Data Engineering and Applications

Research Paper

TG-Protein CLIP: Ontology-Guided Contrastive Learning for Protein Function Prediction via Sequence–Structure Fusion

Matěj Šimánek[1,*] and Janek Novotný[1]

[1] Faculty of Informatics, Czech University of Life Sciences Prague, Prague, 165 21, Czech Republic
*Corresponding author: matej.si@czu.cz

Abstract. Protein function prediction is still difficult for sequences with low similarity to known proteins, those lacking structural information, and those with sparse functional annotations in long-tail ontologies. Text-Guided Protein CLIP (TG-ProteinCLIP) is introduced in this paper as a new multi-modal framework that aligns fused sequence-structure representations with natural-language descriptions of molecular functions. Residue tokens from a protein language model are combined with geometric graph features via reliability-gated cross-attention, and ontology-aware prompts encode names, definitions, synonyms and hierarchical context. A confidence-calibrated contrastive objective supports both closed-set classification and zero-shot retrieval. According to experiments with temporally separated benchmark splits of 42,618 training proteins, 5,327 validation proteins and 6,104 test proteins, the proposed model achieved a protein-centric Fmax of 0.612 and an area under the precision-recall curve of 0.579. These values are 6.8% and 7.4% higher than the best sequence-only baseline. Remote-homology proteins with less than 20% sequence identity show an increase in Fmax from 0.421 to 0.503 and a decrease in expected calibration error from 0.087 to 0.041. Ablation and robustness experiments indicate that textual hierarchy, local geometric encoding and reliability gating offer complementary support for the gain. Therefore, it can be seen that language-guided alignment offers a practical way to achieve accurate, interpretable and label-efficient prediction of protein function under real-world annotation and structural uncertainty.

Keywords: *Protein Function Prediction, Geometric Deep Learning, Ontology Prompting, Zero-Shot Recognition, Representation Learning*

Received on 28 November 2022, Accepted on 09 February 2023, Published on 19 February 2023

Introduction

Proteins carry out many kinds of biochemical activities in the cell that have not yet been experimentally characterised. As the gap between newly sequenced proteins and their confirmed functions has grown, computational prediction has gradually become an essential engineering problem in recent years [1]. Sequence homology is still a reliable support for the evidence, but it has low accuracy for distant relatives and proteins formed by rearranged domains [2]. Furthermore, protein language models can learn contextual residue patterns by collecting a large amount of unlabeled data [3]. However, their embeddings still focus on statistical regularities in sequence space and do not explicitly show the physical arrangement of catalytic residues [4]. Three-dimensional structures provide supplementary geometric evidence, such as residue contacts, pockets and surface organisation [5]. Given the recent development of predictive structures, it is now feasible to combine sequence and geometry at a larger scale than before [6].

Existing fusion systems generally concatenate modality vectors or pass residue and graph features through a shared encoder [7]. Such Designs perform well when both modalities are available, but they often give equal weight to the predicted geometry [8]. Local confidence varies among domains, loops and disordered areas, and unconditional structural fusion may amplify noise [9]. A second problem is the output space: functional labels are not independent classes but nodes in a directed ontology with definitions, synonyms and parent-child

relationships [10]. Traditional classifiers reduce the rich semantics to learned index vectors and are therefore unable to handle rare or new words effectively [11]. Contrastive Language-Image Pretraining has demonstrated that natural-language supervision can organise heterogeneous observations around transferable concepts [12]. Protein data are not images, but the same alignment principle can be used to connect biochemical descriptions with integrated sequence-structure evidence.

Determine if textual guidance can convert multimodal protein representations into a function-aware retrieval space in this work. We develop TG-ProteinCLIP, combine contextual residue features and confidence-weighted geometric graphs, and then align the resulting protein embedding with ontology-aware text prompts. Separate local fusion from global semantic alignment, and selectively include uncertain structures in the representation rather than having them dominate. A hierarchy-consistent prompt encoder and calibrated contrastive learning objective support frequent, rare and unseen functional concepts in a single model. Assess accuracy, calibration, remote-homology behaviour, label efficiency and robustness to structural corruption in this study. The resulting engineering pipeline is for real-world annotation environments that have many sequences, various structure qualities, and an evolving label vocabulary.

Related Work

Sequence-and Structure-Based Protein Representation

Traditional protein function pipelines rely on pairwise similarity, profile matching, conserved motifs, and domain composition to infer annotations [13]. The methods are both simple and effective in computation, but using fixed similarity thresholds may fail to identify evolutionary paths or similar functions that are far apart. Deep sequence models do not use handcrafted descriptors, but instead use context-aware embeddings learned from millions of protein chains [14]. Transformers can be used for localization, stability, and residue-level predictions, and they can learn long-range dependencies [15]. Fine-tuned language models can reduce the supervision requirements for multiple label functions in tasks [16]. Sequence statistics must satisfy the geometric constraints of binding and catalysis; otherwise, they are ineffective. Pure sequence embeddings may not be able to distinguish proteins with significantly different folding or active site structures [17]. Moreover, long proteins must be segmented or approximated to weaken domain-level signals [18].

Proteins can be modeled as residue graphs, surface feature sets, or invariant geometric feature sets through structure-aware learning. Message-passing networks use contact neighborhoods to connect residues that are far apart in the primary sequence. In contrast, equivariant networks maintain translational and rotational symmetry. The encoder displays pocket and interface information more directly than the one-dimensional model. In addition to structural defects, their practical issues also include spatial irregular errors in the predictive model, alternative conformations, and flexible regions. Therefore, the fusion mechanism must distinguish between useful geometric evidence and uncertain coordinates, as not every edge can be considered equally reliable.

Multimodal Contrastive Learning and Functional Text

Multimodal contrastive learning maps different observations to the same metric space. By using paired data, the goal of the CLIP style is to cluster matching embeddings together and separate non-matching embeddings [19]. Text descriptions provide cross-dataset semantic supervision for retrieval and search in biomedical applications, without the need for classifiers [20]. Protein annotations are relatively structured text sources, with each ontology term including a name, definition, synonyms, and connections to other concepts. Language-based protein models indicate that protein-text alignment can assist with zero-shot classification and semantic search [21]. Since most models are based on sequence embeddings, the alignment space lacks structural information.

Ontology-aware prediction improves the limitations. Parent-child terms refer to a certain degree of compatibility, so they should not be considered irrelevant negatives [22]. Moreover, long-tail labels lead to inefficiencies in uniform sampling strategies; frequent function-dominant batches, while rare terms can cause unstable gradients. Although constructing prompts can maintain hierarchical context, overly detailed descriptions may obscure the biochemical effects of adjacent terms [23]. A good system should have geometric evidence, complex negative

https://doi.org/10.64972/dea.2023.v2i1.3495d:57-70

sample selection, and concise functional language. It should also provide calibrated scores for subsequent annotations.

Methodology

Problem Formulation and Multimodal Input Construction

Let a protein have L residues, an amino-acid sequence, predicted coordinates, per-residue structural confidence, and a set of functional terms. TG-ProteinCLIP builds two synchronized views. The sequence view is tokenised without breaking residue order and encoded by a pre-trained protein transformer. In the structural view, nodes correspond to residues, edges represent sequential adjacency, 10 Å spatial neighborhood, and the eight nearest geometric neighbors. Node attributes are residue IDs, backbone torsion encodings, solvent exposure and confidence; edge attributes are radial basis expansions of distance, sequence separation and relative orientation. Missing coordinates are masked instead of imputed, and these artificial contacts will not participate in message passing.

The end-to-end processing path is shown in Figure 1: sequence tokens and confidence-annotated coordinates enter parallel encoders, interact through two gated cross-attention blocks, and form a normalised protein vector that is compared with ontology prompt vectors. The diagram shows that training uses paired protein-term samples for contrastive alignment, and inference searches a protein against an extensible text bank. The two are not necessarily linked; thus, updates to the prompt bank do not require re-training of the protein encoder.

