BioGraphX: Bridging the Sequence-Structure Gap via Physicochemical Graph Encoding for Explainable Subcellular Localization Prediction

Abubakar Saeed[1,*], Waseem Abbas[2,*]

[1]Department of Computer Science, Government College University Faisalabad, Chiniot Campus
Corresponding author: waseem.abbas@gcuf.edu.pk

Abstract: Computational approaches for protein subcellular localization prediction are important for understanding cellular mechanisms and developing treatments for complex diseases. However, a critical limitation of current methods is their lack of interpretability: while they can predict where a protein localizes, they fail to explain why the protein is assigned to a specific location. Moreover, traditional approaches rely on Anfinsen's principle, which assumes that protein behavior is determined by its native three-dimensional structure, requiring costly and time-consuming process. Here, we propose BioGraphX, a novel encoding framework that constructs protein interaction graphs directly from protein sequences using biochemical rules. This approach eliminates the need for three-dimensional structure determination by encoding 151 interpretable features grounded in biophysical principles. Building upon this representation, BioGraphX-Net demonstrates superior performance on the DeepLoc benchmarks by integrating ESM-2 embeddings with the proposed features via a gating mechanism. Gating analysis shows that although ESM-2 embeddings provide strong contributions, BioGraphX features function as high-precision filters. SHAP analysis reveals class-specific feature importance, uncovering a negative selection mechanism in which sequence profiles associated with the Golgi and plastid act as repellors that suppress sequence mimics, thereby requiring confirmation from physical anchoring signals. It has the additional advantage of promoting Green AI in bioinformatics, achieving performance comparable to the state-of-the-art while maintaining a minimal parameter count of 12.2 million. In summary, BioGraphX not only provides accurate predictions but also offers new insights into the language of life.

Keywords: subcellular localization prediction, graph networks, explainable AI, encoding, neural network, constraint graph, Anfinsen principle, Green AI

1 Introduction

Proteins are the most important part of our understanding of living things. Figuring out where they are in the cell is a key part of cellular processes, and it is very important for making sure precise drug discovery and comprehension of disease pathogenesis [Thul et al., 2017, Thumuluri et al., 2022]. In recent years, high-throughput sequencing technologies have advanced rapidly, generating enormous volumes of genetic data. This has led to an explosive increase in the number of discovered protein sequences. The immediate outcome of this imbalance is a vast and growing "dark matter" of uncharacterized proteins, with over 60% of sequences in UniProt lacking experimental data. To illuminate this unknown proteomic space, the field increasingly relies on accurate computational prediction, which has rapidly evolved from a supplementary tool to a central methodology in structural biology and proteomics [Makrodimitris et al., 2020].

Historically, identifying a protein's subcellular destination has required direct experimental observation using techniques like fluorescence microscopy [Kohnhorst et al., 2016, Orbán et al., 2008]. These methods remain the gold standard for real-time, high-quality protein distribution analysis but are limited by high cost and specialized equipment requirements. In the post-genomic era, a widening gap between reviewed and unreviewed entries in the UniProt database has driven the development of computational methods to bridge this gap. [Consortium et al., 2023].

Over time, computational approaches to protein localization have evolved from simple motif-based methods to more advanced deep learning models. Initial studies in this field relied on conserved targeting signals, before gradually incorporating hand-crafted physicochemical features [Lobley et al., 2008]. Several methods have been introduced to enhance localization predictions by integrating different types of biological data. YLoc+, for instance, predicts multiple locations by combining sorting signals [Briesemeister et al., 2010], PROSITE [Sigrist et al., 2013] patterns, and Gene Ontology (GO) terms, whereas Fuel-mLoc [Wan et al., 2016] uses only GO terms from the ProSeq-GO

database to identify protein locations in diverse organisms.

A major milestone was achieved with DeepLoc 1.0 [Almagro Armenteros et al., 2017], which was among the first methods to apply recurrent neural networks (RNNs) to protein sequence modeling. More recently, there has been growing interest in protein language models (pLMs), such as ESM-2 [Lin et al., 2023] and ProtT5 [Elnaggar et al., 2022]. Although most protein language models depend on extensive self-supervision, LAProtT5 [Stärk et al., 2021] introduced a sequence-focused deep learning approach for localization prediction. This transition led to the development of DeepLoc 2.0 [Thumuluri et al., 2022], which delivered significant gains in global accuracy. More recently, LocPro [Zhang et al., 2025] has emerged as a promising method that overcomes several limitations of sequence-only embeddings. LocPro introduces a dual-channel architecture that integrates global protein representations from the ESM-2 language model with 1,484 expert-driven descriptors from PROFEAT. By restructuring these expert features into similarity-based images (PF-IMG) and processing them through a hybrid CNN-BiLSTM network, LocPro has achieved new state-of-the-art results in multi-label prediction across 91 hierarchical levels of granularity.

However, despite the success of sequence-driven hybrids like LocPro, there remains a critical gap in capturing the explicit physical topology of the protein. While LocPro relies on statistical sequence descriptors, BioGraphX-Net shifts the paradigm toward structural graph encoding. By representing proteins as interaction networks, our framework captures invariant "Structural Proxies" that provide a more efficient and physically grounded resolution of localization determinants, particularly in evolutionarily distant proteomes where primary sequence identity is low.

While these pLMs achieve impressive accuracy on benchmark datasets, they suffer from three critical limitations: (1) they function as "computational black boxes" [Hunklinger and Ferruz, 2025], hindering mechanistic interpretation essential for biological hypothesis generation; (2) they often exploit phylogenetic artifacts, species-specific evolutionary patterns rather than universal biophysical principles, leading to poor generalization on evolutionarily distant proteins with $< 30\%$ sequence; and (3) they fail to explicitly model the sequence-structure relationship central to Anfinsen's dogma, where primary sequence determines three-dimensional structure [Anfinsen, 1973]. Although graph representation learning offers a promising way to model protein structures [Gligorijević et al., 2021], it is still constrained by the availability of high-quality 3D data.

Recent work has increasingly focused on incorporating biochemical knowledge into graph-based representations for protein-related tasks. One representative method is PSICHIC [Koh et al., 2024], which predicts binding affinities by constructing a bipartite interaction graph from protein sequences and ligand SMILES. PSICHIC is based on modeling inter-molecular recognition and leverages predicted protein contact maps together with covalent ligand bonds to build a relational scaffold for simulating binding dynamics using a Graph Neural Network (GNN). In contrast, BioGraphX focuses on intramolecular structure and is driven by structural proxies. BioGraphX constructs a homogeneous constraint graph directly from the protein sequence. This construction relies on explicit biophysical rules, such as hydrophobic packing and salt bridges. Importantly, it does not depend on learned structural priors or external ligand information. While PSICHIC focuses on inter-molecular interactions, BioGraphX emphasizes the structural constraints encoded within a protein sequence that determine its cellular localization.

This work addresses three specific questions: (1) Can biophysically-grounded graph encodings bridge the sequence-structure gap without requiring 3D structures? (2) Do such encodings generalize better than pLMs to evolutionarily diverse proteins? (3) Can explicit biophysical rules enable interpretable predictions that reveal mechanistic insights? We introduce BioGraphX, a framework that constructs multi-scale interaction graphs directly from sequences using biochemical constraint rules as illustrated in Figure 1. In contrast to existing hybrid methods that implement biochemical features in a post-hoc manner, training embeddings first and then appending features, the proposed framework implements integration natively at the encoding stage.

