computation

MDPI

Article

Embedding-Based Alignments Capture Structural and Sequence Domains of Distantly Related Multifunctional Human Proteins

Gabriele Vazzana [1], Matteo Manfredi [1], Castrense Savojardo [1], Pier Luigi Martelli [1,2,3,*] and Rita Casadio [1,2,3,*]

[1] Biocomputing Group, University of Bologna, 40126 Bologna, Italy; gabriele.vazzana2@unibo.it (G.V.); matteo.manfredi4@unibo.it (M.M.); castrense.savojardo2@unibo.it (C.S.)
[2] Alma Climate Institute, University of Bologna, 40126 Bologna, Italy
[3] Institute of Biomembrane and Bioenergetics, Italian National Research Council (IBIOM-CNR), 70126 Bari, Italy
* Correspondence: pierluigi.martelli@unibo.it (P.L.M.); rita.casadio@unibo.it (R.C.)

Abstract

Protein embedding is a protein representation that carries along the information derived from filtering large volumes of sequences stored in large archives. Routinely, the protein is represented by a matrix in which each residue is a context-specific vector whose dimensions reflect the size of the large architectures of neural networks (transformers) trained with deep learning algorithms on large volumes of sequences. A recently introduced method (Embedding-Based Alignment, EBA) is particularly suited for pairwise embedding comparisons and, as we report here, allows for remote homolog detection under specific constraints, including protein sequence length similarity. Multifunctional proteins are present in different species. However, particularly in humans, the problem of their structural and functional annotation is urgent since, according to recent statistics, they comprise up to 50% of the human reference proteome. In this paper we show that when EBA is applied to a set of randomly selected multifunctional human proteins, it retrieves, after a clustering procedure and rigorous validation on the reference Swiss-Prot database, proteins that are remote homologs to each other and carry similar structural and functional features as the query protein.

Keywords: embedding; EBA; protein structure alignments; transfer of knowledge; functional annotation; distantly related homologs; clustering

1. Introduction

The standard method for the functional annotation of protein sequences relies on the detection of sequence and/or structural similarities. By this method, the protein is assigned to a family/superfamily whose members, through evolution, diverged in sequence similarities, thus conserving the same function and structure and/or structural domains [1].

Many standard tools are reliable for sequence similarity detection; however, in the so-called twilight zone (<30%) [2], proteins of the same family may retain across long evolutionary timescales structural similarity more than sequence homology [3]. These proteins are referred to as remote homologs.

Routinely, structural protein domains are conserved across families/superfamilies. They have been modeled with Hidden Markov models and collected in databases such as Pfam [4,5]. Presently, Pfam is hosted by InterPro [6]. InterPro (https://www.ebi.ac.uk/interpro/, accessed on 1 December 2025) integrates different predictive models, known

Received: 11 December 2025
Revised: 2 January 2026
Accepted: 15 January 2026
Published: 20 January 2026

as signatures that are routinely derived after multiple sequence alignments, to classify proteins into families and/or superfamilies. The integration of both structural and sequence features allows the characterization of a protein family/superfamily, which in turn makes it possible to assign function/s to a protein with the assumption that proteins of the same family conserve function and structure through evolution and proteins in the superfamily conserve function with different structures [6]. InterPro signatures and domains are also integrated in the ARBA rule system (https://www.uniprot.org/help/arba, accessed on 1 December 2025) for protein automatic annotation, adopted by Uniprot [7], the largest collection of protein sequences (about 250 millions in the last release (2025_4, https://www.uniprot.org/uniprotkb/, accessed on 1 December 2025)).

Protein homologs can therefore be searched either by structural alignment tools or by domain conservation.

When protein structures are available in the PDB database [8] (https://www.rcsb.org, accessed on 1 December 2025), structural alignment tools such as TM-align [3], DALI [9], FAST [10], Mammoth [11], and Foldseek [12] can compute structural similarities by superimposing protein structures. More recently, tools based on deep learning were developed for the same purpose [13–15]. The introduction of AlphaFold [16] for protein structure predictions expanded the number of proteins to be included in the structural space for distant homology search.

An alternative to AlphaFold is ESMFold [17], a protein structure prediction method based on a different form of protein representation called embedding [17,18]. Protein embeddings are produced by transformer-based [19] protein language models (pLMs) [18], which encode every residue in the sequence in a high-dimensional vector. Collectively, these vectors form a matrix that encapsulates the complete representation of the protein [18]. In this way, by considering it a problem of matrix alignment, the search for evolutionarily distant proteins was successfully solved [20–22]. Briefly, the methods encode the two sequences to be compared with pLMs and compute a per-residue embedding similarity matrix with pairwise residue distances based on the Euclidean distance [21,22] or the cosine similarity [20] of the embedding vectors. Dynamic programming procedures based on the gapped Needleman–Wunsch or Smith–Waterman algorithms are then applied to compute global or local alignments, and thresholds on similarity scores are determined to assess the level of homology. Despite the implementation differences, when compared using the same datasets, two of the methods [21,22] perform very similarly. Protein embeddings have also been adopted for different, although related, problems [23], such as the search for remote homologs in large databases [24].

We recently applied embeddings and protein matrix pairwise alignment with Embedding-Based Alignment (EBA, [21]) to classify glutathione S-transferase (GST) proteins for their functional annotation [25]. We found that when constraining protein length, remote homologs within a family can be optimally detected with EBA [25].

The problem of remote homology detection is particularly relevant for the annotation of multifunctional proteins. A recently released version of MultifacetedProtDB [26] (https://multifacetedprotdb.biocomp.unibo.it/, accessed on 1 December 2025), a database of human multifunctional proteins, supports the notion that over 50% of the human Swiss-Prot reference proteome includes proteins endowed with at least two different functional annotations. In this context, it is important to assess whether remote homology detection can be adopted to transfer functional annotation when a protein carries multiple functions. Here, we describe a prototype analysis to assess whether the procedure based on the alignment of embeddings derived from pLMs can be successfully applied to detect remote homologs in a set of randomly selected multifunctional proteins extracted from MultifacetedProtDB [26] (full list available in Supplementary Tables S1 and S2). Among

the different methods available [20–22], we stick to EBA, which does not require training and/or parameter optimization. Notably, in this work we adopt the general-purpose ESM2 pLM, which has 5120-dimensional input vectors, without applying any of the pLM fine-tuning techniques that have been recently introduced to improve performance in different tasks (e.g., [27,28]).

