" [18], where each score of a residue pair is normalized to the scores of all residue pairs of the two aligned proteins. We adopt this enhanced similarity matrix to score the pairwise global alignment obtained with the Needleman–Wunsch (NW) method [18]. Following the procedure, we normalize the alignment similarity score s_{align} by the length l of the longer sequence in the pair (l_{max}), according to the following:

$$EBA_{min} = \frac{s_{align}}{l_{max}}$$

Following [18], the length normalization is an important factor on the final score when the proteins being compared are very different in length. Whenever the difference is large, a high EBA_{max} score, obtained by normalizing by the length of the shortest protein, reflects the fact that the shorter sequence is entirely contained in the longest [18]. In this case, EBA_{min} is much lower since the longer sequence is only partially aligned. Following the author's suggestion and considering template structure conservation as an essential element of knowledge transfer [43], we adopted the EBA_{min} to score any two protein sequences during our procedure. In any case, each protein after embedding is aligned with all of the other ones. For a given query protein, we compute EBA_{min} by aligning to all of the proteins in the reference set. The query protein assumes the class annotation of the best scoring protein among the references and then classification (annotation) is transferred. By this a bait can "fish" a prey (Figure 1).

The main difference of our method, as compared to [18], is the adoption of ESM2-15b with an embedding vector dimension of 5120 and a different procedure for the output selection (ProstT5 [4] with a vector dimension of 1024 was adopted in the original implementation [18]). We analyze results considering that the transfer of knowledge requires structure template conservation [43] and for this reason, present the results as a function of the protein length.

3.3. Computational Time

After downloading EBA in house, the time required to align 100,000 proteins with our reference set (284) was one week with a machine endowed with 80 CPUs and a 754 gigabyte RAM.

4. Conclusions and Perspectives

In this paper, we exploit the capabilities of an embedding-based alignment method [EBA, 18] to annotate proteins like the UniProt ARBA system of automatic annotation. For this, we focused on the GST protein superfamily, given the complex relationship among sequence and structure in the different protein classes which in different taxa characterize the group. GST main characteristics include sequences of different lengths, sharing the same folding when active in the same cellular compartments (Table 1). This blurs the classification of GST proteins in newly sequenced proteomes. The UniProt ARBA automatic annotation system annotates into GST classes proteins of any length, provided that InterPro motifs and/or domains are conserved, without taking into consideration the fold conservation, which obviously sets a limit for the protein length. We find that EBA performance compares well with the ARBA rule annotation system (over 99% of accuracy), and from error analysis (Table 2), we derive as a rule of thumb that classification is optimal when fold is also conserved. We find that at least 46% of GSTs of a selected subset of classified TrEMBL GST proteins do not conserve the length of the reference protein folding typical of the class. These proteins, beyond the conservation of the typical GST domains, are often endowed with other domains, possibly suggesting new folds, not yet experimentally available. EBA routinely does not misclassify protein fold, and as to misclassification within a class, it may happen that when the bait is contained or contains the prey, matrix alignment is not sufficient to recognize the subdomain. In this case, sequence alignment to another close class can prevail over the ARBA annotation. The EBA assignment in the fold conservation region is able to recover remote homologs in four classes (mu, sigma, omega and kappa), confirming the capability of the system to also assign to a class/family proteins sharing low sequence identity [18]. We also classify proteins not yet classified in UniProt, releasing a list of proteins for experimental validation. This can encourage experiments to further cluster GST proteins in more functional classes, for a better definition of their role in cell complexity. We propose the EBA classification procedure as a valid complement to the ARBA rule

classification system for the GST superfamily, considering that sequence embeddings carry along information on structural templates, motifs and domains of the family.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules29194616/s1, Supplementary S1: The GST superfamily; Table S1: Interclass sequence identity between proteins in the reference set; Table S2: Embedding based GST classification on the ARBA Test set.

Author Contributions: Conceptualization, G.V., C.S., P.L.M. and R.C.; methodology, G.V., C.S., P.L.M. and R.C.; formal analysis, G.V., C.S., P.L.M. and R.C.; data curation, G.V., C.S., P.L.M. and R.C.; writing—original draft preparation, G.V, C.S., P.L.M. and R.C.; writing—review and editing, G.V., C.S., P.L.M. and R.C.; supervision, P.L.M. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the European Union- NextGenerationEU through the Italian Ministry of University and Research under the projects "Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics" (ElixirxNextGenIT)" (Investment PNRRM4C2-I3.1, Project IR_0000010, CUP B53C22001800006) and "HEAL ITALIA" (Investment PNRR-M4C2-I1.3, Project PE_00000019, CUP J33C22002920006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data underlying this article are available in the article, in its online Supplementary Material, and at https://bar.biocomp.unibo.it/GST_Datasets/index.htm (accessed on 19 September 2024).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
2. Bepler, T.; Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **2021**, *12*, 654–669.e3. [CrossRef]
3. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127. [CrossRef] [PubMed]
4. Heinzinger, M.; Weissenow, K.; Sanchez, J.G.; Henkel, A.; Mirdita, M.; Steinegger, M.; Rost, B. Bilingual Language Model for Protein Sequence and Structure. *bioRxiv* **2023**. [CrossRef]
5. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016239118. [CrossRef]
6. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130. [CrossRef]
7. Kandathil, S.M.; Lau, A.M.; Jones, D.T. Machine learning methods for predicting protein structure from single sequences. *Curr. Opin. Struct. Biol.* **2023**, *81*, 102627. [CrossRef] [PubMed]
8. Jha, K.; Saha, S.; Singh, H. Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.* **2022**, *12*, 8360. [CrossRef]
9. Manfredi, M.; Savojardo, C.; Martelli, P.L.; Casadio, R. ISPRED-SEQ: Deep Neural Networks and Embeddings for Predicting Interaction Sites in Protein Sequences. *J. Mol. Biol.* **2023**, *435*, 167963. [CrossRef]
10. Heinzinger, M.; Littmann, M.; Sillitoe, I.; Bordin, N.; Orengo, C.; Rost, B. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinf.* **2022**, *4*, lqac043. [CrossRef]
11. Yeung, W.; Zhou, Z.; Li, S.; Kannan, N. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. *Brief. Bioinform.* **2023**, *24*, bbac599. [CrossRef]
12. Kaminski, K.; Ludwiczak, J.; Pawlicki, K.; Alva, V.; Dunin-Horkawicz, S. pLM-BLAST: Distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* **2023**, *39*, btad579. [CrossRef]
13. Yeung, W.; Zhou, Z.; Mathew, L.; Gravel, N.; Taujale, R.; O'Boyle, B.; Salcedo, M.; Venkat, A.; Lanzilotta, W.; Li, S.; et al. Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Brief. Bioinform.* **2023**, *24*, bbac619. [CrossRef]
14. Hie, B.L.; Yang, K.K.; Kim, P.S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **2022**, *13*, 274–285.e6. [CrossRef] [PubMed]

15. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V.P.; Ashford, P.; Scholes, H.M.; Pang, C.S.M.; Woodridge, L.; Rauer, C.; Sen, N.; et al. CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **2021**, *49*, D266–D273. [CrossRef] [PubMed]
16. Hamamsy, T.; Morton, J.T.; Blackwell, R.; Berenberg, D.; Carriero, N.; Gligorijevic, V.; Strauss, C.E.M.; Leman, J.K.; Cho, K.; Bonneau, R. Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.* **2024**, *42*, 975–985. [CrossRef] [PubMed]
17. Kabir, A.; Moldwin, A.; Shehu, A. A Comparative Analysis of Transformer-based Protein Language Models for Remote Homology Prediction. In Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Houston, TX, USA, 3–6 September 2023; pp. 1–9. [CrossRef]
18. Pantolini, L.; Studer, G.; Pereira, J.; Durairaj, J.; Tauriello, G.; Schwede, T. Embedding-based alignment: Combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. *Bioinformatics* **2024**, *40*, btad786. [CrossRef]
19. Mazari, A.M.A.; Zhang, L.; Ye, Z.-W.; Zhang, J.; Tew, K.D.; Townsend, D.M. The Multifaceted Role of Glutathione S-Transferases in Health and Disease. *Biomolecules* **2023**, *13*, 688. [CrossRef]
20. Allocati, N.; Federici, L.; Masulli, M.; Di Ilio, C. Glutathione transferases in bacteria. *FEBS J.* **2009**, *276*, 58–75. [CrossRef]
21. Meux, E.; Prosper, P.; Ngadin, A.; Didierjean, C.; Morel, M.; Dumarçay, S.; Lamant, T.; Jacquot, J.-P.; Favier, F.; Gelhaye, E. Glutathione Transferases of Phanerochaete chrysosporium. *J. Biol. Chem.* **2011**, *286*, 9162–9173. [CrossRef]
22. Huang, C.; Chen, M.; Pang, D.; Bi, D.; Zou, Y.; Xia, X.; Yang, W.; Luo, L.; Deng, R.; Tan, H.; et al. Developmental and Activity-Dependent Expression of LanCL1 Confers Antioxidant Activity Required for Neuronal Survival. *Dev. Cell* **2014**, *30*, 479–487. [CrossRef]
23. Kumar, S.; Trivedi, P.K. Glutathione S-Transferases: Role in Combating Abiotic Stresses Including Arsenic Detoxification in Plants. *Front. Plant Sci.* **2018**, *9*, 751. [CrossRef]
24. Morel, F.; Aninat, C. The glutathione transferase kappa family. *Drug Metab. Rev.* **2011**, *43*, 281–291. [CrossRef]
25. Oakley, A. Glutathione transferases: A structural perspective. *Drug Metab. Rev.* **2011**, *43*, 138–151. [CrossRef] [PubMed]
26. Bresell, A.; Weinander, R.; Lundqvist, G.; Raza, H.; Shimoji, M.; Sun, T.; Balk, L.; Wiklund, R.; Eriksson, J.; Jansson, C.; et al. Bioinformatic and enzymatic characterization of the MAPEG superfamily. *FEBS J.* **2005**, *272*, 1688–1703. [CrossRef] [PubMed]
27. Zhuge, X.-L.; Xu, H.; Xiu, Z.-J.; Yang, H.-L. Biochemical Functions of Glutathione S-Transferase Family of *Salix babylonica*. *Front. Plant Sci.* **2020**, *11*, 364. [CrossRef]
28. Sonu Koirala, B.K.; Moural, T.; Zhu, F. Functional and Structural Diversity of Insect Glutathione S-transferases in Xenobiotic Adaptation. *Int. J. Biol. Sci.* **2022**, *18*, 5713–5723. [CrossRef]
29. Konishi, T.; Kato, K.; Araki, T.; Shiraki, K.; Takagi, M.; Tamaru, Y. A new class of glutathione S-transferase from the hepatopancreas of the red sea bream Pagrus major. *Biochem. J.* **2005**, *388*, 299–307. [CrossRef] [PubMed]
30. Munyampundu, J.-P.; Xu, Y.-P.; Cai, X.-Z. Phi Class of Glutathione S-transferase Gene Superfamily Widely Exists in Nonplant Taxonomic Groups. *Evol. Bioinform.* **2016**, *12*, 59–71. [CrossRef]
31. Shehu, D.; Abdullahi, N.; Alias, Z. Cytosolic Glutathione S-transferase in Bacteria:A Review. *Pol. J. Environ. Stud.* **2018**, *28*, 515–528. [CrossRef]
32. Garcerá, A.; Barreto, L.; Piedrafita, L.; Tamarit, J.; Herrero, E. *Saccharomyces cerevisiae* cells have three Omega class glutathione S-transferases acting as 1-Cys thiol transferases. *Biochem. J.* **2006**, *398*, 187–196. [CrossRef]
33. Xun, L.; Belchik, S.M.; Xun, R.; Huang, Y.; Zhou, H.; Sanchez, E.; Kang, C.; Board, P.G. S-Glutathionyl-(chloro)hydroquinone reductases: A novel class of glutathione transferases. *Biochem. J.* **2010**, *428*, 419–427. [CrossRef]
34. Blisnick, T.; Vincensini, L.; Barale, J.C.; Namane, A.; Braun Breton, C. LANCL1, an erythrocyte protein recruited to the Maurer's clefts during Plasmodium falciparum development. *Mol. Biochem. Parasitol.* **2005**, *141*, 39–47. [CrossRef]
35. Ladner, J.E.; Parsons, J.F.; Rife, C.L.; Gilliland, G.L.; Armstrong, R.N. Parallel Evolutionary Pathways for Glutathione Transferases: Structure and Mechanism of the Mitochondrial Class Kappa Enzyme rGSTK1-1. *Biochemistry* **2004**, *43*, 352–361. [CrossRef] [PubMed]
36. Morgenstern, R.; Zhang, J.; Johansson, K. Microsomal glutathione transferase 1: Mechanism and functional roles. *Drug Metab. Rev.* **2011**, *43*, 300–306. [CrossRef] [PubMed]
37. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **1999**, *12*, 85–94. [CrossRef] [PubMed]
38. UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
39. Ketterman, A.J.; Saisawang, C.; Wongsantichon, J. Insect glutathione transferases. *Drug Metab. Rev.* **2011**, *43*, 253–265. [CrossRef]
40. Scian, M.; Le Trong, I.; Mazari, A.M.A.; Mannervik, B.; Atkins, W.M.; Stenkamp, R.E. Comparison of epsilon- and delta-class glutathione S -transferases: The crystal structures of the glutathione S -transferases DmGSTE6 and DmGSTE7 from Drosophila melanogaster. *Acta Crystallogr. D Biol. Crystallogr.* **2015**, *71*, 2089–2098. [CrossRef]
41. Park, H.; Ahn, I.-Y.; Kim, H.; Lee, J. Glutathione S-transferase as a biomarker in the Antarctic bivalve. *Laternula elliptica* after exposure to the polychlorinated biphenyl mixture Aroclor 1254. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **2009**, *150*, 528–536. [CrossRef]
42. Thom, R.; Cummins, I.; Dixon, D.P.; Edwards, R.; Cole, D.J.; Lapthorn, A.J. Structure of a Tau Class Glutathione S- Transferase from Wheat Active in Herbicide Detoxification. *Biochemistry* **2002**, *41*, 7008–7020. [CrossRef] [PubMed]
43. Lesk, A.M. *Introduction to Protein Science*, 3rd ed.; Oxford University Press: Oxford, UK, 2016.

44. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
45. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* **2020**, *2020*, baaa062. [CrossRef] [PubMed]
46. MacDougall, A.; Volynkin, V.; Saidi, R.; Poggioli, D.; Zellner, H.; Hatton-Ellis, E.; Joshi, V.; O'Donovan, C.; Orchard, S.; Auchincloss, A.H.; et al. UniRule: A unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* **2020**, *36*, 4643–4648. [CrossRef]
47. Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B.L.; Salazar, G.A.; Bileschi, M.L.; Bork, P.; Bridge, A. Colwell InterPro in 2022. *Nucleic Acids Res.* **2023**, *51*, D418–D427. [CrossRef]
48. Manfredi, M.; Savojardo, C.; Iardukhin, G.; Salomoni, D.; Costantini, A.; Martelli, P.L.; Casadio, R. Alpha&ESMhFolds: A Web Server for Comparing AlphaFold2 and ESMFold Models of the Human Reference Proteome. *J. Mol. Biol.* **2024**, *436*, 168593. [CrossRef]

Journal of Computational Biophysics and Chemistry

Research

Cite This: *J. Comput. Biophys. Chem.* **2026**, *25* (4), 627–640

Can Human Canonical Glutathione S-Transferases Act as RNA-Binding Proteins?

Gabriele Vazzana *, Pier Luigi Martelli *,† and Rita Casadio †,‡,§

*Department of Pharmacy and Biotechnology, Biocomputing Group, University of Bologna, Italy
‡AlmaClimate Interdepartmental Center, Biocomputing Group, University of Bologna, Italy
†Institute of Biomembrane and Bioenergetics, Italian National Research Council (IBIOM-CNR), Italy
§Corresponding author. E-mail: *rita.casadio@unibo.it*

ABSTRACT: Glutathione *S*-transferase (GST) is an enzyme superfamily of particular interest for human health with many functional roles, and it is involved in several cancer types. Under cell stress conditions, their concentrations can increase by up to 10% of cell protein content. Recently, a study describing the landscape of RNA-binding proteins in mammalian spermatogenesis reported evidence of canonical GST–RNA interactions in three mouse mu GSTs. Prompted by this, we searched for available databases and found that RBP2GO, which collects candidate RNA-binding proteins (RBPs) detected in recent human proteomic studies, also lists a few human GST–RNA interactions, without any molecular details. To highlight the molecular features of the GST–RNA interaction, we applied recently developed predictors of RNA-binding sites and validated the results with AutoDock Vina, a docking program that computes binding affinity. Overall, our findings support the notion that a GST–RNA interaction can exist and suggest a potential overlap between RNA binding sites and residues responsible for binding glutathione (GSH), which is the most common GST substrate. Our computational analysis supports the notion that human GSTs can bind RNA that shares the binding region with the glutathione-binding pocket.

KEYWORDS: Glutathione *S*-transferases; RNA-binding proteins; moonlighting proteins; molecular docking.

1. INTRODUCTION

In a crowded cell environment, RNA, comprising 10–20% of the cell's dry weight (approximately 20–40 mg/ml), plays a crucial role in cellular physiology[1] and the expression of genetic information.[2] To accomplish these diverse functions, RNA interacts with specific proteins known as RNA-binding proteins (RBPs), which perform numerous biological processes.[3,4] However, the recognition of RNA binding sites on RBPs is not always straightforward.[5]

Proteomic studies on human immortalized cell lines, mainly involving RNA–interactome capture techniques (RIC),[6] have identified new proteins that bind RNA without a previously annotated RNA-binding behavior. The results were recently collected from the RBP2GO database[7] (available at https://rbp2go.dkfz.de/, last release August 2023), which collects information on 22,552 RBP candidates from 13 organisms (6100 are from *Homo sapiens,* 1% of the total proteome). Such evidence suggests that certain proteins, already known for their important biological roles, can perform multiple functions by "moonlighting" as RBPs,[8] supporting the hypothesis of a "riboregulation," in which protein activity can be modulated through interactions with RNA.[9] Additionally, a novel study,[10] addressing the role of RBPs in mammalian spermatogenesis, has identified new candidate RBPs in mouse male germ cells (mMGCs). This is interesting, as

Received: 24 February 2025
Accepted: 24 April 2025
Published: 9 July 2025

DOI: 10.1142/S2737416525500607

the mammalian testis transcribes the majority of the genome and thus harbors one of the most complex tissue transcriptomes. We focused on three candidate RBPs, identified in the work,[10] belonging to the glutathione S-transferase (GST; EC 2.5.1.18) superfamily and specifically members of the Mu class with a canonical GST fold, GSTM1, GSTM2 and GSTM5, corresponding to UniProt[11] IDs P10649, P15626 and P48774, respectively (https://www.uniprot.org/). GST proteins are particularly intriguing because they are crucial for human health and are already known to be multifunctional.[12–14] The superfamily is composed of a diverse group of enzymes organized into classes that vary in both structure and cellular localization.[15] Canonical GSTs are relatively small cytosolic enzymes (200–250 residues) with a conserved distinct fold (within a maximum root mean square deviation (RMSD) of 3 Å), characterized by an N-terminal thioredoxin-like domain ($\beta\alpha\beta\alpha\beta\beta\alpha$) and a C-terminal all-alpha domain.[12] Despite sharing this common fold, canonical GSTs have highly variable sequences, posing challenges for annotation through standard alignment methods.[16] As a consequence, they are routinely clustered in subclasses, indicated with Greek letters and introduced on the basis of sequence similarity, structural properties and chromosomal organization in different organisms.[12,16] In humans, there are a total of seven cytosolic canonical GSTs: mu, sigma, alpha, pi, theta, omega and zeta. These enzymes play primary roles in cellular detoxification and the stress response by catalyzing the conjugation of glutathione to toxic compounds, thereby promoting their export from the cell.[12] Glutathione (GSH) is a tripeptide (γ-L-glutamyl-L-cysteinylglycine) present in all mammalian tissues at 1–10 mM concentrations, with the highest expression level in the liver and whose increased transcription is associated with oxidative stress.[17] Similarly, GST enzymes are widely expressed in nearly all human tissues and are overexpressed in several human pathologies[18] when their concentrations can rise to as much as 10% of the cytosolic protein content under stress conditions.[19] Pi-class cytosolic GSTs are of particular interest as they are most widely expressed in human tissues, and their overexpression in cancer cells is highly correlated with chemoresistance.[18,20]

In cytosolic GSTs, GSH binds within a specific pocket, known as the G-site, which largely consists of residues from the N-terminal domain. The H-site, which is responsible for binding with toxic compounds, involves residues of the C-terminal domain.[21] Canonical GST proteins are functionally active upon dimerization.[12]

To our knowledge, there is no annotation record of interaction between GSTs and RNA in online databases such as UniProt (https://www.uniprot.org/), notwithstanding the evidence of such interaction in a 2013 study reporting the binding between a plant GST and the Bamboo mosaic virus RNA at the 3'-UTR region.[22]

Based on this, we searched for human proteins homologous to mouse RBP candidates of the GST mu-class in the RBP2GO database (release August 2023) and found two mu-class enzymes homologous to mouse proteins (sequence identity ≥ 65%): GSTM2 and GSTM3 (UniProt ID P28161 and P21266, respectively). In addition, we found three more cytosolic GSTs belonging to the pi, omega and zeta classes, GSTP1, GSTO1 and MAAI (GSTZ1), respectively, (UniProt ID P90211, P78417 and O43708, respectively). With this support, we conducted a bioinformatics and computational analysis of the RNA-binding sites on human GSTs by leveraging recent machine learning algorithms and docking programs.

Experimental determination of RBPs provides room for the development of computational methods to predict RNA-binding proteins. Two recent reviews[23,24] list RNA-binding site predictors, both sequence- and structure-based. Because the latter methods perform routinely better than the former,[25] we selected two recent and state-of-the-art structure-based methods[23] that are online and available as web servers: PST-PRNA[26] (http://www.zpliulab.cn/PSTPRNA) and GraphBind[27] (http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/).

We further validate the hypothesis that GSTs can bind RNA using docking algorithms. Molecular docking employs computational methods to approximate a ligand–receptor complex with an energetically favorable geometry, aiming to (ideally) replicate an experimental binding mode.[28] Early docking algorithms treat both binding partners as rigid bodies; however, more recent methods, such as AutoDock Vina, allow flexibility at the level of the ligand or both the ligand and receptor.[29,30]

Our results, all together, suggest that a GST–RNA interaction is physically reliable and suggests a potential overlap between RNA-binding sites and residues responsible for binding with glutathione, the common GST substrate. We demonstrated that RNA and glutathione bind to the same protein region and discuss the possible role of both molecules in cell physiology.

2. MATERIALS AND METHODS

2.1. Materials

2.1.1. *RBP2GO database*

The five human canonical GSTs were identified as candidate RBPs in the RBP2GO database (release August 2023),[7] which ranks proteins based on their occurrence

DOI: 10.1142/S2737416525500607

in recent RNA proteomic studies. In total, the human GSTs appear in nine studies,[31–39] leveraging both polyA+ and polyA- capture strategies, that is, experimental methodologies that adopt (polyA+) and do not adopt (polyA-) oligo(dT) probes to capture RNA molecules. Oligo-A-sequences are commonly present in eukaryotic mRNAs; therefore, polyA- methods are supposed to be a more general approach for the identification of RBPs.[39] The database defines an RBP2GO score that ranges from 0 to 100, reflecting the probability of a protein to be an RBP. The score is calculated considering two indicators of RBP propensity: (i) the frequency of a given protein to be listed as RBP in the different datasets stored in the database and (ii) the average frequency of interactions with RBP proteins according to STRING (https://string-db.org/) (up to the top ten interaction partners). These two values were combined with an equal weight in the RBP2GO-score, which was normalized to the number of experimental datasets for each species (43 for *Homo sapiens*).

2.2. Methods

2.2.1. *Prediction of RNA binding sites*

Two RNA-binding prediction algorithms were adopted and used through the respective web servers to obtain predictions of RNA binding sites on the candidate RBPs GSTs. Both adopt protein structures (PDB format) as inputs and are briefly described below.

PST-PRNA[26] predicts RNA binding residues on protein structures by encoding residue descriptors with physicochemical properties, evolutionary features (exploiting position-specific scoring matrix (PSSM) with PSI-BLAST),[40] secondary structure and solvent accessibility information (by adopting DSSP),[41] and hidden Markov Model profiles (with HHblits).[42] The architecture of the model is a convolutional network highly inspired by the classical ResNet18 basic architecture.[43] The model was trained on a low-homology dataset of high-resolution structures gathered from the Protein Data Bank,[44] comprising RNA-binding proteins documented in the Nucleic Acid Database.[45] PST-PRNA, which shows reliable and effective performance,[26] takes into consideration solvent-accessible residues on the protein surface and excludes buried residues. GraphBind[27] represents protein structures as graphs, where nodes are residues and edges are defined according to the spatial relationships among residues. Node representations are extracted from both protein structure and sequence information: pseudo-positions, atomic features of residues, DSSP,[41] PSI-BLAST,[40] and HHblits.[42] The model architecture employs a hierarchical graph neural network (HGNNs). The training dataset was collected from the BioLip database,[46] and comprises a set of biologically relevant ligand–protein interactions that are structurally well-solved in complexes.

2.2.2. *Surface charge characterization*

RNA molecules are polyanions due to the presence of negatively charged phosphate groups; consequently, most protein–RNA interfaces prefer positively charged residues.[47] We characterized the electrostatic properties of protein surfaces using two tools, PDB2PQR and APBS,[48] directly from the web server (https://server.poissonboltzmann.org/). PDB2PQR automates common tasks of preparing structures for continuum solvation calculations and computes the protonation state of residues, and the adaptive Poisson-Boltzmann solver (APBS) computes the electrostatics of a protein by solving the equations of continuum electrostatics (https://www.poissonboltzmann.org/). The programs were run using default parameters.

2.2.3. *Visualization*

We adopted FoldMason[49] from the respective web server (https://search.foldseek.com/foldmason) to compute the multiple structural alignment and relative multiple sequence alignment of GST proteins. Protein structures were visualized using the PyMol molecular visualization tool (PyMOL Molecular Graphics System, Version 3.1.1 Schrödinger, LLC). The same tool was also adopted to highlight residues in the GSH-containing pocket (G-site) and miRNA-binding regions. We used LigPlot+[50] to visualize 2D interaction diagrams between proteins and substrates/RNAs. The protein–ligand connectivity (both hydrogen-bonds and non-bonded interactions) generated by the program is discussed in detail in the original paper.[51]

2.2.4. *Molecular docking*

AutoDock Vina[29,30] (here referred to as Vina) is a molecular docking program that computes non-covalent interactions between a rigid receptor, routinely a protein and a flexible ligand aimed at determining bound conformations and binding affinity while balancing accuracy and computational speed. To predict binding affinity, Vina employed a scoring function to approximate chemical potentials tuned from the PDBbind database[52]: the

DOI: 10.1142/S2737416525500607
J. Comput. Biophys. Chem. **2026**, *25* (4), 627–640

function approach is more of "machine learning" than directly physics-based in its nature as discussed in the original paper.[29] The scoring function calculates the sum of pairwise interactions between movable atoms and accounts for steric, hydrophobic and hydrogen-bonding interactions. The predicted free energy of binding was derived from the lowest-scoring conformation. Vina used an iterated local search algorithm combined with the BFGS[53] method for local optimization.