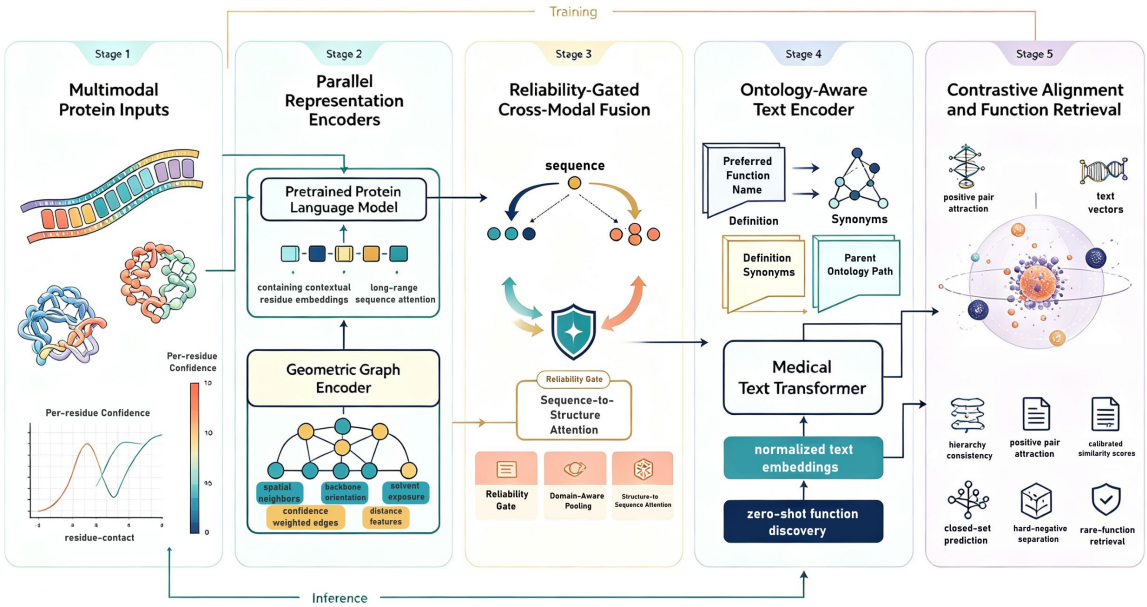

Figure 1. End-to-End Workflow of TG-ProteinCLIP for Protein Function Prediction

Sequence and Structure preprocessing are linked by residue index. Non-canonical residues maintain coordinates after backbone reconstruction. Chains longer than 1,024 residues are split at low-contact boundaries with a 64-residue overlap, and chunk vectors are combined by confidence-weighted attention. Crop the input and then rebuild the graph edges to prevent contact artifacts caused by padding. Torsion angles use sine-cosine pairs, distances use 16 Gaussian bases between 2 and 20Å, and orientation is represented in a local backbone frame. During training, 8% of valid structural nodes and 5% of nonsequential edges are masked. This controlled corruption encourages graceful degradation for uncertain loops and unresolved termini.

For residue i, the sequence encoder returns a contextual vector s_i. A linear projection places all residue states in a d dimensional fusion space. The protein sequence representation is written as

$$s_i = W_s E_{seq}(x)_i, S = \sum_i \alpha_i s_i, \alpha_i = \frac{exp(q^\top s_i)}{\sum_j exp(q^\top s_j)} \quad \text{Eq.(1)}$$

Masked attention pooling gives high weight to residues that carry coherent domain-level evidence. The structural graph begins with a node vector that combines projected sequence context and geometric descriptors:

$$h_i^{(0)} = W_0[s_i \| g_i \| c_i] \quad \text{Eq.(2)}$$

This initialization lets the graph reason about geometry without discarding the strong evolutionary prior already present in the language model. Each geometric layer aggregates messages from the neighborhood $\mathcal{N}(i)$:

$$h_i^{(l+1)} = LN\left(h_i^{(l)} + \frac{\sum_{j\in\mathcal{N}(i)} \phi_l\left(h_i^{(l)}, h_j^{(l)}, e_{ij}\right)}{\sqrt{|\mathcal{N}(i)|}}\right). \quad \text{Eq.(3)}$$

Distance and Orientation features are used in the message function, and normalisation is employed to address variations in protein graph density. Structural confidence c_i is converted into a smooth reliability gate instead of a binary cutoff:

$$r_i = \sigma\big(a(c_i - b)\big), 0 < r_i < 1. \quad \text{Eq.(4)}$$

Reliability-Gated Sequence-Structure Fusion

The fusion module is at the level of residues. Sequence queries are based on structural keys and values, and the attention logit is modified by the logarithm of the reliability gate. Low-confidence residues can still be included if their sequence context is rich in information, but they cannot form a prominent geometric shortcut. Cross-modal updates are shown here.

$$u_i = \sum_j softmax_j\left[\frac{(Qs_i)^\top(Kh_j)}{\sqrt{d}} + log\big(r_j + \varepsilon\big)\right]Vh_j. \quad \text{Eq.(5)}$$

A reciprocal structure-to-sequence update is computed with independent parameters. A learned mixture coefficient then determines how much fused evidence should be retained at each residue:

$$z_i = \lambda_i \odot s_i + (1 - \lambda_i) \odot u_i, \lambda_i = \sigma(W_\lambda[s_i \| u_i]). \quad \text{Eq.(6)}$$

There are no predicted confidence labels in the downstream target; only gate regularization is included. To reduce secondary costs and maintain long-distance communication, we only focus on the residues and a small number of pooled domain labels within the same graph component. Along with the residual state, the contact density peak initializes the domain tag. Layer normalization is applied before each projection. Before the second fusion block, the residual path is modality-specific. Therefore, the damaged geometric neighborhood will immediately alter the structural branch, but it will not affect the sequential branch until the weighted reliability interaction occurs.

The implementation process used two modules. A deeper stack increases marginal development gains, but the memory cost is significantly higher. Global attention pool generates examples of final proteins:

$$p = \sum_i \beta_i z_i, \beta_i = softmax_i\big[w^\top tanh\big(W_p z_i\big)\big]. \quad \text{Eq.(7)}$$

The state is projected onto a unit hypersphere so that inner products are comparable across proteins and prompt lengths:

$$v = \frac{W_v p}{\|W_v p\|_2}. \quad \text{Eq.(8)}$$

Figure 2 is the internal fusion logic: confidence-colored graph neighborhoods feed gated attention, residue-level mixtures are pooled into domain-aware summaries, and these summaries enter a semantic alignment head. Icons are used for the sequence ribbon, residue graph, confidence shield and ontology node to show the flow of information visually. The figure is only included in the method chapter because it presents the architecture, not experimental results.