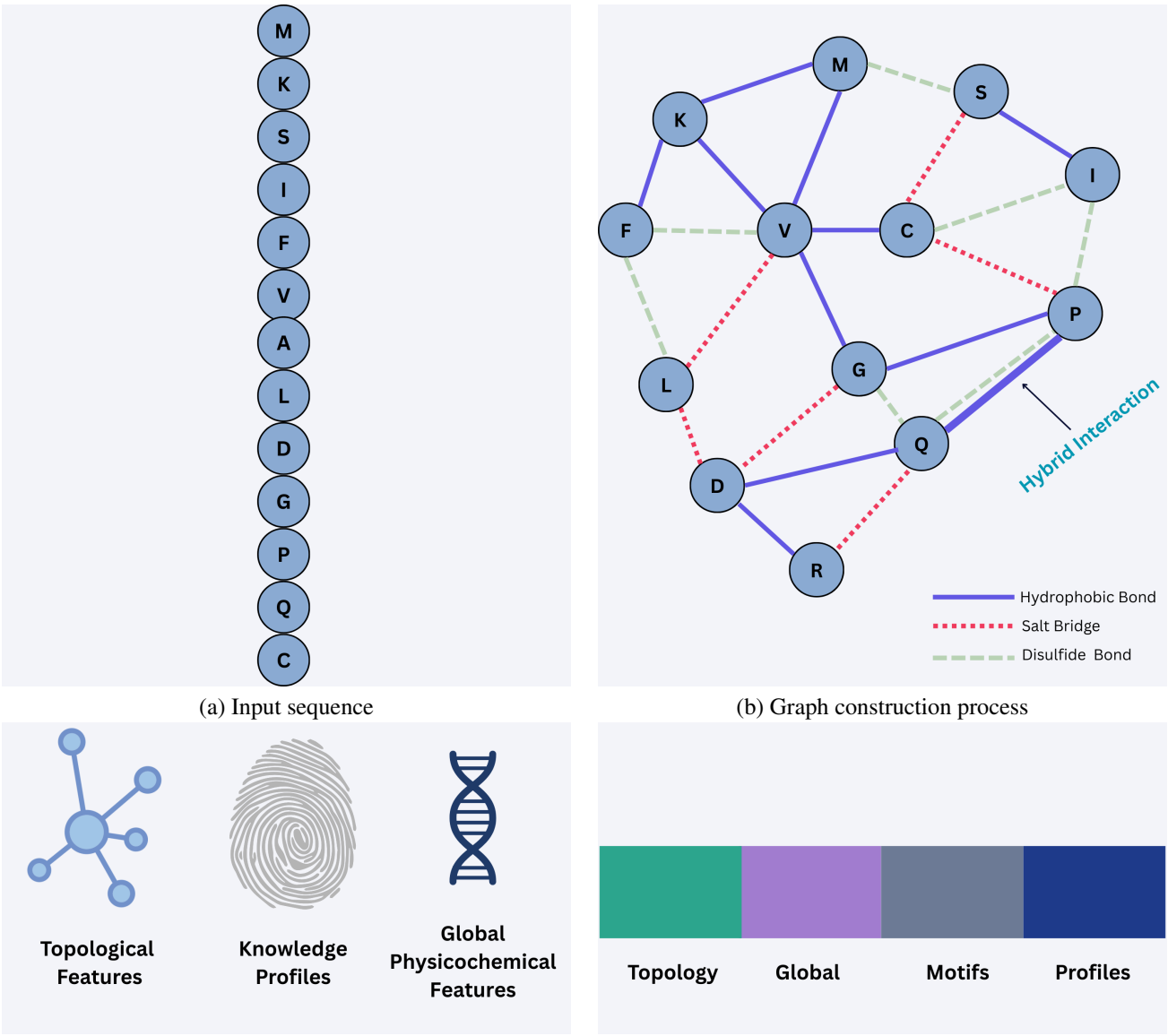
BioGraphX-Net combines these encodings with ESM-2 embeddings through an interpretable gating mechanism that quantifies reliance on evolutionary vs. biophysical signals, enabling transparent decision-making. Unlike models requiring end-to-end fine-tuning of all parameters, our framework maintains the ESM-2 backbone frozen and trains only 12.2M specialized parameters, reducing trainable complexity by two orders of magnitude while maintaining competitive accuracy. This represents a knowledge-driven approach to model efficiency for proteome-scale prediction. Our contributions include: (1) the BioGraphX encoding algorithm; (2) the hybrid BioGraphX-Net architecture; (3) comprehensive benchmarking with superior generalization; and (4) explainable analysis revealing biophysical predictors of localization.

With its reusable, explainable, and structure-aware representation, BioGraphX takes an important step toward modeling the relationship between sequence, structure, and function.

2 Methods

2.1 Datasets and Preprocessing

This study relies on data from two key sources: the DeepLoc 1.0 [Almagro Armenteros et al., 2017] and DeepLoc 2.0 [Thumuluri et al., 2022] benchmark datasets. To ensure data integrity, a series of filtering steps was applied. The primary inclusion criterion required sequences to contain only the 20



(a) Input sequence

(b) Graph construction process

(c) Multi-scale Feature Extraction

(d) Feature Vector

Figure 1: Detailed visualization of BioGraphX framework. (a) Input protein sequences for encoding. (b) Construction of protein residue interaction graphs with edges representing biochemical relationships. (c) Multi-scale feature extraction integrating residue-level, protein-level, and knowledge-profile features. (d) Feature vector representation ready for downstream classification and explainable analysis.

standard amino acids. Accordingly, any entries with unknown or non-standard residues, such as 'X', were excluded from the analysis. Following this filtering process, 13,824 sequences were retained from the initial 13,858 in the DeepLoc 1.0 set, while 28,251 of the original 28,303 sequences in the DeepLoc 2.0 dataset met the inclusion criteria.

Evaluation followed the official DeepLoc 2.0 protocol using 5-fold cross-validation with a strict constraint of $\leq 30\%$ sequence identity between training and test sets, predicting ten eukaryotic subcellular compartments (Cytoplasm, Nucleus, Extracellular, Cell Membrane, Mitochondrion, Plastid, Endoplasmic Reticulum, Lysosome/Vacuole, Golgi Apparatus and Peroxisome). This homology-reduced partitioning ensures that performance differences reflect model capability rather than sequence memorization.

2.2 Overview of BioGraphX-Net Architecture

The BioGraphX-Net architecture is composed of three integrated stages. The first section, BioGraphX encoding, involves the construction of biochemically-calibrated constraint

graphs from protein sequences to facilitate the extraction of multi-scale topological, knowledge-guided, and physicochemical features. Subsequently, ESM-2 sequence embedding is utilized to capture high-dimensional evolutionary context through attention-pooled residue representations. In the final stage, an interpretable gated fusion mechanism dynamically integrates physics-based and evolutionary signals via learned gating weights prior to multi-class classification. A schematic overview of this pipeline is presented in Figure 2.

2.3 Sequence Embedding and Feature Extraction

2.3.1 ESM-2 Embeddings with Attention Pooling

The present study utilizes the ESM-2 model [Lin et al., 2023] to extract residue-level embeddings, providing a dense numerical representation for each sequence. To aggregate these variable-length embeddings into a fixed-length protein representation, an attention pooling mechanism was implemented [Bahdanau et al., 2015]. This technique was employed to selectively weight the importance of specific residues, thereby ensuring that the most relevant biological signals for localization are captured. The raw ESM embeddings $E_{seq} \in \mathbb{R}^{L\times 2560}$ are transformed into a weighted sequence representation E_{vec} according to the following operations:

$$\begin{aligned} A &= \tanh(W_1 E_{\text{seq}} + b_1) \\ \alpha &= \text{softmax}(W_2 A + b_2) \\ E_{\text{vec}} &= \sum_i \alpha_i \, E_{\text{seq},i} \end{aligned} \tag{1}$$

In this model, A represents the attention transformation, while α denotes the masked attention weights assigned to each residue. Following this aggregation, the weighted representation E_{vec} undergoes a bottleneck transformation to reduce dimensionality and facilitate fusion. This reduction to 1024 dimensions is achieved through a fully connected layer integrated with Batch Normalisation (BN) and a Rectified Linear Unit (ReLU) activation:

$$E_{\text{feat}} = \text{ReLU}(\text{BN}(W_3 E_{\text{vec}} + b_3)) \tag{2}$$

where E_{vec} denotes the original high-dimensional embedding generated by the ESM model, W_3 and b_3 are the learnable weight matrix and bias of the bottleneck layer, respectively, $\text{BN}(\cdot)$ represents batch normalization, and $\text{ReLU}(\cdot)$ is the rectified linear unit activation function. The resulting vector E_{feat} corresponds to the compressed semantic feature representation.

The rationale for this bottleneck reduction is to extract the most salient features from the high-dimensional ESM space, preventing the model from overfitting to noise while maintaining a compatible latent space for subsequent integration with biophysical features.

2.4 BioGraphX Encoding Framework

To support proteome-scale analyses, the encoding pipeline implements parallel batch processing using Joblib, significantly improving computational efficiency.

2.4.1 Foundational Biophysical Principles

The BioGraphX encoding framework facilitates the transformation of protein sequences into multi-scale interaction graphs through the application of biochemically-calibrated constraint rules. While pLMs learn statistical patterns from large datasets, BioGraphX uses explicit physical laws supported by experimental data, resulting in deterministic and structure-informed features.

The framework relies on two main biophysical principles. The first, known as Anfinsen's Principle, holds that a protein's tertiary structure is determined by its amino acid sequence

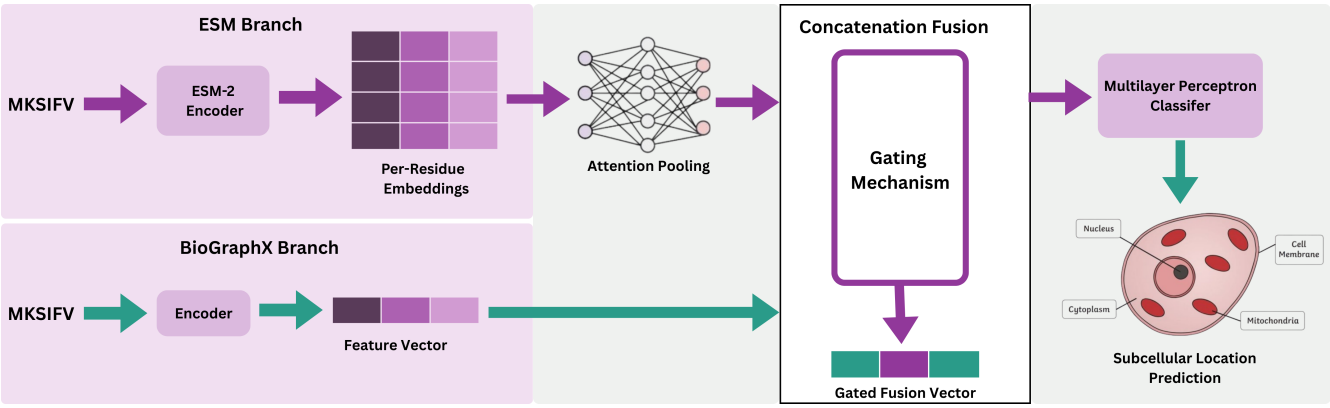


Figure 2: Schematic representation of the BioGraphX-Net computational pipeline. The diagram shows the integration of ESM sequence embeddings and BioGraphX physics-based features through a central gating mechanism, producing the combined Gated Fusion Vector (f_{gated}) for downstream classification.

through encoded structural constraints [Anfinsen, 1973]. BioGraphX implements these constraints explicitly via interaction rules between compatible residues, based on the understanding that specific residue pairs preferentially interact to stabilize fold topology. Secondly, the Linear Distance Assumption for Folded Domains states that within folded protein domains, residues that are proximal in the primary sequence are typically also proximal in three-dimensional space as a result of chain connectivity and folding kinetics [Yan et al., 2013]. Maximum interaction distances were calibrated from the literature, enabling the construction of interaction graphs that serve as structural proxies without the requirement for explicit 3D coordinates.