We installed Vina (version 1.2.5) locally by following the documentation (https://autodock-vina.readthedocs.io/en/latest/).

During data preparation, the user can define a confined search space with a grid box on the receptor molecule to help the program dock the ligand on a preferred region, which is useful when there is prior knowledge about putative binding sites. In our experiments, we adopted a sufficiently wide grid box to accommodate both GSH and miRNA molecules. For pri-miR-18a-oligo1 (PDB: 6UV4,[54] seven nucleotides), the grid dimensions were defined as follows: size_x, size_y and size_z equal to 30.00, 24.75 and 28.5 Å, respectively. For the pri-miRNA-18a terminal loop (PDB: 6DCL,[55] 11 nucleotides), the following grid dimensions were set: size_x, size_y and size_z equal to 41.25, 36.00 and 39.75 Å, respectively. The remaining parameters were left as default.

3. RESULTS AND DISCUSSION

3.1. Human GSTs in RBP2GO

The RBP2GO database[7] (release August 2023) ranks proteins from 13 different organisms (including *Homo sapiens*) based on their occurrence in recent RNA proteomic studies. We searched for human proteins in the RBP2GO database (release August 2023) homologous to mouse GSTs found as candidate RBPs in[10] (GSTM1, GSTM2 and GSTM5, corresponding to UniProt IDs P10649, P15626 and P48774, respectively). We identified two homologous human GSTs (with sequence identity>65%), GSTM2 and GSTM3 (UniProt ID P28161 and P21266, respectively). As all canonical cytosolic GSTs share the same fold ([16] and references therein), we also included three more human GSTs present in the RBP2GO database belonging to the pi, omega and zeta classes: GSTP1, GSTO1 and MAAI (GSTZ1) (UniProt ID P90211, P78417 and O43708, respectively).

All five GSTs shown in Table 1 are multifunctional enzymes in UniProt and are reported in MultifacetedProtDB[56] (release 2023-03, https://multifacetedprotdb.biocomp.unibo.it/). The best-ranking human GSTs were GSTP1 and GSTO1, with scores of 10.3 and 6.5, respectively. Even with low scores, proteins appear in multiple RNA–proteomic studies: seven for GSTP1[31,33–36,38,39] and four for GSTO1.[34,37–39] GSTP1 is of particular interest as it is a well-studied protein that is highly involved in several aspects of human health,[57,58] including its potential role as a therapeutic target in ovarian cancer.[58]

3.2. Prediction of RNA binding sites in GST proteins

We adopted two recently released structure-based machine learning methods (described in the "Methods" section) to predict the RNA binding sites of the eight GST proteins (three mouse GSTs[10] and five human GSTs of Table 1). Figure 1 shows the

Table 1. Candidate RNA binding human GSTs in RBP2GO.

Gene name	UniProt ID	Protein name	[1]Class	[2]PDB	[3]RBP2GO score	[4]Pub no
GSTP1	P09211	Glutathione S-transferase P	Pi	8GSS chain A	10.3	7
GSTO1	P78417	Glutathione S-transferase omega-1	Omega	5YVN chain A	6.5	4
*MAAI (GSTZ1)	O43708	Maleylacetoacetate isomerase	Zeta	1FW1 chain A	3.8	2
GSTM2	P28161	Glutathione S-transferase Mu 2	Mu	1XW5 chain A	3.1	2
GSTM3	P21266	Glutathione S-transferase Mu 3	Mu	°3GTU chain B	2	1

[1]Class refers to the class or family of GST proteins in UniProt.

[2]PDB: identifiers of the PDB structures adopted to represent the GST proteins; monomers were derived from dimers, which are reported to be the protein functional unit.[12] Among available structures, PDB representatives were selected according to three criteria: (i) crystallographic structures in complex with substrate glutathione (ii) the highest resolution and (iii) coverage to the protein sequence.

[3]RBP2GO-score: entries are ranked according to the RBP2GO-score.[7] Briefly, it ranges from 0 to 100 and is computed by weighting the occurrence of the protein in the RBP datasets and the average frequency of protein interactors according to STRING (see Methods for details).

[4]Pub No: number of publications (listing counts in RBP2GO) in which the protein was reported.

*MAAI: GSTZ1 human protein, which in UniProt has a primary function as maleylacetoacetate isomerase (EC:5.2.1.2).

°GSTM3 has been crystallized as a heterodimer, and the GST function is confined to chain B in the 3GTU file.

DOI: 10.1142/S2737416525500607

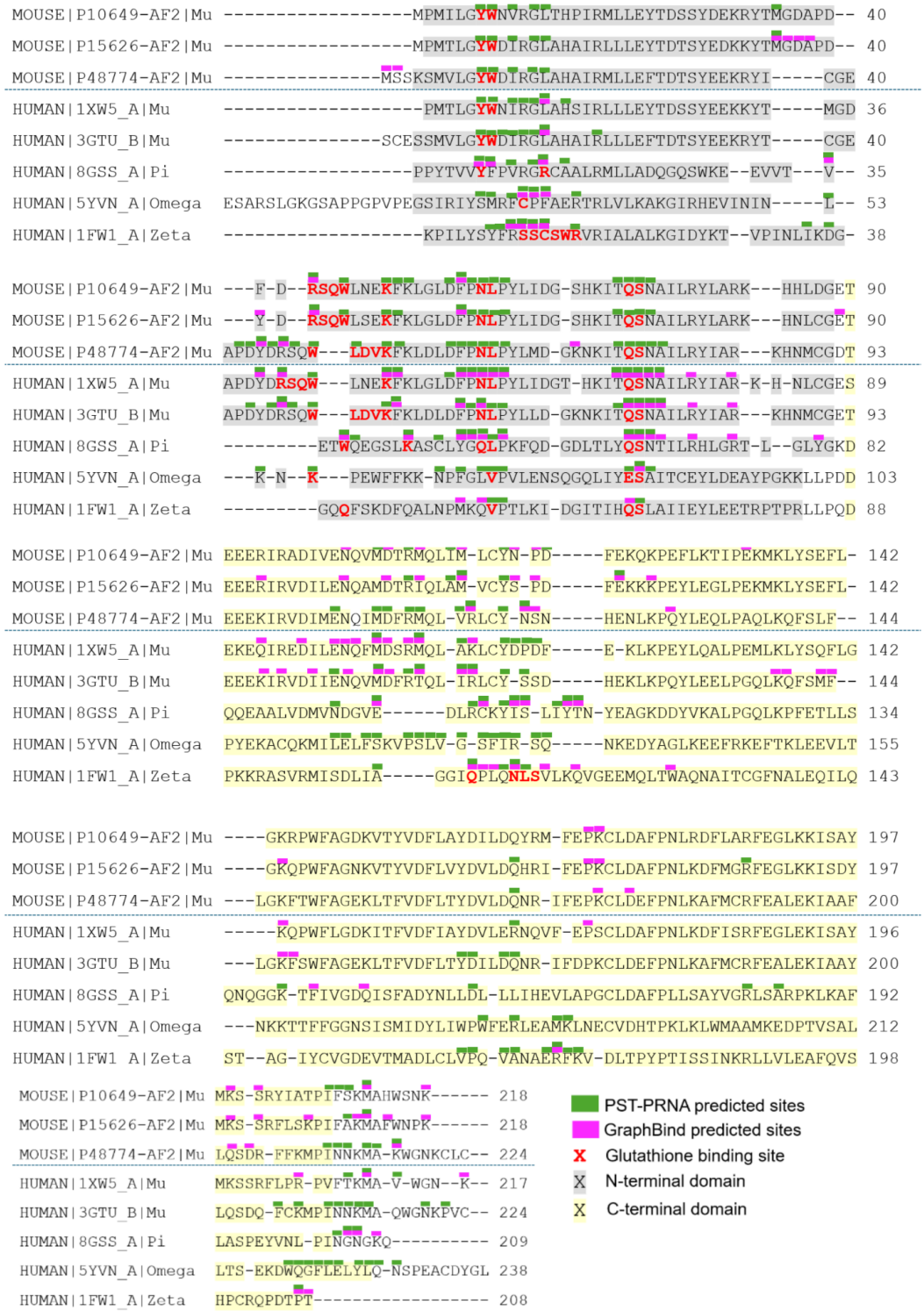


Fig. 1. (Color online) Sequence alignment of mouse and human GSTs after structural alignment. Multiple sequence alignments have been obtained after multiple structural alignments (root mean square deviation at the backbone level < 3Å) with FoldMason[49] (https://search.foldseek.com/foldmason) and includes the five humans GST structures and three mouse GST high-quality AlphaFold2 models (predicted local distance difference test (pLDDT)[59] > 90). PDB representatives were selected according to three criteria: (i) crystallographic structures in complex with substrate glutathione, (ii) the highest resolution and (iii) coverage to the protein sequence. A dotted line separates mouse from human proteins. Glutathione binding residues are red-colored according to UniProt annotation. RNA binding residues predicted with the two structure-based methods, PST-PRNA and GraphBind, are colored in green and magenta, respectively. Residues belonging to the N- and C-terminal domains (according to UniProt annotation) are grey- and yellow-shaded, respectively.

DOI: 10.1142/S2737416525500607
J. Comput. Biophys. Chem. **2026**, *25* (4), 627–640

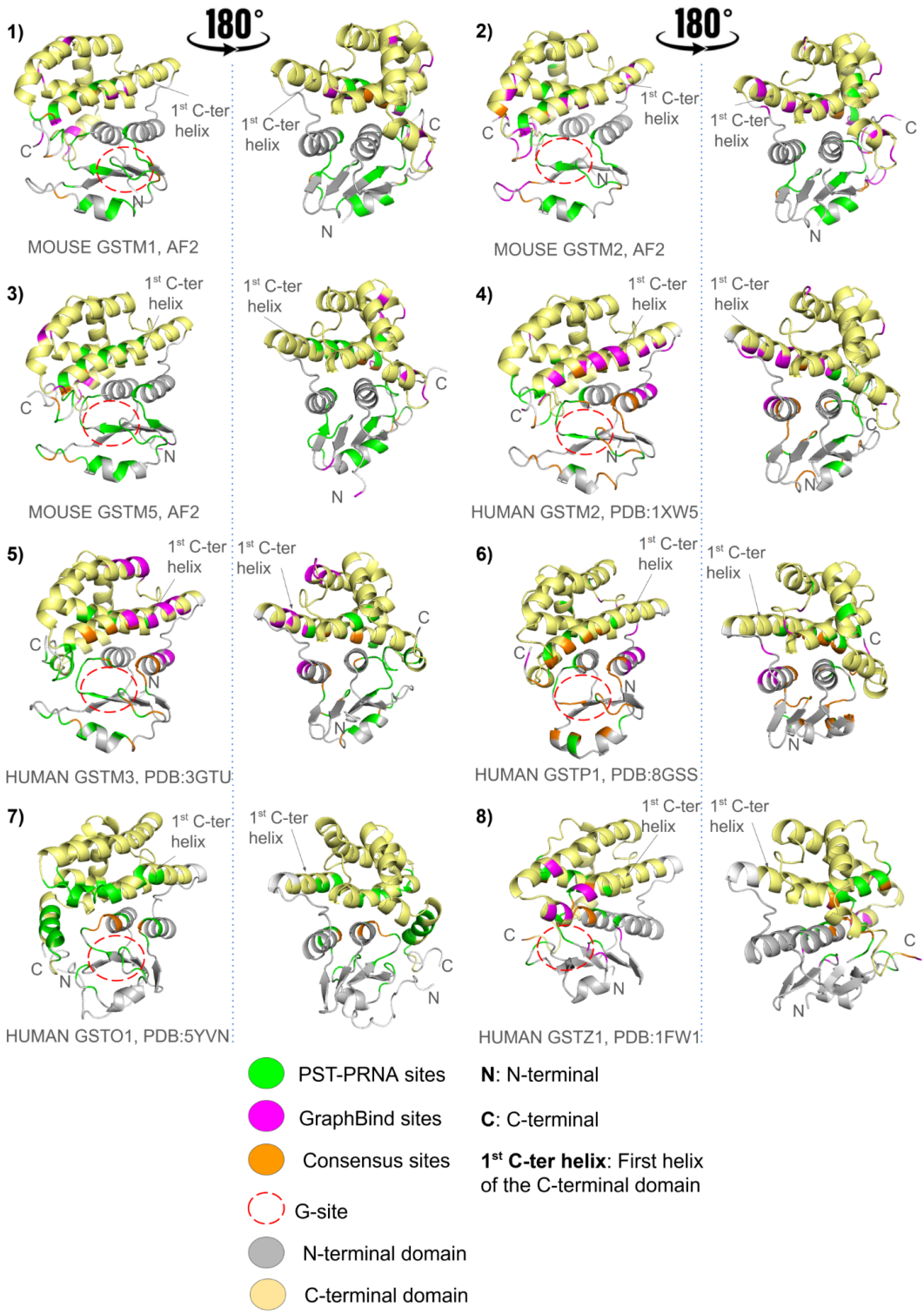
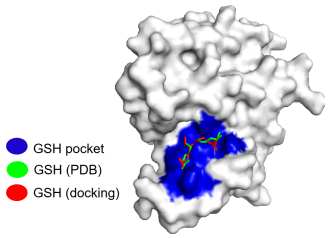
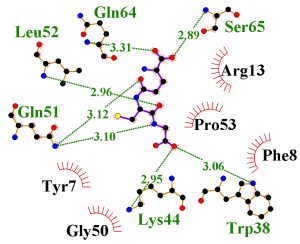
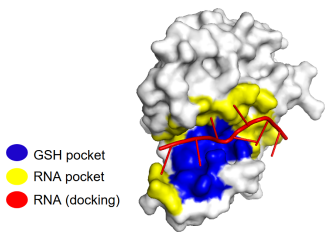
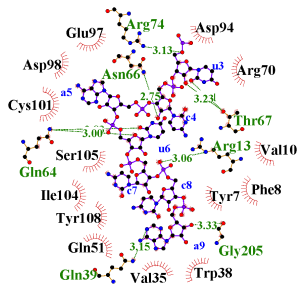
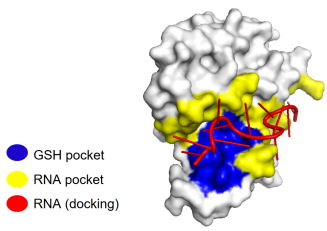
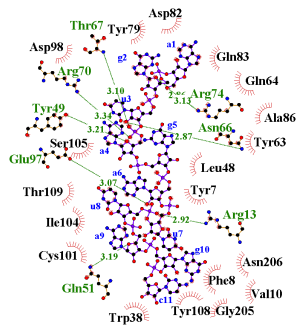
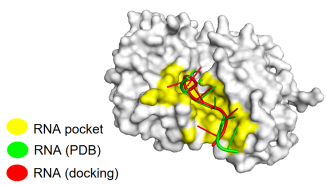
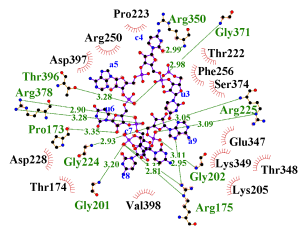


Fig. 2. (Color online) Predictions of Fig. 1 projected on the AlphaFold2 models of the three mouse GSTs and the three-dimensional (3D) structures of the five human GSTs. PST-PRNA, GraphBind and consensus-predicted RNA-binding residues are colored in green, magenta and orange, respectively. The backbone is color-coded in grey (N-terminus domain) and yellow (C-terminus domain). Protein structures are shown from two perspectives: (i) facing the side that displays the G-site (the glutathione, GSH binding pocket), highlighted with a dashed red circle (left), and after a 180° rotation with respect to the G-site face (right). RNA binding residues predicted by the two methods are mainly located on the side displaying the G-site, including the first helix of the C-terminal domain (first C-ter helix).