2. Materials and Methods

2.1. Workflow

The workflow of the methods adopted in the search for remote homologs is shown in Figure 1. We compare the matrices of protein embedding (from the Protein Language Model (M)) with the Embedding-Based Alignment (EBA) algorithm (see Section 2.2). EBA outputs are retained for each protein according to the EBA threshold score for reliability ($\geq$3.5). Protein pairs are then clustered to group pairs with shared proteins and analyzed to validate the procedure. This can be done by structure superimposition (when possible), by recognizing InterPro shared signatures, and by establishing whether proteins belong to the same family/superfamily.

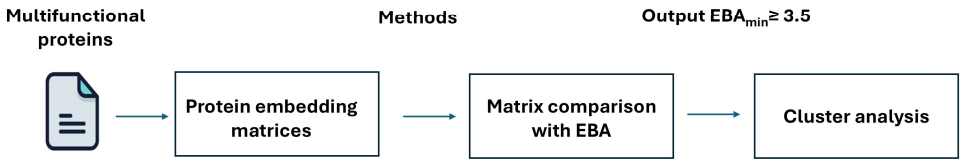

Figure 1. Main steps of the embedding, Embedding-Based Alignment (EBA), and cluster analysis procedures adopted for remote homology detection.

2.2. Protein Embeddings

For protein sequence encoding, we adopted the Meta ESM2 pLM, which has been used to train the protein structure predictor ESMFold [17]. We selected the largest ESM2 model—ESM2-15b—which has 15 billion parameters and was trained with a deep learning approach on 65 million proteins from UniRef50 [17]. The pLM was downloaded and run locally. Given an input protein sequence of length l, the model produces, as output, meaningful and context-aware distributed vector representations for each residue. The dimension 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) [18]. ESM2-15b outputs vectors with D = 5120. The final encoding of the protein is therefore a matrix $M \in R^{l \times 5120}$ (where l is the protein length) that is routinely referred to as per-residue protein embedding (Figure 2).

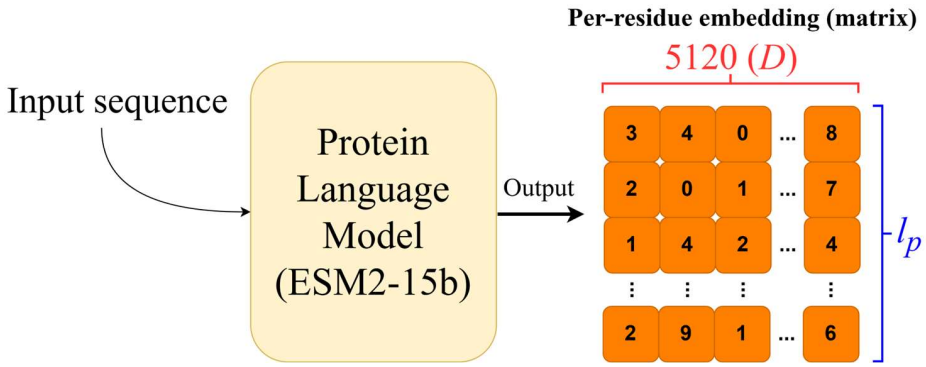

Figure 2. A protein sequence is processed adopting the Protein Language Model, which outputs a per-residue embedding $M \in R^{l \times D}$. In the case of ESM2-15b, this procedure yields a matrix of dimension $l \times 5120$.

2.3. Embedding-Based Alignment (EBA)

We compared per-residue protein embeddings by employing the Embedding-Based Alignment (EBA) method [21]. The algorithm (available at https://git.scicore.unibas.ch/schwede/EBA, accessed on 20 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 also includes an optimizing intermediate step, called "signal enhancement" [21], where each score of a residue pair is normalized to the scores of all the 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. 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 Equation (1):

$$EBA_{min} = \frac{S_{align}}{l_{max}} \quad (1)$$

As stated by [21], length normalization is an important factor in the final score when the proteins being compared are very different in length. Whenever the difference is large, a high EBA_{max}, obtained by normalizing by the length of the shortest protein, reflects the fact that the shorter sequence is entirely contained in the longest. In this case, EBA_{min} is much lower, since the longer sequence is only partially aligned. We analyze results considering that the transfer of knowledge for protein annotation requires structure template conservation [1]. Following the author's suggestion, we selected EBA_{min} to score any two protein sequences during this procedure.

2.4. Clustering

Hierarchical clustering algorithms need a square matrix containing the distances, taken pairwise, between the elements of a set. In our case, these are the EBA values of all of the protein pairs. In order to comply with the constraints of the algorithm, we normalized the EBA values from 0 to 1. In this scale, EBA = 3.5 equals a distance of 0.9, which is the diameter of the circles shown in Figure 3.

As EBA_{min} has no upper bound [21], we normalized the scores with the following procedure. Let x be the EBA_{min} of two proteins, i and j, and min_{score} and max_{score} the minimum and maximum EBA_{min} scores in the similarity matrix, respectively. The normalized distance value of x will be:

$$x_{distance} = 1 - \frac{x - min_{score}}{max_{score} - min_{score}} \quad (2)$$

Lower distance values indicate high embedding similarities. The resulting distance matrix has been given as input to three different hierarchical linkage methods: single, average, and complete [29]. Since it appears that the same protein can be part of different pairs, we adopt a clustering procedure for finding groups containing pairs with similar proteins.

We selected groups from complete-linkage clustering as it represents the most stringent method for conserving the same proteins in the same group.

