BioGraphX: Bridging the Sequence-Structure Gap via Physicochemical Graph Encoding for Interpretable 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 158 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 shows that BioGraphX-Net encodes a sophisticated biophysical logic: sequence profiles act as universal exclusion filters, while organelle-specific combinations of biophysical features enable precise compartment discrimination. Notably, Frustration features help resolve targeting ambiguities in complex compartments, reflecting evolutionary constraints while preventing mislocalization from sequence mimicry. 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 13.46 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 and and immunoelectron microscopy [Xiao et al., 2024]. 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 [Gillani and Pollastri, 2024]. 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., 2017] 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.

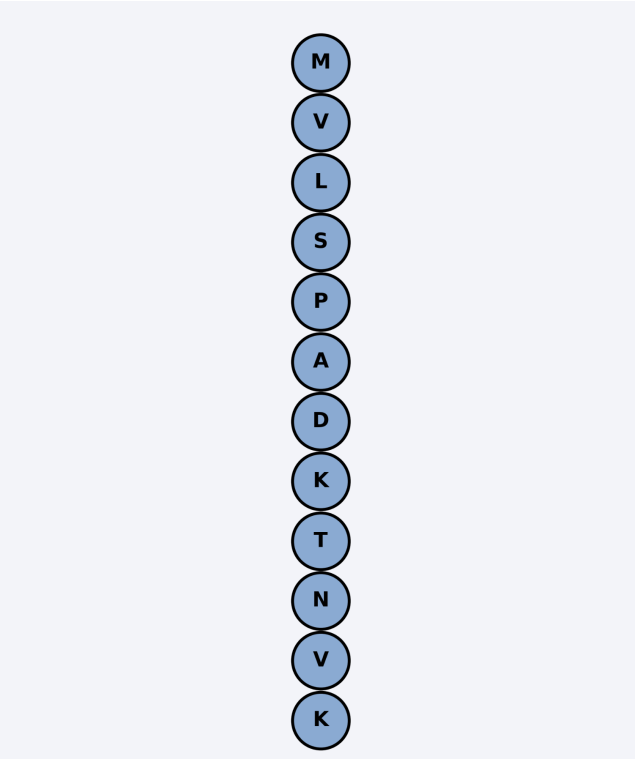
BioGraphX-Net combines these encodings with ESM-2 embeddings through an interpretable gating mechanism that quantifies reliance on evolutionary vs. biophysical signals, enabling transparent decision-making. Unlike models requiring end-to-end fine-tuning of all parameters, our framework maintains the ESM-2 backbone frozen and trains only 13.46M 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.