DOI: 10.1142/S2737416525500607
J. Comput. Biophys. Chem. **2026**, *25* (4), 627–640

Table 2. Validating GST–RNA interaction with Vina.

[1]Docking partners	[2]AutoDock Best pose and Mean Affinity (kcal/mol)	[3]Literature Affinity (kcal/mol)	[4]Docking results	[5]LigPlot+
*GST Pi (8GSS) +GSH (8GSS)	−5.1 (Mean −4.7±0.2)	−5.3[60] Δ −4.6[61] ω		
°GST Pi (8GSS)+miRNA (6UV4)	−8.5 (Mean −8.0±0.2)	Not available		
°GST Pi (8GSS)+miRNA (6DCL)	−8.6 (Mean −8.1±0.2)	Not available		
*DDX17 (6UV4)+miRNA (6UV4)	−9.6 (Mean −9.5 ± 0.1)	−8.8[54]		

[1]Docking partners: structures used for docking experiments. PDB identifiers of the protein are between the round brackets.

*Docking partners supported by structural evidence from the PDB.

°Docking partners with no structural support in the PDB.

[2]Affinity of the best pose (among 20 different ones) and the mean affinity (with standard deviation) across poses computed using AutoDock.

[3]Affinity computed from experimental K_d values reported in the literature.

[Δ]Affinity values computed on the dimer.[60]

[ω]Affinity values computed on the monomer.[61]

[4]Docking results of the best poses. The interacting residues were derived from the corresponding LigPlot+ 2D diagram and defined the blue and yellow pockets of GSH and RNA, respectively. Ligands of the original PDB, when available, are colored green and docked ligands are colored red. The interacting residues of the docked structure with structural evidence show a high overlap with the PDB counterpart.

[5]LigPlot+ 2D diagram of the best pose. A total of six interacting residues over 11 were shared between GST Pi+GSH and GST+both miRNAs: Tyr7, Phe8, Arg13, Trp38, Gln51 and Gln64.

DOI: 10.1142/S2737416525500607
J. Comput. Biophys. Chem. **2026**, *25* (4), 627–640

results of this analysis. PST-PRNA[26] (http://www.zpliulab.cn/PSTPRNA) and GraphBind[27] (http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/) were the two methods adopted (see Methods for details). As discussed in Ref. 26, PST-PRNA performs better than GraphBind on an independent test dataset of 61 RBPs by means of several scoring measures, including the Matthews correlation coefficient (MCC, 0.634 and 0.591, respectively). In Fig. 1, it appears that glutathione-binding residues (in red) correspond to the alignment and are consistently predicted as RNA-binding sites by the two methods in the N-terminal domain of the proteins (which contains the GSH binding site, G-site). Predictions were also present in the C-terminal domain, with a somewhat higher propensity in the case of GraphBind.

Three-dimensional projections of RNA-binding site predictions in the two protein domains are shown in Fig. 2. The results show that RNA-binding site predictions were mainly found close to the G-site (the region of the N-terminal domain responsible for binding with glutathione). Glutathione-binding residues annotated in UniProt, as derived from PDB structures, were consistently predicted as RNA-binding sites. Moreover, sites predicted in the C-terminal domain were mainly located in the first helix (of the C-terminal domain, first C-ter helix in Fig. 2), which is close to the G-site of the enzymes in the structure.

As no clear identification of RNA sequences interacting with RBPs has been reported in RNA proteomic studies, we selected an RNA-binding protein to calibrate the docking procedure. We tuned the performance of our docking method to the structure of the human miRNA–interacting protein DEAD-box helicase (DDX17, PDB ID: 6UV4)[54] cocrystallized with seven nucleotide-containing miRNAs. DDX17 is involved in other functions regarding RNA metabolism, such as pre-mRNA alternative splicing, ribosome biogenesis, mRNA export and coregulation of transcription and miRNA processing, with the latter being the main focus of the structure-related paper.[54] The DEAD-box domain is mainly responsible for the interaction with RNA whose contacts are dominated by hydrogen bonds with phosphate and sugar moieties.[47] We also considered that all the selected GSTs were found as candidate RBP in at least one polyA-capture experiment (RBP2GO,[7] Table 1). This prompted us to focus our docking experiments on miRNAs as short RNA molecules to simulate protein–RNA interactions. Docking was performed on a monomer of GSTP1 as extracted from the dimeric structure contained in the 8GSS PDB file. GSTP1 is the highest-scoring human GST in the RBP2GO database (UniProt ID: P9211; PDB ID: 8GSS). The docking procedure considers both GSH, a miRNA molecule with seven nucleotides (pri-miR-18a-oligo1, PDB ID: 6UV4),[54] and a second miRNA molecule (pri-miRNA-18a terminal loop) with 11 nucleotides (PDB ID: 6DCL)[55]. The AutoDock Vina experiments are listed in Table 2, which includes a LigPlot+ 2D diagram for each docking experiment. The bidimensional map indicates that six of the 11 binding sites of GSH are also present in the RNA-binding pockets. As for the control, it appears that Vina docks both GSH and the miRNA molecule in a position similar to that reported in the original PDB structure (red versus green structures in Table 2). Root mean square deviations to the original GSH and miRNA protein interacting molecules are 0.5 and 1 Å, respectively (computed with PyMol). The computed binding affinity was similar to that reported in the literature (Table 2). For GSH interaction, we report two values: the slightly highest value was found for the GST dimer and the lowest value for the monomer[61] (our simulation in any case deals with half of the GSTP1 dimer molecule contained in the PDB file). Focusing on GSTP1–miRNA interactions, we found that miRNA molecules are stabilized by the interaction of the backbone phosphate groups with the positively

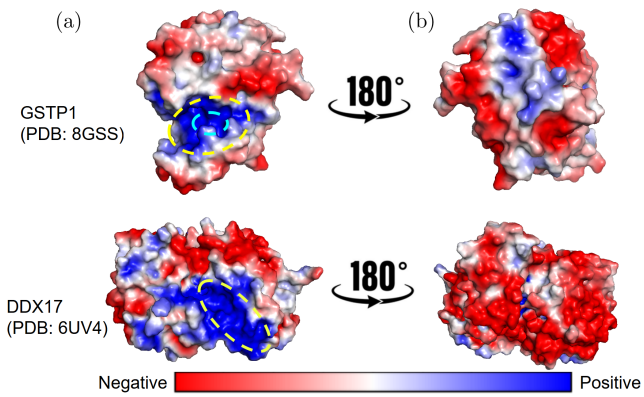


Fig. 3. (Color online) Structure surface charges comparison between GSTP1 (PDB ID 8GSS chain A) and DDX17 (PDB ID 6UV4 chain A). Electrostatic potentials, computed with PDB2PQR and APBS,[48] are shown on a [–3, 3] red–white–blue colormap in units of kJ/mol. (a) On the left side, cyan and yellow circles highlight the G-site pocket and RNA-binding surface, respectively, of GSTP1 and DDX17. Both the GSTP1 G-site pocket and DDX17 RNA-binding region show a positively charged surface, suggesting that the two regions have similar electrostatic characteristics. (b) The right side shows the two structures rotated by 180°. Interestingly, positive charges appear only on the GSTP1 accessible surface, consistent with the residues predicted as RNA binding sites by the predictors (Figs. 1 and 2).

DOI: 10.1142/S2737416525500607

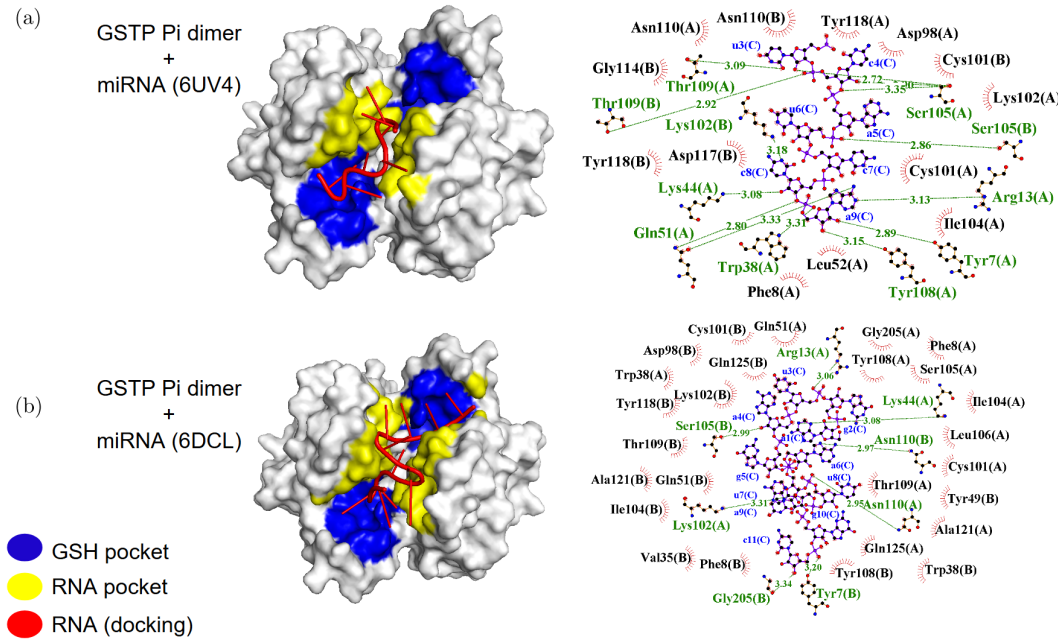


Fig. 4. (Color online) Docking obtained with AutoDock Vina on the GSTP1 dimer (PDB: 8GSS). LigPlot+ 2D diagrams are reported on the right side of the respective docking results. The interacting residues are derived from the corresponding LigPlot+ 2D diagram and define the blue and yellow pockets of the GSH and RNA, respectively; miRNA molecules are colored in red. (a) Best pose docking results with the seven-nucleotide-containing miRNA. Best pose: Affinity –9.3 kcal/mol. Mean affinity among poses: –8.8±0.2 kcal/mol. Overlapping residues with glutathione binding sites: Tyr7, Phe8, Arg13, Trp38, Lys44, Gln51 and Leu52 (Chain A). (b) Best pose docking results with the 11 nucleotide-containing miRNA. Best pose: Affinity –9.6 kcal/mol. Mean affinity among poses: –9.2±0.2 kcal/mol. Overlapping residues with glutathione binding sites: Phe8, Arg13, Trp38, Lys44 and Gln51 (Chain A) and Tyr7, Phe8, Trp38 and Gln51 (Chain B).

charged region of the protein, including the GSH pocket. GSTP1–miRNA docking LigPlot+ 2D diagrams revealed that hydrogen bonds involved in the interaction were formed with the RNA molecule backbone, either phosphate groups or ribose sugar. A total of six interacting residues over 11 were shared between GST Pi+GSH and GST+both miRNAs: Tyr7, Phe8, Arg13, Trp38, Gln51 and Gln64. Computed binding affinities were similar for both miRNA molecules and similar to the affinity of DDX17 with its miRNA.[54] We therefore propose GST–miRNA interaction as a reliable molecular structure of a putative GST–RNA complex.

