Ran *et al. Genome Biology* (2025) 26:407
https://doi.org/10.1186/s13059-025-03886-y

Genome Biology

RESEARCH Open Access

A scalable equivariant graph network framework for precise protein function prediction

Zixu Ran[1†], Xudong Guo[1,2†], Tong Pan[2,3†], Yue Bi[3], Yi Hao[1], Heyun Sun[2], Jiangning Song[3*] and Fuyi Li[1,2*]

[†]Zixu Ran, Xudong Guo and Tong Pan contributed equally to this work.

*Correspondence:
Jiangning Song
Jiangning.Song@monash.edu
Fuyi Li
fuyi.li@nwafu.edu.cn
[1]College of Information Engineering, Northwest A&F University, Yangling 712100, China
[2]South Australian immunoGENomics Cancer Institute (SAiGENCI), The University of Adelaide, Adelaide 5005, Australia
[3]Department of Biochemistry and Molecular Biology, Monash University, Melbourne 3168, Australia

Abstract

Background Protein function research helps in understanding the complex biological processes that occur within cells. However, the intricate nature of protein structures and functions, along with the rapid growth of protein sequence data, presents a pressing challenge to develop efficient computational methods for accurate protein annotation.

Results In this study, we propose ENGINE, a multi-channel deep learning framework designed for robust protein function prediction. ENGINE integrates an equivariant graph convolutional network model to capture geometric features from protein 3D structures, leverages the large language model ESM-C to encode evolutionary and sequence-derived information, and combines an innovative 3D sequence representation that unifies spatial and sequential signals. We demonstrate that ENGINE consistently surpasses current state-of-the-art methods across diverse protein function prediction benchmarks, demonstrating robust generalisation and high predictive accuracy. Beyond performance, ENGINE provides interpretable insights into key sequence features and structural motifs, enabling the identification of functionally critical residues and substructures within proteins. This facilitates a deeper mechanistic understanding of protein function annotation outcomes and supports hypothesis generation for downstream biological studies.

Conclusion By offering reliable predictions with biological interpretability, ENGINE contributes to advancing research into cellular processes and disease mechanisms. The model is available at GitHub (https://github.com/ABILiLab/ENGINE) and Zenodo (https://doi.org/10.5281/zenodo.17221153), serving as a valuable tool for the broader scientific community.

Background

Proteins serve as the central mediators of biological processes in living organisms, orchestrating various functions, including the catalysis of biochemical reactions, the regulation of signalling pathways and gene expression, and the maintenance of cellular structural integrity [1, 2]. While many proteins contain intrinsically disordered regions that confer functional flexibility, most amino acid sequences inherently encode the

information required to fold into unique three-dimensional structures, enabling them to carry out a wide range of cellular functions [3, 4]. Accurate protein function annotation is essential for elucidating the molecular mechanisms underlying cellular behaviour, identifying disease-associated proteins, and accelerating therapeutic discovery [5–7]. To systematically categorise protein function, several classification frameworks have been developed. Among these, the Gene Ontology (GO) has become the widely adopted standard, providing a structured and hierarchical vocabulary to describe protein functions across three interrelated domains: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) [8, 9].

The rapid development of high-throughput, low-cost sequencing technologies has led to a surge in protein sequence data [10]. As of February 2025, the UniProtKB database contains over 250 million protein sequences, while only 0.2% have been functionally annotated in the manually curated Swiss-Prot database [11]. This disparity stems largely from the intrinsic low throughput of experimental validation, underscoring the urgent need for scalable, high-throughput, and accurate approaches to protein function annotation [12, 13]. While experimental techniques such as X-ray crystallography [14], Nuclear Magnetic Resonance (NMR) spectroscopy [15], and cryo-electron microscopy [16] provide gold-standard insights into protein structure and function, their high cost and time-consuming nature preclude their use for large-scale annotation, compelling the development of robust and reliable computational predictive strategies.

With the advancements in computational biology and artificial intelligence, numerous machine learning (ML) and deep learning (DL) approaches have been developed to facilitate protein functional annotation. For one of the earliest tools, Das et al. [17] proposed a machine learning (ML) based method that leverages protein structural domains. This was followed by tools such as GOLabeler [18] and NetGO [19], which employed logistic regression (LR) and K-nearest neighbours (KNN) algorithms for protein function classification. However, traditional machine learning (ML) methods often struggle to handle high-dimensional data and lack the capacity to capture the complex nonlinear relationships between protein sequences and their functions. In recent years, the emergence of deep learning has driven a wave of research leveraging neural network-based models trained on curated "ground truth" annotations from public databases to predict protein function. These models, empowered by advanced neural architectures, can capture intricate patterns within protein data for protein function prediction. Current methods, such as TALE [20], DeepGOPlus [21], DeepFRI [22], and PFresGO [23], primarily harness sequence features derived from pre-trained protein language models to enhance GO term prediction [24–27]. A summary of these methods, including algorithmic frameworks, feature encoding strategies, evaluation metrics, and input types, is provided in Additional file 1: Table S1. While existing methods have yielded valuable insights into protein function, fully harnessing the rich information embedded in three-dimensional structures to model the intricate sequence-structure–function relationship remains a persistent challenge. Recognising the potential of integrating both structural and sequential features to advance protein function prediction and deepen our understanding of protein biology, we propose a novel approach that goes beyond traditional sequence-based analyses and static structural descriptors. Specifically, we introduce an innovative transformation of protein 3D structures into 3Di token sequences. This unique and complementary structural representation enables the model to derive

structure-aware embeddings through standard sequence-based processing, thereby facilitating more accurate and insightful functional annotation by effectively harnessing multi-level protein information.

We present ENGINE, a multi-channel neural network designed for accurate protein function annotation and identification of functionally relevant regions within proteins. ENGINE integrates information from both protein sequences and structures through multiple complementary channels. The first channel leverages the pre-trained protein language model ESMFold [28] to predict the three-dimensional (3D) structure of proteins. A graph is then constructed using a K-nearest neighbours (KNN) algorithm to capture spatial relationships between residues, which is subsequently processed by Equivariant Graph Neural Networks (EGNNs). The second channel captures evolutionary and contextual information from protein sequences via residue-level embeddings derived from the large language model ESM-C [29]. The third channel incorporates a discrete structural representation derived from Foldseek's 3Di alphabet [30], which encodes tertiary residue interactions into a sequence format. Information from these channels is subsequently fused, enabling ENGINE to assign confidence scores to GO terms. We demonstrate that ENGINE outperforms current state-of-the-art models on protein function annotation benchmarks. Furthermore, we conduct a series of interpretable analyses to elucidate the model's predictions and to localise functionally important residues within protein sequences.

Results

Overview of the ENGINE framework

The ENGINE framework is shown in Fig. 1. To integrate information from protein sequences and structures, we designed ENGINE as a multi-channel deep learning model incorporating diverse biological features, such as evolutionary signals and structural conformations, to enable efficient and accurate protein function prediction. As illustrated in Fig. 1b, the model comprises three main components: the Structural Channel, the 3Di Sequence Channel, and the Sequence Channel. These complementary feature extraction channels harness distinct types of protein information to assign GO terms. The Structural Channel transforms the three-dimensional protein structure into a graph representation, capturing its topological features and spatial arrangements of residues. The 3Di Sequence Channel leverages Foldseek, a deep learning-based protein language model that encodes information about 3D structural conformations into the feature representation of the entire protein sequence. This results in sequence embeddings enriched with structural context, thereby facilitating the functional representation of proteins. The Sequence Channel utilises a pretrained protein language model, ESM-C, to extract contextual features from the primary amino acid sequence, generating high-dimensional sequence embeddings that capture residue-level information and their interdependencies within the full protein context. During the feature fusion stage, we process the outputs from these three channels—encoding complementary sequence and structural information, using a Multilayer Perceptron (MLP) network to generate confidence scores for GO terms. Empowered by effectively unifying both structural and sequential representations into a comprehensive, protein-wise embedding, the ENGINE framework significantly improves the accuracy and interpretability of protein function annotation.

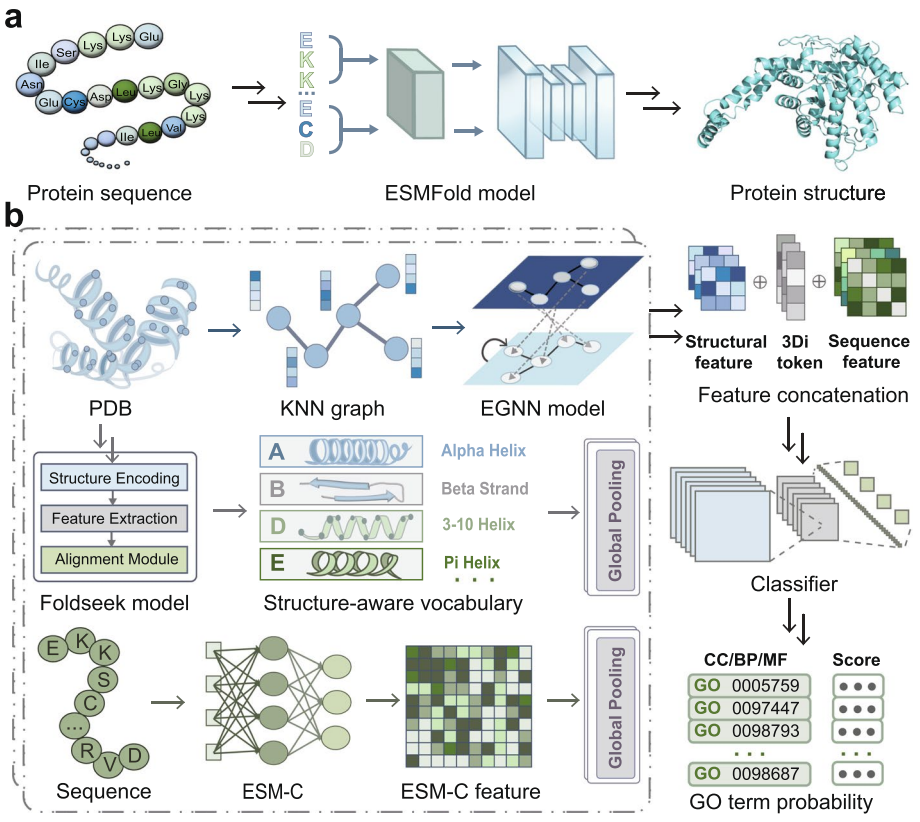

Fig. 1 An overview of the ENGINE framework. **a** Data Collection. Protein sequences in FASTA format are processed with ESMFold to generate predicted structures in PDB format. **b** The ENGINE model architecture. The framework consists of two main stages: (1) *Multi-channel feature extraction*, which leverages EGNN for structural representation learning, ESM-C for sequence embedding, and Foldseek for capturing 3Di token features; (2) *Feature fusion and classification*, where features from all channels are concatenated and passed through a classification layer to produce confidence scores for each GO term