2.4.2 Graph Construction Algorithm

For a protein sequence S of length L, we construct an undirected weighted graph $G(V, E)$ where vertices $V = \{v_1, v_2, ..., v_L\}$ represent residues and edges E represent biochemical interactions. The complete algorithm is formalized in Algorithm 1.

Algorithm 1 BioGraphX Encoding Pipeline

Require: Protein sequence S of length L
Ensure: Feature vector $F \in \mathbb{R}^{151}$

```
Initialize empty graph G(V, E) with |V| = L
for each residue pair (i, j) with i < j do
    d_ij ← j − i                                  ▷ Linear sequence distance
    for each interaction type t ∈ T (Table 1) do
        if d_ij ≤ d_max,t and residues are compatible for t then
            w_e ← w_t × (1 − d_ij / d_max,t × 0.3)
            Add edge (i, j) with weight w_e and type t
            Store: Interaction t for hybrid detection
        end if
    end for
end for
Add backbone edges (i, i + 1) with w = 1.0, t = backbone
Detect hybrid interactions H (Equation 4)
Extract topological features → F_topo ∈ ℝ^85
Extract hybrid features → F_hybrid ∈ ℝ^27
Compute knowledge-guided scores → F_know ∈ ℝ^20
Calculate global physicochemical properties → F_phys ∈ ℝ^19
return F = [F_topo; F_hybrid; F_know; F_phys]
```

2.4.3 Biochemical Interaction Rules

The identification of the twelve interaction types and the calibration of their corresponding parameters were based on established biophysical literature and physical chemistry principles. It is important to clarify that all distance thresholds specified in this framework refer to linear sequence separation rather than three-dimensional spatial proximity. As Table 1 shows, these rules cover many types of forces, from the main stabilizing effects, like hydrophobic interactions [Chothia, 1974, Dill, 1990], to specific polar–aromatic motifs, including sulfur–π interactions [Valley et al., 2012]. The chosen distance thresholds and strength weights (w_t) help the graph capture the hierarchical structure of protein folding and stability based on the sequence. Specifically, distance thresholds were set to capture the maximum sequence separation over which these forces typically act to maintain structural integrity, while weights were assigned to represent the relative contribution of each interaction type to biochemical stability. By applying these deterministic rules, BioGraphX creates a representation grounded in biophysics, serving as a robust stand-in for 3D structural features.

To enhance these representations, we assigned a strength weight $w_t \in [0, 1]$ to each interaction type t and introduced a distance decay function to account for the reduced impact of interactions across longer sequence separations. The weight w_e of a given edge E is then computed as follows:

$$w_e = w_t \times \left(1 - \frac{d_{ij}}{d_{\max,t}} \times 0.3\right) \tag{3}$$

In this model, d_{ij} is the sequence distance between residues i and j, and $d_{max,t}$ is the maximum allowed separation for the interaction type t. A constant factor of 0.3 was applied to ensure that, even at the maximum threshold, a baseline interaction strength of approximately 70% is retained. Ablation studies show that this choice matters: removing decay reduced accuracy by 1.44%, and increasing the factor to 0.5 caused a 0.5% drop. This approach preserves a detectable signal for functional contacts despite increased linear separation, effectively balancing long-range constraint detection with noise reduction. This helps preserve functional contact signals across long sequence separations while mitigating noise. The weighting approach makes the graph sensitive to both the presence of a bond and how close the interacting residues are, providing a strong stand-in for structural features even without 3D coordinates.

2.4.4 Hybrid Interaction Detection

A key innovation of the framework is the detection of hybrid interactions, where two interaction types occur simultaneously between the same residue pair. These interactions serve as high-fidelity indicators of specialized structural motifs.

$$H = \{(t_1, t_2) \mid t_1, t_2 \in T,\ t_1 \neq t_2\}. \tag{4}$$

We define nine hybrid interaction types, assigning them increased contributions based on their biological significance, as summarized in Table 2.

Table 1: **Biochemical Constraint Rules Used in BioGraphX**

Constraint Type	Participating Residues	Strength	Max Dist	Key References
Hydrophobic	A, V, L, I, M, F, W, Y, C	1.0	20	[Chothia, 1974]
Hydrogen Bond	Donors: N, Q, S, T, Y, H, K, R, W Acceptors: D, E, N, Q, S, T, Y	0.8	12	[Baker and Hubbard, 1984]
Salt Bridge	Positive: K, R, H Negative: D, E	0.7	35	[Kumar and Nussinov, 1999]
Disulfide Bond	C–C	0.9	2000	[Thornton, 1981]
π-Interaction	Aromatic: F, Y, W Positive: R, K, H	0.6	15	[Burley and Petsko, 1985, Dougherty, 1996]
Cation-π	Positive: K, R, H Aromatic: F, Y, W	0.65	10	[Gallivan and Dougherty, 1999]
Van der Waals	A, V, L, I, M, F, W, Y, C, P	0.4	6	[Dill, 1990]
CH-π	Donors: A, V, L, I, M, C Acceptors: F, Y, W	0.3	8	[Nishio et al., 2014]
NH-π	Donors: N, Q, S, T, H, K, R, W Acceptors: F, Y, W	0.5	8	[Nishio et al., 2014]
Carbonyl–Carbonyl	D, E, N, Q, S, T, Y	0.35	8	[Fufezan et al., 2010]
Sulfur-π	Sulfur: M, C Aromatic: F, Y, W	0.45	8	[Valley et al., 2012]
Backbone Interaction	Peptide backbone atoms	1.0	1	[Pauling et al., 1951]

Table 2: **Hybrid Interaction Types**

Hybrid Interaction	Biological Significance
SaltBridge–HBond	Common in enzyme active sites
Hydrophobic–Pi	Membrane protein anchoring
Disulfide–HBond	Extracellular protein stability
Cation–Pi–HBond	Nuclear localization signals
Hydrophobic–VDW	Hydrophobic core packing
Pi–Cation–HBond	DNA-binding interfaces
CH–Pi–Hydrophobic	Lipid bilayer interactions
Sulfur–Aromatic	Redox-active sites
Carbonyl–Charge	Electrostatic clusters

2.4.5 Multi-Scale Feature Extraction

BioGraphX generates 151 features in four groups: topological, hybrid, knowledge-guided, and global physicochemical. Topological features show network structure, hybrid features capture overlapping biochemical interactions, knowledge-guided scores use known targeting motifs, and global physicochemical features describe general protein characteristics. Hybrid interactions, in which two interaction types occur simultaneously between the same residue pair, are assigned weight multipliers according to their functional significance, as summarized in Table 3.The complete set of extracted features, along with their definitions, is documented in Supplementary Material S1.

2.4.6 Sequence Adaptive Processing

Variable-length proteins were handled using a three-tier adaptive processing strategy. This method balances computational efficiency with preservation of biologically important information and is categorized by sequence length (L) as follows

- **Short sequences ($L \leq 2,000$):** These proteins were processed using their full sequences, as the computation required was considered negligible.
- **Medium sequences ($2,000 < L \leq 10,000$):** The sequences were truncated smartly, keeping the first 40% , middle 20%, and last 40% of each sequence. The rationale for this specific split is that the N- and C-termini typically contain the most critical signalling and targeting motifs, respectively.
- **Long sequences ($L > 10,000$):** A sliding window approach was employed. Feature aggregation for these sequences was carried out using an information content-weighted mechanism to prioritize high-entropy regions and ensure that functionally dense segments are represented.

The biochemical constraint graph encoding algorithm introduced in this study differs fundamentally from existing approaches in several key respects. Whereas traditional methods are generally limited to simple compositional features or require the acquisition of external structural data, contemporary approaches typically rely on "black-box" learned embeddings. BioGraphX distinguishes itself from these models by implementing explicit biochemical rules validated by experimental data, as shown in Table 4.