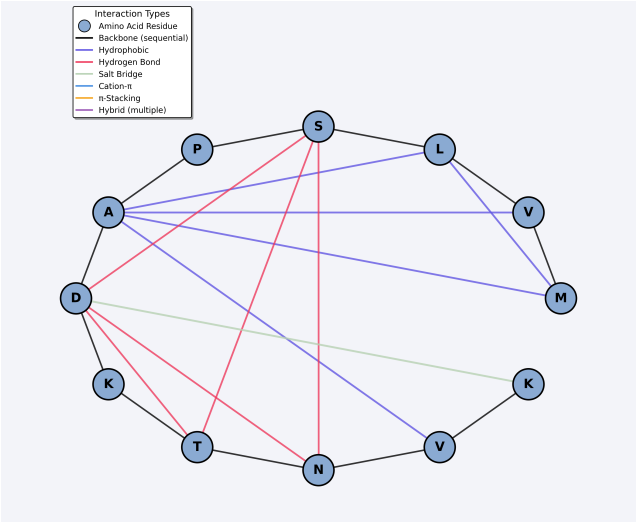
2 Methods

2.1 Datasets and Preprocessing

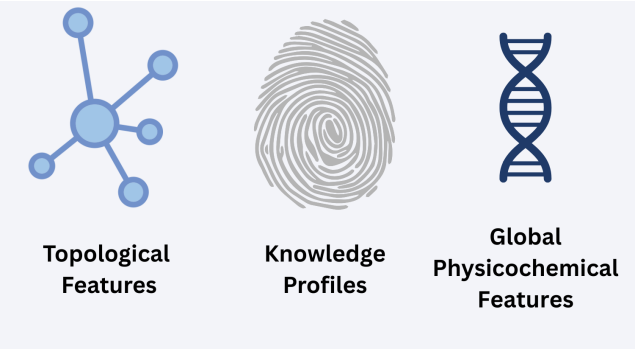
This study relies on data from two key sources: the DeepLoc 1.0 [Almagro Armenteros et al., 2017] and DeepLoc 2.0 [Thumuluri et al., 2022] benchmark datasets. To ensure data integrity, a series of filtering steps was applied. The primary inclusion criterion required sequences to contain only the 20



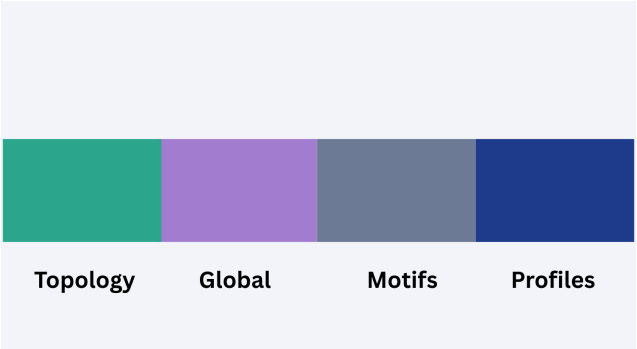
(a) Input sequence



(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, global-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 stands for the attention transformation, while α refers to the masked attention weights that is assigned to each residue. After this aggregation step, the weighted representation, E_{vec}, is passed through a bottleneck transformation. This step reduces its dimensionality and makes it easier to combine with other features. Specifically, the reduction to 1024 dimensions is performed by a fully connected layer, which is paired with Batch Normalisation (BN) and a ReLU (Rectified Linear Unit) activation function:

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

Here, E_{vec} is the original high-dimensional embedding produced by the ESM model. W_3 and b_3 are the learnable weight matrix and bias for the bottleneck layer. After this transformation, the resulting vector, E_{feat}, represents the compressed semantic feature representation.

The purpose of this bottleneck reduction is to capture the most important features from the high-dimensional ESM embedding. This helps the model avoid overfitting to noise, while keeping the latent space compatible for later integration with biophysical features.

2.4 BioGraphX Encoding Framework

To support proteome-scale analyses, the encoding pipeline provides parallel batch processing using Joblib, which significantly improves computational efficiency.

2.4.1 Foundational Biophysical Principles

The BioGraphX encoding framework converts protein sequences into multi-scale interaction graphs by applying 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 through encoded structural constraints [Anfinsen, 1973]. BioGraphX implements these constraints explicitly via interaction rules between compatible residues, based on the un-