ENGINE outperforms the state-of-the-art tools for protein function annotation

To comprehensively evaluate ENGINE's performance in protein function prediction, we benchmarked it against seven publicly available state-of-the-art methods on a benchmark set. This enabled a robust comparison of ENGINE's predictive capabilities with existing approaches. Among these approaches, FunFams [17] employed domain-based methods for protein function prediction, while DeepFRI [22] and TAWFN [23] employed Graph Convolutional Networks (GCN) to leverage graph structures for feature learning. TALE [20] applied sequence information through a Transformer-based architecture. DeepGOZero [25] and PFresGO [23] incorporated ontology-based embeddings to improve protein function prediction, whereas DeepGO [24] incorporated the hierarchical structure of the GO through an ontology-aware classification architecture. Performance was assessed using multiple evaluation metrics, including F_{max} (Maximum F-score), AUC (Area Under the ROC Curve), and AUPRC (Area Under the Precision-Recall Curve), by comparing the predicted protein functions with experimentally validated annotations. As summarised in Fig. 2a, ENGINE achieved competitive AUPRC scores across the three branches of GO terms: 0.4395 for MF, 0.3709 for BP, and 0.4496 for CC. While the MF score is slightly lower than that of DeepGOZero (0.6144), the overall results underscore ENGINE's strong performance on protein function annotation. Further analysis of the AUC metric reveals that ENGINE consistently

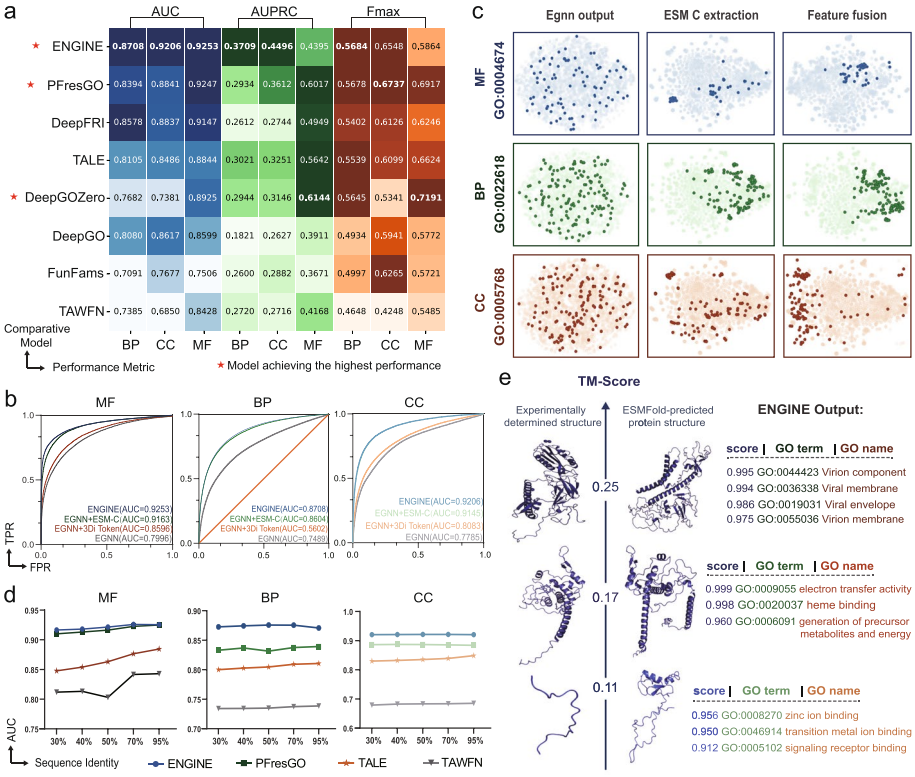

Fig. 2 Performance comparison of the ENGINE on different methods and designs. **a** Comparison of the performance of ENGINE with existing protein function prediction methods in terms of AUC, AUPRC, and F_{max} metrics. **b** Feature distribution visualisations using t-SNE for different layers in the ENGINE model. For each functional category, a representative GO term is selected for visualisation. Samples correctly annotated with the corresponding GO term (i.e., positive samples) are highlighted. As the feature fusion progresses across model layers, positive samples become increasingly clustered. **c** Comparison of AUC across different ENGINE variants with varying input configurations across MF, BP, and CC categories. **d** Performance comparison of AUC on varying sequence identity across MF, BP, and CC categories among different methods: ENGINE, TAWFN, TALE, and PFresGO. **(e)** Assessing structural similarity with TM-Score and corresponding ENGINE predictions

outperforms all baseline methods, achieving AUC scores of 0.9253, 0.8708, and 0.9206 for MF, BP, and CC, respectively. These predictions were significantly better than those achieved using the existing tools, demonstrating the overall superiority of the proposed approach. We showed that ENGINE achieved the best performance among all methods for the BP ontology and ranked within the top three for CC in terms of F_{max} metric. While DeepGOZero obtained the highest F_{max} score for the MF ontology prediction; PFresGO exhibited the best performance on the CC ontology. These findings highlight the effectiveness of ENGINE as a multi-channel integration framework capable of capturing multi-level protein information. Its robust generalisation across diverse function ontologies highlights its potential as a powerful tool for comprehensive protein function annotation.

To gain deeper insights into the factors underlying ENGINE's superior performance in protein function annotation, we conducted a series of ablation studies to assess the contribution of individual model components. Specifically, we systematically disabled key modules to construct several ENGINE variants and quantified their impact on overall model performance, thereby assessing the importance of each channel. The 'EGNN' variant includes only the structural channel, in which the 3D protein structure is converted into a graph representation, excluding both the sequence channel and the 3Di

sequence channel. The 'EGNN + ESM-C' variant retains the structural and sequence channels (with ESM-C embeddings) but removes the 3Di sequence channel. Conversely, the 'EGNN + 3Di Token' variant incorporates the structural and 3Di sequence channels, while omitting the ESM-C sequence embeddings. The experimental results shown in Additional file 2: Fig. S1 demonstrate that removing any channel leads to a significant drop in predictive performance, underscoring the critical role of each channel. These findings highlight the complementary nature of ENGINE's multi-channel architecture, with each component contributing synergistically to the model's overall effectiveness. The results also support the rationale for using ESM-C, Foldseek, and EGNN: ESM-C captures global sequence semantics, Foldseek provides structural semantic tokens, and EGNN precisely models 3D spatial relationships, with their combination significantly enhancing predictive performance.

As shown in Fig. 2b, the AUC comparison underscores the critical role of high-dimensional, residue-level sequence embeddings derived from the pretrained protein language model ESM-C in enhancing ENGINE's performance in protein function prediction. When only the structural and ESM-C-based sequence channels are retained, excluding the 3Di sequence channel, the AUC values for protein function predictions across the MF, BP, and CC ontologies remain relatively high, at 0.9163, 0.8604, and 0.9145, respectively, only slightly lower than those achieved by the original ENGINE model. This result underscores that protein sequence embeddings extracted from the pretrained protein language model ESM-C effectively capture informative contextual features within protein sequences, preserving essential functional signals for accurate function prediction. In addition, we compared different ESM model scales and sequence features. The results showed that ESM-C 6B consistently achieved the best performance, particularly in the BP and CC categories, and was therefore selected as the sequence feature extractor in our final framework. Moreover, Foldseek's 3Di token features outperformed ProteinBERT embeddings, offering more stable and structurally informative representations that complement sequence-based models (Additional file 2: Figs. S2 and S3). In summary, the ablation study confirms the strength of the proposed ENGINE's multi-channel integration strategy for protein function prediction. The structural, ESM-C-based sequential, and 3Di sequence channels contribute synergistically, enabling the model to capture function-relevant information comprehensively. Their combined use significantly enhances prediction accuracy and generalizability across diverse branches of protein function annotation. Additional ablation results and detailed analyses are provided in Additional file 2: Figs. S4 and S5.