4. CONCLUSIONS

In this study, we conducted a computational investigation of the possible GST–RNA interactions. An increasing number of proteins have been discovered to interact with RNA,[8] and a recent investigation has indicated that GSTs in mice are RNA interactors. Other studies reporting large-scale experiments in humans found evidence of RNA binding in GSTs as stored in RBP2GO, a recent database collecting experiments of RNA–protein interactions in 13 organisms, including *Homo sapiens*.[7] However, structure and sequence reference databases, including UniProt, although collecting many functional roles for GSTs, still do not annotate this important feature. Here, we investigated whether this role can be recognized by computational tools and validated by classical docking computations. The predicted RNA-binding residues consistently overlapped with the glutathione binding pocket (G-site) in all proteins tested. Docking validation suggested that RNA–protein interactions are due to RNA backbone interactions with the G-site. The RNA-binding solvent-accessible surface is, on average, more positively charged than that of the protein. Interestingly, although our docking focused on the G-site pocket contained in the N-terminal protein domain, other predicted RNA-binding sites were present in the C-terminal protein domain (Figs. 1 and 2). This is consistent with the possible role of the positively charged accessible surface with a further stabilizing RNA–protein interaction area (Fig. 3). We tested two different miRNA molecules of different lengths (7 and 11 nucleotides) and found that in both cases, the binding affinities were similar. Thus, we can

DOI: 10.1142/S2737416525500607
J. Comput. Biophys. Chem. **2026**, *25* (4), 627–640

support the notion that RNA–protein interactions are mainly due to electrostatic stabilizing interactions that can occur, when necessary, in the cytoplasmic environment. It has been noticed that in the cytoplasm, most of the GSTs are dimers.[61] For this reason, we docked the two miRNAs on the protein dimer (Figs. 4(a) and 4(b)), finding binding affinities similar to those reported in Table 2, where a monomer derived from the dimer was adopted. Data taken from the literature[1,62,63] estimate that a typical mammalian cell may contain nanomolar to micromolar concentrations of total RNA. If this is the case, it is reasonable to assume that, under stress conditions, the increase in the concentration of glutathione (up to millimolar concentration)[17] and the increased expression of GSTs[18] favor the protein role as glutathione-S transferases.[61]

STATEMENT OF USAGE OF ARTIFICIAL INTELLIGENCE

None.

DATA AVAILABILITY

All data regarding this research are included in the paper.

AUTHOR CONTRIBUTIONS

Gabriele Vazzana: Data Curation, Methodology, Software, Investigation, Writing- Original draft preparation. **Pier Luigi Martelli**: Conceptualization, Validation, Writing – Review & Editing. **Rita Casadio**: Conceptualization, Supervision, Investigation, Validation, Writing – Review & Editing.

CONFLICT OF INTEREST

The authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this paper.

FUNDING INFORMATION

This work was supported by ELIXIR-IT, the Italian node of the research infrastructure for life science data, and by the European Union Next Generation EU through the Italian Ministry of University and Research under the projects "Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics" (ELIXIRxNext-GenIT)" (Investment PNRRM4C2-I3.1, Project IR_0000010, CUP B53C22001800006) and "National Centre for HPC, Big Data and Quantum Computing" (Investment PNRR-M4C2-I1.4, Project CN_00000013).

ORCID

Rita Casadio
https://orcid.org/0000-0002-7462-7039
Gabriele Vazzana
https://orcid.org/0009-0007-5669-0249
Pier Luigi Martelli
https://orcid.org/0000-0002-0274-5669

References

1. Zacco, E.; Broglia, L.; Kurihara, M.; Monti, M.; Gustincich, S.; Pastore, A.; Plath, K.; Nagakawa, S.; Cerase, A.; Sanchez De Groot, N.; Tartaglia, G. G. RNA: The Unsuspected Conductor in the Orchestra of Macromolecular Crowding. *Chem. Rev.* **2024**, *124* (8), 4734–4777, https://doi.org/10.1021/acs.chemrev.3c00575
2. Berry, S.; Pelkmans, L. Mechanisms of Cellular mRNA Transcript Homeostasis. *Trends Cell Biol.* **2022**, *32* (8), 655–668, https://doi.org/10.1016/j.tcb.2022.05.003
3. Ule, J.; Blencowe, B. J. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol. Cell* **2019**, *76* (2), 329–345, https://doi.org/10.1016/j.molcel.2019.09.017
4. Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G. RNA-binding Proteins and Post-transcriptional Gene Regulation. *FEBS Lett.* **2008**, *582* (14), 1977–1986, https://doi.org/10.1016/j.febslet.2008.03.004
5. Schwarzl, T.; Sahadevan, S.; Lang, B.; Miladi, M.; Backofen, R.; Huber, W.; Hentze, M. W.; Tartaglia, G. G. Improved Discovery of RNA-Binding Protein Binding Sites in eCLIP Data Using DEWSeq. *Nucleic Acids Res.* **2024**, *52* (1), e1, https://doi.org/10.1093/nar/gkad998
6. Castello, A.; Horos, R.; Strein, C.; Fischer, B.; Eichelbaum, K.; Steinmetz, L. M.; Krijgsveld, J.; Hentze, M. W. Comprehensive Identification of RNA-Binding Proteins by RNA Interactome Capture. In *Post-Transcriptional Gene Regulation*; Dassi, E., Ed.; Methods in Molecular Biology; Springer, New York, NY, 2016; Vol. 1358, pp. 131–139, https://doi.org/10.1007/978-1-4939-3067-8_8
7. Caudron-Herger, M.; Jansen, R. E.; Wassmer, E.; Diederichs, S. RBP2GO: A Comprehensive Pan-Species Database on RNA-Binding Proteins, Their Interactions

and Functions. *Nucleic Acids Res.* **2021**, *49* (D1), D425–D436, https://doi.org/10.1093/nar/gkaa1040

8. Curtis, N. J.; Jeffery, C. J. The Expanding World of Metabolic Enzymes Moonlighting as RNA Binding Proteins. *Biochem. Soc. Trans.* **2021**, *49* (3), 1099–1108, https://doi.org/10.1042/BST20200664
9. Hentze, M. W.; Castello, A.; Schwarzl, T.; Preiss, T. A Brave New World of RNA-Binding Proteins. *Nat. Rev. Mol. Cell Biol.* **2018**, *19* (5), 327–341, https://doi.org/10.1038/nrm.2017.130
10. Li, Y.; Wang, Y.; Tan, Y.-Q.; Yue, Q.; Guo, Y.; Yan, R.; Meng, L.; Zhai, H.; Tong, L.; Yuan, Z.; Li, W.; Wang, C.; Han, S.; Ren, S.; Yan, Y.; Wang, W.; Gao, L.; Tan, C.; Hu, T.; Zhang, H.; Liu, L.; Yang, P.; Jiang, W.; Ye, Y.; Tan, H.; Wang, Y.; Lu, C.; Li, X.; Xie, J.; Yuan, G.; Cui, Y.; Shen, B.; Wang, C.; Guan, Y.; Li, W.; Shi, Q.; Lin, G.; Ni, T.; Sun, Z.; Ye, L.; Vourekas, A.; Guo, X.; Lin, M.; Zheng, K. Landscape of RNA Binding Proteins in Mammalian Spermatogenesis. *Science* **2024**, *386* (6720), eadj8172, https://doi.org/10.1126/science.adj8172
11. The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Garmiri, P.; Da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; De Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523–D531, https://doi.org/10.1093/nar/gkac1052
12. Mazari, A. M. A.; Zhang, L.; Ye, Z.-W.; Zhang, J.; Tew, K. D.; Townsend, D. M. The Multifaceted Role of Glutathione S-Transferases in Health and Disease. *Biomolecules* **2023**, *13* (4), 688, https://doi.org/10.3390/biom13040688
13. Aloke, C.; Onisuru, O. O.; Achilonu, I. Glutathione S-Transferase: A Versatile and Dynamic Enzyme. *Biochem. Biophys. Res. Commun.* **2024**, *734*, 150774, https://doi.org/10.1016/j.bbrc.2024.150774
14. Allocati, N.; Masulli, M.; Di Ilio, C.; Federici, L. Glutathione Transferases: Substrates, Inhibitors and Pro-Drugs in Cancer and Neurodegenerative Diseases. *Oncogenesis* **2018**, *7* (1), 8, https://doi.org/10.1038/s41389-017-0025-3
15. Oakley, A. Glutathione Transferases: A Structural Perspective. *Drug Metab. Rev.* **2011**, *43* (2), 138–151, https://doi.org/10.3109/03602532.2011.558093
16. Vazzana, G.; Savojardo, C.; Martelli, P. L.; Casadio, R. Testing the Capability of Embedding-Based Alignments on the GST Superfamily Classification: The Role of Protein Length. *Molecules* **2024**, *29* (19), 4616, https://doi.org/10.3390/molecules29194616
17. Lu, S. C. Glutathione Synthesis. *Biochim. Biophys. Acta, Gen. Subj.* **2013**, *1830* (5), 3143–3153, https://doi.org/10.1016/j.bbagen.2012.09.008
18. Lv, N.; Huang, C.; Huang, H.; Dong, Z.; Chen, X.; Lu, C.; Zhang, Y. Overexpression of Glutathione S-Transferases in Human Diseases: Drug Targets and Therapeutic Implications. *Antioxidants* **2023**, *12* (11), 1970, https://doi.org/10.3390/antiox12111970
19. Mukanganyama, S.; Bezabih, M.; Robert, M.; Ngadjui, B. T.; Kapche, G. F. W.; Ngandeu, F.; Abegaz, B. The Evaluation of Novel Natural Products as Inhibitors of Human Glutathione Transferase P1-1. *J. Enzyme Inhib. Med. Chem.* **2011**, *26* (4), 460–467, https://doi.org/10.3109/14756366.2010.526769
20. Townsend, D. M.; Tew, K. D. The Role of Glutathione-S-Transferase in Anti-Cancer Drug Resistance. *Oncogene* **2003**, *22* (47), 7369–7375, https://doi.org/10.1038/sj.onc.1206940
21. Oakley, A. J.; Bello, M. L.; Battistoni, A.; Ricci, G.; Rossjohn, J.; Villar, H. O.; Parker, M. W. The Structures of Human Glutathione Transferase P1-1 in Complex with Glutathione and Various Inhibitors at High Resolution. *J. Mol. Biol.* **1997**, *274* (1), 84–100, https://doi.org/10.1006/jmbi.1997.1364
22. Chen, I.; Chiu, M.; Cheng, S.; Hsu, Y.; Tsai, C. The Glutathione Transferase of *Nicotiana benthamiana* NbGSTU4 Plays a Role in Regulating the Early Replication of *Bamboo Mosaic Virus*. *New Phytol.* **2013**, *199* (3), 749–757, https://doi.org/10.1111/nph.12304

23. Jia, P.; Zhang, F.; Wu, C.; Li, M. A Comprehensive Review of Protein-Centric Predictors for Biomolecular Interactions: From Proteins to Nucleic Acids and Beyond. *Brief. Bioinform.* **2024**, *25* (3), bbae162, https://doi.org/10.1093/bib/bbae162
24. Zuo, Y.; Chen, H.; Yang, L.; Chen, R.; Zhang, X.; Deng, Z. Research Progress on Prediction of RNA-Protein Binding Sites in the Past Five Years. *Anal. Biochem.* **2024**, *691*, 115535, https://doi.org/10.1016/j.ab.2024.115535
25. Wei, J.; Chen, S.; Zong, L.; Gao, X.; Li, Y. Protein–RNA Interaction Prediction with Deep Learning: Structure Matters. *Brief. Bioinform.* **2022**, *23* (1), bbab540, https://doi.org/10.1093/bib/bbab540
26. Li, P.; Liu, Z.-P. PST-PRNA: Prediction of RNA-Binding Sites Using Protein Surface Topography and Deep Learning. *Bioinformatics* **2022**, *38* (8), 2162–2168, https://doi.org/10.1093/bioinformatics/btac078
27. Xia, Y.; Xia, C.-Q.; Pan, X.; Shen, H.-B. GraphBind: Protein Structural Context Embedded Rules Learned by Hierarchical Graph Neural Networks for Recognizing Nucleic-Acid-Binding Residues. *Nucleic Acids Res.* **2021**, *49* (9), e51–e51, https://doi.org/10.1093/nar/gkab044
28. Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Des.* **2011**, *7* (2), 146–157, https://doi.org/10.2174/157340911795677602
29. Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461, https://doi.org/10.1002/jcc.21334
30. Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61* (8), 3891–3898, https://doi.org/10.1021/acs.jcim.1c00203
31. Beckmann, B. M.; Horos, R.; Fischer, B.; Castello, A.; Eichelbaum, K.; Alleaume, A.-M.; Schwarzl, T.; Curk, T.; Foehr, S.; Huber, W.; Krijgsveld, J.; Hentze, M. W. The RNA-Binding Proteomes from Yeast to Man Harbour Conserved enigmRBPs. *Nat. Commun.* **2015**, *6* (1), 10127, https://doi.org/10.1038/ncomms10127
32. Brannan, K. W.; Jin, W.; Huelga, S. C.; Banks, C. A. S.; Gilmore, J. M.; Florens, L.; Washburn, M. P.; Van Nostrand, E. L.; Pratt, G. A.; Schwinn, M. K.; Daniels, D. L.; Yeo, G. W. SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes. *Molecular Cell* **2016**, *64* (2), 282–293, https://doi.org/10.1016/j.molcel.2016.09.003
33. Milek, M.; Imami, K.; Mukherjee, N.; Bortoli, F. D.; Zinnall, U.; Hazapis, O.; Trahan, C.; Oeffinger, M.; Heyd, F.; Ohler, U.; Selbach, M.; Landthaler, M. DDX54 Regulates Transcriptome Dynamics during DNA Damage Response. *Genome Res.* **2017**, *27* (8), 1344–1359, https://doi.org/10.1101/gr.218438.116
34. Mullari, M.; Lyon, D.; Jensen, L. J.; Nielsen, M. L. Specifying RNA-Binding Regions in Proteins by Peptide Cross-Linking and Affinity Purification. *J. Proteome Res.* **2017**, *16* (8), 2762–2772, https://doi.org/10.1021/acs.jproteome.7b00042
35. Bao, X.; Guo, X.; Yin, M.; Tariq, M.; Lai, Y.; Kanwal, S.; Zhou, J.; Li, N.; Lv, Y.; Pulido-Quetglas, C.; Wang, X.; Ji, L.; Khan, M. J.; Zhu, X.; Luo, Z.; Shao, C.; Lim, D.-H.; Liu, X.; Li, N.; Wang, W.; He, M.; Liu, Y.-L.; Ward, C.; Wang, T.; Zhang, G.; Wang, D.; Yang, J.; Chen, Y.; Zhang, C.; Jauch, R.; Yang, Y.-G.; Wang, Y.; Qin, B.; Anko, M.-L.; Hutchins, A. P.; Sun, H.; Wang, H.; Fu, X.-D.; Zhang, B.; Esteban, M. A. Capturing the Interactome of Newly Transcribed RNA. *Nat. Methods* **2018**, *15* (3), 213–220, https://doi.org/10.1038/nmeth.4595
36. Huang, R.; Han, M.; Meng, L.; Chen, X. Transcriptome-Wide Discovery of Coding and Noncoding RNA-Binding Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (17), E3879–E3887, https://doi.org/10.1073/pnas.1718406115
37. Queiroz, R. M. L.; Smith, T.; Villanueva, E.; Marti-Solano, M.; Monti, M.; Pizzinga, M.; Mirea, D.-M.; Ramakrishna, M.; Harvey, R. F.; Dezi, V.; Thomas, G. H.; Willis, A. E.; Lilley, K. S. Comprehensive Identification of RNA–Protein Interactions in Any Organism Using Orthogonal Organic Phase Separation (OOPS). *Nat. Biotechnol.* **2019**, *37* (2), 169–178, https://doi.org/10.1038/s41587-018-0001-2
38. Backlund, M.; Stein, F.; Rettel, M.; Schwarzl, T.; Perez-Perri, J. I.; Brosig, A.; Zhou, Y.; Neu-Yilik, G.; Hentze, M. W.; Kulozik, A. E. Plasticity of Nuclear and Cytoplasmic Stress Responses of RNA-Binding Proteins. *Nucleic Acids Res.* **2020**, *48* (9), 4725–4740, https://doi.org/10.1093/nar/gkaa256
39. Urdaneta, E. C.; Vieira-Vieira, C. H.; Hick, T.; Wessels, H.-H.; Figini, D.; Moschall, R.; Medenbach, J.; Ohler, U.; Granneman, S.; Selbach, M.; Beckmann, B. M. Purification of Cross-Linked RNA-Protein Complexes by Phenol-Toluol Extraction. *Nat. Commun.* **2019**, *10* (1), 990, https://doi.org/10.1038/s41467-019-08942-3
40. Altschul, S. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402, https://doi.org/10.1093/nar/25.17.3389
41. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded