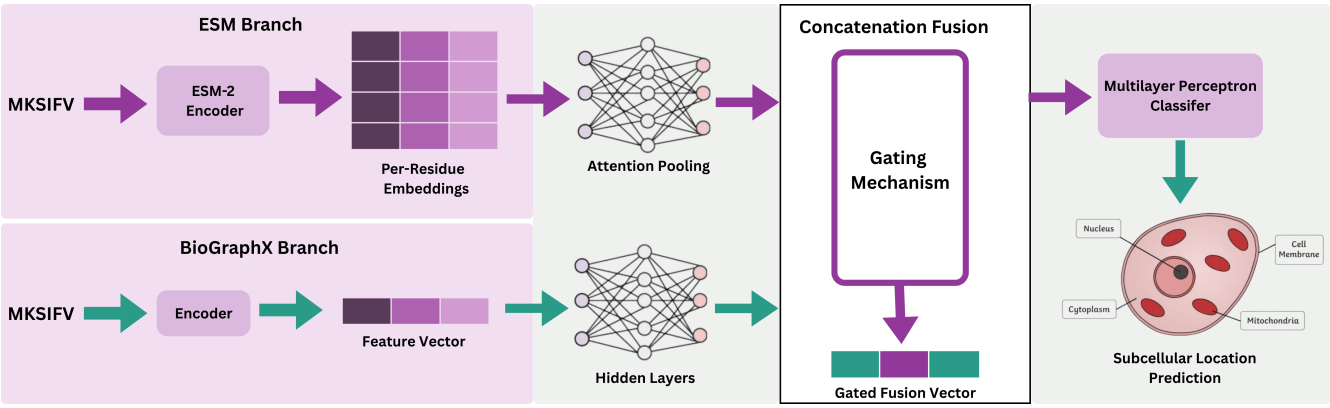


Figure 2: Schematic representation of the BioGraphX-Net computational pipeline. The diagram shows the integration of ESM sequence embeddings and BioGraphX physics-based features through a central gating mechanism, producing the combined Gated Fusion Vector (f_{gated}) for downstream classification.

derstanding 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 $F \in \mathbb{R}^{158}$

```
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^23
Compute knowledge-guided scores → F_know ∈ R^20
Calculate global physicochemical properties → F_phys ∈ R^19
Compute constraint frustration features → F_frust ∈ R^11
return F = [F_topo.F_hybrid.F_know.F_phys.F_frust]
```

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 distance thresholds and the strength weights (w_t) used by the graph enable the representation of the hierarchy of protein folding and stability using the protein sequence. This is achieved by using the distance thresholds to represent the maximum sequence distance over which these forces act to maintain structural integrity and the weights to represent the relative contribution of each type of interaction to biochemical stability. By using these deterministic rules, BioGraphX is able to generate a representation that is grounded in the principles of biophysics and acts as a robust proxy for 3D structural features.

In order to further enhance the representation, we have assigned a weight $w_t \in [0, 1]$ to each type of interaction t and have proposed a distance decay function to account for the weakening of the influence of the interaction as the sequence distance increases. The weight w_e of a given edge E is given by:

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

In this model, d_{ij} represents the sequence distance between two residues i and j, while $d_{max,t}$ represents the maximum allowed separation for a particular type of interaction t. In addition, the constant factor of 0.3 is used to guarantee that a base level of interaction strength of approximately 70% remains even when the maximum threshold is reached. Results of ablation studies have shown that the choice of this factor has a significant impact: removing the decay reduced the accuracy by 1.44%, while increasing the factor to 0.5 leads to a 0.5% decrease in accuracy. 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,

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	[Trachsel et al., 2015]
Carbonyl–Carbonyl	D, E, N, Q, S, T, Y	0.35	8	[Bartlett 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]

as summarized in Table 2.

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 158 features in five groups: topological, hybrid, knowledge-guided, global physicochemical, and constraint frustration. Topological features show network structure, hybrid features capture overlapping biochemical interactions, knowledge-guided scores use known targeting motifs, global physicochemical features describe general protein characteristics, and constraint frustration features quantify conflicting interaction energies. 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	23	Co-occurrence patterns of interaction types	Salt-bridge / H-bond frequency, regional hybrid density
Knowledge-Guided	20	Features derived from known biological motifs	Regular expressions for NLS and targeting motifs
Physicochemical	19	Whole-protein physicochemical properties	Isoelectric point, GRAVY score, entropy
Constraint Frustration	11	Per-residue conflicting interaction energies	Signal–structure contrast, frustration localization, hotspot count
Total	**158**	**Multi-scale protein representation**	**Integrated sequence–structure features**

Note: 81.6% of BioGraphX features (129/158) are graph-based and derived from biochemical interaction graphs rather than raw sequence patterns, providing structural proxies that complement evolutionary embeddings.