Table 3: **Feature Scales Used in BioGraphX**

Feature Scale	# Features	Description	Key Metrics
Topological	85	Graph metrics quantifying network architecture	Node/edge counts, centrality, path metrics, modularity, communities
Hybrid	27	Co-occurrence patterns of interaction types	SaltBridge-HBond frequency, regional hybrid density
Knowledge-Guided	20	Features derived from known biological motifs	Regular expressions for NLS, targeting motifs
Physicochemical	19	Whole-protein physicochemical properties	Isoelectric point, GRAVY, entropy

2.5 Hybrid Fusion Architecture

2.5.1 Dual-Branch Network Design

The hybrid architecture is composed of two specialized input branches: the first processes ESM-2 embeddings, while the second handles BioGraphX features. These branches are merged with an interpretable gating mechanism (Figure 2). Meanwhile, the BioGraphX feature vector $B \in \mathbb{R}^{151}$ is projected into the same 1,024-dimensional space as the sequence embeddings using the following transformation:

$$B_{\text{feat}} = \text{ReLU}(\text{BN}(W_4 B + b_4)) \tag{5}$$

The projection aligns evolutionary and biophysical features on the same scale, helping maintain balanced information flow and stable training.

2.5.2 Interpretable Gated Fusion Mechanism

The fusion stage uses a gating mechanism to measure how much the model depends on either the language model or the biophysical features. The features are merged into a joint vector $C \in \mathbb{R}^{2048}$ the gating signal G is computed using the equations below:

$$\begin{aligned} G &= \sigma(W_5 C + b_5) \\ C_{\text{gated}} &= C \odot G \end{aligned} \tag{6}$$

G represents the learned gating weights, and $\odot$ means element-wise multiplication. This lets the model dynamically choose which features to emphasize for each protein, helping it generalize across different structural types. By implementing this selective fusion, the framework avoids the information redundancy often associated with simple concatenation, allowing for a more nuanced interpretation of how different biophysical signals contribute to localization accuracy.

2.5.3 Multi-layer Classifier

The final classification stage comprises a multi-layer perceptron (MLP) that maps the gated features to the subcellular compartment labels. The architecture consists of a 1024-dimensional hidden layer with Batch Normalisation, followed by a 512-dimensional layer and an output layer representing the 10-class logits:

$$\begin{aligned} H_1 &= \text{ReLU}(\text{BN}(W_6 C_{\text{gated}} + b_6)) \\ H_2 &= \text{ReLU}(W_7 H_1 + b_7) \\ \hat{Y} &= W_8 H_2 + b_8 \end{aligned} \tag{7}$$

where C_{gated} is the integrated feature vector, W_i and b_i are learnable weights and biases, $\text{BN}(\cdot)$ denotes batch normalization, $\text{ReLU}(\cdot)$ is the rectified linear unit, H_1 and H_2 are hidden layer activations, and $\hat{Y}$ is the final prediction. The resulting model contains approximately **12.2 million param-**

Table 4: **Encoding Strategy Comparison for Protein Function Prediction**

Encoding Type	Key Characteristics	Advantages	Limitations
Traditional AAC	Simple amino acid frequency statistics	Interpretable, fast, low cost	Limited expressiveness, ignores order
PSSM-based	Evolutionary conservation from MSAs	Captures evolutionary info	Requires MSAs, expensive
Learned embeddings (ESM, ProtBERT)	Deep sequence representations	High accuracy	Black-box, poor generalization
GNN (structure-based)	Graph networks on 3D structures	Captures spatial relationships	Requires 3D structures, intensive
BioGraphX (our method)	**Biochemically-calibrated constraint graphs**	**Interpretable, sequence-only, multi-scale**	**Assumes linear proximity**

eters, achieving a balance between high representational capacity and computational efficiency.

2.6 Training Protocol and Optimization

The classification thresholds were optimized by maximising the Matthews Correlation Coefficient (MCC) [Gorodkin, 2004] for each individual class. The optimization procedure employed binary cross-entropy loss with class weighting to mitigate class imbalance and was carried out using the AdamW optimiser with an initial learning rate of $\eta = 1 \times 10^{-5}$. Class weights were computed as the inverse of the class frequencies to ensure balanced contributions from rare compartments, such as peroxisomes ($< 5\%$ of the training data), relative to abundant compartments, such as the cytoplasm ($\sim 40\%$).

The training protocol comprised the following procedures: (1) five-fold cross-validation was employed while maintaining $\leq 30\%$ sequence identity between folds; (2) an early stopping strategy with a patience of 10 epochs was implemented; (3) a batch size of 64 was used; and (4) the maximum number of training epochs was set to 70. The final model was selected based on the highest observed validation performance.

2.7 Explainability Framework

BioGraphX makes its predictions interpretable through four mechanisms, letting researchers analyze decisions at the level of each residue.

- **Gate Analysis:** Looking at the gating weights G shows how much each protein relies on evolutionary versus biophysical features, making it easy to see which type dominates.
- **SHAP Feature Importance:** We used the SHAP framework with a GradientExplainer to measure how the 151 BioGraphX features contribute to different organelle classes. For the BioGaphX branch, a "Wrapper" setup was used to isolate the BioGraphX features' contribution while keeping the ESM embeddings fixed. This method lets us isolate the effect of biophysical constraints from evolutionary information. For each protein i, feature j, and class k, SHAP values (ϕ_{ijk}) were calculated to show how much each feature affects the prediction for that class. The mean absolute SHAP values were then calculated across 5,442 validation proteins using the following formula:

$$\text{Importance}_{jk} = \frac{1}{N} \sum_{i=1}^{N} |\phi_{ijk}| \tag{8}$$

  Furthermore, the directional impacts of these features were determined based on the sign of the mean SHAP values. Positive SHAP values show features that help proteins go to a particular organelle, and negative values show features that push them away. This two-way analysis gives a clear picture of how biophysical features guide localization decisions.
- **Knowledge-Guided Score:** Directly mapping model features to known targeting motifs, such as nuclear localization signals, mitochondrial targeting peptides, or signal peptides, enables biologically grounded hypothesis generation about the mechanisms underlying the predictions.
- **Hybrid Pattern Visualization:** Areas with clusters of specific interaction patterns, like SaltBridge–HBond motifs near active sites, are highlighted to show where key structural features occur.

2.8 Evaluation Metrics

We assessed the model using standard multi-label metrics to measure how accurate and robust the framework is. The following parameters were used:

- **Accuracy:** The fraction of cases where the predicted labels exactly match the true labels..
- **Jaccard Index:** This looks at how similar the predicted and true label sets are, helping check accuracy for secondary localizations.
- **Micro F1:** This measures the overall balance of precision and recall across all protein–label pairs.
- **Macro F1:** This measures the average per compartment so that even rare classes affect the total score fairly.
- **Matthews Correlation Coefficient (MCC):** This measures performance for each location so that class imbalance does not skew the results [Gorodkin, 2004].

3 Results

3.1 Baseline Validation and Ablation Analysis

To test the encoding without the neural network, we ran an ablation study using XGBoost with just BioGraphX features and no ESM embeddings. This gave an accuracy of 0.6515, showing that the graph encoding alone is highly informative. Then we tested BioGraphX-Net on the DeepLoc 1.0 dataset to compare it with earlier sequence-based models. BioGraphX-Net showed a clear improvement, achieving a global accuracy of 0.8382 and a Gorodkin score of 0.8059, compared with 0.7797 and 0.7347 reported for the original model. Looking at MCC for each organelle, the model does especially well on hard-to-predict locations with sparse data. MCC almost doubled for Lysosome/Vacuole (0.40 vs. 0.19) and Peroxisome (0.62 vs. 0.32), showing that multi-scale interaction graphs

Table 5: **Performance Comparison Across Localization Prediction Methods**

Metric	YLoc+	DeepLoc 2.0 (ESM1b)	DeepLoc 2.0 (ProtT5)	LocPro	BioGraphX-Net (Ours)
Prediction Type	Multi	Multi	Multi	Multi	Multi
Num. Labels	1.57 ± 0.02	1.27 ± 0.02	1.26 ± 0.02	-	1.44 ± 0.02
Accuracy	0.32 ± 0.02	0.53 ± 0.02	$\mathbf{0.55 \pm 0.02}$	-	$\mathbf{0.55 \pm 0.02}$
Jaccard	0.50 ± 0.01	0.68 ± 0.01	0.69 ± 0.01	-	$\mathbf{0.72 \pm 0.01}$
Micro F1	0.56 ± 0.01	0.72 ± 0.01	0.73 ± 0.01	-	$\mathbf{0.77 \pm 0.01}$
Macro F1	0.42 ± 0.01	0.64 ± 0.01	0.66 ± 0.01	-	$\mathbf{0.68 \pm 0.01}$
MCC per location (↑)					
Cytoplasm	0.38 ± 0.02	0.61 ± 0.01	0.62 ± 0.01	0.612 ± 0.002	$\mathbf{0.63 \pm 0.01}$
Nucleus	0.42 ± 0.02	0.66 ± 0.02	0.69 ± 0.01	$\mathbf{0.718 \pm 0.002}$	0.70 ± 0.02
Extracellular	0.61 ± 0.05	0.85 ± 0.03	0.85 ± 0.04	$\mathbf{0.892 \pm 0.003}$	0.86 ± 0.03
Cell membrane	0.44 ± 0.02	0.64 ± 0.01	$\mathbf{0.66 \pm 0.01}$	$\mathbf{0.662 \pm 0.006}$	$\mathbf{0.66 \pm 0.02}$
Mitochondrion	0.47 ± 0.02	0.73 ± 0.03	0.76 ± 0.02	0.754 ± 0.003	$\mathbf{0.77 \pm 0.02}$
Plastid	0.72 ± 0.02	0.88 ± 0.01	$\mathbf{0.90 \pm 0.01}$	0.816 ± 0.003	$\mathbf{0.90 \pm 0.01}$
Endoplasmic reticulum	0.17 ± 0.04	0.52 ± 0.01	0.56 ± 0.03	0.533 ± 0.007	$\mathbf{0.57 \pm 0.02}$
Lysosome / Vacuole	0.07 ± 0.03	0.24 ± 0.03	0.28 ± 0.04	$\mathbf{0.341 \pm 0.027}$	0.30 ± 0.07
Golgi apparatus	0.11 ± 0.04	0.36 ± 0.06	0.34 ± 0.05	0.373 ± 0.003	$\mathbf{0.42 \pm 0.07}$
Peroxisome	0.05 ± 0.02	0.48 ± 0.05	0.56 ± 0.08	0.524 ± 0.017	$\mathbf{0.63 \pm 0.05}$