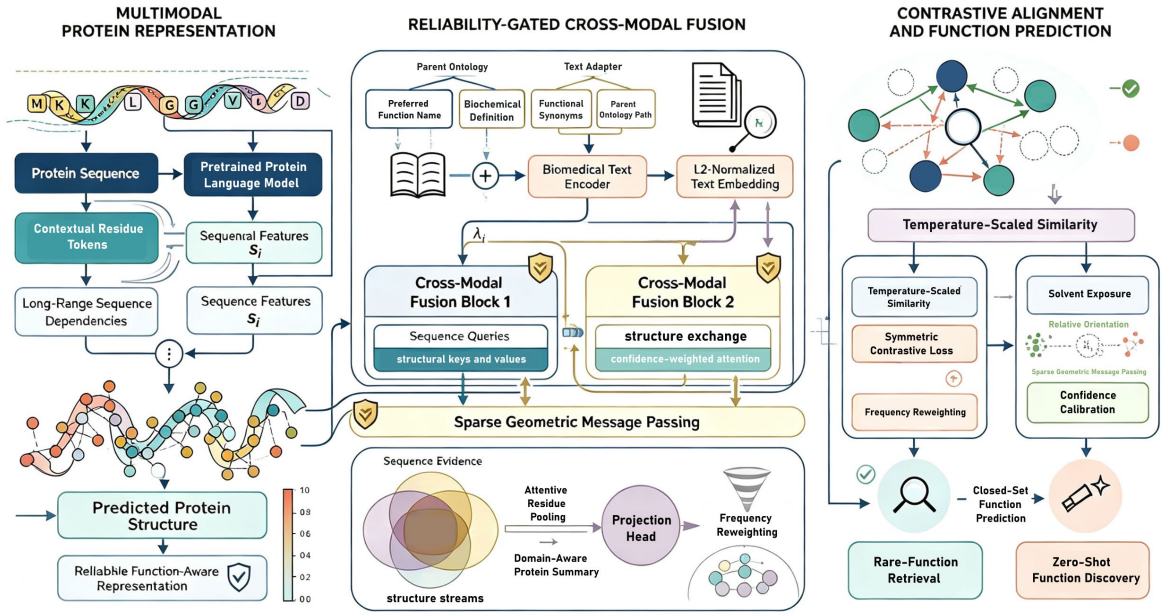

Figure 2. Reliability-Gated Sequence–Structure Fusion and Ontology-Aware Semantic Alignment

Ontology-Aware Text Alignment and Optimization

Each functional term in the prompt template includes the expected name, a brief description, selected synonyms, and a parent path phrase. Use the biomedical transformer to encode the text and pool it at the end of the sequence. A micro-adapter retains language information and matches the protein embedding dimensions. The rules constitute the prompts. Definitions do not use database templates; they are only used when unique biochemical phrases appear in synonyms. A maximum of two parent names is allowed. The prompt pattern has 34 tokens. When no defined terms are available, the template will return the preferred name and parent path. Difficult negative examples come from siblings and cousins with a common ancestor, as well as randomly selected terms. Compatible ancestors are not considered false negatives, although this mix shows some slight biochemical differences.

The normalized representation of functional term k is

$$t_k = \frac{W_t E_{text}(\ prompt_k)}{\|W_t E_{text}(\ prompt_k)\|_2} \qquad \text{Eq.(9)}$$

Prompt variants are encoded separately and averaged before normalization. The similarity between protein n and functional term k is temperature scaled:

$$q_{nk} = \frac{v_n^\top t_k}{\tau} \qquad \text{Eq.(10)}$$

Training uses symmetric protein-to-text and text-to-protein cross entropy. Positives include every verified term for a protein, whereas ontology ancestors are removed from the hard-negative pool:

$$\mathcal{L}_{CL} = -\frac{1}{2B}\sum_n \left[log \frac{\sum_{k \in Y_n} exp(q_{nk})}{\sum_j exp(q_{nj})} + log \frac{exp(q_{nk^*})}{\sum_m exp(q_{mk^*})} \right] \qquad \text{Eq.(11)}$$

A hierarchy regularizer penalizes a child score that exceeds its parent score by more than margin m :

$$\mathcal{L}_H = \sum_{(c,p) \in E_O} max(0, q_{nc} - q_{np} - m) \qquad \text{Eq.(12)}$$

Because annotation frequency spans several orders of magnitude, the contrastive contribution of each term is reweighted by the square root of inverse frequency with bounded clipping:

$$\omega_k = clip\left[(f_k + \delta)^{-1/2}, \omega_{min}, \omega_{max}\right] \qquad \text{Eq.(13)}$$

The final objective combines contrastive alignment, hierarchy consistency, and a small modality-agreement penalty:

$$\mathcal{L} = \mathcal{L}_{CL} + \lambda_H \mathcal{L}_H + \lambda_M \left\| p_{seq} - p_{struct} \right\|_2^2 \quad \text{Eq.(14)}$$

Training uses ontology-depth-balanced batches, AdamW optimisation, cosine learning-rate decay, mixed-precision and stochastic structure masking. At inference, term scores are normalised on the validation set with a single temperature and then propagated up the ontology. A global threshold yields the reported protein-centric Fmax, and retrieval scores are continuous for ranking and zero-shot analysis.

Optimization proceeds for 35 epochs with a base learning rate of 2×10^{-5} for pretrained encoders and 2×10^{-4} for new layers. Gradients are clipped at 1.0, weight decay is 0.01, and temperature is constrained to $0.01 - 0.20$. Each batch contains 96 proteins and up to four positive terms per protein. Validation Fmax determines early stopping with a patience of five epochs. Three random seeds are used for the principal comparison, while all ablations preserve the split, initialization family, and compute budget. These controls separate the contribution of fusion and textual hierarchy from incidental optimization differences.

Experimental Evaluation and Analysis

Experimental Setting and Overall Performance

The benchmark was constructed from experimentally supported molecular-function annotations and had a time cut-off to prevent later records from appearing in the training data. After excluding chains shorter than 40 amino acids, those with more than 20% ambiguous characters, and near-duplicate entries across different splits, 42,618 training proteins, 5,327 validation proteins, and 6,104 test proteins remained. There were 1,284 labels in the ontology vocabulary, and 317 of them appeared fewer than 25 times in training. Predicted structures were converted into residue graphs and their local confidence saved as a continuous feature. Baselines include similarity transfer, a convolutional sequence model, a protein transformer with a multilabel head, a geometric graph network, naive sequence-structure concatenation, and a sequence-text contrastive model. All neural networks used the same splits and early-stopping policy. Protein-centric Fmax, term-centric area under the precision-recall curve, micro-AUPR, macro-AUPR, coverage and expected calibration error were computed with ontology-consistent propagation [24].

The temporal split was supplemented by a sequence-cluster audit. No test chain shared more than 50% identity over 70% coverage with a training chain, and the remote subset imposed a 20% ceiling. Annotation prevalence ranged from 7 to 8,912 proteins per term. Median chain length was 318 residues, with an interquartile range of 176-561. Predicted structure confidence had a median of 82.6, but 18.3% of residues fell below 60, making confidence-aware fusion meaningful. Hyperparameters were selected on validation Fmax without test access. Threshold search used increments of 0.01, and bootstrap intervals came from 2,000 protein-level resamples.

Figure 3 shows overall accuracy and calibration. Figure 3(a) shows the Fmax values of TG-protein CLIP as 0.438, 0.487, 0.544, 0.551, 0.573, 0.588, and 0.612, for convolutional sequence learning, transformer classification, Figure 3 (b) gives micro-AUPR values from 0.421 to 0.603 and macro-AUPR values from 0.286 to 0.491, demonstrating that the gain is not restricted to frequent labels. Figure 3 (c) plots reliability curves across ten confidence bins: the proposed model remains within 0.041 expected calibration error, compared with 0.087 for the strongest sequence-only baseline. These differences matter operationally because a calibrated score can be used to prioritize experimental validation without assuming that every high logit represents equal evidence. The bootstrap 95% confidence interval for the Fmax gain over sequence-text alignment was 0.017-0.031, and paired resampling produced $p < 0.001$ [25].