2.5. Computational Time

The time required for embedding generation and EBA runs on the 90 proteins of the multifunctional dataset was 20 and 5 min, respectively, on a machine endowed with 80 CPUs and 755 gigabytes of RAM.

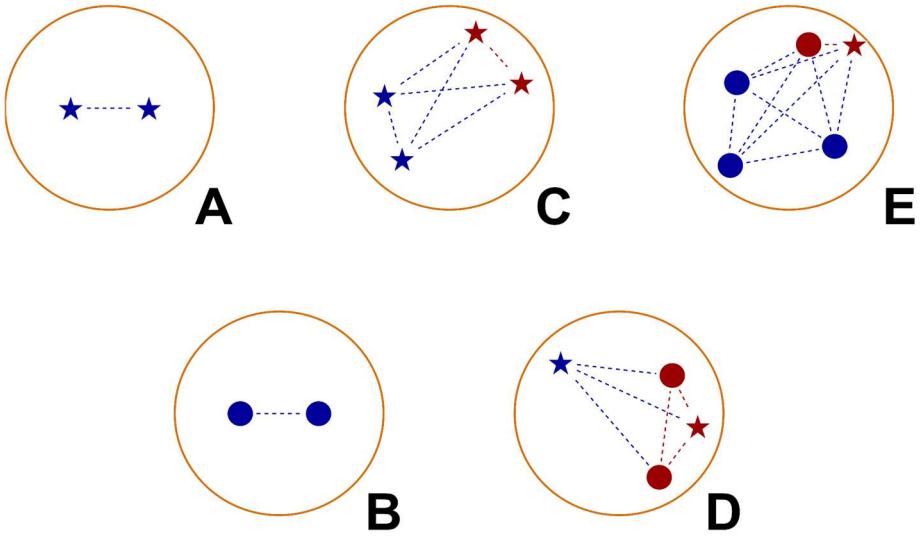

Figure 3. Types of topologies obtained from our dataset using a complete-linkage clustering procedure. Each protein is represented as a point in a distance space; red proteins represent pairs of known homologs (sequence identity $\geq$ 30%) and blue proteins are proteins with low sequence similarity; proteins with a PDB structure are represented with a star. If two proteins are close in space (EBA $\geq$ 3.5, distance < 0.9), a dotted line links them. In this example, we depict the five different observed scenarios analyzed in Tables 1–3. (**A**,**B**) represent unique pairs that do not cluster with any other sequence. In (**A**), both have an available PDB structure, allowing for the validation of remote homology via structural superimposition; in (**B**), neither has a PDB, and functional annotation of InterPro signatures can assist in validation. (**C**–**E**) represent groups of proteins that are all similar to each other. Inside groups, pairs of proteins may have a high similarity (red points), and the only significant pairs are those among distant proteins (blue dotted lines). In (**C**), all proteins have a PDB; similar to (**A**), a multiple structural superimposition can validate that they retain the same structure. In (**D**), some proteins have a PDB structure, and some do not; in this case, only some pairs can be validated through the structure superimposition. In (**E**), only one protein has a PDB structure; similar to (**B**), validation can be derived only considering shared signatures (InterPro).

2.6. Validation

Validation of the results was performed by comparing the available protein structures of the multifunctional proteins and by analyzing shared structural and sequence features. InterPro signatures and UniProt family and superfamily annotations were used to identify structural and functional similarities among the proteins. In addition, we compared the Enzyme Commission (EC) numbers of the protein pairs, considering only the first digit, which is indicative of the seven classes in which enzymes are routinely grouped: 1 oxidoreductase, 2 translocase, 3 hydrolase, 4 lyase, 5 isomerase, 6 ligase, and 7 translocase (https://iubmb.qmul.ac.uk/enzyme/ accessed on 1 December 2025).

3. Results

We consider 90 proteins randomly selected from a database of multifunctional proteins [26]. Following the methodology described in Figure 1, we identify 147 pairs of proteins with an EBA $\geq$ 3.5, which is the EBA reliability threshold [21]. Amongst them, 71 pairs have a sequence identity $\leq$ 30% and are candidates to be remote homologs; this set includes 63 proteins, of which 38 are remote homologs in the reference families (marked in bold in Tables 1–3). The results of the remote homology search are validated by considering two groups of data. First, we consider 8 pairs of proteins that are coupled only with one another. Then, we analyze groups with at least 3 proteins, as generated by the procedure described in Section 2.4.

3.1. Validation of Unique Protein Pairs

We focus on protein pairs that, after the clustering procedure, are paired only with each other and whose sequence identity is ≤30%. Table 1 lists eight pairs of remote homologs (types A and B in Figure 3). For validating the distantly related homology, we compute, when possible, their structural alignment. It is reported above that structural superimposition is a signature of remote homology when the sequence identity is low (≤30%) [1,2,14]. Our goal is to validate the methodology adopted for distant homology search considering the available structural and functional features present in the corresponding PDB and UniProt files, when the structure is available (Table 1). Six of the eight protein pairs are endowed with a PDB structure, which covers some 50% of the protein sequence. We adopt the Pairwise Structure Alignment tool hosted on the RCSB PDB website (https://www.rcsb.org/alignment accessed on 1 December 2025). With this tool, we can visualize the superimposition, and for each structure pair, we can compute a TM-score, which is a metric representing the global similarity of the two structures [30]. Considering the results in Table 1, we observe three different possibilities.

- Pair #1 (Figure 4A). The embedded sequences have an EBA score (15.48) much higher than the reliability threshold (3.5) (see Section 2.3 for details), and the two sequences have similar length. In this case, the two structures are almost perfectly superimposed (TM-score = 0.86), supporting the notion that they are indeed remote homologs, notwithstanding the computed low sequence identity. The two proteins are multifunctional, with the same two Enzyme Commission (EC) numbers (4 and 5), indicating that they can act as lyases (4) and isomerases (5).