ENGINE driven by the dynamic fusion of protein representations

ENGINE integrates three complementary feature extraction channels to capture multi-level protein features for function prediction. We demonstrate that protein representations in the latent space progressively evolve from an initially disordered distribution during the feature fusion process into well-separated clusters aligned with their functional categories. To gain deeper insights into the contribution of features from each channel and to examine how these features evolve throughout the fusion process to support improved function classification performance, we employed t-distributed stochastic neighbour embedding (t-SNE) to visualise feature distributions at different stages of the model across the three GO branches. Specifically, we selected representative GO

terms from the MF (GO:0004674), BP (GO:0022618), and CC (GO:0005768) ontologies, and extracted the protein features of EGNN from distinct stages: (1) the final layer of the EGNN module, (2) the sequence embeddings generated by ESM-C, and (3) the integrated feature representation at the fusion layer. As shown in Fig. 2c, protein feature embeddings in the pre-fusion stage display irregular and diffuse distributions, with samples scattered and lacking distinct cluster boundaries. In contrast, following feature fusion, positive samples from the same functional category become more tightly grouped, forming well-defined clusters. This transformation highlights the role of the fusion process in enhancing the discriminative power of the learned protein representations. These results suggest that integrating features from all three channels (structural, sequence, and 3Di information) prior to classification enables the construction of comprehensive and biologically meaningful protein representations, thereby facilitating accurate functional annotation. Moreover, this analysis further underscores the strength of ENGINE's multi-channel fusion strategy in capturing critical information from multimodal biological inputs, leading to more discriminative embeddings and improved prediction accuracy and generalisation across diverse functional ontologies.

ENGINE shows superior performance in annotating protein function with different sequence identities and structural similarity

It is crucial to understand how well protein function prediction models perform on sequences with varying similarity to known data to evaluate their generalisability. In this context, we initially evaluated ENGINE performance on protein sequences grouped by their levels of sequence identity to those in the training set. We focused particularly on novel sequences with low similarity to those in the training dataset. The test set was partitioned into five groups according to maximum sequence identity thresholds relative to the training set: 30%, 40%, 50%, 70%, and 95%. ENGINE was benchmarked against several baseline models, including TAWFN, TALE, and PFresGO, using AUC as the evaluation metric across these identity-defined subsets. As shown in Fig. 2d, AUC scores for all methods generally increased with higher sequence identity across the three GO categories. Notably, ENGINE outperformed all competing methods at every identity threshold. In particular, ENGINE retained a marked predictive advantage on sequences with low similarity to the training data, especially in regions of very low homology, such as the 30% and 40% identity groups. Its consistently high accuracy across all identity levels underscores ENGINE's strong generalisation capability.

To further verify ENGINE's ability to annotate proteins with low structural similarity, we conducted a structural similarity analysis. Specifically, we downloaded experimentally resolved structures of the test set proteins in bulk from the RCSB PDB database [31]. We used the TM-align tool [32] to compare each predicted structure to its corresponding true structure using the TM-score metric. The TM-score, which ranges from 0 to 1, is a widely used indicator of structural similarity; scores below 0.3 generally indicate significant differences in global folding.

To evaluate ENGINE's robustness under structural perturbations, we examined its performance on proteins with low TM-scores — scenarios in which accurate function prediction is particularly challenging. As shown in Fig. 2e, we selected three representative proteins from the test set, spanning different species and functional categories, to illustrate that ENGINE can still deliver reliable predictions even when structural

similarity is low. The first case, PDB ID 3DUZ (chain A) [33], originates from Autographa californica nucleopolyhedrovirus and corresponds to a domain of the viral envelope fusion protein GP64. Its predicted structure has a TM-score of only 0.25 when compared to the experimentally resolved structure, indicating a pronounced difference in overall conformation. Nonetheless, ENGINE successfully predicted several high-confidence Gene Ontology (GO) terms, including cellular component (GO:0044423), structural molecule activity (GO:0036338), viral process (GO:0019031), and protein transport (GO:0055036). These functions are closely aligned with the known roles of GP64 in viral entry, suggesting that ENGINE can capture essential functional features despite structural inaccuracies. The second case, PDB ID 6GIQ (chain D) [34], is a component of the mitochondrial respiratory supercomplex III_2IV from Saccharomyces cerevisiae, with an even lower TM-score of 0.17. Despite this extreme structural discrepancy, ENGINE correctly predicted several key functions, such as electron transfer activity (GO:0009055), heme binding (GO:0020037), and energy metabolic process (GO:0006091). These functions are typically associated with evolutionarily conserved sequence motifs and local spatial environments, further supporting ENGINE's capability to integrate multimodal inputs and capture signals relevant to functional essence. The third case, PDB ID 4A0K(chain B) [35], comes from the human DNA repair complex DDB1-DDB2-CUL4A-RBX1. It exhibits the lowest TM-score among the examples (0.11), representing a scenario with the most pronounced structural deviation. Remarkably, ENGINE accurately identified metal ion binding (GO:0046914) and zinc ion binding (GO:0008270), and additionally predicted a previously unannotated function, signalling receptor binding (GO:0005102), demonstrating the model's potential in novel function discovery. These case studies collectively confirm that ENGINE is capable of producing high-confidence GO term predictions even for proteins with low structural similarity. This underscores the model's robustness to structural errors and its effectiveness in function annotation under challenging conditions.

Building upon these case studies, we further stratified the entire test set based on TM-score into four categories: random similarity (TM-score < 0.3), low similarity (0.3 ≤ TM-score < 0.5), medium similarity (0.5 ≤ TM-score < 0.8), and high similarity (TM-score ≥ 0.8), and computed performance metrics for each category. As shown in Additional file 3: Table S2, ENGINE maintained robust predictive performance even for proteins in the random similarity and low similarity groups, with only modest decreases compared to higher similarity groups. This indicates that ENGINE can effectively capture functional signals even under substantial structural deviations, highlighting its broad applicability and robustness for protein function annotation across structurally divergent proteins.

ENGINE uncovers functional substructures essential for protein activity

In this study, we conducted an interpretability analysis of the protein annotation model using GNNExplainer [36]. GNNExplainer is a model-agnostic interpretability method designed to identify the most influential subgraphs, node features, or edge attributes contributing to a graph neural network's prediction. Specifically, we computed importance scores for eight distinct protein secondary structure types (as detailed in Table 1) to quantify their contributions to protein function prediction. By averaging

Table 1 Mapping of DSSP Secondary Structure Types to 8-Dimensional Vectors

Secondary structure type	Description	Numerical vector
H	Alpha helix	[1, 0, 0, 0, 0, 0, 0, 0]
B	Beta bridge	[0, 1, 0, 0, 0, 0, 0, 0]
E	Extended strand	[0, 0, 1, 0, 0, 0, 0, 0]
G	3–10 helix	[0, 0, 0, 1, 0, 0, 0, 0]
I	Pi helix	[0, 0, 0, 0, 1, 0, 0, 0]
T	Turn	[0, 0, 0, 0, 0, 1, 0, 0]
S	Bend	[0, 0, 0, 0, 0, 0, 1, 0]
-	Loop or irregular	[0, 0, 0, 0, 0, 0, 0, 1]

these importance scores across proteins in the dataset, we derived a global ranking of secondary structure types, enabling us to assess their relative influence on the model's predictions.

Figure 3a demonstrates that the Loop regions played an essential role in protein function. Their inherent flexibility and active involvement in molecular interactions drive their contribution to protein functional activity [37–39]. Loops are also commonly enriched in functionally important residues and mutation hotspots, further supporting their significance in predictive modelling [40]. Turn structures ranked second, reflecting their role in maintaining protein stability and their facilitation of conformational transitions essential for function [41–43]. The Alpha helices structure followed closely, in line with their well-established roles in molecular recognition, structural integrity, and catalytic activity [44, 45]. In contrast, the Extended strands showed a moderate importance score, aligning with their primarily structural, rather than functional roles [46, 47]. Bend structures and Beta bridge structures were associated with lower importance, suggesting a limited role primarily in maintaining structural organisation [48]. Finally, the 3–10 helix and Pi helix showed limited functional relevance, likely attributable to their rarity [49–51]. These results offer valuable insights into the structural underpinnings of protein function. To further elucidate these relationships, we analysed the key substructures in-depth and examined their potential associations with known functional domains.

We identify functionally important substructures in three-dimensional space by computing edge importance scores and locating the nodes associated with high-scoring edges. These substructures are defined as subgraphs composed of the selected nodes and their connecting edges, representing local structural fragments with high functional relevance. In contrast, functional domains refer to annotated functional units, which may encompass multiple such substructures [52–54]. To further characterise the identified substructures, BLAST comparisons were performed, enabling the detection of analogous sequences within existing protein databases. Functionally relevant substructures identified in the test set were ranked by frequency, are shown in Fig. 3b, with the HHHHH substructure appearing most frequently. We also visualised the corresponding protein structures with a detailed analysis of these most frequently occurring substructures. As illustrated in Fig. 3c for PDB ID 5J13 (chain A), the HHHHH substructure spans amino acid sites 4 to 8 of the protein [55]. A BLAST search using its amino acid sequence revealed that residues 1 to 24 in this protein are annotated in the Gene3D database as part of the acetylglutamate kinase-like superfamily [56]. Acetylglutamate kinase plays a central role in metabolic regulation, particularly in energy and amino acid metabolism [57]. As shown in Fig. 3d, amino acid residues 4 through 9 of this protein form the SHHHHHH substructure, which is also commonly observed. This suggests