Note: Performance values are taken from the DeepLoc 2.0 publication [Thumuluri et al., 2022], while results for LocPro are reported as presented in the LocPro study [Zhang et al., 2025]. All models were evaluated on the DeepLoc 2.0 benchmark dataset under identical experimental conditions. Bold values indicate the best performance.

help capture structural details that sequence-only models often miss.

3.2 Benchmarking Against State-of-the-Art

BioGraphX-Net outperforms DeepLoc 2.0 and LocPro (2025) on key multi-label metrics (Table 5). It achieves a Micro-F1 of 0.77, higher than DeepLoc 2.0 (0.73) and LocPro (0.76). Higher Jaccard (0.72) and Macro-F1 (0.68) scores also show better handling of multi-label predictions.

While LocPro uses ESM-2 together with 1,484 sequence-based PROFEAT descriptors, BioGraphX-Net achieves similar overall performance and better results on hard organelles using just 151 structural graph features. Importantly, BioGraphX-Net trains only 12.2M parameters, compared with over 3B parameters in typical full fine-tuning approaches.

3.2.1 Organelle-Specific Performance and Sensitivity

A detailed per-location analysis of the Matthews Correlation Coefficient (MCC) highlights the distinct mechanistic advantages of structural graph encoding over sequence-only and expert-descriptor models. In particular, BioGraphX-Net shows superior resolution for organelles that are traditionally difficult to classify due to sparse data and dynamic interactions.

Challenging Compartments: For the Golgi apparatus, the framework achieved an MCC of 0.42, surpassing both DeepLoc 2.0 (0.34) and LocPro (0.373). Similarly, the model reached an MCC of 0.63 for the Peroxisome, which is significantly higher than the 0.524 reported for LocPro and 0.56 for DeepLoc.

Targeting Signal Identification: In mitochondrial resolution, BioGraphX-Net maintains a state-of-the-art advantage with an MCC of 0.77, exceeding the LocPro ESM2+PROFEAT hybrid (0.754).

Standardized Benchmarks: Performance on the Nucleus (0.70) and Lysosome/Vacuole (0.30) remains highly competitive on standardized DeepLoc benchmarks.

Regarding the multi-label prediction, BioGraphX-Net identified an average of 1.44 locations per protein, compared to a benchmark value of 1.27. The significantly higher Jaccard Index (0.72 vs. 0.69) and Micro F1 score (0.77 vs. 0.73) indicate that the framework is substantially more successful at capturing the full set of co-localizations when the primary location is correctly identified. These findings suggest that BioGraphX-Net is capable of identifying secondary localizations that are frequently masked by the conservative thresholds of pure sequence-only models.

3.2.2 Generalization on Independent HPA Dataset

To assess the generalizability of the framework beyond the training distribution, BioGraphX-Net was benchmarked on an independent test set from the Human Protein Atlas (HPA; $N = 1{,}717$). This dataset was utilized as a rigorous blind test, as it is composed exclusively of proteins with $< 30\%$ sequence identity to the SwissProt training data, the critical threshold at which pure sequence-based models typically exhibit sharp performance degradation. Table 6 provides a performance comparison on the HPA dataset. While

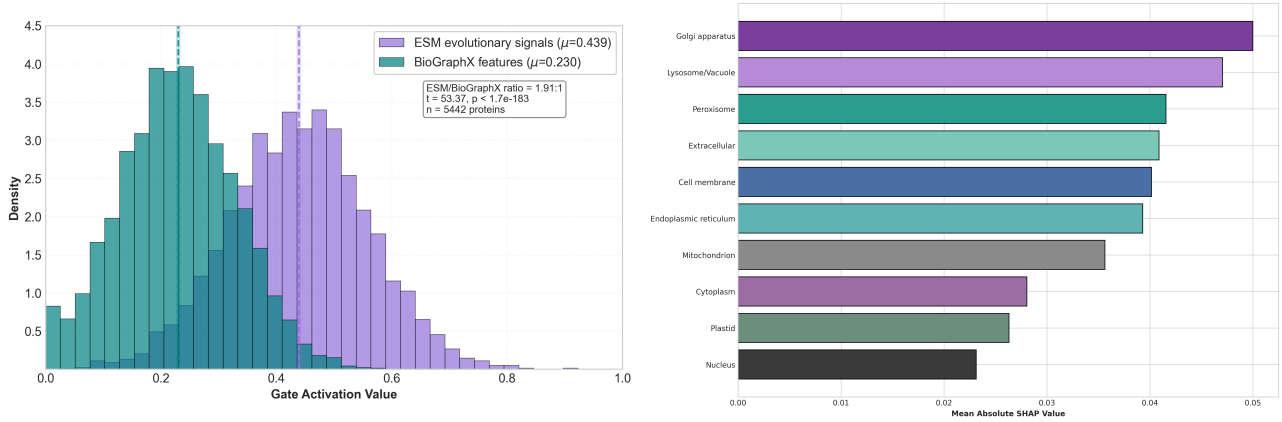


Figure 3: (A)Distribution of dynamic gating weights. The gating mechanism assigns consistently higher activation to sequence embeddings while maintaining a substantial contribution from BioGraphX features, demonstrating complementary feature fusion. (B)The hierarchy of reliance on BioGraphX features, calculated using the mean absolute SHAP attribution.

protein language models (pLMs) are a dominant feature of sequence-motif tasks, BioGraphX appears to provide a necessary correction for structure-driven localization. BioGraphX-Net demonstrates competitive robustness, achieving a Micro F1 score of 0.55 and predicting a conservative average of 1.09 labels per protein, which represents a more biologically realistic profile than over-predicting baselines.

Crucially, the performance of BioGraphX-Net was found to be comparable to the ESM1b-based DeepLoc 2.0 on the mitochondrial localization task (MCC = 0.60), while outperforming both ProtT5-based DeepLoc 2.0 (MCC = 0.56) and LAProtT5 (MCC = 0.59). This compartment is notoriously difficult to resolve because mitochondrial targeting signals are highly variable across species, rendering simple memorization ineffective. These results show that BioGraphX captures key physicochemical principles behind organelle recognition. This Physics-First Advantage suggests that explicit biophysical encoding can match the performance of models that are much larger in size. By learning biophysical principles, such as transit-peptide hydropathy, topological constraints, and charge distribution, rather than relying solely on sequence statistics, BioGraphX generzlizes effectively to evolutionarily distant proteins.

3.2.3 Explainability Analysis

3.2.4 Gating Mechanism: Evolutionary versus Biophysical Signal Weighting

Analysis of gate activations across a subset of 5442 validation proteins revealed a significant preference for evolutionary features ($t = 53.37, p < 1.7 \times 10^{-183}$). Specifically, the ESM branch accounts for 65.6% of the gated information flow, whereas the physics branch contributes 34.4%, resulting in an ESM-to-physics gate ratio of 1.91:1. These results indicate that the framework consistently prioritizes evolutionary representations during the feature fusion process while maintaining a substantial biophysical input as a complementary signal ,as illustrated in Figure 3A.

Table 6: Performance comparison on the HPA independent test set.

Location	YLoc+	DeepLoc 1.0	Fuel-mLoc	LAProtT5	DeepLoc 2.0 ESM1b	DeepLoc 2.0 ProtT5	BioGraphX-Net (Ours)
Pred. Num. Labels (Actual: 1.22)	1.44	0.89	1.00	0.94	1.15	1.21	1.09
Accuracy	0.23	0.37	0.38	0.45	0.34	0.39	0.37
Jaccard	0.41	0.42	0.46	0.52	0.48	0.53	0.50
MicroF1	0.51	0.46	0.52	0.56	0.57	0.60	0.55
MacroF1	0.34	0.35	0.39	0.43	0.44	0.46	0.43
MCC per location (↑)							
Cytoplasm	0.14	0.23	0.23	0.33	0.29	**0.36**	0.30
Nucleus	0.20	0.28	0.41	**0.45**	0.41	0.44	0.41
Cell membrane	0.20	0.23	0.32	0.30	0.34	**0.36**	0.34
Mitochondrion	0.37	0.39	0.33	0.59	**0.60**	0.56	**0.60**
Endoplasmic reticulum	0.12	**0.23**	0.14	0.22	0.20	0.17	0.14
Golgi apparatus	0.08	0.10	0.24	0.26	0.17	**0.31**	0.27

Note: Performance values for comparison methods are taken from DeepLoc2.0 [Thumuluri et al., 2022]) . All models were evaluated on the same Human Protein Atlas (HPA) independent test set. Bold values indicate the best score.