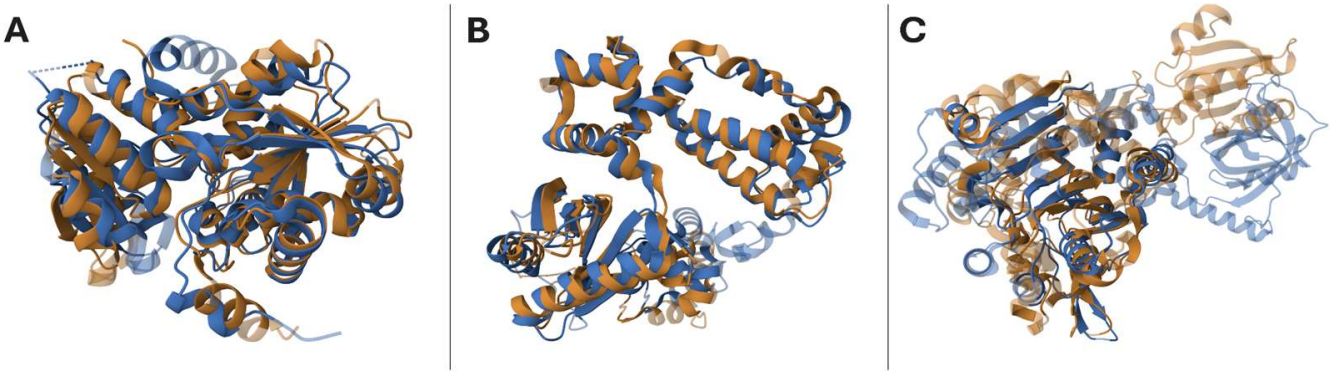

Figure 4. The figure shows superimpositions of the PDB structures of three pairs, as indicated in Table 1. Panels (**A**–**C**) are relative to pair #1, #2, and #4, respectively. The superimpositions are made from the RCSB PDB website (https://www.rcsb.org/alignment accessed on 1 December 2025). Panel (**A**) shows an example of an almost perfect superimposition (TM-score = 0.86); Panel (**B**) depicts the good superimposition of the whole P06746 protein with the available fragment of the protein Q9UGP5 (TM-score = 0.80); and, finally, Panel (**C**) displays a poor general superimposition of two paired proteins (TM-score = 0.36): The two structures have a superimposed region (176 residues) that contains the two shared InterPro signatures. In all panels, the orange backbone represents the protein listed in the column "Protein 1" ot Table 1, while the blue backbone represents the protein listed in the column "Protein 2".

- Pairs #2 (Figure 4B) and #3. The embedded sequences of both pairs have an EBA score higher than the threshold (11.27 and 4.22, respectively), and the two sequences in the pairs have very different lengths. In both cases, the PDB of the longer sequence covers only a fragment of the protein, which superimposes well (TM-score > 0.6) with the other member of the pair. Despite the two proteins being very different in terms of their number of residues, the shortest protein assumes the same conformation

as a domain of the longer protein. Validation indicates that the proteins in pair #2 belong to the DNA polymerase type-X family, sharing 11 InterPro signatures, and that they act as transferases (2) and lyases (4). Proteins in pair #3 belong to the protein kinase superfamily, sharing an InterPro signature, and they act as transferases (2) and hydrolases (3).

- Pairs #4 (Figure 4C), #5, and #6. The embedded sequences of all the pairs have an EBA score above the threshold (4.41, 5.96, and 5.63, respectively), and the three sequences are different in length (the difference ranges from 63 to 132 residues). In all cases, the structures in the pairs superimpose only partially (TM-score < 0.6). However, we observe a significant overlap of specific segments, suggesting that they have common structural features. Validation indicates that the proteins in pair #4 belong to the class-II aminoacyl-tRNA synthetase family, sharing two InterPro signatures (Figure 4C), and that they both act as transferases (2) and ligases (6). Proteins in pair #5 belong to the HhH-GPD superfamily, sharing three InterPro signatures, and they act as hydrolases (3) and lyases (4). Proteins in pair #6 belong to the DNA polymerase families, share two InterPro signatures, and act as transferases (2) and hydrolases (3), and one of the two also acts as a lyase (4).

Table 1. Validation of remote homology in unique protein pairs.

Entries			Validation	
#	**Protein 1**	**Protein 2**	**Comparison**	**Shared Annotation**
1 (Figure 4A)	**Q96GA7** **Len**: 329 **PDB**: 2RKB **Cov**: 318	**Q9GZT4** **Len**: 340 **PDB**: 3L6B **Cov**: 322	**EBA**: 15.48 **Seq. Id.**: 29% **Len diff**: 11 **TM-score**: 0.86 **Aln res**: 286	**InterPro**: IPR000634, IPR001926, IPR036052 **Family**: serine/threonine dehydratase family **E.C.**: (4, 5)
2 (Figure 4B)	**P06746** **Len**: 335 **PDB**: 8VFG **Cov**: 327	**Q9UGP5** **Len**: 575 **PDB**: 7M09 **Cov**: 331	**EBA**: 11.27 **Seq. Id.**: 21% **Len diff**: 240 **TM-score**: 0.80 **Aln res**: 262	**InterPro**: IPR002054, IPR019843, IPR010996, IPR028207, IPR018944, IPR027421, IPR037160, IPR022312, IPR002008, IPR043519, IPR029398 **Family**: DNA polymerase type-X family **E.C.**: (2, 4)
3	**Q96S44** **Len**: 253 **PDB**: 7SZC **Cov**: 229	**Q9BRS2** **Len**: 568 **PDB**: 4OTP **Cov**: 236	**EBA**: 4.22 **Seq. Id.**: 12% **Len diff**: 315 **TM-score**: 0.67 **Aln res**: 153	**InterPro**: IPR011009 **Superfamily**: protein kinase superfamily **E.C.**: (2, 3)
4 (Figure 4C)	**P41250** **Len**: 703 **PDB**: 2ZT5 **Cov**: 530	**Q15046** **Len**: 597 **PDB**: 6ILD **Cov**: 501	**EBA**: 4.41 **Seq. Id.**: 20% **Len diff**: 106 **TM-score**: 0.36 **Aln res**: 176	**InterPro**: IPR006195, IPR045864 **Family**: class-II aminoacyl-tRNA synthetase family **E.C.**: (2, 6)
5	**O15527** **Len**: 345 **PDB**: 2XHI **Cov**: 316	**P78549** **Len**: 282 **PDB**: 7RDS **Cov**: 232	**EBA**: 5.96 **Seq. Id.**: 21% **Len diff**: 63 **TM-score**: 0.37 **Aln res**: 120	**InterPro**: IPR011257, IPR003265, IPR023170 **Superfamily**: HhH-GPD superfamily **E.C.**: (3, 4)
6	**P54098** **Len**: 1239 **PDB**: 4ZTU **Cov**: 1222	**P28340** **Len**: 1107 **PDB**: 9EKB **Cov**: 1107	**EBA**: 5.63 **Seq. Id.**: 18.33 **Len diff**: 132 **TM-score**: 0.25 **Aln res**: 87	**InterPro**: IPR043502, IPR012337 **Superfamily**: DNA polymerase families **E.C.**: (2, 3); (2, 3, 4)