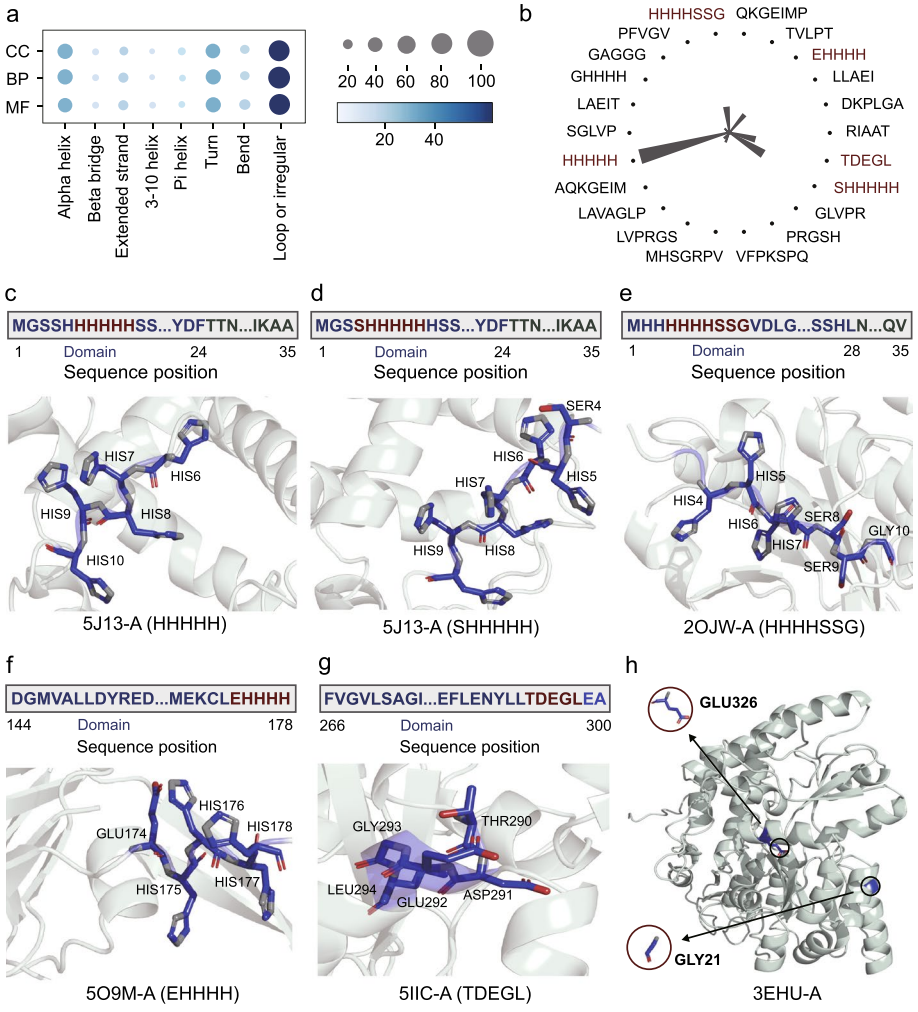

Fig. 3 Analysis of the importance of secondary structure features and substructures. **a** Contributions of different secondary structure types to protein function prediction. The graph shows the average importance score of secondary structures, with colour intensity and circle size indicating relative importance. **b** Frequency of functionally relevant substructures within protein structural domains. **c**-**g** Representative examples and atomograms of proteins associated with the top-ranked substructures. The blue intervals represent structural domains, and the red intervals indicate the position of the substructure on the structural domain. **h** An example of a protein mutated in an important substructure. The arrow points to the mutation sites

that the HHHHH and SHHHHH substructures may contribute to the functional or mechanistic properties of this enzyme family. Figure 3e presents the human flavoprotein enzyme (PDB ID: 2OJW, chain A), in which residues 1–28 form a region critical for binding flavin cofactors, such as FMN and FAD[58]. The substructure HHHHSSG, which our model identified, is located precisely within this region, suggesting its potential involvement in cofactor binding and enzymatic function. Figure 3f shows PDB ID: 5O9M (chain A), in which the important substructure EHHHH lies within the TCTP domain and is associated with functions such as apoptosis and tumourigenesis [59]. Figure 3g displays the *Escherichia coli* maltose-binding periplasmic protein (PDB ID: 5IIC, chain A), a key component of the ABC transport system responsible for substrate recognition and delivery [60]. The substructure TDEGL, identified by our model as highly important, is located within residues 277–298, annotated by the PRINTS database as the Maltose/Cyclodextrin ABC transporter substrate-binding domain. Structural studies

confirm that this domain forms the core of the maltose-binding pocket [61], highlighting its direct role in substrate-specific recognition. Overall, by identifying frequent and functionally important substructures and aligning them with known protein families, we offer potential functional insights into complex or previously uncharacterised proteins [62].

Building on this, we further focused on potential mutation events occurring within important substructures. Figure 3h shows human galactokinase (PDB ID: 3EHU, chain A), which is one of the key enzymes in the galactose metabolism pathway [63]. We identified functionally relevant substructures and located two mutation sites—GLU326 and GLY21—within these critical regions. The GLU326 is situated in the ATP-binding region at the C-terminus; its side-chain carboxyl group coordinates with the catalytic Mg^{2+} in, and mutations at this site may directly impair catalytic efficiency [64]. The GLY21 is located within a loop structure in the substrate-binding region; its substitution may induce local conformational alterations, potentially impairing galactose binding [65]. These findings suggest that such mutations may potentially disrupt the function of GALK1 and could have biomedical implications for metabolic disorders such as galactosemia. By integrating known mutation events with key substructures identified by our model, our functional annotation framework offers a novel approach for predicting the effects of protein variants and investigating disease mechanisms.

ENGINE locates key residues underlying protein function

Protein function is governed not only by its amino acid sequence but also by the specific spatial arrangement of key residues. Functionally important residues often cluster in three-dimensional space to form active or binding sites essential for biological activity [66]. In the EGNN model, each amino acid is represented as a node in a graph, allowing us to evaluate the contribution of individual residues to function prediction. To do this, we compute attention weights for edges connecting node i to node j, quantifying the information passed between them. The functional importance of a residue is then defined as the sum of attention weights across all its connected edges, including both incoming and outgoing interactions.

$$\mathrm{Attention} = \sigma(\mathrm{weight}(\mathrm{edge\_mlp}(\mathrm{x_i}, \mathrm{x_j}, \mathrm{edge\_attr})))$$

$$\mathrm{N} = \mathrm{Sum}(\mathrm{Attention}, \mathrm{edge\_index}, \mathrm{node\_num})$$

Figure 4a and b demonstrate the residues related to 'fructose-6-phosphate binding' (GO:0070095) identified by ENGINE in *Escherichia coli* 6-phosphofructokinase (PDB: 1PFK, chain A). In Fig. 4a, black asterisks mark the functional residues associated with GO:0070095 retrieved from the BioLiP database [67], while circles represent the importance score calculated by ENGINE across residues. This alignment indicates ENGINE's ability to recover most of the key residues involved in fructose-6-phosphate binding. Figure 4c further illustrates ENGINE's ability to accurately identify residues involved in "GTP binding" (GO:0005525) for GTP Cyclohydrolase I (PDB: 1A9C, chain A), again confirming its precision in detecting functional sites. Figure 4d features triose-phosphate isomerase (TIM) (PDB: 1KV5, chain A) from Trypanosoma brucei [68], annotated with the catalytic function "triose-phosphate isomerase activity" (GO: 0004807). ENGINE successfully identifies catalytic residues associated with this catalytic function

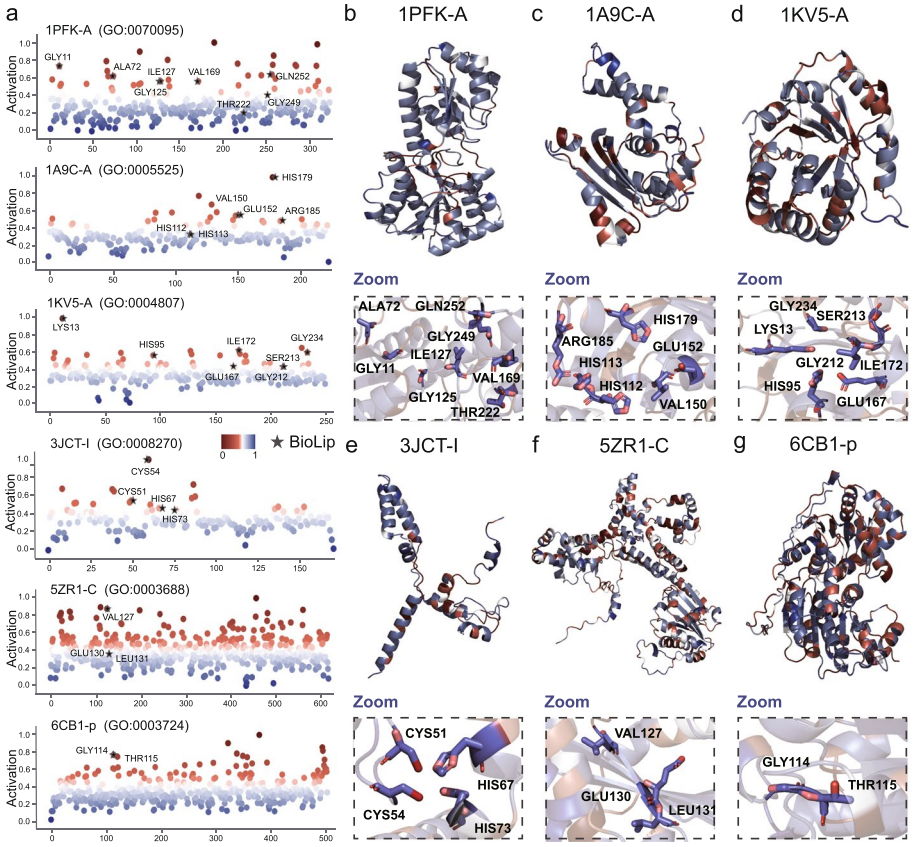

Fig. 4 ENGINE locates functional residues based on residue-level attention weights. **a** The functional site prediction results for six representative proteins (PDB: 1PFK, Chain A; PDB: 1A9C, Chain A; PDB: 1KV5, Chain A; PDB: 3JCT, Chain I; PDB: 5ZR1, Chain C; PDB: 6CB1, Chain P), each associated with distinct functional annotations (GO:0070095; GO:0005525; GO:0004807; GO:0008270; GO:0003688; GO:0003724). In the residue functional site score plots, black stars indicate experimentally annotated functional residues obtained from the BioLiP database, and red, gray, and blue circles represent the attention-based scores computed by ENGINE, illustrating the variation in residue-level functional relevance across the sequence. **b**-**g** Tertiary structures for six representative proteins, with structural renderings highlighting residue-level attention weights derived from ENGINE

in TIM. Figure 4e originates from *Saccharomyces cerevisiae* (PDB: 3JCT, chain I) and is annotated with the function of metal ion binding (GO: 0008270) [69]. The high-score region predicted by our model overlaps with known metal-coordinating residues, demonstrating strong sensitivity to local functional sites. Moreover, Fig. 4f shows the Origin Recognition Complex Subunit 3 (ORC3) (PDB: 5ZR1, chain C) [70], where our model's prediction aligns well with the experimentally characterised DNA-binding interface (GO: 0003688). Figure 4g presents the ribosome biogenesis protein RLP7 (PDB: 6CB1, chain P) [71], in which ENGINE accurately identifies residues involved in "RNA binding" (GO:0003724). Together, these examples highlight that ENGINE effectively identifies residues relevant to distinct biological activities by quantitatively estimating the functional importance of individual amino acids, offering valuable insights into the structural basis of protein function and enabling residue-level functional annotation.