3.2.5 SHAP Analysis

To elucidate the biophysical rules governing the predictions of BioGraphX-Net, a comprehensive SHapley Additive exPlanations (SHAP) analysis was conducted on a held-out validation set of 5,442 proteins. This analytical approach quantifies the directional impact of each of the 151 BioGraphX features across the ten eukaryotic compartments, thereby revealing compartment-specific biophysical "codes". This analysis also helps explain protein localization by showing which biochemical rules influence predictions for different compartments. The hierarchy of physics-based feature usage is shown in Figure 3B.

3.2.6 Directional Impact of Biophysical Features

The directional impact matrix shown in Figure 5A highlights the core exclusion principles captured by the framework. An interesting result from this analysis is that the Plastid Profile descriptor repels its own class but strongly attracts the cell membrane. This result suggests that the model learns to correct for evolutionary mimicry. Since plastids originated from cyanobacterial endosymbiosis [El-Kafafi et al., 2005], their membrane proteins resemble bacterial-like plasma membrane proteins [McFadden, 2014]. The model reduces false positives by penalizing generic sequence similarity and requiring specific transit peptide signals to predict plastid localization. It also identifies Hydrophobic External Edges as a key signal for extracellular proteins, supporting the idea that high surface hydrophobicity drives secretion [Loo et al., 2014].

3.2.7 Characterisation of Biophysical Switch Features

An important result of this analysis is the discovery of switch features that help separate cellular compartments by promoting localization to some organelles while blocking others (Figure 5B). The key descriptor, Isoelectric Point (pI), helps proteins localize extracellularly while preventing their import into peroxisomes [González-Gordo et al., 2022]. This makes sense biologically: secreted proteins are usually acidic to stay soluble outside the cell [Loo et al., 2014], while peroxisomal proteins are basic to allow efficient import by Pex5. Likewise, the Profile Golgi Apparatus Residue feature attracts nuclear proteins but repels plastid proteins, hinting at an evolutionary split between secretion and chloroplast import. Overall, these switch features help prevent mislocalization by acting as a "structural veto," enhancing targeting specificity. This allows the model to represent the relationship between sequence, structure, and function more accurately than simple attraction signals alone.

3.2.8 Golgi Apparatus Targeting Determinants

Looking closely at Golgi targeting, the framework uses a two-layered Physics-Vetting strategy to accurately handle this challenging compartment. Figure 4A shows how this mechanism helps separate the complex localization signals of the secretory pathway.

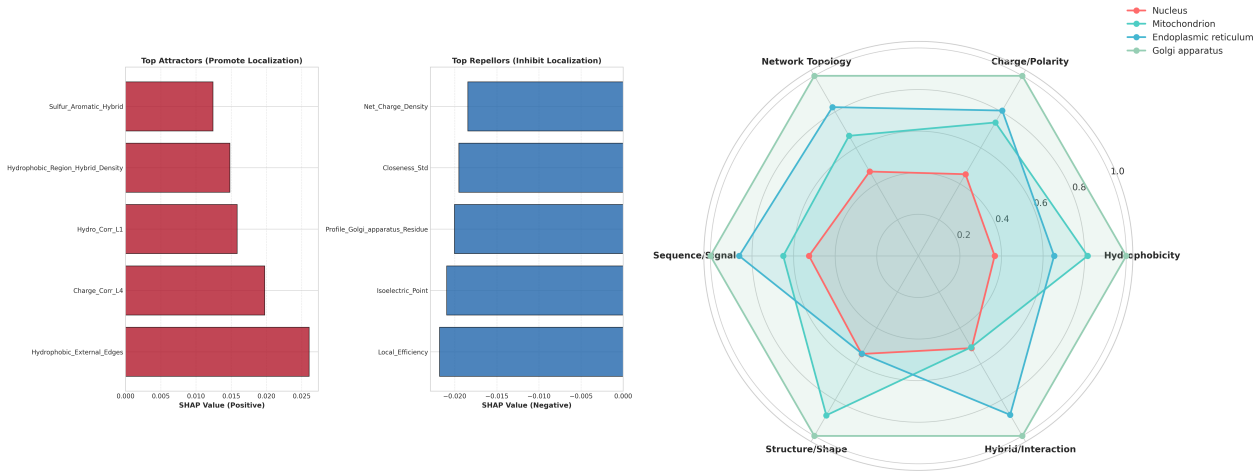


Figure 4: (A) Top Attractors and Repellors for the Golgi Apparatus. Bar chart ranking the five strongest attractors (positive SHAP) and repellors (negative SHAP) for Golgi apparatus localization (B)Radar chart summarizing the mean absolute SHAP contribution of six biophysical feature categories (Hydrophobicity, Charge, Network, Sequence motifs, Structural compactness, Hybrid interactions) across four representative organelles.

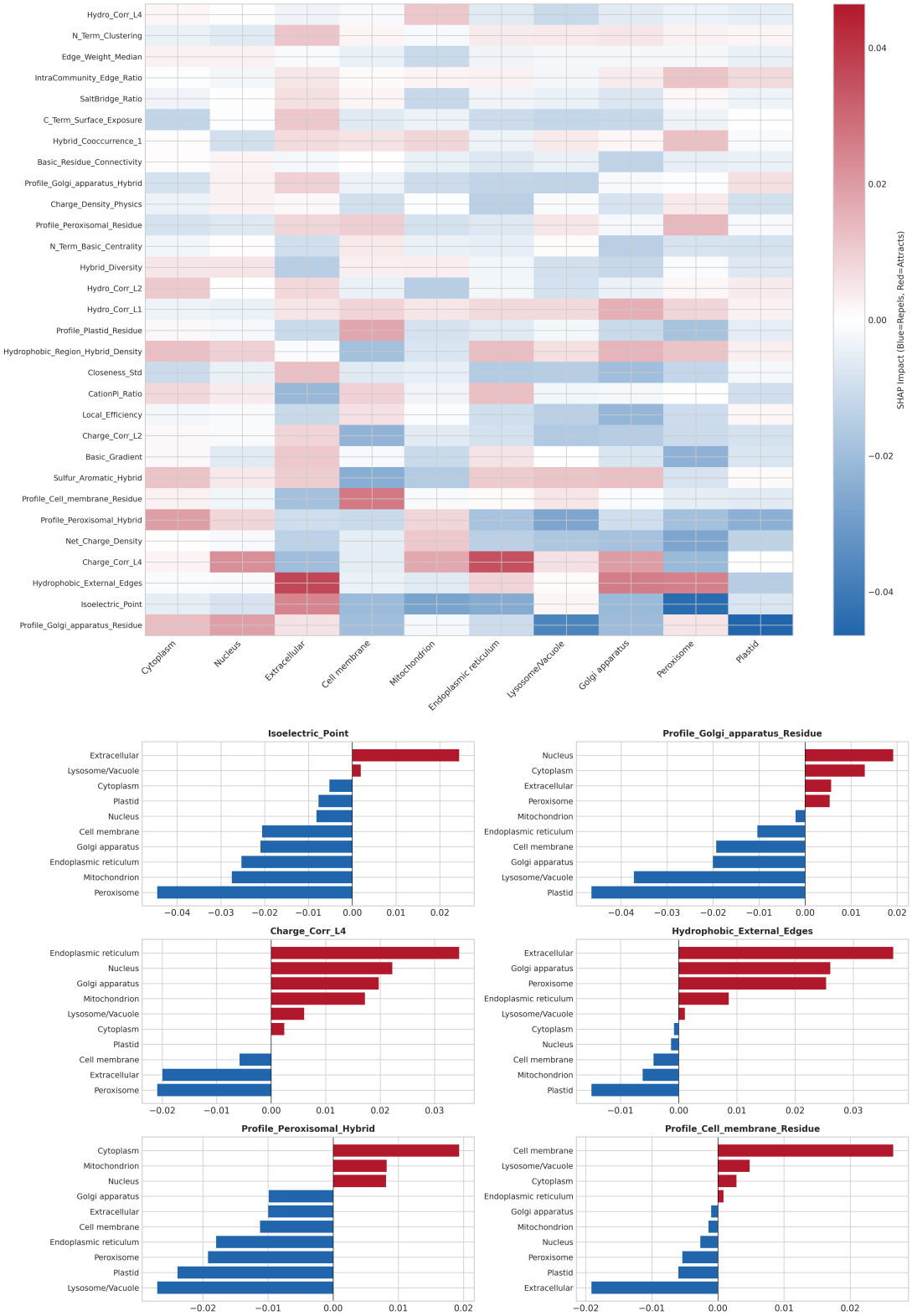


Figure 5: (A) Directional Impact Heatmap showing the mean SHAP value (attraction: red; repulsion: blue) for the top 30 BioGraphX features across ten eukaryotic compartments.
(B) Bar chart highlighting eight switch features that exert opposite SHAP impacts in different organelles.