Table 1. *Cont.*

Entries			Validation	
#	**Protein 1**	**Protein 2**	**Comparison**	**Shared Annotation**
7	**Q9NST1** **Len**: 481 **PDB**: NO	**Q9UP65** **Len**: 541 **PDB**: NO	**EBA**: 3.95 **Seq. Id.**: 16% **Len diff**: 60	**InterPro**: IPR016035 **Superfamily**: FabD/lysophospholipase-like superfamily **E.C.**: (2, 3)
8	**Q14032** **Len**: 418 **PDB**: NO	**Q99487** **Len**: 392 **PDB**: NO	**EBA**: 7.05 **Seq. Id.**: 20% **Len diff**: 26	**InterPro**: IPR029058 **Superfamily**: α/β hydrolase superfamily **E.C.**: (2, 3)

Protein pairs with an EBA $\geq$ 3.5 and sequence identity $\leq$ 30%, considering unique proteins that are paired with only one partner. For each protein, we report its length. For 5 out of the 7 pairs, a PDB structure is available. We select X-ray PDB entries, prioritizing those with the highest coverage and the best resolution. We report the PDB ID, the chain, and the number of modeled residues in the structure ("Cov"). In the column comparison, we report the EBA score, the sequence identity of the two proteins obtained with the Align resource from UniProt [31], and, when applicable, the TM-score [30] and the number of aligned residues obtained after structurally superimposing the PDBs (with the TM-Align method [3] selected from the RCSB PDB Pairwise Structure Alignment web page: https://www.rcsb.org/alignment accessed on 1 December 2025). The column "Shared InterPro" reports a list of InterPro entries that are annotated in both proteins.

Finally, for pairs #7 and #8, we cannot make any structure-based comparison. Still, they share common features, as in pair #7, where both proteins belong to the FabD/lysophospholipase-like superfamily, share an InterPro signature, and act as transferases (2) and hydrolases (3). Proteins in pair #8 belong to the α/β hydrolase superfamily, share an InterPro signature, and act as transferases (2) and hydrolases (3). According to the UniProt annotation, they apparently conserve an InterPro motif indicative of a given family and can therefore be attributed to the given family for the transfer of further annotation.

3.2. Remote Homology Validation in Protein Groups Detected with Complete-Linkage Clustering

After complete-linkage clustering, we obtain 7 groups containing 33 proteins. Groups have two topologies (C and D in Figure 3). As in Table 1, we manually verify whether proteins in the same group share overlapping structures (TM $\geq$ 0.6) and common annotations in UniProt. Groups #1 and #2 have similar lengths and overlapping structures, with an average TM-score of 0.75 and 0.96, respectively. Furthermore, the proteins in the two groups share 4 and 6 InterPro signatures, respectively.

It appears that in 6 out of 7 groups, homologous proteins (sequence identity > 30%, color-coded in red in Table 2) are also present due to clustering. The homologous proteins that are within the same group are distantly related homologs for all other proteins in the group (see Table S1 in the Supplementary Materials for details). These proteins are important for the family/superfamily annotation. Groups #3–#7 are of the type D depicted in Figure 3 (mixed groups with protein endowed with structure and sequences).

Overlapping protein structures validate remote homology even when two homologous proteins are present in the group (Group #1, Figure 5A). If a sequence without a structure is in the group with folded proteins, the AlphaFold2 model of the sequence can be structurally aligned to the other protein structures of the group (Group #3, Figure 5B; the AlphaFold model of P24557 is color-coded in red). We find that all the groups in Table 2 have in common at least one InterPro signature, confirming the remote homology of all the proteins in the group. Group #6 contains proteins with very different lengths (P40939, Q08426, and P30084, with lengths of 727, 723, and 263 residues, respectively). Proteins P40939 and P30084 have a PDB structure, and they superimpose with a low TM-score (0.27). However, the shortest protein completely superimposes to the N-terminus region of the longer protein,

as shown in Figure 6. The superimposed region contains the InterPro family signature (IPR001753, enoyl-CoA hydratase/isomerase family) shared by the proteins in the group.

Table 2. Validation of remote homology in protein groups after clustering.