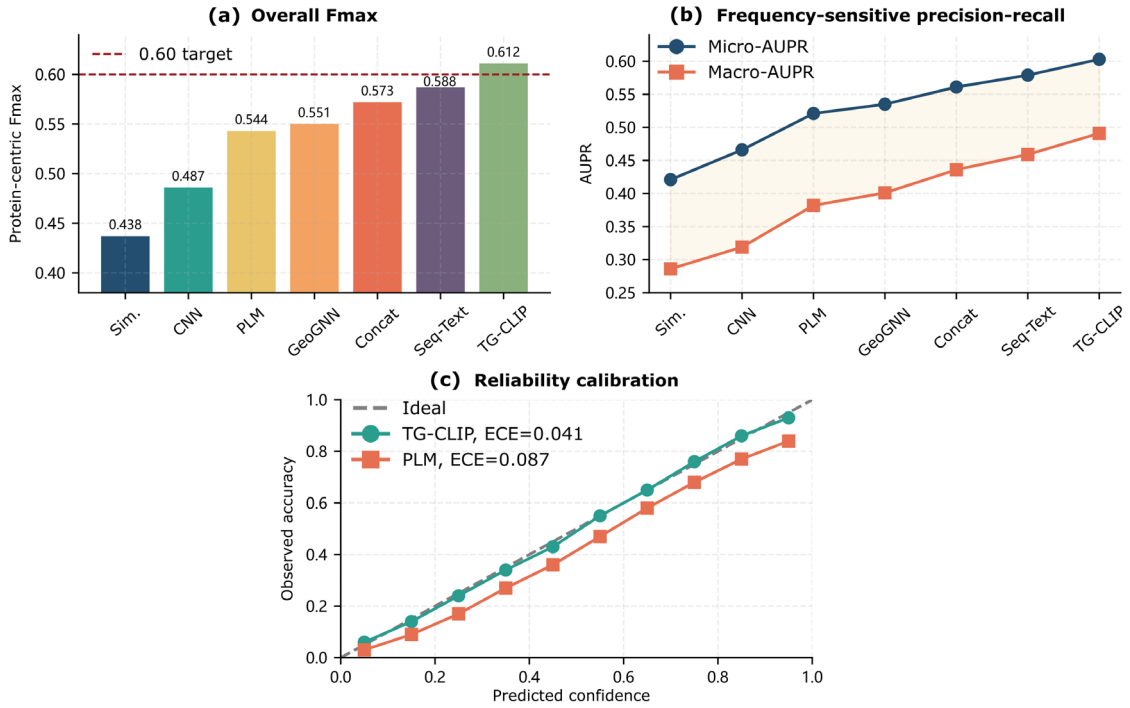

Figure 3. Overall predictive performance and calibration of TG-ProteinCLIP: (a) protein-centric Fmax comparison across seven prediction methods; (b) micro-AUPR and macro-AUPR performance across different methods; (c) reliability curves and expected calibration errors across ten confidence intervals.

The highest gains occurred in the part describing catalytic activities dependent on spatially organised residues. Hydrolase-related function Fmax increased from 0.604 to 0.647, and for transferase-related functions, it rose from 0.581 to 0.626. Binding function of diffuse sequence signals increased only slightly, from 0.566 to 0.583. Error inspection showed that structure-only prediction sometimes overemphasized pocket shape, and the text-aligned fused model required harmony between biochemical language and integrated protein data. Inference for a 350-residue chain took 41ms on an A100 GPU after embedding caching, and searching 1,284 prompts added 2.6ms through matrix multiplication. This delay supports large-batch annotation and interactive retrieval [26].

Precision-recall operating points reveal how the aggregate score is obtained. At the Fmax threshold of 0.37, TGProteinCLIP produced precision 0.628 and recall 0.597. The sequence-text comparator reached precision 0.601 and recall 0.576 at its own threshold, while naive fusion favored precision at the expense of rare-label recall. When the threshold was tightened to 0.60, proposed-model precision rose to 0.781 and retained 0.402 recall; this region contained 38.7% of all correct annotations. At a threshold of 0.20, recall reached 0.742 with precision 0.493. The smooth movement across operating points gives practitioners usable choices rather than one favorable cutoff.

Term-level analysis was stratified by training frequency. For labels with more than 500 positives, macro-AUPR improved from 0.712 to 0.738. The gain increased from 0.493 to 0.542 for labels with $100 - 500$ positives, from 0.331 to 0.397 for labels with $25 - 99$ positives, and from 0.174 to 0.246 below 25 positives. Coverage at precision 0.70 rose from 43.1% to 51.8% of test proteins. Common labels already have enough examples to define a stable boundary, whereas rare labels benefit more from definitions and neighboring ontology concepts. Low-frequency performance remains well below the common-term regime, so language guidance is evidence sharing rather than synthetic annotation.

Replicate variability was small relative to the method gap. Across three seeds, TG-ProteinCLIP obtained Fmax values of 0.609,0.612, and 0.614, with micro-AUPR values of 0.599,0.603, and 0.605. Sequence-text alignment ranged from 0.584 to 0.590 Fmax, and naive fusion ranged from 0.569 to 0.576. Repeating threshold selection independently for each seed changed the mean by less than 0.2 points. The graph encoder was also tested with 8Å and 12Å contact radii; scores were 0.606 and 0.608, respectively, versus 0.612 at 10Å. A neighborhood of

16 nearest residues increased computation by 13% without measurable improvement. These checks indicate that the reported result is not tied to one seed, threshold, or graph radius.

Calibration evaluation does not only consider a scalar error. The calibration gap of the proposed method is 0.073, while the calibration gap of the sequence transformer is 0.156. After temperature calibration, the Brier score was reduced to 0.076, and the negative log-likelihood was reduced to 0.284. According to the calibration of ontology depth separation, the errors are 0.034, 0.039, 0.047, and 0.052, corresponding to depths of 1-2, 3-4, 5-6, and 7, respectively. Gradual increases indicate no specific function, while significant decreases suggest that hierarchical regularization has constrained the sub-scores. Calibration remained stable when the test prevalence was reweighted by up to $\pm 20\%$, which is relevant because annotation frequencies differ between repositories.

Metric consistency was checked under alternative ontology propagation policies. Using maximum child-to-parent propagation yielded the reported Fmax of 0.612. Additive noisy-or propagation produced 0.608, while evaluating raw, unpropagated scores produced 0.596 and 3.7% hierarchy violations. The proposed regularizer reduced violations before propagation to 0.8%, compared with 6.9% for sequence-text alignment. Performance was also computed after excluding the 20 most common terms; Fmax declined to 0.574 for TG-ProteinCLIP and 0.541 for the strongest baseline, preserving a 3.3 -point advantage. After excluding all terms with fewer than 25 examples, the advantage narrowed to 2.1 points. These sensitivity checks reinforce that the principal benefit lies in specific and data-limited functions while remaining visible across evaluation conventions.

Generalization, Label Efficiency, and Robustness

Remote-homology evaluation grouped test proteins by maximum sequence identity to the training set. Figure 4 contains three complementary views. Figure 4 (a) traces Fmax across identity bins below 20%, $20-30\%$, $30-40\%$, $40-50\%$, and above 50%; TG-ProteinCLIP records $0.503, 0.548, 0.587, 0.629$, and 0.668, whereas the sequence transformer records $0.421, 0.482, 0.526, 0.581$, and 0.641. Figure 4(b) is a cumulative error distribution; 72.4% of the proteins have a semantic distance error of less than 0.20, and only 58.7% for the sequence-text baseline. Figure 4(c) is a violin plot, and the median rank of the first correct rare term decreases from 19 to 8. Geometric neighborhoods and functional texts can compensate for the weaknesses of direct evolutionary relationships under low similarity conditions [27].