Layer 1: Negative Selection via Sequence Profiles At the first stage, the model uses Negative Selection on the sequence profile, and the Profile Golgi Residue feature acts as a repellent. This seems to act as a filter to block "nuclear impostors." Many nuclear proteins and transcription factors have disordered regions that look statistically similar to Golgi enzymes [Munro, 1998]. By downweighting generic sequence similarities, the model lowers the chance of false positives that frequently affect sequence-only methods.

Layer 2: Physical Confirmation via Structural Proxies To bypass the initial Negative Selection, the model depends on structural cues. Hydrophobic External Edges and Sulfur–Aromatic Stacking interactions play a major role as attractors [Navarro and Cheeseman, 2022]. These features help the model identify single transmembrane anchors and stabilizing motifs needed for Golgi localization. As a result, it captures physical topology information that sequence-only models often overlook.

Structural Refinement and Flexibility Constraints Conversely, when Local Efficiency is high, it repels proteins, showing the model's attention to Golgi enzyme biophysical constraints. Proteins with flexible stalk regions have less rigid interaction graphs. This shows that BioGraphX-Net doesn't use the same rules for all proteins, it learns specific biophysical features for each compartment to make accurate predictions.

3.2.9 Organelle-Specific Biophysical "Fingerprints"

Radar plots of organelle affinities show that every compartment has its own distinct biophysical "fingerprint.". While the Golgi apparatus functions as a biophysical "generalizt," leveraging a diverse array of hydrophobicity and charge signals, the profile of the Nucleus is dominated by Network Topology features, as illustrated in Figure 4B.

These results indicate that nuclear localization is identified primarily through the connectivity patterns and globular density characteristic of DNA-binding domains, rather than through specific chemical motifs. Specifically, the prioritization of structural hubs and packing compactness suggests that the model has internalized abstract principles of protein architecture associated with the nuclear environment. This structural approach provides a more robust resolution of localization determinants compared to pure sequence-motif recognition, particularly when sequence identity is low.

4 Discussion

4.1 Toward Green AI in Proteomics

Beyond predictive accuracy, **BioGraphX-Net** demonstrates that explicit biochemical knowledge can dramatically reduce parameter dependency. Contemporary protein language models (pLMs) achieve performance primarily through scale, requiring energy-intensive fine-tuning of billions of parameters[Schwartz et al., 2020]. Our framework presents an alternative paradigm: using a fixed, general-purpose sequence encoder (ESM-2) coupled with a small, trainable network of interpretable biophysical features. This approach lowers the number of trainable parameters by over 99% compared with traditional end-to-end fine-tuning..

Practically, this makes high-resolution localization predictions accessible on normal computers, without needing GPU clusters, because ESM inference is only done once per sequence. BioGraphX therefore moves from scale-focused AI to knowledge-driven AI, showing how to build efficient and interpretable models in computational biology.

4.2 Biological Implications: Capturing Complexity

BioGraphX-Net excels at capturing the complexity of protein localization. It combines evolutionary information with biophysical rules in a gated graph, preserving subtle relationships between residues and domains. This lets the model detect delicate localization cues, including motifs and structural constraints

BioGraphX-Net shows that protein localization isn't just about sequence motifs; spatial and physicochemical context also matters. This highlights the value of considering topology and biophysics in computational studies and can help design experiments for proteins with context-dependent localization.

4.3 Resolving the Eukaryotic Secretory Pathway

BioGraphX-Net can better handle complex secretory-pathway localizations. Its predictions often include multiple compartments, reflecting moonlighting proteins that work in different places in the cell.

The gating mechanism in BioGraphX-Net is sensitive to overlapping signals from the ER, Golgi, and lysosome. Unlike transformer models that might force exclusive predictions when sequences are similar, BioGraphX-Net uses graph-based features to untangle complex localization signals.

4.4 The Negative Selection'Hypothesis

A notable finding emerging from the SHAP analysis is the role of sequence profiles as *"Negative Selection"* filters. While the existing literature typically characterizes sequence profiles as positive indicators of localization, the BioGraphX-Net framework appears to treat the Golgi profile as an inhibitory signal for the Golgi apparatus itself. We hypothesize that this represents a learned correction for *"sequence mimicry."*

Specifically, many nuclear and cytoplasmic proteins contain disordered regions that exhibit statistical similarities to

the *tails* of Golgi proteins. A high profile score in isolation may therefore be an unreliable predictor. The model tends to ignore the sequence profile unless it's backed by strong physical anchors, like hydrophobic edges. This two-step check ensures that structural rules take precedence over simple sequence patterns for accurate targeting

4.5 Biological Significance of the Physics-Gated Fusion

The gating system balances evolutionary embeddings (ESM-2) with BioGraphX features to reduce noise. Evolutionary signals are informative but can be misleading due to conservation across different functions. BioGraphX physics features act as a check, ensuring that predictions make biophysical sense.

The framework knows that a protein won't reside in the Golgi membrane without the right hydrophobic profile, even if the evolutionary sequence suggests otherwise. This two-step check helps the model generalize to new sequences where evolutionary data is limited. By combining sequence, structure, and function, BioGraphX-Net provides a biologically explainable approach to protein prediction.

4.6 Comparison with Alternative Approaches

4.6.1 Graph Neural Networks on Protein Structures (GNN-3D)

Traditional structure-based GNNs [Gligorijević et al., 2021] use high-quality 3D protein structures from AlphaFold2 [Jumper et al., 2021] or experiments like X-ray crystallography. While they perform well for well-folded proteins, they have important limitations. First, running AlphaFold2 at the proteome scale is very computationally expensive. Second, predictions for intrinsically disordered regions are often low-confidence, which reduces reliability. Third, because these models depend on structure quality, errors in predicted coordinates can directly affect the learning results. BioGraphX bypasses these issues by learning approximate structures straight from amino acid sequences, so it doesn't rely on full 3D coordinates. Like PSICHIC [Koh et al., 2024], our approach favors knowledge-based graph encodings over black-box embeddings. But the two differ: PSICHIC uses chemistry-guided interactions to guide attention, while BioGraphX builds a physics-driven graph where the physicochemical rules define structure on their own. Graphs can either simulate processes like binding or directly represent structural properties like folding constraints. BioGraphX takes the second approach, using a sequence-based graph to capture structural determinants of localization, effectively applying Anfinsen's postulate in a predictive and explainable way.

4.6.2 Contact Map–Based Methods

Contact prediction approaches, such as PSICOV [Jones et al., 2011] and MetaPSICOV [Jones et al., 2014], are primarily based on statistical inference of residue–residue contacts derived from multiple sequence alignments (MSAs) [Thompson et al., 1994]. While effective when deep and diverse MSAs are available, these methods are limited in their applicability to evolutionarily isolated proteins for which high-quality alignments cannot be constructed. Furthermore, contact maps encode predominantly binary interaction information and fail to capture higher-order physicochemical constraints governing protein organization. In contrast, the constraint-based formulation of BioGraphX employs deterministic interaction rules grounded in physical chemistry, enabling robust generalization to proteins with sparse homologous information.

4.6.3 Hybrid Feature Fusion Approaches

The majority of existing hybrid localization frameworks combine protein language model (pLM) representations with external biological features through post-hoc feature concatenation. In such architectures, the learned pLM embeddings frequently dominate the decision-making process, resulting in feature asymmetry and limited interpretability of auxiliary signals. BioGraphX fundamentally departs from this paradigm by incorporating physics-based structural descriptors as first-class features within the model architecture. Through an interpretable gated fusion mechanism, balanced contributions from both evolutionary and biophysical sources are explicitly enforced. This architectural design accounts for the ability of BioGraphX to match the predictive performance of billion-parameter models while employing substantially smaller and computationally efficient networks.

4.7 Universality of the BioGraphX Representation

Although the current implementation of BioGraphX is primarily concerned with the eukaryotic proteome, the underlying physicochemical graph encoding logic appears to be inherently domain-agnostic. The framework's reliance on node-aligned topological descriptors suggests a natural extensibility to other biological macromolecules, such as RNA and DNA. By redefining node identities as nucleotides and edge constraints as base-pairing or stacking interactions, it is possible that BioGraphX could provide a unified 'Structural Proxy' for the entire Central Dogma. Specifically, the topological constants identified in this study, may represent universal physical requirements for the stability and localization of any linear biopolymer. Future research should be undertaken to explore the deployment of BioGraphX in predicting mRNA half-life and lncRNA subcellular trafficking, thereby positioning this framework as a universal sequence-to-physics encoder for the broader 'omics' landscape.

4.8 Limitations and Future Work

BioGraphX-Net works well, but it has some limitations. Currently, it assumes that residues close in sequence are also close in 3D space. This usually works for folded domains, but it can miss long-range contacts. Future versions will use predicted contact maps from AlphaFold2 to better capture complex interactions.