ENGINE uncovers unannotated protein functions and accurately predicts rare GO terms

To further highlight the utility of ENGINE in protein function prediction, we demonstrate that ENGINE not only consistently produces high-confidence predictions for known functional annotations but also identifies Gene Ontology (GO) terms absent from

current ground truth datasets, underscoring its potential to uncover previously unannotated biological functions. A compelling example is Secale cereale chitinase GH19 [72], ENGINE, which not only accurately identified the annotated ground truth GO terms but also assigned high confidence to GO:0005615 (extracellular space), as indicated in Fig. 5a. While this term is currently absent from its curated annotations, prior research has indeed reported the involvement of *Secale cereale* chitinase GH19 in extracellular processes [73]. As shown in Fig. 5b, in the case of the pseudouridine synthase RluD from *Escherichia coli*, ENGINE confidently identified two previously unannotated functions: GO:0006399 (tRNA metabolic process) and GO:0008033 (tRNA processing). Although these GO terms are not present in the current ground-truth annotations, it is noteworthy that RluD belongs to the broader pseudouridine synthase family. Structurally, RluD shares significant homology with known tRNA-modifying enzymes such as RluA and TruA, and several members of this enzyme family have been experimentally validated to act on tRNA [74].

In Fig. 5c, the Ferric Enterobactin Receptor (PDB: 1FEP, Chain A) is assigned the GO term GO:0098796 (integrin transport) by the model with notably high confidence [75]. Although this function has not yet appeared in its existing ground-truth annotations, the prediction is biologically plausible: the Ferric Enterobactin Receptor is an outer membrane protein involved in iron uptake in Gram-negative bacteria, particularly in the binding and translocation of siderophores such as enterobactin. Its function entails highly specific ligand recognition and energy-dependent translocation mechanisms, which

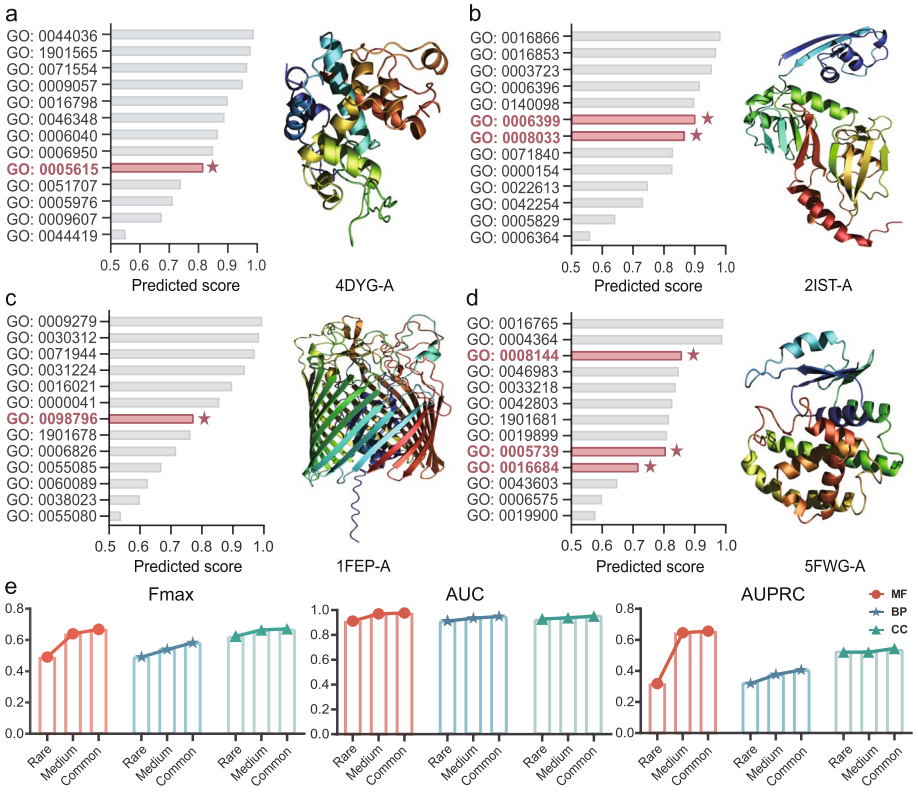

Fig. 5 GO term analysis predicted by ENGINE and identification of rare functions. **a**-**d** Solid lines represent GO terms that are present in the ground truth annotations, whereas asterisks indicate high-confidence GO terms predicted by ENGINE that are not included in the ground truth. **e** F_{max}, AUC, and AUPRC performance of ENGINE on Rare, Medium, and Common GO term frequency levels across MF, BP, and CC categories

significantly overlap with the molecular transport processes described by GO:0098796. Similarly, in Fig. 5d, we further examined a genetically engineered mutant of glutathione S-transferase GSTM1 (PDB: 5FWG, Chain A) [76]. ENGINE not only identified its inherent catalytic activity associated with glutathione but also confidently predicted several GO terms potentially reflecting novel functionalities, including GO:0008144 (drug binding), GO:0005739 (mitochondrion), and GO:0016684 (oxidoreductase activity). These results uncover ENGINE's ability to generate biologically meaningful predictions that align with experimental evidence, underscoring its potential to support large-scale computational annotation of proteins with unknown functions.

In view of the results obtained above, we further explore the ability of ENGINE in identifying rare GO terms to evaluate its robustness and generalisation performance in solving the data sparsity problem. Specifically, we stratify the labels based on the frequency of occurrence of each GO term in the training set. GO terms are classified into three categories based on their frequency of occurrence: those appearing fewer than 50 times are defined as 'Rare', those with 50 to 100 occurrences as 'Medium', and those with more than 100 occurrences as 'Common'. This classification reflects the natural distribution pattern of the labels. The distribution of these categories is shown in Additional file 2: Fig. S6.

In terms of evaluation metrics, we used F_{max}, AUC and AUPRC to measure the performance of ENGINE at each frequency level, respectively, and the results are shown in Fig. 5e. It is evident from the analysis of both the BP and CC functional ontologies that the ENGINE demonstrates enhanced consistency in all three metrics across the rare, medium and common frequency intervals. Even within the rare category, no significant decline in performance is observed. In the MF ontology, although F_{max} and AUPRC metrics are slightly lower for rare GO terms compared to medium and common categories, the overall differences are minimal, and the AUC remains nearly unchanged. This demonstrates that ENGINE maintains strong robustness in identifying low-frequency functions. The consistent accuracy of ENGINE on rare GO terms underscores its promise in addressing the long-tailed distribution challenge inherent in protein function annotation, enabling biologically meaningful and generalised predictions even for GO terms with limited training data.

Discussion

Although computational approaches have achieved notable progress in protein function prediction, their performance is often limited when handling proteins with low homology or incomplete structures. ENGINE innovatively converts protein 3D structures into 3Di token sequences. Even for proteins with low sequence homology, low structural similarity, and rare GO terms, the model maintains high predictive capability. This robustness is primarily attributed to the complementary nature of the multi-channel architecture, where each component acts synergistically. BLAST alignment results validate that ENGINE can identify biologically significant key substructures of known proteins, implying that ENGINE can provide potential functional insights for unannotated proteins. Concurrently, the incorporation of attention mechanisms enables ENGINE to accurately identify key residues involved in diverse biological activities, achieving functional annotation at the residue level.

It should also be noted that there are limitations in this study. Firstly, the input provided by ENGINE is dependent on protein structural information, with longer proteins requiring greater computational resources. Secondly, the present framework does not fully leverage evolutionary information, such as conserved sites in multiple sequence alignments. Future work could integrate these evolutionary signals to deepen our understanding of functional conservation and the evolutionary relationships between protein functions across species.

Conclusions

We present a multi-channel deep learning model, ENGINE, that integrates protein 3D structural data and sequence information to annotate protein function. ENGINE combines a structural channel, which utilises the Equivariant Graph Neural Networks to learn graph-based features from protein structures, with a sequence channel that leverages the large protein language model ESM-C embedding to capture evolutionary and contextual information from protein sequences. Additionally, ENGINE incorporates an innovative structural encoding strategy by integrating Foldseek to extract 3Di tokens from protein structure data, enabling the modelling of spatial contextual relationships between residues. These complementary multi-source features are effectively fused to enhance protein function prediction.