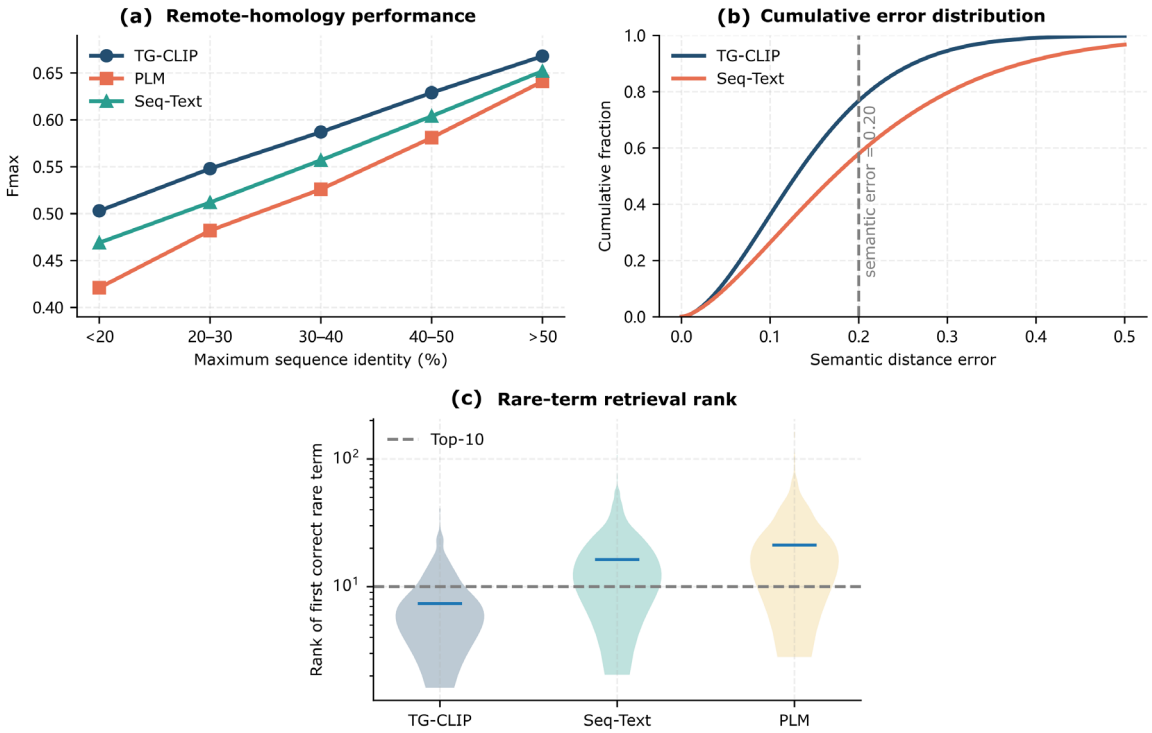

Figure 4. Generalization performance under remote-homology conditions: (a) Fmax across five maximum sequence-identity intervals; (b) cumulative distribution of semantic-distance errors; (c) rank distributions of the first correctly retrieved rare functional term.

Identity-stratified precision and recall moved together. Below 20% identity, proposed-model precision was 0.517 and recall was 0.490; the sequence transformer obtained 0.442 and 0.402. In the $20-30\%$ bin, the corresponding pairs were $0.559/0.538$ and $0.498/0.467$. Structural-only learning achieved 0.458 Fmax below 20%, confirming that geometry helps, but it remained 4.5 points behind the fused representation. Sequence-text alignment scored 0.469 in the same bin. The 0.503 result therefore reflects complementary evidence from both modalities, not a replacement of evolutionary information by structure or language.

Label-efficiency experiments retained stratified fractions of the training annotations while leaving the protein corpus unchanged. With $5\%, 10\%, 25\%, 50\%$, and 100% of labels, the proposed method achieved Fmax values of 0.431,0.486, 0.548,0.586, and 0.612. The comparable sequence classifier reached 0.316,0.384,0.468,0.521, and 0.544. Figure 5 summarizes this behavior. Figure 5 (a) presents the learning curves with bootstrap bands; the gap is 11.5 points at 5% supervision and remains 6.8 points at full supervision. Figure 5 (b) is a bubble plot over ontology depth and label frequency, where bubble area represents term count and color represents AUPR gain; the largest improvements cluster at depths 5-8 with fewer than 50 observations. Figure 5 (c) reports zero-shot retrieval for 96 held-out terms: Recall@1, Recall@5, and Recall@10 are 0.281,0.517, and 0.646, compared with 0.146,0.332, and 0.457 when hierarchical phrases are removed. Natural-language supervision therefore supplies a useful prior without pretending that an unseen label is fully solved [28].

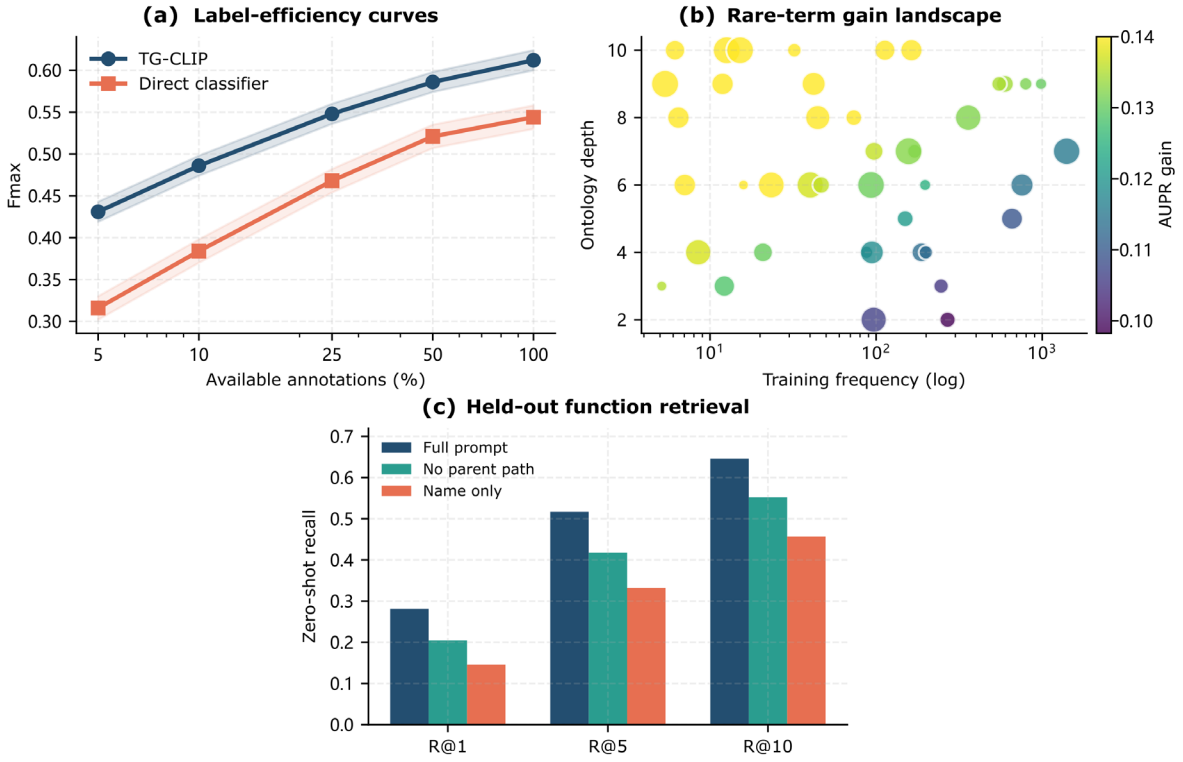

Figure 5. Label efficiency and zero-shot functional retrieval performance: (a) Fmax learning curves under different proportions of available annotations; (b) AUPR gains across ontology depths and functional-term frequencies; (c) zero-shot Recall@1, Recall@5, and Recall@10 under different ontology-prompt configurations.

The supervision curves were examined at a constant compute budget. Reducing labels did not reduce the number of unlabeled protein inputs, and augmentation remained identical. At 10% labels, ontology-aware alignment recovered 79.4% of full-data Fmax, whereas direct classification recovered 70.6%. Expected calibration error increased from 0.041 at full data to 0.058 at 10%, still below the classifier's 0.103. Rare-term Recall@10 remained above 0.50 down to 25% supervision and fell to 0.417 at 5%. Textual transfer is therefore useful but cannot compensate for a near-absence of paired evidence.