2.5 Hybrid Fusion Architecture

2.5.1 Dual-Branch Network Design

The hybrid architecture consists of two specialized input branches designed to process complementary protein representations. The first branch operates on ESM-2 embeddings, applying attention pooling followed by bottleneck compression to extract evolutionary sequence information.

The second branch processes BioGraphX features through an enhanced three-layer nonlinear transformation, substantially increasing representational capacity for modeling biophysical constraints. Formally, the BioGraphX transformation is defined as:

$$\begin{aligned} \mathbf{B}_{\text{hidden1}} &= \text{ReLU}\left(\text{BN}\left(\mathbf{W}_4^{(1)}\mathbf{B} + \mathbf{b}_4^{(1)}\right)\right) \\ \mathbf{B}_{\text{hidden2}} &= \text{ReLU}\left(\text{BN}\left(\mathbf{W}_4^{(2)}\mathbf{B}_{\text{hidden1}} + \mathbf{b}_4^{(2)}\right)\right) \\ \mathbf{B}_{\text{feat}} &= \text{ReLU}\left(\text{BN}\left(\mathbf{W}_4^{(3)}\mathbf{B}_{\text{hidden2}} + \mathbf{b}_4^{(3)}\right)\right) \end{aligned} \tag{5}$$

This expanded transformation projects the original 158-dimensional BioGraphX feature vector into a 1,024-dimensional latent space, matching the dimensionality of the ESM-2 embeddings. The successive nonlinear mappings enable the model to capture complex and higher-order relationships among biophysical constraints.

Finally, the outputs of the ESM and BioGraphX branches are fused using an adaptive gating mechanism (Figure 2), allowing the network to dynamically balance evolutionary and biophysical information during prediction.

2.5.2 Interpretable Gated Fusion Mechanism

The fusion stage employs an adaptive gating mechanism that dynamically balances the contributions of evolutionary embeddings and biophysical features on a per-protein basis. After obtaining the transformed feature representations $\mathbf{E}_{\text{feat}} \in \mathbb{R}^{1024}$ (Equation 2) and $\mathbf{B}_{\text{feat}} \in \mathbb{R}^{1024}$ (Equation 5), the two vectors are concatenated to form a joint representation:

$$\mathbf{C} = [\mathbf{E}_{\text{feat}}; \mathbf{B}_{\text{feat}}] \in \mathbb{R}^{2048}. \tag{6}$$

A two-layer gating controller then computes protein-specific weighting signals as follows:

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**

$$
\begin{aligned}
\mathbf{G}_{\text{hidden}} &= \mathrm{ReLU}\left(\mathbf{W}_5^{(1)}\mathbf{C} + \mathbf{b}_5^{(1)}\right), \\
\mathbf{G}_{\text{raw}} &= \mathbf{W}_5^{(2)}\mathbf{G}_{\text{hidden}} + \mathbf{b}_5^{(2)}, \\
\mathbf{G} &= \sigma\left(\mathbf{G}_{\text{raw}}\right) = \left[g_{\text{physics}},\ g_{\text{ESM}}\right],
\end{aligned} \tag{7}
$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, producing normalized, per-protein gating coefficients.

The gated feature vector is then computed via element-wise modulation of each feature block:

$$
\mathbf{C}_{\text{gated}} = \left[\mathbf{E}_{\text{feat}} \odot g_{\text{ESM}}\ ;\ \mathbf{B}_{\text{feat}} \odot g_{\text{physics}}\right], \tag{8}
$$

where $\odot$ denotes element-wise multiplication, with each scalar gate value broadcast across the corresponding 1024-dimensional feature subspace. This formulation allows the model to independently regulate the influence of evolutionary and biophysical signals for each protein. The use of a two-layer gating controller provides greater expressive capacity than a single linear transformation, enabling the learning of complex and context-dependent fusion strategies.