Compared to current state-of-the-art models, ENGINE demonstrates superior performance in protein function prediction. Through extensive ablation studies and feature visualisation analyses, we demonstrate that the multi-modal biological features captured by the proposed multi-channel fusion strategy enhance ENGINE's feature representation capabilities. Furthermore, we conduct multiple interpretability analyses, offering deeper insights into functional annotation results and enabling the effective identification of functionally relevant residues within protein sequences. To facilitate protein function annotation research, we have open-sourced ENGINE and provided a graphical user interface (GUI) at https://github.com/ABILiLab/ENGINE, making it easy for researchers to reproduce and build upon our work. We envision ENGINE as a powerful and practical tool for protein function prediction, with broad applications in protein function characterisation, target discovery, drug discovery, and related areas.

Methods

Data collection

We adopted the dataset from the DeepFRI study, which contains 36,641 protein sequences with protein function annotations across three ontologies [22]. To ensure high data quality and reduce redundancy, the CD-HIT tool [77] was employed with a 95% sequence identity threshold to eliminate similar sequences, thereby avoiding overlap between the training and test sets. Furthermore, all proteins included in the test set were required to have at least one experimentally validated GO term, ensuring the credibility of the data.

Subsequently, we used the ESMFold model to predict the tertiary structures of the protein sequences, which were then saved in PDB format files, as shown in Fig. 1a. The dataset was divided into training, validation, and test sets (Table 2) with the proportions of 80%, 10%, and 10%, respectively, covering a total of 2,752 GO terms: 489 from

Table 2 Statistical summary of the training and test datasets in this study

Data	MF	BP	CC
Train	29,902	29,902	29,902
Validation	3,323	3,323	3,323
Test	3,416	3,416	3,416
GO terms	489	1943	320

Molecular Function (MF), 1,943 from Biological Process (BP), and 320 from Cellular Component (CC).

Protein representations

To effectively capture the spatial arrangement, structural conformation, and functional characteristics of proteins, we represent the targeted protein as a graph in which amino acids serve as nodes and the edges quantify connectivity of nodes/residues. Specifically, we construct the protein graph, denoted as $\mathrm{G} = (\mathrm{X},\ \mathrm{A},\ \mathrm{E})$, utilising the KNN algorithm based on the 3D coordinates of residues. In this representation, X represents node features, A is the adjacency matrix reflecting residue connectivity, and E denotes edge attributes. This graph-based structure efficiently encodes the complex topological organisation of protein residues, thereby facilitating downstream tasks such as functional prediction. We employ the KNN algorithm to construct the protein graph because it efficiently captures the local spatial environment of each residue while preserving the geometric topology of the protein, providing structured and meaningful input for deep learning models.

Node representations

In this study, we enhance protein function prediction by incorporating secondary structure features as node attributes within Equivariant Graph Neural Networks. Secondary structures, such as alpha helices, beta bridges, and loops, are formed through local folding patterns stabilised by hydrogen bonds and other non-covalent interactions. These structural patterns are crucial for maintaining protein stability and activity, as well as mediating biological processes, making their accurate representation essential for functional annotation tasks (see Table 2).

We employ the Dictionary of Secondary Structure of Proteins (DSSP) algorithm to systematically capture protein secondary structure information, which assigns a secondary structure label to each amino acid residue [78]. These categorical labels are then transformed into 8-dimensional one-hot encoded vectors, where each dimension corresponds to a specific structural class. By embedding this structural context into the feature representation of each residue, our model gains deeper insight into local conformational patterns, ultimately improving the precision of function prediction.

Edge representations

Each edge in the graph denotes an interaction between amino acids, with the edge feature matrix E encoding pairwise residue relationships. The edge features encode various essential structural and topological information, including 15 high-dimensional distance metrics, 12 relative spatial positions, and 66 relative sequence distances, collectively capturing structural and topological information [79].

For every connected node pair (*i*, *j*), their pairwise distance is transformed using a Gaussian radial basis function (RBF) to generate 15 distance-based attributes:

$$\mathrm{E}_{\mathrm{r}}^{\mathrm{rbf}}(\mathrm{x_i}, \mathrm{x_j}) = \exp(-\frac{||\mathrm{x_j} - \mathrm{x_i}||^2}{2\sigma_{\mathrm{r}}^2})$$

where σ_r denotes the scale parameter for each distance-based feature.

Additionally, the 12 relative spatial positions are derived from the heavy-atom coordinates of residues, captured fine-grained local interactions [80]. To further encode sequence order, 66 binary features are computed based on the relative sequential positions d(i,j) =|s_i-s_j|, where s_i and s_j denote the absolute positions of residues i and j within the primary sequence [81]. This comprehensive encoding scheme effectively preserves both local structural interactions al relationships in the graph representation.

Graph construction

To construct the KNN-based protein graph, each amino acid (node) is connected to up to K nearest neighbours in Euclidean space, provided their pairwise distance falls below a predefined threshold. This approach preserves local structural connectivity, allowing the graph to capture the microenvironment of each residue effectively. By integrating geometric and sequential information, this graph representation provides a rich and structured input for DL models in protein function prediction.

E(n)-equivariant graph neural network

Proteins possess complex 3D structures that are not fixed in position or orientation within a coordinate system. Instead, they can undergo arbitrary translations, rotations, or reflections. To ensure robust protein function prediction, it is essential that the model exhibits SE(3) equivariance—that is, the feature representation remains consistent under spatial transformations such as translation, rotation, and reflection [82]. However, traditional Graph Neural Networks (GNNs), which operate primarily on topological relationships in non-Euclidean spaces, inherently lack such symmetry-awareness [83, 84]. Therefore, we use the EGNN as the structural channel backbone, as it directly leverages 3D coordinates to capture spatial relationships among residues while preserving SE(3) equivariance. This ensures that the learned representations remain consistent under spatial transformations, which is crucial for protein function prediction since biological activity often depends on relative rather than absolute spatial arrangements. To address this, we construct a 3D graph that adopts SE(3) equivariant graph convolution layers (EGC) for feature learning. This 3D graph enables the model to capture detailed structural characteristics of proteins while preserving geometric equivariance.

Within each EGC layer, we perform three core operations. First, we calculate the edge-wise feature embeddings via a message passing mechanism, allowing the model to aggregate information from neighbouring nodes:

$$\mathrm{m_{ij}} = \varnothing_{\mathrm{e}}(\mathrm{h_i^l}, \mathrm{h_j^l}, ||\mathrm{x_i^l} - \mathrm{x_j^l}||^2, \mathrm{e_{ij}})$$

Here, $\mathrm{h_i^l}$ and $\mathrm{h_j^l}$ represent the hidden features of neighbouring nodes i and j, respectively. The term $||\mathrm{x_i^l} - \mathrm{x_j^l}||^2$ refers to the squared Euclidean distance between nodes i and j. $\mathrm{e_{ij}}$ denotes the edge attributes, and $\varnothing_{\mathrm{e}}$ is a multi-layer perceptron (MLP) used to compute

the edge features. The MLP is trained to fuse the features of neighbouring nodes with spatial cues to generate enhanced edge representations. This approach enables the construction of richer edge features, allowing nodes to more effectively aggregate information from their neighbours. These features encompass not only topological information but also geometric distance relationships, enabling the network to capture structural variations.

In the next step, we update the 3D coordinates of each node, enabling the network to capture spatial relationships within the local neighbourhood:

$$\mathrm{x_i^{l+1}} = \mathrm{x_i^l} + \frac{1}{\mathrm{n}} \sum_{\mathrm{j} \neq \mathrm{i}} (\mathrm{x_i^l} - \mathrm{x_j^l}) \cdot \varnothing_{\mathrm{x}}(\mathrm{m_{ij}})$$

where $\varnothing_{\mathrm{x}}(\mathrm{m_{ij}})$ represents the MLP that maps the edge feature $\mathrm{m_{ij}}$ to scalar weights. The term $\mathrm{x_i^l} - \mathrm{x_j^l}$ denotes the relative position vector between the node i and its neighbouring node j. The product $(\mathrm{x_i^l} - \mathrm{x_j^l}) \cdot \varnothing_{\mathrm{x}}(\mathrm{m_{ij}})$ is used to update $\mathrm{x_i^l}$ based on the information from neighbouring nodes. This allows for dynamic adjustment of the node coordinates, optimising their relative positions in spatial space. Consequently, this enables the network to effectively encode spatial relationships within the 3D structure, rather than relying solely on topological connectivity.

With the edge messages and coordinates updated, we then update the hidden features of the node , allowing it to aggregate information from its neighbouring nodes effectively:

$$\mathrm{h_i^{l+1}} = \varnothing_{\mathrm{h}}(\mathrm{h_i^l}, \sum_{\mathrm{j} \neq \mathrm{i}} \mathrm{m_{ij}})$$

The term $\sum_{\mathrm{j} \neq \mathrm{i}} \mathrm{m_{ij}}$ aggregates the edge feature information transmitted from all neighbouring nodes. The function $\varnothing_{\mathrm{h}}$ is a neural network that updates node features through nonlinear transformations, enhancing the expressive power of the model.

The EGC layer maintains both rotational and translational equivariance of 3D node coordinates, while being permutation-invariant to node ordering, ensuring that predictions remain unaffected by the input sequence. These properties make it well-suited for modelling the 3D structure of proteins.

Embedding of 3Di tokens extracted from Foldseek

To incorporate spatial contextual information into sequence-based protein function annotation, we introduce an innovative structural encoding strategy by employing Foldseek to extract 3Di tokens from protein structure data. Foldseek is an efficient protein structure alignment tool that encodes 3D protein conformations into a sequence-like format. We adopt Foldseek as it robustly captures both local and global 3D residue information, preserving key structural features while integrating them with sequence-based representations. This representation effectively maps the spatial context of the protein onto a one-dimensional token sequence. Specifically, for each amino acid residue, both its local and distal spatial environment are captured and discretised into a structural alphabet (SA), thereby yielding a transformed sequence that encapsulates the underlying 3D structural characteristics:

$$\mathrm{S_{3Di}} = (\mathrm{t_1}, \mathrm{t_2}, ..., \mathrm{t_n})$$

where t_i represents the Foldseek-encoded token for residue i, encapsulating the 3D environmental information of the amino acid.