Groups		Validation	
#	**Group Members**	**Comparison**	**Shared Annotation**
1 (Figure 5A)	P08263 * (222); P78417 * (241); Q9H4Y5 * (243); Q03013 * (218); O60760 * (199); O43708 * (216);	**Mean EBA**: 11.29 **Mean TM-score**: 0.75	**InterPro**: IPR036249; IPR010987; IPR004045; IPR036282 **Superfamily**: GST superfamily **E.C.**: (1, 2, 5); (1, 2); (1, 2); (2, 4); (2, 5); (2, 5)
2	P11586 * (935); P13995 * (315); Q9H903 * (347)	**Mean EBA**: 10.21 **Mean TM-score**: 0.96	**InterPro**: IPR020867; IPR000672; IPR020630; IPR046346; IPR020631; IPR036291 **Family**: tetrahydrofolate dehydrogenase/cyclohydrolase family **E.C.**: (1, 3, 6); (1, 3); (1, 3)
3 (Figure 5B)	Q96SQ9 (504); P04798 * (512); P05177 * (516); Q16678 * (543); P51589 (502); P24557 (533); Q16647 * (500)	**Mean EBA**: 16.89 **Mean TM-score**: 0.84	**InterPro**: IPR036396; IPR001128 **Family**: cytochrome P450 family **E.C.**: (1, 4, 5); (1, 4); (1, 4); (1, 4); (1, 5); (4, 5); (4, 5)
4	O14880 (152); Q99735 * (147); O14684 * (152); Q16873 * (150)	**Mean EBA**: 9.71 **Mean TM-score**: 0.79	**InterPro**: IPR001129; IPR023352 **Family**: MAPEG family **E.C.**: (1, 2, 4); (1, 2, 4); (1, 2, 5); (2, 4)
5	P16118 * (471); Q16875 * (520); Q16877 (469); Q96T60 (521); O60825 * (505)	**Mean EBA**: 13.50 **Mean TM-score**: 0.92	**InterPro**: IPR027417 **Superfamily**: P-loop containing nucleoside triphosphate hydrolases **E.C.**: (2, 3); (2, 3); (2, 3); (2, 3); (2, 3)
6 (Figure 6)	P40939 * (727); Q08426 (723); P30084 * (263)	**Mean EBA**: 11.85 **Mean TM-score**: 0.27	**InterPro**: IPR029045; IPR018376; IPR001753 **Family**: enoyl-CoA hydratase/isomerase family **E.C.**: (1, 2, 4); (1, 4, 5); (4, 5)
7	P53816 * (162); Q9HDD0 (168); Q9NWW9 * (162); Q96KN8 (279); Q9UL19 * (164)	**Mean EBA**: 10,16 **Mean TM-score**: 0.88	**InterPro**: IPR007053; IPR051496 **Family**: H-rev107 family **E.C.**: (2, 3); (2, 3); (2, 3); (2, 3)

Complete-linkage groups with at least three members identified in our dataset. Proteins having a sequence identity higher than 30% among themselves are marked in red. A group contains only proteins with an EBA score of 3.5 or higher. We mark with a star (*) proteins for which a PDB structure is available. The column "Shared Annotation" lists entries annotated in all members of the cluster. Finally, in the last column, we mark whether proteins in the group belong to the same family/superfamily.

The groups in Table 3 belong to the topology type E, depicted in Figure 3, including only one protein with a PDB structure among sequences. In this case, validation of remote homology relies on sharing InterPro features that are signatures of the different protein families and superfamilies. For protein modeling, the user may rely on AlphaFold2, and remote homology should be validated by structure overlapping. Table S2 in the Supplementary Materials contains a detailed list of the scores of all the pairs in the groups.

Table 3. Validation of remote homology in protein groups that contain only one structure.

Groups			Validation
#	**Group Members**	**Comparison**	**Shared Annotation**
8	Q14191 * (1432); Q9H8H2 (1009); Q14527 (1009)	**Mean EBA**: 5.92	**InterPro**: IPR014001; IPR001650; IPR027417 **Superfamily**: helicase superfamily **E.C.**: (3, 5); (3, 5); (2, 3)
9	Q8TAT5 (605); Q96FI4 * (390); Q969S2 (332)	**Mean EBA**: 6.80	**InterPro**: IPR015886; IPR012319; IPR010979 **Family**: FPG family **E.C.**: (3, 4); (3, 4); (3, 4)
10	A6NGU5 (568); P19440 * (569); P36269 (586); Q6P531 (493); Q9UJ14 (662)	**Mean EBA**: 17.26	**InterPro**: IPR029055; IPR043137 **Family**: gamma-glutamyltransferase family **E.C.**: (2, 3); (2, 3); (2, 3); (2, 3); (2, 3)
11	P30043 * (206); P14060 (373); P26439 (372)	**Mean EBA**: 9.48	**InterPro**: IPR036291 **Superfamily**: NAD(P)-binding domain superfamily **E.C.**: (1, 2); (1, 5); (1, 5)

Complete-linkage groups with at least three members identified in our dataset. Proteins having a sequence identity higher than 30% among themselves are marked in red. A group contains only proteins with an EBA score of 3.5 or higher. We mark with a star (*) proteins for which a PDB structure is available. The column "Shared Annotation" lists entries annotated in all members of the cluster. Finally, in the last column, we mark whether proteins in the group belong to the same family/superfamily.