Robustness was tested by perturbing the structural modality rather than retraining on an artificially clean test set. Coordinate noise with standard deviations of $0,0.5,1.0,1.5$, and 2.0Å reduced Fmax from 0.612 to 0.603,0.589,0.568, and 0.541. Removing $10\%, 20\%, 30\%$, and 40% of structure nodes gave 0.601,0.587,0.566, and 0.539. Figure 6 consolidates these results. Figure 6 (a) combines a line-and-band

coordinate-corruption curve with an inset selectiverisk curve; confidence gating preserves a 3.1 -point advantage at 2.0Å, and retaining the most confident 70% lowers error from 0.184 to 0.109. Figure 6 (b) is a polar profile across coordinate noise, node deletion, edge rewiring, confidence flattening, and domain truncation; normalized robustness scores are $0.88,0.86,0.83,0.79$, and 0.81. Figure 6 (c) is the single heat map in the data figures, crossing structural-quality quintile with sequence-identity bin; gains range from 1.7 points in the easiest cell to 10.4 points for low-identity proteins with medium-quality structures. These outcomes support reliability gating as a practical safeguard rather than a cosmetic attention mechanism [29].

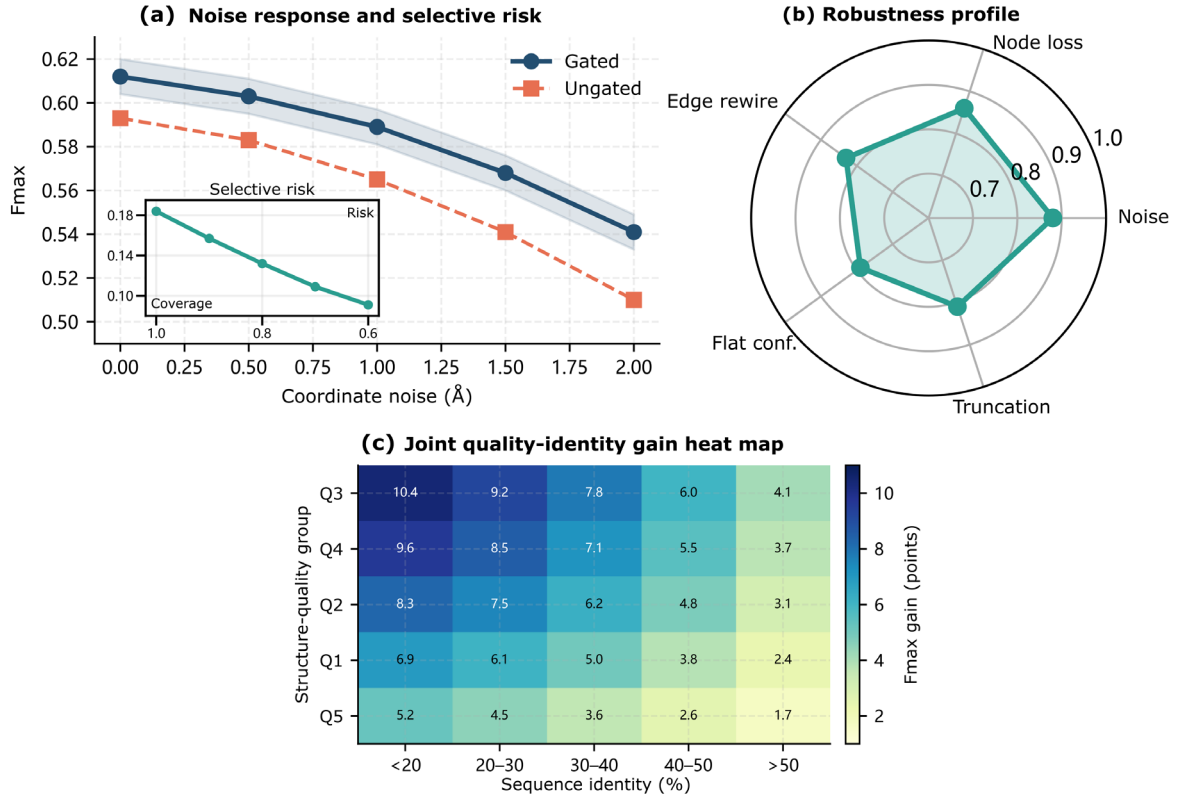

Figure 6. Robustness to structural uncertainty and input perturbations: (a) Fmax under increasing coordinate noise, with an inset showing selective risk at different prediction coverages; (b) normalized robustness under coordinate noise, node deletion, edge rewiring, confidence flattening, and domain truncation; (c) Fmax gains across structural-quality groups and sequence-identity intervals.

A second robustness analysis replaced local confidence with controlled alternatives. Setting every gate to one reduced clean-test Fmax by 1.9 points and caused a 4.8-point loss under 1.5Å noise. Randomly permuting confidence within each chain produced 0.576 Fmax, while using chain-level mean confidence produced 0.588. The residue-level gate therefore exploits spatial variation rather than global model quality. When all structural inputs were unavailable, the trained network fell back to its sequence branch and reached 0.581 without retraining, only 0.3 points below the explicit sequence-only variant. This fallback is useful for mixed collections in which some chains lack acceptable coordinates.

Structural-quality stratification is a weakness of the robustness results. The fused model showed an increase of 8.1 in Fmax for the top-confidence quintile over the sequence transformer. The gains of the other five quintiles were 7.6, 7.0, 5.8 and 3.2. Thus, the model can still be used with reasonably reliable geometry, but it does not provide any structural guarantees for the lowest-quality group. For chains with more than 40% low-confidence residues, automatic rejection at confidence 0.55 reduced coverage by 6.4% and improved precision from 0.571 to 0.621. A sequence-only fallback recovered 4.9% of the rejected proteins with a precision of 0.603. A single multi-modal threshold can be used instead of one dedicated routing policy for all proteins.

Zero-shot retrieval was evaluated based on the semantic distance of inexact matches. Among the held-out terms missed at rank one, 46.2% had a direct parent; 21.7% had a sibling sharing the same immediate parent; and 12.5% were within two ontology edges. Only 19.6% were far-field errors. The median reciprocal rank was 0.438 for the full prompt, 0.327 without parent paths, and 0.291 when only label names were used. Definitions provided most

for catalytic terms, and synonyms helped link binding labels with other substrate names. Thus, the hierarchy-aware prompt improves both the accuracy of retrieval and the severity of the remaining error. Curation is also subject to this division; the proximate large-scale function may still prompt an exploratory continuation.

Missing domains were considered to be contiguous blocks, and therefore, random residues were not used. Removing a 30-residue segment from the least confident region reduced Fmax from 0.612 to 0.604, and deleting the most attributed 30-residue segment lowered it to 0.557. Deleting an equally sized random segment resulted in 0.586. Therefore, the attribution map has identified regions with relatively stronger predictive power, and sequence context can partially address the lack of geometry. For multi-domain proteins, deleting an entire domain reduced the recall by 12.4 percentage points and the precision by 4.1 percentage points; thus, the missing domain was likely not functionally essential but merely omitted. This method is more suitable for conservative annotations, but the calibrated scores still cannot cover the entire structure.

Ablation, Interpretability, and Error Analysis

The same data partitioning and optimization budget are used for component ablation. Removal of structural features reduced Fmax to 0.584 and, upon removal of sequence context, to 0.559. Unconditional sequence-structure fusion reached 0.593; thus, confidence gating added 1.9 percentage points to the simple multimodal integration result. Remove parent-path text, hierarchy regularization, synonym augmentation and frequency reweighting to get 0.598, 0.601, 0.604 and 0.606, respectively. Replace bidirectional cross-attention with vector concatenation and further reduce Fmax to 0.581 and increase expected calibration error to 0.069. The above results show that geometric evidence offers the largest individual gain, and ontology-aware text and reliability gating are used to determine how consistently this evidence is transferred to sparse functions [30].