and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637, https://doi.org/10.1002/bip.360221211

42. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* **2012**, *9* (2), 173–175, https://doi.org/10.1038/nmeth.1818
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing, Cham, 2016; Vol. 9908, pp. 630–645, https://doi.org/10.1007/978-3-319-46493-0_38
44. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242, https://doi.org/10.1093/nar/28.1.235
45. Coimbatore Narayanan, B.; Westbrook, J.; Ghosh, S.; Petrov, A. I.; Sweeney, B.; Zirbel, C. L.; Leontis, N. B.; Berman, H. M. The Nucleic Acid Database: New Features and Capabilities. *Nucleic Acids Res.* **2014**, *42* (D1), D114–D122, https://doi.org/10.1093/nar/gkt980
46. Yang, J.; Roy, A.; Zhang, Y. BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand–Protein Interactions. *Nucleic Acids Res.* **2012**, *41* (D1), D1096–D1103, https://doi.org/10.1093/nar/gks966
47. Corley, M.; Burns, M. C.; Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell* **2020**, *78* (1), 9–29, https://doi.org/10.1016/j.molcel.2020.03.011
48. Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; Chun, M.; Li, P.; Gohara, D. W.; Dolinsky, T.; Konecny, R.; Koes, D. R.; Nielsen, J. E.; Head-Gordon, T.; Geng, W.; Krasny, R.; Wei, G.; Holst, M. J.; McCammon, J. A.; Baker, N. A. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci.* **2018**, *27* (1), 112–128, https://doi.org/10.1002/pro.3280
49. Gilchrist, C. L. M.; Mirdita, M.; Steinegger, M. Multiple Protein Structure Alignment at Scale with FoldMason. BioRxiv, August 1, 2024, https://doi.org/10.1101/2024.08.01.606130
50. Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *51* (10), 2778–2786, https://doi.org/10.1021/ci200227u
51. Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: A Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng. Des. Sel.* **1995**, *8* (2), 127–134, https://doi.org/10.1093/protein/8.2.127
52. Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48* (12), 4111–4119, https://doi.org/10.1021/jm048957q
53. Nocedal, J.; Wright, S. J. *Numerical Optimization*; Springer Series in Operations Research; Springer Verlag, Berlin, 1999.
54. Ngo, T. D.; Partin, A. C.; Nam, Y. RNA Specificity and Autoregulation of DDX17, a Modulator of MicroRNA Biogenesis. *Cell Reports* **2019**, *29* (12), 4024-4035.e5, https://doi.org/10.1016/j.celrep.2019.11.059
55. Kooshapur, H.; Choudhury, N. R.; Simon, B.; Mühlbauer, M.; Jussupow, A.; Fernandez, N.; Jones, A. N.; Dallmann, A.; Gabel, F.; Camilloni, C.; Michlewski, G.; Caceres, J. F.; Sattler, M. Structural Basis for Terminal Loop Recognition and Stimulation of Pri-miRNA-18a Processing by hnRNP A1. *Nat. Commun.* **2018**, *9* (1), 2479, https://doi.org/10.1038/s41467-018-04871-9
56. Bertolini, E.; Babbi, G.; Savojardo, C.; Martelli, P. L.; Casadio, R. MultifacetedProtDB: A Database of Human Proteins with Multiple Functions. *Nucleic Acids Res.* **2024**, *52* (D1), D494–D501, https://doi.org/10.1093/nar/gkad783
57. Goto, S.; Kawakatsu, M.; Izumi, S.; Urata, Y.; Kageyama, K.; Ihara, Y.; Koji, T.; Kondo, T. Glutathione S-Transferase π Localizes in Mitochondria and Protects against Oxidative Stress. *Free Radical Biol. Med.* **2009**, *46* (10), 1392–1403, https://doi.org/10.1016/j.freeradbiomed.2009.02.025
58. Simic, P.; Pljesa, I.; Nejkovic, L.; Jerotic, D.; Coric, V.; Stulic, J.; Kokosar, N.; Popov, D.; Savic-Radojevic, A.; Pazin, V.; Pljesa-Ercegovac, M. Glutathione Transferase P1: Potential Therapeutic Target in Ovarian Cancer. *Medicina* **2022**, *58* (11), 1660, https://doi.org/10.3390/medicina58111660
59. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589, https://doi.org/10.1038/s41586-021-03819-2
60. Lo Bello, M.; Oakley, A. J.; Battistoni, A.; Mazzetti, A. P.; Nuccetelli, M.; Mazzarese, G.; Rossjohn, J.; Parker, M. W.; Ricci, G. Multifunctional Role of Tyr 108 in the Catalytic Mechanism of Human Glutathione Transferase P1-1. Crystallographic and Kinetic Studies on the Y108F Mutant Enzyme. *Biochemistry* **1997**, *36* (20), 6207–6217, https://doi.org/10.1021/bi962813z

61. Fabrini, R.; De Luca, A.; Stella, L.; Mei, G.; Orioni, B.; Ciccone, S.; Federici, G.; Lo Bello, M.; Ricci, G. Monomer–Dimer Equilibrium in Glutathione Transferases: A Critical Re-Examination. *Biochemistry* **2009**, *48* (43), 10473–10482, https://doi.org/10.1021/bi901238t
62. Van Treeck, B.; Protter, D. S. W.; Matheny, T.; Khong, A.; Link, C. D.; Parker, R. RNA Self-Assembly Contributes to Stress Granule Formation and Defining the Stress Granule Transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (11), 2734–2739, https://doi.org/10.1073/pnas.1800038115
63. Ho, V.; Baker, J. R.; Willison, K. R.; Barnes, P. J.; Donnelly, L. E.; Klug, D. R. Single Cell Quantification of microRNA from Small Numbers of Non-Invasively Sampled Primary Human Cells. *Commun. Biol.* **2023**, *6* (1), 458, https://doi.org/10.1038/s42003-023-04845-8

Computational and Structural Biotechnology Journal 27 (2025) 461–466

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

AlphaFold2 and ESMFold: A large-scale pairwise model comparison of human enzymes upon Pfam functional annotation

Matteo Manfredi [a,b], Gabriele Vazzana [a,b], Castrense Savojardo [a,b], Pier Luigi Martelli [a,b,*], Rita Casadio [a,c,*]

[a] Biocomputing Group, University of Bologna, Italy
[b] Dept. of Pharmacy and Biotechnology, University of Bologna, Italy
[c] the Alma Climate Institute, University of Bologna, Italy

ARTICLE INFO

Keywords:
Human enzyme structural and functional annotation
Pfam domains
Enzyme Active site
AlphaFold2
ESMFold
Human Reference Proteome

ABSTRACT

AlphaFold2 predicts protein structures from structural and functional knowledge. Alternatively, ESMFold does the same adopting protein language models. Here, we map available Pfam domains on pairs of models of the human reference proteome computed with both procedures and we compare the mapped regions relevant for functional annotation. We find that, rather irrespectively of the global superimposition of the pairwise models, Pfam-containing regions overlap with a TM-score above 0.8 and a predicted local distance difference test (pLDDT) which is higher than the rest of the modeled sequence. This indicates that both methods are similarly performing in modeled regions that overlap Pfam domains, carrying structural and functional information, with pLDDT values slightly higher for AlphaFold2. The mapping of 9834 Pfam domains also allows the location of 2578 active sites in 3382 enzymes of the human proteome, including 807 proteins for which the active site is not reported in UniProt.

1. Introduction

Automated sequencing techniques generate a huge and increasing gap between the number of protein sequences deposited in databases, and their available three-dimensional structures ([1,2] and references therein). Recently, AlphaFold2 from DeepMind has been proposed as a useful tool for filling this gap in UniProt files ([3], https://www.uniprot.org). AlphaFold2 is a deep machine learning method trained on protein multiple sequence alignments, protein contact maps, correlated mutations, and protein family templates to infer protein structures [4,5]. However, when basic structural and functional information is missing, convincing and complete models are still poorly predicted. As a recent alternative, ESMFold adopts a protein "embedded" representation, derived from protein language models generated after filtering hundreds of millions of protein sequences, to compute the protein structure ([6] and references therein). Basically, both methods rely on an enormous computational power to extract evolutionary information at the base of concepts such as protein families and superfamilies, which routinely allowed the development of very successful methods for protein structure prediction including that of building by comparison [7]. AlphaFold2-based methods have been proven superior to ESMFold in the international benchmark of CASP15 [8], where however a limited number of structures were tested. For the sake of comparing both approaches, we recently generated a database of models for the human reference proteome, in which each human protein is endowed with AlphaFold2 and ESMFold models [9]. We commented before on the relatively better performance in structure prediction of AlphaFold2 when protein family templates are known [9]. Here, we focus on human enzymes and their functional annotations. In UniProt, human enzymes derive their functional annotation from the Association-Rule-Based Annotator (ARBA) rule system (https://www.uniprot.org/help/arba), "a multiclass learning system trained on expertly annotated entries" that unifies both InterPro signatures [10] and Pfam models [11], relying on the notion that proteins in families and superfamilies conserve functional and structural domains [12].

A protein domain is a compact three-dimensional region of the folded polypeptide chain that is self-stabilizing and that can fold independently from the rest [12]. Many proteins consist of several domains, and a domain may be shared by different proteins so that they can be considered as building blocks that molecular evolution adopted to

* Corresponding authors at: Biocomputing Group, University of Bologna, Italy.
E-mail addresses: pierluigi.martelli@unibo.it (P.L. Martelli), rita.casadio@unibo.it (R. Casadio).

https://doi.org/10.1016/j.csbj.2025.01.008
Received 21 October 2024; Received in revised form 8 January 2025; Accepted 13 January 2025
Available online 14 January 2025

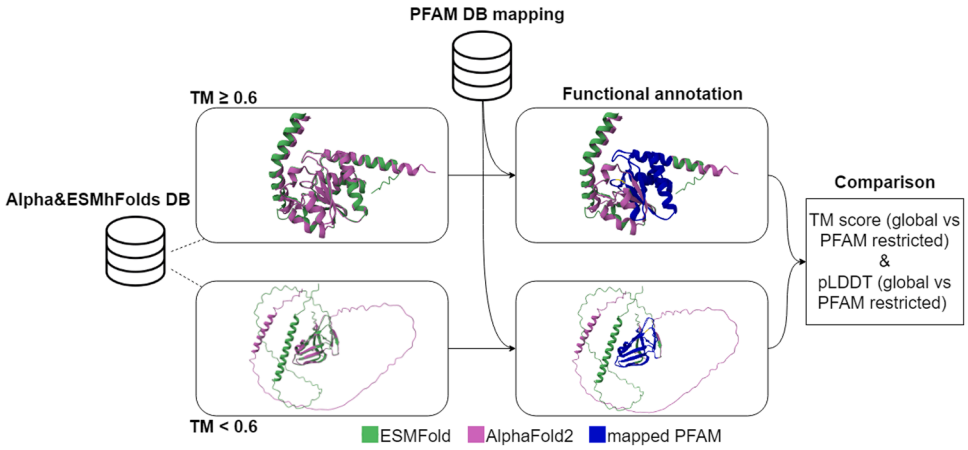


Fig. 1. Workflow of the annotation procedure adopted in this work. Alpha&ESMhFolds DB (https://alpha-esmhfolds.biocomp.unibo.it/); Pfam DB (https://pfam-docs.readthedocs.io). For TM-score and pLDDT definitions see the Material and Methods section.

generate proteins with different functions [12]. Domains can be detected with protein multiple structural and sequence alignments [13]. Traditionally, they have been modeled with hidden Markov models (HHMs) [14], and the Pfam database of protein-conserved domains is available ([11], https://pfam-docs.readthedocs.io). In the enzyme world, Pfam domains can include active sites, superfamily or family signatures, and/or structural characteristics ([10], https://www.ebi.ac.uk/interpro/). The procedure that we apply here is therefore based on mapping Pfam domains carrying along their annotation on our pairwise AlphaFold2 and ESMFold models of human enzymes out of the human reference proteome, focusing on those containing an active site and their comparison. We find that independently of the pairwise model superimposition, mapped Pfam regions are structurally well-predicted and superimposed. This indicates that both predictors in the Pfam regions are similarly effective in grasping functional and structural features.