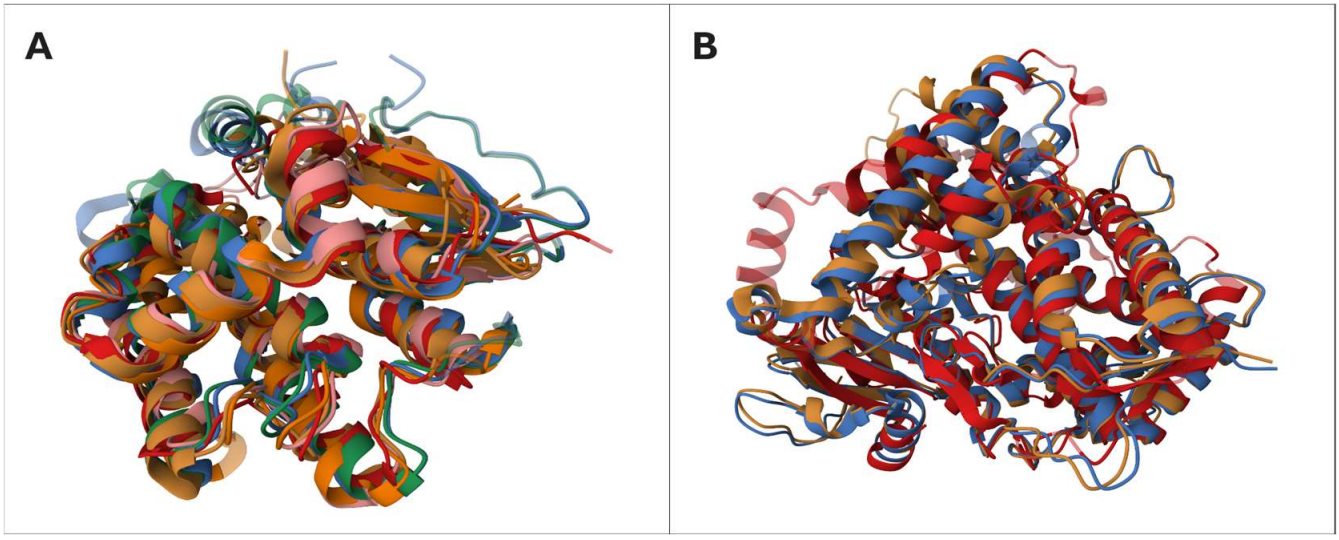

Figure 5. Example of multiple structural superimpositions of PDB structures for proteins belonging to the same group. Panel (**A**) shows the superimposition of 6 PDB structures belonging to group #2 (GST superfamily; following the order in which they appear in the table, the proteins are colored in orange, blue, green, pink, red, and light orange respectively). Panel (**B**) shows the superimposition of 2 PDB structures from group #3 (cytochrome P450 family; protein P04798 with an orange backbone and protein P05177 with a blue backbone) and an AlphaFold2 model (backbone in red) for the protein P24557 in the same group, which lacks a PDB.

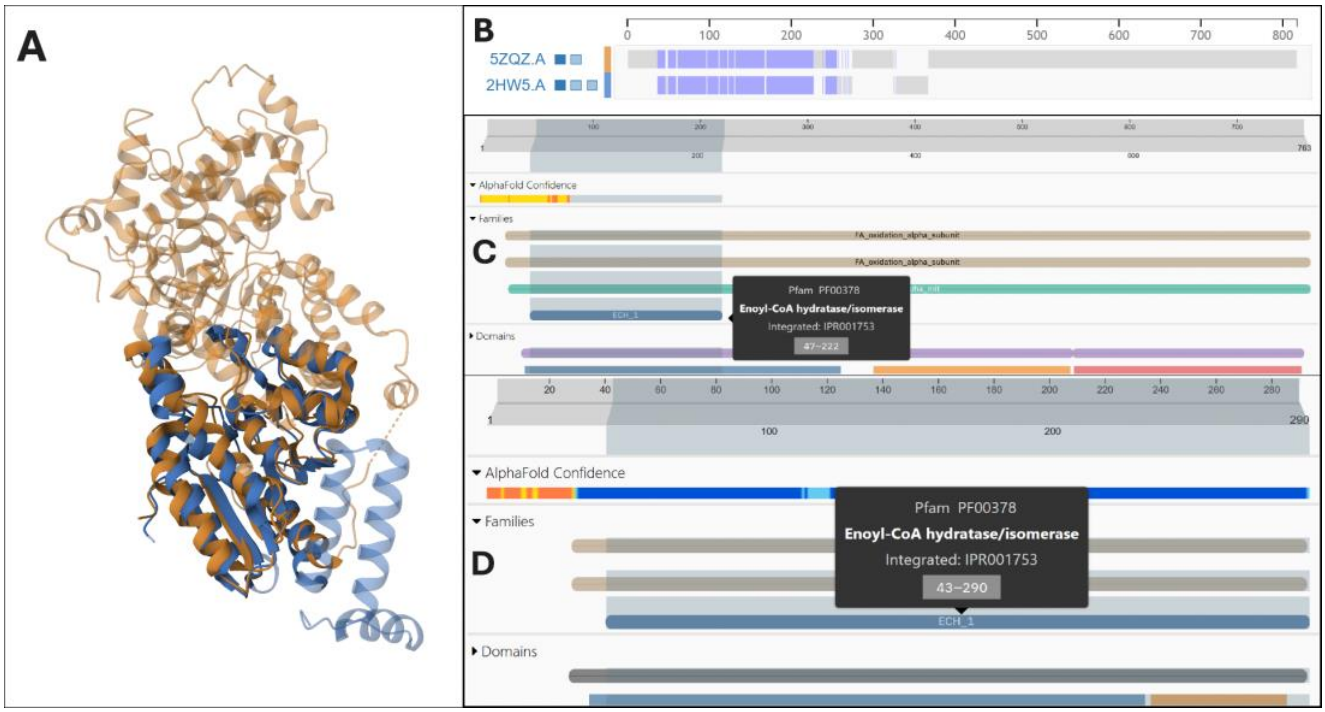

Figure 6. Panel (**A**) shows the superimposition of the PDB structures of cluster #6, P40939 (in orange) and P30084 (blue). Panel (**B**) highlights the superimposed region on the two sequences. It can be observed that the superimposition happens in the N-terminus of the longer protein and in the entirety of the shortest protein. Panels (**C**,**D**) display the InterPro pages of proteins P40939 and P30084, respectively, highlighting the shared family signature (InterPro IPR001753) corresponding to the superimposed regions of the structures.