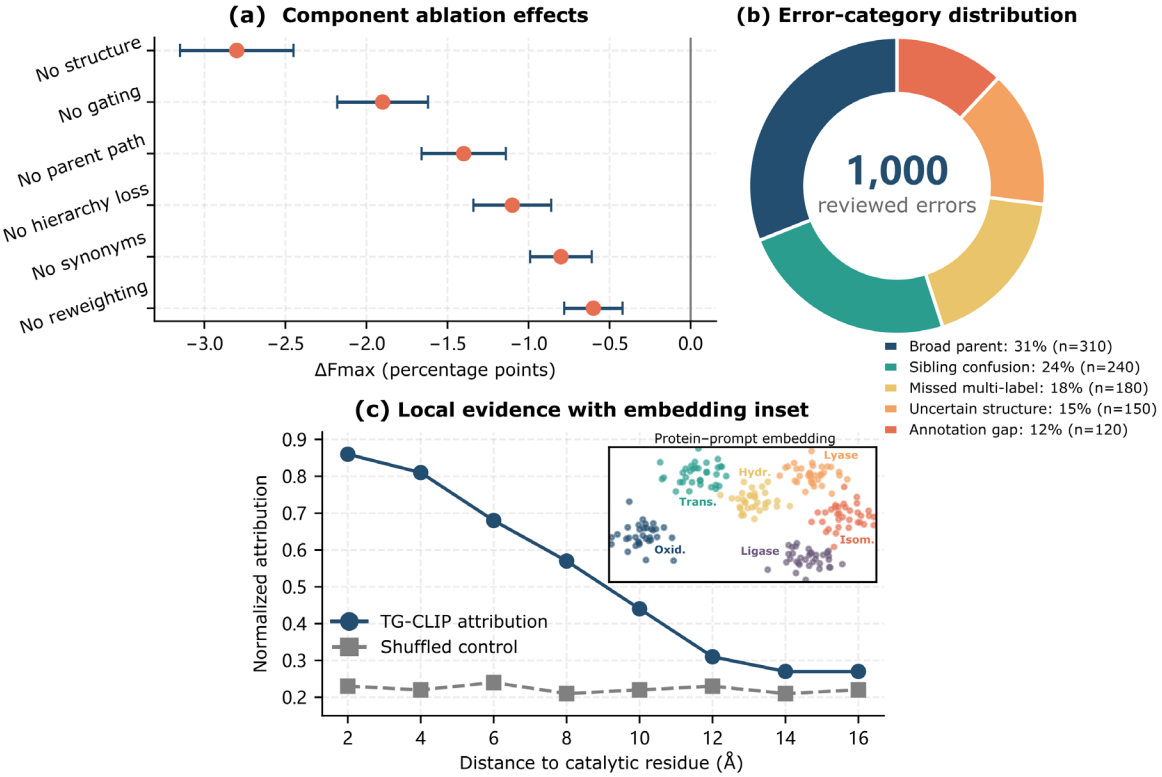

Figure 7. Component ablation, error distribution, and functional interpretation: (a) Fmax changes and bootstrap intervals after removing individual model components; (b) distribution of broad-parent predictions, sibling confusions, missed multifunctional labels, uncertain-structure errors, and annotation mismatches; (c) residue attribution as a function of distance from catalytic residues, with an inset showing the joint protein–prompt embedding distribution.

Figure 7 presents the principal diagnostic results. Figure 7 (a) reports the Fmax changes and bootstrap intervals for six component removals; excluding structure, confidence gating, and parent-path prompts produces decreases of 2.8,1.9, and 1.4 percentage points. Figure 7 (b) partitions 1,000 sampled errors into overly broad

parent predictions (31%), sibling confusions (24%), missed multifunctional labels (18%), uncertain structures (15%), and annotation mismatches (12%). Figure 7 (c) relates residue attribution to catalytic-site distance. Median attribution decreases from 0.81 at 4Å to approximately 0.27 beyond 12Å, whereas the shuffled control remains near 0.22. Its embedding inset also shows partial separation among six major enzyme-function groups, with multifunctional proteins concentrated near shared boundaries [31].

The distribution of the error has two defects. Broad-parent and sibling errors are mainly due to insufficient semantic specificity; the competing prompts only differ by one substrate or reaction qualifier. Structural errors occur more frequently in multi-domain proteins and low-confidence terminal regions. In a remote-homology example with 17% sequence identity, the model correctly identified the metal-dependent activity because three spatially close acidic residues received high attribution. However, due to the relevant domain being located in an undetermined tail, the regulatory protein failed to recognize a rare subterm. A few predictions were biologically plausible but not included in the benchmark record, so they were considered errors and did not need to be reclassified based on post hoc interpretation [32].

The interpretation results are the same under medium structural occlusion. The average Spearman correlation coefficient for residue attribution ranking across the five masks was 0.74; after removing confidence gating, this value dropped to 0.52. Among the proteins with annotated catalytic residues, 63.8% of the ten highest-attribution residues were within 8Å of a catalytic site; the corresponding values for the ungated model and random selection were 54.1% and 29.6%, respectively. The above associations are not biochemical mechanisms; they only indicate whether a prediction is consistent with an adjacent local structure.

Computational profiling showed that the sequence encoder accounted for 68% of training time, the graph encoder for 19%, cross-modal fusion for 9%, and text alignment for 4%. A full run required 18.6 hours on one 40 GB accelerator and 5.3 hours on four devices. Freezing the first half of the sequence transformer reduced peak memory from 31.6 to 23.4 GB with a 0.7 -point Fmax loss. Prompt vectors can be cached, so new ontology terms require text encoding and score calibration rather than complete protein-model retraining. Due to ontology revisions, calibration drift, and changes in structural confidence potentially reducing the accuracy of predictive functions, deployment should continue to monitor these changes [33].

Conclusion

TG-ProteinCLIP is a text-guided contrastive framework for protein function prediction based on fused sequence and structure data introduced in this paper. The model integrates contextual residue representations, confidence-aware geometric message passing, bidirectional gated fusion and ontology-informed prompt encoding in a shared retrieval space. In addition, the other kinds of assessments were general accuracy, remote homology, sparse supervision, structural perturbation, calibration and interpretability, etc., to provide a comprehensive view of the engineering choices in multimodal alignment.

Some Deficiencies Remain. Functional descriptions inherit the ambiguity and lack of specific details in the ontology, and predicted structures may not cover all conformational states or ligand-induced changes. Currently, only single-chain residue geometry is used to build graphs, and complexes, cofactors and the cellular environment are not directly modelled. The zero-shot scores for highly specialized terminology are insufficient for autonomous annotation, and due to the protein language model, the computational cost remains relatively high.

Future work should include protein complexes, molecular surfaces, interaction partners and experimentally determined dynamics in the aligned representation. Adaptive prompts can differentiate catalytic mechanisms and substrate preferences, etc., of biological functions without indiscriminately expanding the label set. Continuous calibration and expert-in-the-loop validation can be employed to further guarantee the reliability of the deployed ontology as new structural evidence is continuously added.