The core of this transformation is a predefined structural alphabet (SA), comprising discrete symbols that represent distinct local 3D environments. For each residue, Foldseek assigns a unique token t_i, on its spatial configuration, resulting in a 3Di token sequence that serves as a compact and informative structural fingerprint of the protein. To make the 3Di token sequence amenable to DL models, we employ an embedding layer that projects the discrete tokens into a high-dimensional space. Specifically, each token is mapped to a 128-dimensional embedding vector during the model's forward pass, enabling the model to learn spatially aware representations for downstream functional annotation tasks.

ESM-C protein representations

Accurate representation and interpretation of the biological characteristics encoded in protein sequences are fundamental for annotating protein function. To this end, we employed the ESM-C model for protein sequence feature extraction, as it efficiently captures residue-level contextual information, evolutionary signals, and physicochemical properties, producing rich embeddings that are highly informative for protein function prediction. The ESM-C model was developed to overcome the computational efficiency and memory limitations of ESM-2, and it serves as a parallel architecture to ESM-3 [85]. While ESM-3 is primarily geared toward controllable protein generation, ESM-C is specifically designed to learn biologically meaningful representations of presentability. The 6B-parameter version of ESM-C offers substantial improvements in computational efficiency over the 3B-parameter ESM-2 model.

The ESM-C 6B model takes a protein's amino acid sequence as input and generates contextualised embedding vectors for each residue, given a sequence $S = (s_1, s_2, ..., s_L)$ of length L, where s_i denotes a discrete amino acid. This sequence is first mapped into a lower-dimensional space via an embedding layer, and subsequently processed by multiple transformer layers to model contextual dependencies. For each input protein sequence S, the ESM-C 6B model produces an output matrix of shape $L \cdot 2560$, where 2560 is the dimensionality of the residue-level embeddings. These embeddings encapsulate evolutionary signals, residue interactions, and physicochemical properties, serving as informative features for downstream tasks such as protein function annotation.

Model training and evaluation

The ENGINE model was implemented in PyTorch and optimised using the Adam optimiser [86] with a learning rate of 0.0001. The learning rate was selected to promote stable convergence and mitigate the risk of overshooting the optimal solution [87, 88]. The Adam optimiser adapts the learning rate and optimises model parameters by considering both the mean and variance of the gradients, ensuring an efficient and stable training process [89]. The training objective was defined using the cross-entropy loss function, which is well-suited for multi-class classification tasks, particularly in the presence of class imbalance or uneven sample distributions. It quantifies the discrepancy between the predicted class probabilities and the ground truth labels. To further reduce the risk of overfitting, validation sets and dropout regularisation were employed during training [90–92]. Model training was conducted on a Linux server with a 32-core CPU, 128 GB

RAM, and two NVIDIA GeForce RTX 3090 Ti GPUs, each with 24 GB of dedicated memory. The configuration and usage guidelines for the ENGINE model are provided in Additional file 4.

In this study, we employ three commonly used evaluation metrics to assess model performance: F_{max}, the area under the precision-recall curve (AUPRC), and the Area Under the ROC Curve(AUC) [93]. Among these, AUPRC is well-suited for imbalanced classification tasks, as it emphasises the penalisation of false positive predictions. AUC quantifies the area under the ROC curve and measures the model's ability to distinguish between positive and negative classes, such as determining whether a protein is associated with a specific GO term. F_{max} is an official evaluation metric adopted by the Critical Assessment of Functional Annotation (CAFA) benchmarking initiative [94], and is widely used to evaluate the overall performance of protein function annotation models. It represents the maximum F-score computed across all possible decision thresholds, thereby reflecting the model's best trade-off between precision and recall under varying confidence levels. The mathematical formulation of F_{max} is given as follows:

$$F_{max} = {}_t{}^{max}\left(\frac{2 \times pr(t) \times rc(t)}{pr(t) + rc(t)}\right)$$

where $pr(t)$ and $rc(t)$ denote the precision and recall at a given threshold t, respectively, and are defined as follows:

$$pr(t) = \frac{1}{h(t)} \sum_{j=1}^{h(t)} \frac{\sum_i 1(S(G_i, P_j) \geq t) \times I(G_i, P_j)}{\sum_i 1(S(G_i, P_j) \geq t)}$$

$$rc(t) = \frac{1}{h(t)} \sum_{N_T}^{h(t)} \frac{\sum_i 1(S(G_i, P_j) \geq t) \times I(G_i, P_j)}{\sum_i I(G_i, P_j)}$$

where $h(t)$ represents the number of proteins with at least one GO term assigned a score no smaller than t. The prediction score between protein P_j and GO term G_i, denoted by $S(G_i, P_j)$, is generated by the model and used for ranking or thresholding. $I(G_i, P_j)$ is a binary indicator, which denotes that protein P_j is annotated with the GO term G_i.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03886-y.

Additional file 1. Overview and Summary of the Methods in Table S1. Summary of reviewed predictors for protein function annotation, including feature encoding schemes, algorithms, evaluation metrics, and input types.

Additional file 2. Design and analysis of ablation experiments and dataset.Design and analysis of ablation experiments, including classifier comparison, ESM model impact, sequence feature evaluation, and an analysis of the GO term frequency distribution in the training dataset. Fig S1. Performance comparison of ENGINE variants with ablated input channels. Fig S2. Performance comparison of different ESM models. Fig S3. Performance comparison of different sequence features. Fig S4. Performance comparison of different classifiers. Fig S5. Comparison of two channel design strategies for incorporating ESM-C features. Fig S6. Distribution of GO term frequencies stratified by category in the training set.

Additional file 3. Performance Evaluation of ENGINE Across TM-score Structural Similarity Groups. Table S2 summarizes ENGINE's prediction performance on proteins stratified by structural similarity based on their TM-scores.

Additional file 4. Default parameters and usage guidelines of the ENGINE model. This file provides detailed documentation on the ENGINE model's configuration, including the default prediction threshold, batch size, learning rate, and the integration strategy of its multi-channel architecture, offering practical guidance for users.

Acknowledgements
This work was supported by the National Natural Science Foundation of China [62202388]; the National Key Research and Development Program of China [2022YFF1000100]; the Qin Chuangyuan Innovation and Entrepreneurship Talent Project [QCYRCXM-2022–230]; and the Australian National Health and Medical Research Council [2041439].

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
F.L. and J.S. conceived the ideas. Z.R., X.G., and T.P. designed the experiments. Z.R. and X.G. performed the experiments. Z.R., X.G., Y.B., and H.S. analysed the data and prepared figures. Z.R., X.G., and H.S. developed the webserver and software. Z.R., F.L., and J.S. wrote the manuscript. Y.B. and Y.H. provided guidance on data analyses. F.L. and J.S. supervised the project. All authors contributed ideas to the work and assisted in manuscript editing and revision.

Data availability
The protein sequence data used for training and testing were obtained from (https://doi.org/10.5281/zenodo.4650027) [22]. ENGINE is available under the MIT license in the GitHub repository (https://github.com/ABILiLab/ENGINE) [95], including model weights and source code. The source code is also available in Zenodo (https://doi.org/10.5281/zenodo.17221153) [96].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 June 2025 / Accepted: 24 November 2025
Published online: 29 November 2025

References

1. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
2. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell. 1998;92(3):291–4.
3. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(4096):223–30.
4. Pan T, et al. SCREEN: a graph-based contrastive learning tool to infer catalytic residues and assess enzyme mutations. Genomics Proteomics Bioinformatics. 2025;22(6):qzae094. https://doi.org/10.1093/gpbjnl/qzae094.
5. Radivojac P, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–7.
6. Paananen J, Fortino V. An omics perspective on drug target discovery platforms. Brief Bioinform. 2020;21(6):1937–53.
7. Zhang C, Freddolino L, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res. 2017;45(W1):W291-9.
8. Ashburner M, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
9. Kulmanov M, et al. Protein function prediction as approximate semantic entailment. Nat Mach Intell. 2024;6(2):220–8.
10. Zhang F, et al. Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics. 2019;19(12):1900019.
11. Consortium, U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506-15.
12. Torres M, et al. Protein function prediction for newly sequenced organisms. Nat Mach Intell. 2021;3(12):1050–60.
13. Eisenberg D, et al. Protein function in the post-genomic era. Nature. 2000;405(6788):823–6.
14. Billeter M. Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. Q Rev Biophys. 1992;25(3):325–77.
15. Carreras H. Cryo Electron Microscopy: Principle, Strengths, Limitations and Applications. Technology Networks; 2023. https://www.technologynetworks.com/analysis/articles/cryo-electron-microscopy-principle-strengths-limitations-and-applications-377080.
16. Eisenstein M. The field that came in from the cold. Nat Methods. 2016;13(1):19–23.
17. Das S, et al. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics. 2015;31(21):3460–7.
18. You R, et al. Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics. 2018;34(14):2465–73.
19. You R, et al. NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res. 2019;47(W1):W379–87.
20. Cao Y, Shen Y. Tale: transformer-based protein function annotation with joint sequence–label embedding. Bioinformatics. 2021;37(18):2825–33.
21. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020;36(2):422–9.

22. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor B C, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. Structure-based protein function prediction using graph convolutional networks. Zenodo. 2021. https://doi.org/10.5281/zenodo.4650027.
23. Pan T, et al. PfresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. Bioinformatics. 2023;39(3):btad094.
24. Meng L, Wang X. TAWFN: a deep learning framework for protein function prediction. Bioinformatics. 2024;40(10):btae571.
25. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–8.
26. Kulmanov M, Hoehndorf R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. Bioinformatics. 2022;38(Supplement_1):i238–45.
27. Xia W, et al. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. Comput Biol Med. 2022;145:105465.
28. Lin Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30.
29. Team E. ESM Cambrian: revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Website; 2024. https://www.evolutionaryscale.ai/blog/esm-cambrian.
30. Van Kempen M, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2):243–6.
31. Berman HM. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
32. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302–9.
33. Kadlec J, et al. The postfusion structure of baculovirus gp64 supports a unified view of viral fusion machines. Nat Struct Mol Biol. 2008;15(10):1024–30.
34. Rathore S, et al. Cryo-EM structure of the yeast respiratory supercomplex. Nat Struct Mol Biol. 2019;26(1):50–7.
35. Fischer ES, et al. The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. Cell. 2011;147(5):1024–39.
36. Ying Z, et al. Gnnexplainer: generating explanations for graph neural networks. Adv Neural Inf Process Syst. 2019;32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.
37. Papaleo E, et al. The role of protein loops and linkers in conformational dynamics and allostery. Chem Rev. 2016;116(11):6391–423.
38. Shehu A, Kavraki LE. Modeling structures and motions of loops in protein molecules. Entropy. 2012;14(2):252–90.
39. Kozome D, Sljoka A, Laurino P. Remote loop evolution reveals a complex biological function for chitinase enzymes beyond the active site. Nat Commun. 2024;15(1):3227.
40. Gunasekaran K, Nussinov R. Modulating functional loop movements: the role of highly conserved residues in the correlated loop motions. Chembiochem. 2004;5(2):224–30.
41. Marcelino AMC, Gierasch LM. Roles of β-turns in protein folding: from peptide models to protein engineering. Biopolymers. 2008;89(5):380–91.
42. Fu H, et al. Increasing protein stability by improving beta-turns. Proteins Struct Funct Bioinform. 2009;77(3):491–8.
43. Ananthanarayanan V, et al. Structural and functional importance of the β-turn in proteins. Studies on proline-containing peptides. J Biosci. 1985;8:209–21.
44. Errington N, Iqbalsyah T, Doig AJ. Structure and stability of the α-Helix: lessons for design. Protein Design Methods Appl. 2006:340:3–26.
45. Hol WG. The role of the α-helix dipole in protein function and structure. Prog Biophys Mol Biol. 1985;45(3):149–95.
46. Murray KA, et al. Extended β-strands contribute to reversible amyloid formation. ACS Nano. 2022;16(2):2154–63.
47. Freire F, et al. Impact of strand length on the stability of parallel-β-sheet secondary structure. Angew Chem Int Ed Engl. 2011;50(37):8735.
48. Daffner C, Chelvanayagam G, Argos P. Structural characteristics and stabilizing principles of bent β-strands in protein tertiary architectures. Protein Sci. 1994;3(6):876–82.
49. Enkhbayar P, et al. 310-helices in proteins are parahelices. Proteins Struct Funct Bioinform. 2006;64(3):691–9.
50. Cooley RB, Arp DJ, Karplus PA. Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. J Mol Biol. 2010;404(2):232–46.
51. Fodje M, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π-helix. Protein Eng. 2002;15(5):353–8.
52. Ponting CP, Russell RR. The natural history of protein domains. Annu Rev Biophys Biomol Struct. 2002;31(1):45–71.
53. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol. 2000;297(1):233–49.
54. Hvidsten TR, et al. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. PLoS One. 2009;4(7):e6266.
55. Burley SK, et al. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019;47(D1):D464–74.
56. Lees JG, et al. Gene3d: multi-domain annotations for protein sequence and comparative genome analysis. Nucleic Acids Res. 2014;42(D1):D240–5.
57. Ramon-Maiques S, et al. Structure of acetylglutamate kinase, a key enzyme for arginine biosynthesis and a prototype for the amino acid kinase enzyme family, during catalysis. Structure. 2002;10(3):329–42.
58. Massey V. The chemical and biological versatility of riboflavin. Biochem Soc Trans. 2000;28(4):283–96.
59. Susini L, et al. TCTP protects from apoptotic cell death by antagonizing bax function. Cell Death Differ. 2008;15(8):1211–20.
60. Raj I, Al Hosseini HS, Dioguardi E, et al. Structural basis of egg coat-sperm recognition at fertilization. Cell. 2017;169(7):1315-1326. e17.
61. Licht A, et al. Structural and functional characterization of a maltose/maltodextrin ABC transporter comprising a single solute binding domain (MalE) fused to the transmembrane subunit MalF. Res Microbiol. 2019;170(1):1–12.
62. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 2001;307(4):1113–43.

63. Pioszak AA, et al. Molecular recognition of corticotropin-releasing factor by its G-protein-coupled receptor CRFR1. J Biol Chem. 2008;283(47):32900–12.
64. Holden HM, et al. Galactokinase: structure, function and role in type II galactosemia. Cell Mol Life Sci. 2004;61(19–20):2471–84.
65. Timson DJ, Reece RJ. Functional analysis of disease-causing mutations in human galactokinase. Eur J Biochem. 2003;270(8):1767–74.
66. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996;257(2):342–58.
67. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res. 2012;41(D1):D1096-103.
68. Kursula I, et al. The importance of the conserved Arg191-Asp227 salt bridge of triosephosphate isomerase for folding, stability, and catalysis. FEBS Lett. 2002;518(1–3):39–42.
69. Wu S, et al. Diverse roles of assembly factors revealed by structures of late nuclear pre-60S ribosomes. Nature. 2016;534(7605):133–7.
70. Li N, et al. Structure of the origin recognition complex bound to DNA replication origin. Nature. 2018;559(7713):217–22.
71. Sanghai ZA, et al. Modular assembly of the nucleolar pre-60S ribosomal subunit. Nature. 2018;556(7699):126–9.
72. Ohnuma T, et al. Crystal structure and chitin oligosaccharide-binding mode of a "loopful" family GH19 chitinase from rye, *Secale cereale*, seeds. FEBS J. 2012;279(19):3639–51.
73. Taira T, et al. Localization, accumulation, and antifungal activity of chitinases in rye (*Secale cereale*) seed. Biosci Biotechnol Biochem. 2001;65(12):2710–8.
74. Addepalli B, Limbach PA. Pseudouridine in the anticodon of *Escherichia coli* tRNATyr (QΨA) is catalyzed by the dual specificity enzyme RluF. J Biol Chem. 2016;291(42):22327–37.
75. Buchanan SK, et al. Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. Nat Struct Biol. 1999;6(1):56–63.
76. Parsons JF, et al. Enzymes harboring unnatural amino acids: mechanistic and structural analysis of the enhanced catalytic activity of a glutathione transferase containing 5-fluorotryptophan. Biochemistry. 1998;37(18):6286–94.
77. Fu L, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
78. Touw WG, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015;43(D1):D364–8.
79. Zhou B, et al. Protein engineering with lightweight graph denoising neural networks. J Chem Inf Model. 2024;64(9):3650–61.
80. Ganea OE, et al. Independent se (3)-equivariant models for end-to-end rigid protein docking. arXiv preprint arXiv:2111.07786. 2021. https://doi.org/10.48550/arXiv.2111.07786.
81. Liu Y, et al. Rotamer-free protein sequence design based on deep learning and self-consistency. Nat Comput Sci. 2022;2(7):451–62.
82. Satorras VG, Hoogeboom E, Welling M. E (n) equivariant graph neural networks. In International conference on machine learning. PMLR; 2021;139:9323–32. https://proceedings.mlr.press/v139/satorras21a.html.
83. Gainza P, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020;17(2):184–92.
84. Ingraham J, et al. Generative models for graph-based protein design. Adv Neural Inf Process Syst. 2019;1417:15820–31. https://doi.org/10.5555/3454287.3455704.
85. Hayes T, et al. Simulating 500 million years of evolution with a language model. Science. 2025. https://doi.org/10.1126/science.ads0018.
86. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014. https://doi.org/10.48550/arXiv.1412.6980.
87. Li F, et al. Digerati–a multipath parallel hybrid deep learning framework for the identification of mycobacterial PE/PPE proteins. Comput Biol Med. 2023;163:107155.
88. Li F, et al. Prosperousplus: a one-stop and comprehensive platform for accurate protease-specific substrate cleavage prediction and machine-learning model construction. Brief Bioinform. 2023;24(6):bbad372.
89. Ran Z, et al. Characterizing secretion system effector proteins with structure-aware graph neural networks and pre-trained language models. IEEE J Biomed Health Inform. 2024. https://doi.org/10.1109/JBHI.2024.3413146.
90. Nitish S. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1.
91. Bi Y, et al. Clarion is a multi-label problem transformation method for identifying mRNA subcellular localizations. Brief Bioinform. 2022;23(6):bbac467.
92. Li F, et al. Advancing mRNA subcellular localization prediction with graph neural network and RNA structure. Bioinformatics. 2024;40(8):btae504.
93. Wu J, et al. HiFun: homology independent protein function prediction by a novel protein-language self-attention model. Brief Bioinform. 2023;24(5):bbad311.
94. Zhou N, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 2019;20:1–23.
95. Ran Z, Guo X, Pan T, Bi Y, Hao Y, Sun H, Song J, Li F. A scalable equivariant graph network framework for precise protein function prediction. Github. 2025. https://github.com/ABILiLab/ENGINE.
96. Ran Z, Guo X, Pan T, Bi Y, Hao Y, Sun H, Song J, Li F.. A scalable equivariant graph network framework for precise protein function prediction. Zenodo. 2025. https://doi.org/10.5281/zenodo.17221153.

Publisher's Note