4. Conclusions and Perspectives

In this paper, we show that protein embedding representations and Embedding-Based Alignment can retrieve remote homologs in specific families/superfamilies in a random set of 90 multifunctional proteins. After clustering, we find that 38 proteins are remote homologs in specific families/superfamilies. Validation highlights that they are also multifunctional and share common InterPro domains. These domains are signatures of the protein family/superfamily. We find that remote homologs share protein structures when available and/or structural domains. In both cases, the EBA score of the embedded protein sequences is high and well above the EBA threshold for reliability. When the structures only partially overlap (TM-score < 0.6), the sharing of protein InterPro signatures can still guarantee a remote homology relationship. All our results, although limited to a small set of proteins for the sake of clarity, suggest that protein embedding, Embedding-Based Alignment, and complete-linkage clustering are useful and fast computational approaches for remote homology detection. A protein can inherit structural and functional features from entering a family/superfamily with enough knowledge useful for the transfer of annotation.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/computation14010025/s1, Table S1: Validation of remote homology in protein groups after clustering; Table S2: Validation of remote homology in protein groups which contain only one structure.

Author Contributions: G.V.: data curation, methodology, software, investigation, and writing—original draft preparation. M.M.: data curation, methodology, software, investigation, and writing—original draft preparation. C.S.: validation, writing—review and editing, and funding acquisition. P.L.M.:

validation, writing—review and editing, and funding acquisition. R.C.: conceptualization, supervision, investigation, validation, and writing—review and editing. 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), "HEAL ITALIA" (Investment PNRR-M4C2-I1.3, Project PE_00000019, CUP J33C22002920006), and "National Centre for HPC, Big Data and Quantum Computing" (Investment PNRR-M4C2-I1.4, Project CN_00000013).

Data Availability Statement: All data used in this research are included in the paper and in the Supplementary Tables.

Conflicts 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 manuscript.

References

1. Lesk, A.M. *Introduction to Protein Science: Architecture, Function, and Genomics*, 3rd ed.; Oxford University Press: Oxford, UK, 2016.
2. Doolittle, R.F. *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*; University Science Books: Mill Valley, CA, USA, 1987.
3. Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309. [CrossRef]
4. Finn, R.D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.E.; Gavin, O.L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; et al. The Pfam Protein Families Database. *Nucleic Acids Res.* **2010**, *38*, D211–D222. [CrossRef]
5. Paysan-Lafosse, T.; Andreeva, A.; Blum, M.; Chuguransky, S.R.; Grego, T.; Pinto, B.L.; Salazar, G.A.; Bileschi, M.L.; Llinares-López, F.; Meng-Papaxanthos, L.; et al. The Pfam Protein Families Database: Embracing AI/ML. *Nucleic Acids Res.* **2025**, *53*, D523–D534. [CrossRef]
6. Blum, M.; Andreeva, A.; Florentino, L.C.; Chuguransky, S.R.; Grego, T.; Hobbs, E.; Pinto, B.L.; Orr, A.; Paysan-Lafosse, T.; Ponamareva, I.; et al. InterPro: The Protein Sequence Classification Resource in 2025. *Nucleic Acids Res.* **2025**, *53*, D444–D456. [CrossRef]
7. 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.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [CrossRef]
8. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
9. Holm, L.; Laiho, A.; Törönen, P.; Salgado, M. DALI Shines a Light on Remote Homologs: One Hundred Discoveries. *Protein Sci.* **2023**, *32*, e4519. [CrossRef] [PubMed]
10. Zhu, J.; Weng, Z. FAST: A Novel Protein Structure Alignment Algorithm. *Proteins* **2005**, *58*, 618–627. [CrossRef]
11. Ortiz, A.R.; Strauss, C.E.M.; Olmea, O. MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison. *Protein Sci.* **2002**, *11*, 2606–2621. [CrossRef]
12. Van Kempen, M.; Kim, S.S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C.L.M.; Söding, J.; Steinegger, M. Fast and Accurate Protein Structure Search with Foldseek. *Nat. Biotechnol.* **2024**, *42*, 243–246. [CrossRef]
13. Bileschi, M.L.; Belanger, D.; Bryant, D.H.; Sanderson, T.; Carter, B.; Sculley, D.; Bateman, A.; DePristo, M.A.; Colwell, L.J. Using Deep Learning to Annotate the Protein Universe. *Nat. Biotechnol.* **2022**, *40*, 932–937. [CrossRef]
14. 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]
15. Radivojac, P. Advancing Remote Homology Detection: A Step toward Understanding and Accurately Predicting Protein Function. *Cell Syst.* **2022**, *13*, 435–437. [CrossRef]
16. 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]
17. 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] [PubMed]
18. 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]

19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
20. Iovino, B.G.; Ye, Y. Protein embedding based alignment. *BMC Bioinform.* **2024**, *25*, 85. [CrossRef] [PubMed]
21. 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]
22. Spicer, R.; Raychawdhary, N.; Danwada, S.; Udomprasert, P.; Seals, C.; Bhattacharya, S. Evaluating the significance of embedding-based protein sequence alignment with clustering and double dynamic programming for remote homology. *Sci. Rep.* **2025**, *15*, 39601. [CrossRef]
23. Kilinc, M.; Jia, K.; Jernigan, R.L. Major advances in protein function assignment by remote homolog detection with protein language models—A review. *Curr. Opin. Struct. Biol.* **2025**, *90*, 102984. [CrossRef] [PubMed]
24. Kilinc, M.; Jia, K.; Jernigan, R.L. Improved global protein homolog detection with major gains in function identification. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2211823120. [CrossRef] [PubMed]
25. 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. [CrossRef]
26. 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*, D494–D501. [CrossRef] [PubMed]
27. Schmirler, R.; Heinzinger, M.; Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nat. Commun.* **2024**, *15*, 7407. [CrossRef]
28. Saadat, A.; Fellay, J. Fine-tuning protein language models to understand the functional impact of missense variants. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 2199–2207. [CrossRef] [PubMed] [PubMed Central]
29. Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. *Cluster Analysis*, 5th ed.; Wiley Series in Probability and Statistics; Wiley: Chichester, UK, 2011.
30. Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* **2004**, *57*, 702–710. [CrossRef]
31. Sievers, F.; Higgins, D.G. Clustal Omega for Making Accurate Alignments of Many Protein Sequences. *Protein Sci.* **2018**, *27*, 135–145. [CrossRef]