Another limitation is that BioGraphX-Net is designed mainly for single-chain proteins. In complex systems, many proteins act as part of multimeric assemblies, and their subcellular localization is influenced by inter-chain contacts and overall quaternary geometry.

5 Conclusion

In this study, we have introduced BioGraphX, a novel framework that bridges the sequence-structure gap by transforming primary protein sequences into multi-scale interaction graphs. By explicitly modeling biophysical constraints based on Anfinsen's Dogma, the framework provides a deterministic structural proxy that eliminates the dependency on high-quality 3D data. The results of our investigation demonstrate that BioGraphX-Net achieves state-of-the-art performance on both DeepLoc benchmarks and LocPro, significantly outperforming billion-parameter language models in difficult-to-resolve organelles such as the Peroxisome and Golgi apparatus.

A key finding of this research is the standalone predictive power of the biophysical encoding, which achieved a global accuracy of 65.15% using an XGBoost classifier. This provides evidence that biochemically-grounded rules capture essential localization signals that remain conserved even when sequence identity falls into the 30% "twilight zone". Furthermore, SHAP analysis reveals that the model employs a sophisticated "Negative Selection" mechanism, treating generic sequence profiles as repellors to filter false positives while relying on specific physical constraints, such as hydrophobicity and sulfur-aromatic stacking, for final verification. This shows that the framework performs well by combining evolutionary information with structural hubs and compact packing in a transparent way.

BioGraphX pioneers a new paradigm of knowledge-integrated artificial intelligence for bioinformatics. By combining frozen language model embeddings with trainable biophysical graph features, we match state-of-the-art accuracy while training two orders of magnitude fewer parameters. This result demonstrates that encoding domain knowledge directly into model architecture is a powerful alternative to parameter scaling. The framework further provides mechanistic interpretation through built-in explainability tools and enables practical deployment, contributing to a more sustainable and interpretable future for machine learning in protein science.

Acknowledgments

The author acknowledges the use of publicly accessible computational resources provided through the Kaggle platform for conducting the experiments in this study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

No competing interest is declared.

Author contributions statement

A.S. conceived the study, developed the methodology, implemented the algorithms, conducted experiments, analyzed the data, and wrote the manuscript.

Data Availability

The datasets analyzed in this study are publicly available. The DeepLoc benchmark dataset can be accessed from its original publication. The source code of the BioGraphX encoding framework used in this study is available at: `https://github.com/Abubakar-Saeed/BioGraphX`

Bibliography

References

[Almagro Armenteros et al., 2017] Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395.

[Anfinsen, 1973] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., Bengio, Y., and LeCun, Y. (2015). Neural machine translation by jointly learning to align and translate.

[Baker and Hubbard, 1984] Baker, E. N. and Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44:97–179.

[Briesemeister et al., 2010] Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Yloc+—an interpretable web server for predicting subcellular localization. *Nucleic acids research*, 38(suppl_2):W497–W502.

[Burley and Petsko, 1985] Burley, S. K. and Petsko, G. A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, 229:23–28.

[Chothia, 1974] Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248:338–339.

[Consortium et al., 2023] Consortium, U. et al. (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531.

[Dill, 1990] Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29:7133–7155.

[Dougherty, 1996] Dougherty, D. A. (1996). Cation-π interactions in chemistry and biology. *Science*, 271:163–168.

[El-Kafafi et al., 2005] El-Kafafi, E.-S., Mukherjee, S., El-Shami, M., Putaux, J.-L., Block, M. A., Pignot-Paintrand, I., Lerbs-Mache, S., and Falconet, D. (2005). The plastid division proteins, ftsz1 and ftsz2, differ in their biochemical properties and sub-plastidial localization. *The Biochemical Journal*, 387(Pt 3):669–676.

[Elnaggar et al., 2022] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2022). Prottrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.

[Fufezan et al., 2010] Fufezan, C. et al. (2010). Role of backbone-backbone interactions in protein folding. *Proteins*, 78:2078–2090.

[Gallivan and Dougherty, 1999] Gallivan, J. P. and Dougherty, D. A. (1999). Cation-π interactions in structural biology. *Proceedings of the National Academy of Sciences*, 96:9459–9464.

[Gligorijević et al., 2021] Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Cho, K., Vatanen, T., Berenberg, D., Taylor, B. C., Fisk, I. M., Xavier, R. J., Knight, R., and Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168.

[González-Gordo et al., 2022] González-Gordo, S., Palma, J. M., and Corpas, F. J. (2022). Peroxisomal proteome mining of sweet pepper (capsicum annuum l.) fruit ripening through whole isobaric tags for relative and absolute quantitation analysis. *Frontiers in Plant Science*, 13:893376.

[Gorodkin, 2004] Gorodkin, J. (2004). Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5–6):367–374.

[Hunklinger and Ferruz, 2025] Hunklinger, A. and Ferruz, N. (2025). Toward the explainability of protein language models for sequence design.

[Jones et al., 2011] Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2011). Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190.

[Jones et al., 2014] Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2014). Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7):999–1006.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[Koh et al., 2024] Koh, H. Y., Nguyen, A. T. N., Pan, S., et al. (2024). Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 6:673–687.

[Kohnhorst et al., 2016] Kohnhorst, C. L., Kyoung, M., Jeon, M., Schmitt, D. L., Kennedy, E. L., Ramirez, J., Bracey, S. M., Luu, B. T., Russell, S. J., and An, S. (2016). Subcellular fractionation of human neutrophils reveals distinct compartmentalization of signaling proteins. *International journal of molecular sciences*, 17(2):189.

[Kumar and Nussinov, 1999] Kumar, S. and Nussinov, R. (1999). Salt bridge stability in proteins. *Journal of Molecular Biology*, 293:1241–1255.

[Lin et al., 2023] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

[Lobley et al., 2008] Lobley, A., Sadowski, M. I., and Jones, D. T. (2008). Phyre: protein homology/analogy recognition engine. *BMC bioinformatics*, 9:1–15.

[Loo et al., 2014] Loo, L.-H., Laksameethanasan, D., and Tung, Y.-L. (2014). Quantitative protein localization signatures reveal an association between spatial and functional

divergences of proteins. *PLoS Computational Biology*, 10(3):e1003504.

[Makrodimitris et al., 2020] Makrodimitris, S., van Ham, R. C., and Reinders, M. J. (2020). Automated protein function prediction using textual and protein sequence data. *Bioinformatics*, 36(15):4223–4231.

[McFadden, 2014] McFadden, G. I. (2014). Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, 6(4):a016105.

[Munro, 1998] Munro, S. (1998). Localization of proteins to the golgi apparatus. *Trends in Cell Biology*, 8(1):11–15.

[Navarro and Cheeseman, 2022] Navarro, A. P. and Cheeseman, I. M. (2022). Identification of a golgi-localized peptide reveals a minimal golgi-targeting motif. *Molecular Biology of the Cell*, 33(12):ar110.

[Nishio et al., 2014] Nishio, M., Umezawa, Y., et al. (2014). Ch/π hydrogen bonds in biological macromolecules. *Physical Chemistry Chemical Physics*, 16:12648–12683.

[Orbán et al., 2008] Orbán, E., Szikora, S., Mihály, J., Pálfi, A., Földi, I., and László, L. (2008). Immunoelectron microscopic localization of the hsp72 chaperone protein in the nuclear matrix of heat-shocked neuronal cells. *Acta Histochemica*, 110(1):42–50.

[Pauling et al., 1951] Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37:205–211.

[Schwartz et al., 2020] Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12):54–63.

[Sigrist et al., 2013] Sigrist, C. J., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at prosite. *Nucleic acids research*, 41(D1):D344–D347.

[Stärk et al., 2021] Stärk, H., Dallago, C., Heinzinger, M., and Rost, B. (2021). Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035.

[Thompson et al., 1994] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680.

[Thornton, 1981] Thornton, J. M. (1981). Disulphide bridges in globular proteins. *Journal of Molecular Biology*, 151:261–287.

[Thul et al., 2017] Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., et al. (2017). A subcellular map of the human proteome. *Science*, 356(6340):eaal3321.

[Thumuluri et al., 2022] Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H., and Winther, O. (2022). Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, 50(W1):W228–W234.

[Valley et al., 2012] Valley, C. C., Cembran, A., Perlmutter, J. D., Lewis, A. K., Labello, N. P., Gao, J., and Sachs, J. N. (2012). The methionine-aromatic motif plays a unique role in stabilizing protein structure. *Journal of Biological Chemistry*, 287(42):34979–34991.

[Wan et al., 2016] Wan, S., Mak, M.-W., and Kung, S.-Y. (2016). f-mloc: sequence-based prediction of protein subcellular localization using effectively optimized information. *Journal of Chemical Information and Modeling*, 56(9):1648–1656.

[Yan et al., 2013] Yan, J., Mizianty, M. J., Filipow, P. L., Uversky, V. N., and Kurgan, L. (2013). Rapid: Fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(8):1671–1680.

[Zhang et al., 2025] Zhang, Y., Zheng, L., You, N., Hu, W., Jiang, W., Lu, M., Xu, H., Dai, H., Fu, T., and Zhou, Y. (2025). Locpro: A deep learning-based prediction of protein subcellular localization for promoting multi-directional pharmaceutical research. *Journal of Pharmaceutical Analysis*, 15(8):101255.