Author Contributions

Matěj Šimánek1 contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Janek Novotný contributes to data collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Saha, S., Chatterjee, P., Basu, S., Nasipuri, M., & Plewczynski, D. (2019). FunPred 3.0: improved protein function prediction using protein interaction network. PeerJ, 7, e6830. https://doi.org/10.7717/peerj.6830

[2] Bonetta, R., & Valentino, G. (2020). Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics, 88(3), 397-413. https://doi.org/10.1002/prot.25832

[3] Törönen, P., & Holm, L. (2022). PANNZER—A practical tool for protein function prediction. Protein Science, 31(1), 118-128. https://doi.org/10.1002/pro.4193

[4] Reijnders, M. J. M. F. (2022). Wei2GO: weighted sequence similarity-based protein function prediction. PeerJ, 10, e12931. https://doi.org/10.7717/peerj.12931

[5] van den Bent, I., Makrodimitris, S., & Reinders, M. (2021). The Power of Universal Contextualized Protein Embeddings in Cross-species Protein Function Prediction. Evolutionary Bioinformatics, 17, 11769343211062608. https://doi.org/10.1177/11769343211062608

[6] Ayoub, R., & Lee, Y. (2021). Protein structure search to support the development of protein structure prediction methods. Proteins: Structure, Function, and Bioinformatics, 89(6), 648-658. https://doi.org/10.1002/prot.26048

[7] The Intelligent Networks and Systems Society (2021). MDPFP-FCNN: Multidomain Protein Function Prediction Using Fuzzy Convolutional Neural Network. International Journal of Intelligent Engineering and Systems, 14(6), 642-655. https://doi.org/10.22266/ijies2021.1231.57

[8] Sengupta, K., Saha, S., Halder, A. K., Chatterjee, P., Nasipuri, M., Basu, S., & Plewczynski, D. (2022). PFP-GO: Integrating protein sequence, domain and protein-protein interaction information for protein function prediction using ranked GO terms. Frontiers in Genetics, 13, 969915. https://doi.org/10.3389/fgene.2022.969915

[9] Dhanuka, R., & Singh, J. P. (2021). Protein function prediction using functional inter-relationship. Computational Biology and Chemistry, 95, 107593. https://doi.org/10.1016/j.compbiolchem.2021.107593

[10] Scheibenreif, L., Littmann, M., Orengo, C., & Rost, B. (2019). FunFam protein families improve residue level molecular function prediction. BMC Bioinformatics, 20(1), 400. https://doi.org/10.1186/s12859-019-2988-x

[11] Kabir, A., & Shehu, A. (2022). GOProFormer: A Multi-Modal Transformer Method for Gene Ontology Protein Function Prediction. Biomolecules, 12(11), 1709. https://doi.org/10.3390/biom12111709

[12] Zhao, C., Liu, T., & Wang, Z. (2022). PANDA2: protein function prediction using graph neural networks. NAR Genomics and Bioinformatics, 4(1), lqac004. https://doi.org/10.1093/nargab/lqac004

[13] Torres, M., Yang, H., Romero, A. E., & Paccanaro, A. (2021). Protein function prediction for newly sequenced organisms. Nature Machine Intelligence, 3(12), 1050-1060. https://doi.org/10.1038/s42256-021-00419-7

[14] TaoLu,, & Huo, S. (2020). Protein function module of deep learning and PPI network prediction. Journal of Physics: Conference Series, 1684(1), 012128. https://doi.org/10.1088/1742-6596/1684/1/012128

[15] Sitani, D., Giorgetti, A., Alfonso-Prieto, M., & Carloni, P. (2021). Robust principal component analysis-based prediction of protein-protein interaction hot spots. Proteins: Structure, Function, and Bioinformatics, 89(6), 639-647. https://doi.org/10.1002/prot.26047

[16] Mansoor, M., Nauman, M., Rehman, H. U., & Omar, M. (2022). Gene Ontology Capsule GAN: an improved architecture for protein function prediction. PeerJ Computer Science, 8, e1014. https://doi.org/10.7717/peerj-cs.1014

[17] Lai, B., & Xu, J. (2022). Accurate protein function prediction via graph attention networks with predicted structure information. Briefings in Bioinformatics, 23(1), bbab502. https://doi.org/10.1093/bib/bbab502

[18] Zhou, J., Xiong, W., Wang, Y., & Guan, J. (2021). Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment. Frontiers in Genetics, 12, 758131. https://doi.org/10.3389/fgene.2021.758131

[19] Giri, S. J., Dutta, P., Halani, P., & Saha, S. (2021). MultiPredGO: Deep Multi-Modal Protein Function Prediction by Amalgamating Protein Structure, Sequence, and Interaction Information. IEEE Journal of Biomedical and Health Informatics, 25(5), 1832-1838. https://doi.org/10.1109/jbhi.2020.3022806

[20] Piovesan, D., & Tosatto, S. C. E. (2019). INGA 2.0: improving protein function prediction for the dark proteome. Nucleic Acids Research, 47(W1), W373-W378. https://doi.org/10.1093/nar/gkz375

[21] Mahapatra, S., & Sahu, S. S. (2022). ANOVA- particle swarm optimization-based feature selection and gradient boosting machine classifier for improved protein–protein interaction prediction. Proteins: Structure, Function, and Bioinformatics, 90(2), 443-454. https://doi.org/10.1002/prot.26236

[22] Wan, C., & Jones, D. T. (2020). Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nature Machine Intelligence, 2(9), 540-550. https://doi.org/10.1038/s42256-020-0222-1

[23] Si, Y., & Yan, C. (2021). Improved protein contact prediction using dimensional hybrid residual networks and singularity enhanced loss function. Briefings in Bioinformatics, 22(6), bbab341. https://doi.org/10.1093/bib/bbab341

[24] Tahzeeb, S., & Hasan, S. (2022). A Neural Network-Based Multi-Label Classifier for Protein Function Prediction. Engineering, Technology & Applied Science Research, 12(1), 7974-7981. https://doi.org/10.48084/etasr.4597

[25] Mansoor, M., Nauman, M., Ur Rehman, H., & Benso, A. (2022). Gene Ontology GAN (GOGAN): a novel architecture for protein function prediction. Soft Computing, 26(16), 7653-7667. https://doi.org/10.1007/s00500-021-06707-z

[26] Ramola, R., Friedberg, I., & Radivojac, P. (2022). The field of protein function prediction as viewed by different domain scientists. Bioinformatics Advances, 2(1), vbac057. https://doi.org/10.1093/bioadv/vbac057

[27] Elhaj-Abdou, M. E. M., El-Dib, H., El-Helw, A., & El-Habrouk, M. (2021). Deep_CNN_LSTM_GO: Protein function prediction from amino-acid sequences. Computational Biology and Chemistry, 95, 107584. https://doi.org/10.1016/j.compbiolchem.2021.107584

[28] Li, S., Cai, C., Gong, J., Liu, X., & Li, H. (2021). A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing. Proteins: Structure, Function, and Bioinformatics, 89(11), 1541-1556. https://doi.org/10.1002/prot.26176

[29] Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., & Atalay, V. (2019). DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. Scientific Reports, 9(1), 7344. https://doi.org/10.1038/s41598-019-43708-3

[30] Schlee, S., Straub, K., Schwab, T., Kinateder, T., Merkl, R., & Sterner, R. (2019). Prediction of quaternary structure by analysis of hot spot residues in protein-protein interfaces: the case of anthranilate phosphoribosyltransferases. Proteins: Structure, Function, and Bioinformatics, 87(10), 815-825. https://doi.org/10.1002/prot.25744

[31] Tsuchiya, Y., & Tomii, K. (2020). Neural networks for protein structure and function prediction and dynamic analysis. Biophysical Reviews, 12(2), 569-573. https://doi.org/10.1007/s12551-020-00685-6

[32] Arsenescu, V., Devkota, K., Erden, M., Shpilker, P., Werenski, M., & Cowen, L. J. (2022). MUNDO: protein function prediction embedded in a multispecies world. Bioinformatics Advances, 2(1), vbab025. https://doi.org/10.1093/bioadv/vbab025

[33] Huang, G. (2019). Computational Models or Methods for Protein Function Prediction. Current Proteomics, 16(5), 352-353. https://doi.org/10.2174/1570164616051905101141117