2. Materials and methods

The dataset adopted for the analysis comprises 6956 human enzymes, all the proteins endowed with an EC number included in the Alpha&ESMhFolds database [9]. Alpha&ESMhFolds is a collection of 42,942 proteins extracted from the human Reference Proteome (UP000005640, available at UniProt [3], release 2023_03 of January 2023), which for each protein provides the respective AlphaFold2 and ESMFold models. AlphaFold2 models are downloaded from the AlphaFoldDB ([5,15] https://alphafold.com, accessed in January 2023), and their ESMFold counterpart is computed in-house [9]. A fraction of the total enzyme database (1314 proteins of the 6956) is endowed with a PDB [16] three-dimensional structure covering at least 70 % of the protein sequence. Among the remaining 5643 proteins, 3037 are included in Swiss-Prot, the manually curated part of UniProt. The remaining 2606 are listed in TrEMBL, the automatically annotated part of UniProt.

For each enzyme, we extract the set of annotations present in Pfam [11] downloading data available at the website (https://pfam-docs.readthedocs.io, version 37.0 accessed in September 2024). We collected 14,122 Pfam entries, including 9834 domains, 3391 families, 769 repeats, 93 short motifs, 19 conserved intrinsically disordered regions, and 16 coiled-coil regions, all documented also in InterPro (https://www.ebi.ac.uk/interpro/). Notably, the Pfams we extracted are the same as those annotated in the UniProt database. Additionally, we ran the PfamScan tool to annotate 2578 Pfam entries with a reported active site. In total, 5684 residues were determined to be part of an active site with an average of 2.2 residues per active site (1642 active sites with 1 residue, 1230 with 2, 461 with 3, 47 with 4, the active site of domain PF00561 in the protein Q9H4I8 which has 5, and the active site

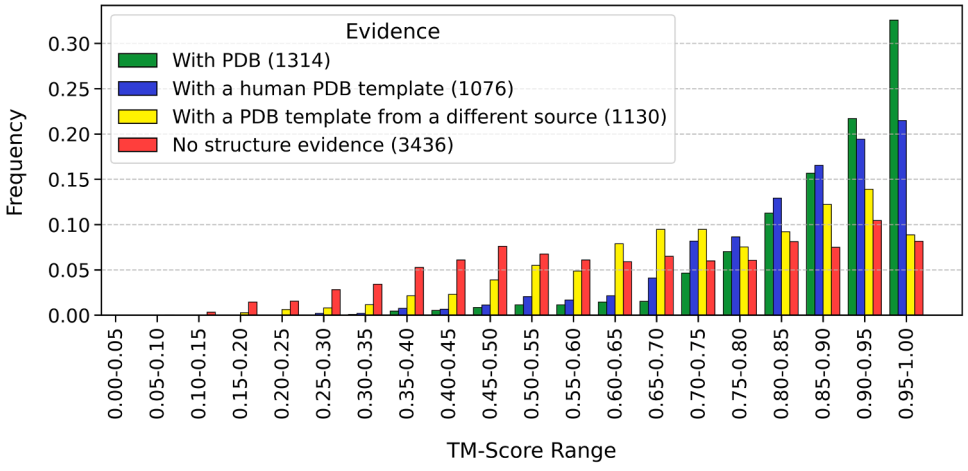


Fig. 2. Distribution of pairwise AlphaFold and ESMFold models of human enzymes in our database ([9], available at https://alpha-esmhfolds.biocomp.unibo.it/)) as a function of their superimposition as evaluated with the TM-score. Colors distinguish models with an underlying PDB structure (green), with a human PDB template (blue), with a PDB template from other organisms (yellow), and without structural evidence (red). It appears that 49.4 % of the 6959 pairwise models do not have PDB structural templates.

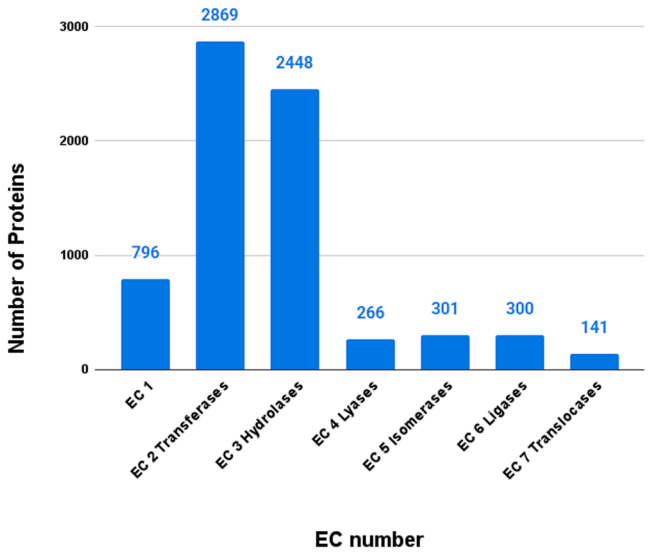


Fig. 3. Histogram showing the number of proteins associated with each EC number. 151 proteins out of the 6956 enzymes in our dataset are associated with more than one EC number. The multi-EC combinations in the dataset and the relative number of proteins is reported here: EC 2–3: 71; EC 2–4: 12; EC 1–2: 9; EC 3–4: 7; EC 4–5: 7; EC 1–3: 6; EC 1–4: 6; EC 1–5: 5; EC 2–6: 5; EC 3–7: 3; EC 3–5: 3; EC 2–5: 2; EC 4–6: 1; EC 1–3–6: 4; EC 1–2–4: 3; EC 1–2–5: 2; EC 2–3–4: 2; EC 1–4–5: 2; EC 1–2–3: 1.

of domain PF01536 in the protein P17707 which has 6).

For each protein, the Alpha&ESMhFolds database [9] provides the per-residue pLDDT of the AlphaFold2 and ESMFold models (a self-assessed measure by both methods of the reliability of the predicted model [17]), as well as the TM-score between the two models (a metric to estimate the similarity between the two 3D models [18]). Here, we focus on computing the local TM-score of the regions of the proteins that are covered by a Pfam entry, obtained by manually extracting the relevant regions from the predicted models. Evaluation is performed with the Foldseek program [13], which computes the structural superimposition of the models and their similarity. The Foldseek program was run using the options "–alignment-type 1" (alignment type set to the TM-align algorithm) and "–prefilter-mode 2" (disabling prefiltering of results). The remaining program parameters were left to default values.

For the sake of reproducibility, we provide a GitHub repository accessible at https://github.com/MatteoManfredi/pfam_models. It contains detailed instructions to run a script that computes the results reported in this manuscript starting from the list of UniProt identifiers of our enzyme dataset. The script can also be used to reproduce our analysis starting from a different list of UniProt identifiers, as long as those are included in our web server Alpha&ESMhFolds which provides the pairwise models.

3. Results and discussion

Proteins Enzyme Commission (EC) numbers are routinely assigned with the UniProt automatic annotations pipeline (https://www.uniprot.org/help/biocuration), which leverages the ARBA Rule system (https://www.uniprot.org/help/arba). In this paper, we are interested in understanding the ability of AlphaFold2 and ESMFold models to capture the functional features of human enzymes. In Fig. 2 we show the distribution of enzyme models as a function of the computed TM-score, color-coded depending on their structural evidence. It appears that 49.4 % of the pairwise models (6956) are without a PDB structural template.

Summing up, human enzyme model pairs can be grouped into three categories: i) models with an underlying PDB structure of the enzyme, ii) models without PDB in the background which superimpose (TM-score ≥ 0.6), or iii) which do not superimpose (TM-score <0.6). Then, after mapping the Pfam database on the models, we compute TM-scores at the Pfam region and we evaluate the local quality of the prediction by averaging the per-residue pLDDT score over the Pfam region (see Materials and Methods for definitions and Fig. 1).

Table 1
The human enzyme set as distributed on the 6 types of Pfam entries.

Pfam Type[a]	# Entries in Pfam database	# Unique Pfam in the HES[b]	# Enzymes with Pfam	# Pfam Occurrences	Range of Pfam lengths[c]
Domains	9147	1517	5204	9834	16–713
Domains with annotated active site	773	249	2459	2577	34–655
Families	11,536	683	2738	3391	11–1444
Repeats	859	78	324	769	14–517
Short Motifs	122	21	77	93	15–61
Intrinsically disordered regions	122	9	19	19	60–165
Coiled-coil regions	193	8	14	16	35–331

= Number of

[a] Pfam classifies its entries into 6 types (see Materials and Methods and https://pfam-docs.readthedocs.io). We additionally distinguish among the "Domains" those containing an active site, as annotated by the PfamScan tool.

[b] HES = Human Enzyme Set comprising 6956 enzymes.

[c] We report the minimum and maximum length of the Pfam types included in our dataset, (number of residues). See Figure S1 for distributions.

Fig. 3 shows the distribution of the human protein enzymes as a function of the seven first levels of EC numbers. The most populated classes are Transferases and Hydrolases. Traditionally, ARBA rules include, among other features, Pfam domains. In the automatic annotation process at UniProt, the biocuration system (https://www.uniprot.org/help/biocuration) filters protein sequences and assigns the functional annotation, inclusive of the EC number, provided that the sequence meets a given set of features. In Table 1 we distribute our enzyme database according to the different Pfam types, as described on the Pfam website (see Materials and Methods). The Domain type is particularly interesting for our analysis since it contains information on the functional annotation and, when known, also on the active sites, which are conserved in the families. In the human enzyme set, 1517 unique Pfam domains are shared by 5204 enzyme proteins for a total of 9834 occurrences. The domain lengths range from 16 up to 713 residues (length distribution is shown in Figure S1). Included in domains are those containing an active site (249) which map into 2459 human enzymes, with lengths ranging from 34 to 655 (See Figure S1 for distribution), and 99 % of these carry structural information. Along with domains, Pfam models are available for family signatures, repeats, short motifs, intrinsically disordered regions, and coiled-coil regions (https://pfam-docs.readthedocs.io/en/latest/summary.html). These Pfam models, in principle, can shed light on functionality, although in a more general way. In Table 1 we list Pfam types, sorted by decreasing number of human enzyme proteins of our data set.

According to the workflow of Fig. 1, after mapping the Pfam database on the pairwise models, we compare the TM-scores of the global models with those evaluated for the Pfam regions. Similarly, we compare the local quality of the prediction of the global models with that restricted to the Pfam regions (by computing the pLDDT score). Results are shown in Fig. 4 and Table S2. TM-scores of the pairwise models are higher when the models have an underlying PDB structure (1047 enzymes); when structural information is lacking, 2766 enzymes have global models with a pairwise TM-score ≥ 0.6, and 1391 enzymes have global models with TM-score < 0.6. Considering the Pfam domains, TM-scores remain high when evaluated on the region of the mapped domain, irrespective of the superimposition of global enzyme

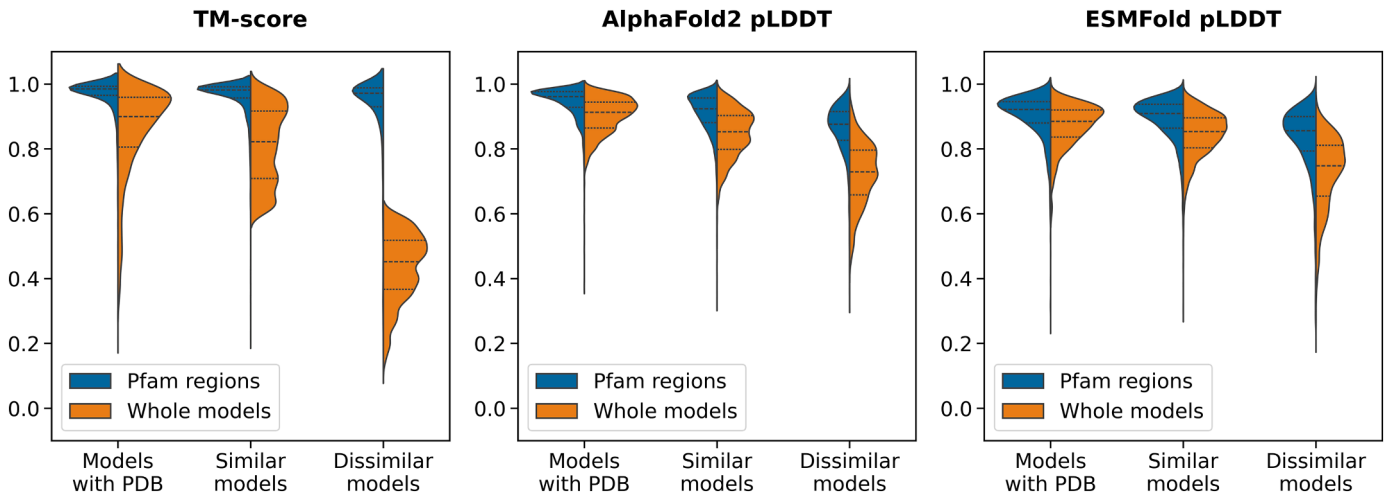


Fig. 4. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score $\geq$ 0.6), and those with dissimilar models (TM-score $<$ 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

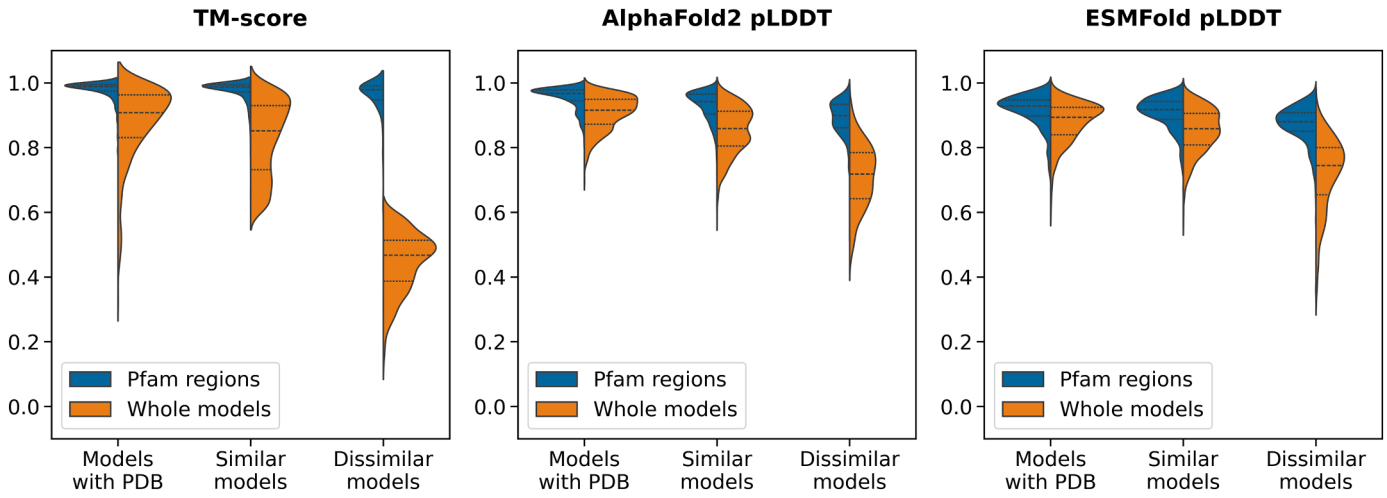


Fig. 5. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam domains with an active site. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score $\geq$ 0.6), and those with dissimilar models (TM-score $<$ 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

pairwise models (Fig. 4). The same holds for the predicted local evaluation of the quality of the models (pLDDT), which is slightly higher in AlphaFold2 than in ESMFold.

When we restrict the analysis to the enzymes with an active site, derived by mapping Pfam domains containing an active site, again we can observe (Fig. 5 and Table S3) that the superimposition of the Pfam regions (TM-score) in the pairwise models increases rather irrespectively of the superimposition and the quality of the global enzyme models. This is particularly evident when the global models diverge (TM-score $<$ 0.6, Fig. 6). Interestingly, our mapping performed with PfamScan allows the annotation of 807 proteins with 858 active sites from 117 Pfam domains. These annotations are lacking in the associated files at UniProt.

According to the Pfam definition, a family includes proteins that share a common evolutionary origin "as reflected in their related functions, sequences or structure" (https://pfam-docs.readthedocs.io/en/latest/summary.html). We therefore mapped Pfam family models into our dataset of enzymes and compared the pairwise models. Noticeably, as reported in Table 1, 509 enzymes map more than one Pfam family (for details see Table S1). We found that the mapped Pfam region overlap (higher TM-score than that of the global models) and the local quality of the self-evaluation of the predictions is higher than average (Fig. 7 and Table S4). Finally, we consider other Pfam models including repeats, short motifs (including metal binding sites), intrinsically disordered regions, and coiled-coiled regions. Even in these cases (with the exception of disordered regions), the shared Pfam regions among the pairwise models have better quality than the overall global models (see Tables S5, S6, S7, and S8). The number of enzymes that map these Pfam types is progressively decreasing (Table 1), indicating that the annotation carried along pertains to a minor fraction of our human enzyme dataset.

4. Conclusions

In this paper, we functionally annotate our dataset of pairwise models of human enzymes derived from the human reference proteome.

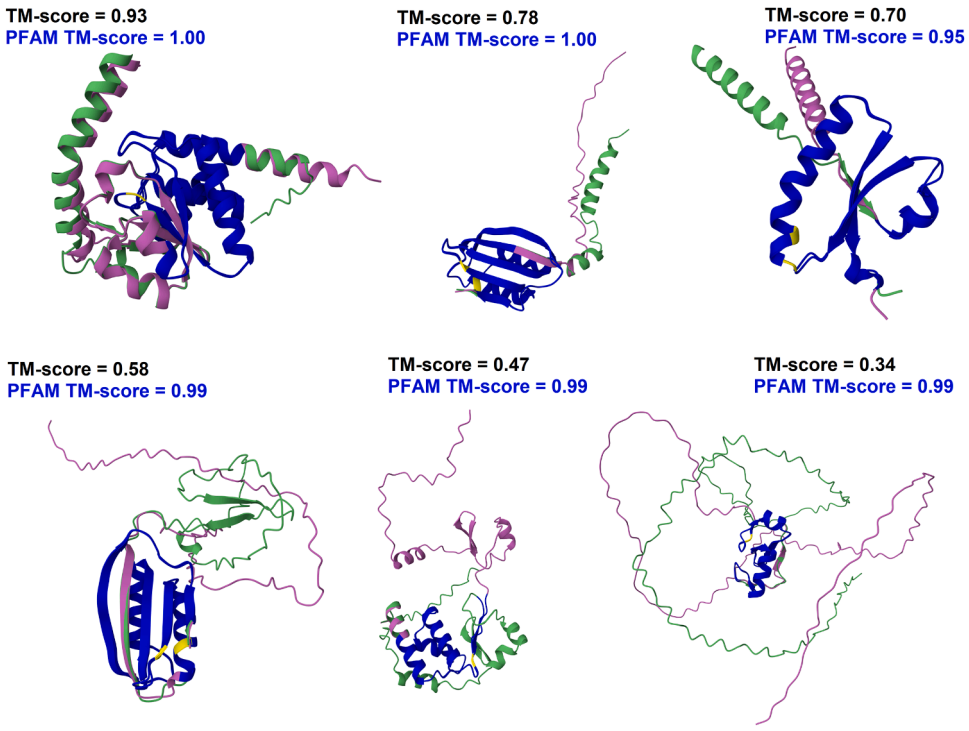


Fig. 6. Examples of enzymes at different levels of model TM-scores showing that the superimposition remains good in the region covered by a Pfam domain. In all images, the green model is obtained with ESMFold and the purple one with AlphaFold2. We highlight the regions covered by a Pfam domain in blue and the active sites in yellow. In the image, we report the TM-score of the AlphaFold2 and ESMFold models (in black) and the TM-score computed on the region covered by the domain (in blue). From left to right, top to bottom, we show: "Phosphatidylglycerophosphatase and protein-tyrosine phosphatase 1" (UniProt accession: Q8WUK0, EC 3.1.3.27), with a "Dual specificity phosphatase, catalytic" domain (Pfam accession: PF00782) covering residues 88–184 (active site in position 132); "Acylphosphatase" (UniProt accession: G3V2U7, EC 3.6.1.7), with a "Acylphosphatase" domain (Pfam accession: PF00708) covering residues 41–127 (active site in positions 54 and 72); "Protein disulfide-isomerase" (UniProt accession: H3BVI1, EC 5.3.4.1), with a "Thioredoxin" domain (Pfam accession: PF00085) covering residues 32–87 (active site in positions 53 and 56); "Acylphosphatase" (UniProt accession: U3KQL2, EC 3.6.1.7), with a "Acylphosphatase" domain (Pfam accession: PF00708) covering residues 94–170 (active site in positions 97 and 115); "Dual specificity protein phosphatase" (UniProt accession: E9PSD4, EC 3.1.3.16), with a "Dual specificity phosphatase, catalytic" domain (Pfam accession: PF00782) covering residues 81–135 (active site in position 87); "E3 ubiquitin-protein ligase RNF4" (UniProt accession: P78317, EC 2.3.2.27), with a "Ring finger" domain (Pfam accession: PF13639) covering residues 131–177 (active site in position 132).

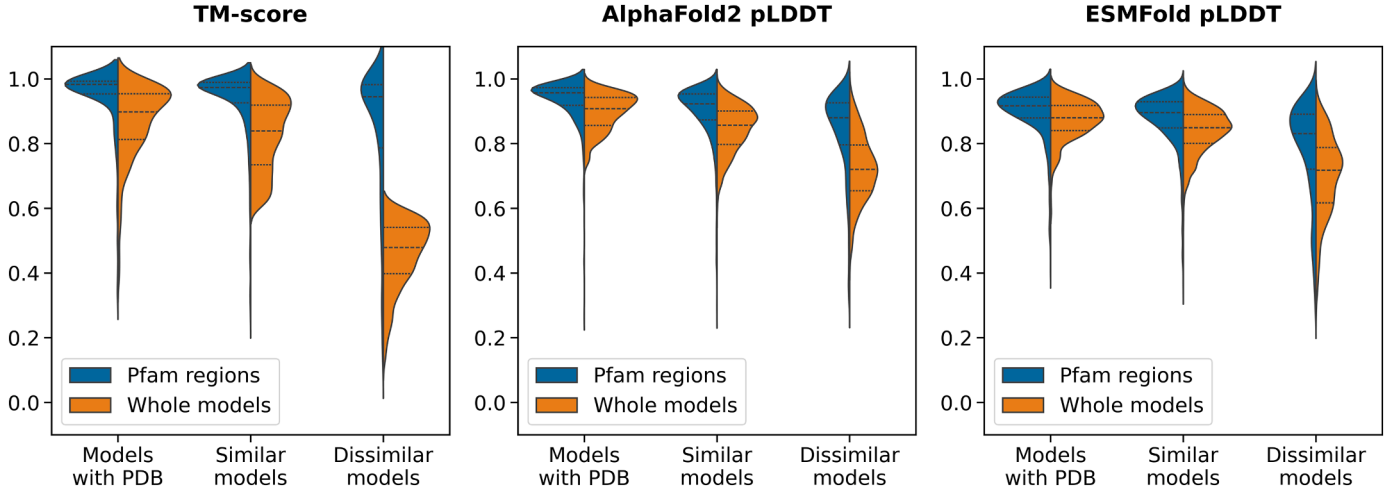


Fig. 7. Violin plots showing the comparison of AlphaFold2 and ESMFold pairwise models of human enzymes including Pfam families. From left to right, on the y-axis we report the TM-score obtained when superimposing the two models, the mean pLDDT for AlphaFold2 models, and the mean pLDDT for ESMFold models. On the x-axis, we report observations for three subsets of enzymes, respectively those with a known PDB structure in the database, those with similar models (TM-score ≥ 0.6), and those with dissimilar models (TM-score < 0.6). Each violin shows the difference between values computed on the Pfam-covered region (blue) and values computed on the whole protein (orange).

Global models are generated with AlphaFold2 and ESMFold, two different methods relying the first on deep learning of different protein features including evolution information derived from multiple sequence alignments, correlated mutations, and contact maps, the second on the protein sequence representation with embeddings derived from protein large language models [5,6,9]. For functional annotation

we took advantage of Pfam models, casting into HMMs local structural and/or functional conservation highlighted by grouping proteins into families and superfamilies (clans) [11]. What we obtain is interesting, particularly when considering pairwise global models of enzymes without a PDB reference structure (81 % of the entire enzyme dataset). Rather independently of the superimposition of the global enzyme models that we evaluate with the TM-score and group in two sets (TM-score higher or lower than 0.6 [9]), Pfam regions overlap, as demonstrated by the Pfam-restricted TM-score values. In these regions, the predicted local evaluation of the quality of the models (pLDDT) is higher than the quality of the global model. Our results suggest that both AlphaFold2 and ESMFold methods are equally good in grasping the information carried out by Pfam models.

Interestingly, our procedure allows mapping structural and functional information in enzyme domains where the active site is present, as detected by PfamScan. For 807 human enzymes (whose list is available as a supplementary file), the functional annotation of the active site is not yet present in the associated UniProt file.

Funding

The work was supported by the European Union- NextGenerationEU through the Italian Ministry of University and Research under the projects “Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics (ElixirxNextGenIT)” (Investment PNRRM4C2-I3.1, Project IR_0000010, CUP B53C22001800006) and "HEAL ITALIA" (Investment PNRR-M4C2-I1.3, Project PE_00000019, CUP J33C22002920006).

CRediT authorship contribution statement

Gabriele Vazzana: Visualization, Investigation, Formal analysis. **Castrense Savojardo:** Writing – review & editing, Software, Methodology, Funding acquisition. **Matteo Manfredi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Pier Luigi Martelli:** Writing – review & editing, Funding acquisition, Conceptualization. **Rita Casadio:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.01.008.

References

[1] Kandathil SM, Lau AM, Jones DT. Machine learning methods for predicting protein structure from single sequences. Curr Opin Struct Biol 2023;81:102627.
[2] Bepler T, Berger B. Learning the protein language: evolution, structure, and function. Cell Syst 2021;12:654–69. e3.
[3] Consortium UniProt. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51. D523–31.
[4] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[5] Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res 2024;52:D368–75.
[6] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[7] Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–6.
[8] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XV. Proteins 2023;91:1539–49.
[9] Manfredi M, Savojardo C, Iardukhin G, Salomoni D, Costantini A, Martelli PL, et al. Alpha&ESMhFolds: a web server for comparing AlphaFold2 and ESMFold models of the human reference proteome. J Mol Biol 2024;436:168593.
[10] Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res 2023;51:D418–27.
[11] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49:D412–9.
[12] Lesk A. Introduction to bioinformatics. 5th ed. London, England: Oxford University Press; 2019.
[13] van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 2024;42:243–6.
[14] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis. Cambridge, England: Cambridge University Press; 2007.
[15] Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature 2021;596:590–6.
[16] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[17] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29:2722–8.
[18] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2007;68. 1020–1020.