2.5.3 Multi-layer Classifier

The last stage of classification consists of a multi-layer perceptron (MLP) which 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:

$$
\begin{aligned}
H_1 &= \text{ReLU}(\text{BN}(W_6 C_{\text{gated}} + b_6)) \\
H_2 &= \text{ReLU}(W_7 H_1 + b_7) \\
\hat{Y} &= W_8 H_2 + b_8
\end{aligned} \tag{9}
$$

where H_1 and H_2 are hidden layer activations, and $\hat{Y}$ is the final prediction. The resulting model contains approximately **13.46 million parameters**, achieving a balance between high representational capacity and computational efficiency.

2.6 Training Protocol and Optimization

The classification thresholds were optimized by maximizing the Matthews Correlation Coefficient (MCC) [Gorodkin, 2004] for each individual class. The optimization procedure employed a Focal Loss with parameters $\alpha = 0.25$ and $\gamma = 2.0$ to address class imbalance and handle difficult examples. Optimization was carried out using the AdamW optimizer with a base learning rate of $\eta_{\text{base}} = 1 \times 10^{-5}$ and weight decay of 1×10^{-2}, together with a cosine annealing learning rate scheduler. Differential learning rates were applied across model components: the physics branch used $\eta = 3 \times 10^{-5}$, the ESM branch used $\eta = 1 \times 10^{-5}$, the gating controller used $\eta = 0.5 \times 10^{-5}$, and the classifier used $\eta = 1 \times 10^{-5}$. During the early stages of training, a physics encouragement loss with coefficient $\lambda = 0.1$ was applied to promote balanced utilization of evolutionary and biophysical information. Gradient clipping was applied in a component-specific manner, with a maximum norm of 5.0 for the physics branch and 1.0 for all remaining components.

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 35. The final model was selected based on the highest observed validation performance.

2.7 Explainability Framework

BioGraphX makes its predictions interpretable via four mechanisms that allow the researcher to analyze decisions at the level of each residue.

- **Gate Analysis:** Looking at the gating weights G makes it easy to see how much the various proteins use evolutionary versus biophysical features, by how much one is more prevalent.
- **SHAP Feature Importance:** Shap analysis [Lundberg and Lee, 2017] framework was applied to gain an insight on the contribution of biophysical features. SHAP values (ϕ_{ijk}) for each protein i, feature j, and class k were calculated using a Kernel Explainer with stratified sampling in order to balance the representation from all subcellular compartments. Feature importance for each class was measured as average absolute SHAP value for all the proteins.

$$
\text{Importance}_{jk} = \frac{1}{N}\sum_{i=1}^{N} |\phi_{ijk}|, \tag{10}
$$

  where N denotes the total number of proteins in the evaluation set. In order to isolate the contribution of biophysical features, while controlling for the information of evolutionary features, a special PhysicsWrapper was built into which ESM embeddings were fixed and BioGraphX features were allowed to vary. This design allows pinpointing localization choices made during development to very specific biophysical constraints, independent of evaluating evolutionary signal strength.
- **Knowledge-Guided Score:** The direct mapping of model properties to known targeting motifs, e.g. nuclear localization signals, mitochondrial targeting peptides, or signal peptides, allows hypothesis to be generated biologically about the mechanism of the predictions.
- **Hybrid Pattern Visualization:** Regions containing clusters of specific interaction patterns, such as Salt-Bridge–HBond motifs near active sites, are highlighted to show where key structural features play a crucial role.

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	-	1.52 ± 0.02
Accuracy	0.32 ± 0.02	0.53 ± 0.02	0.55 ± 0.02	-	$\mathbf{0.55 \pm 0.01}$
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	-	$\mathbf{0.78 \pm 0.01}$
Macro F1	0.42 ± 0.01	0.64 ± 0.01	0.66 ± 0.01	-	$\mathbf{0.69 \pm 0.01}$
MCC per location (↑)					
Cytoplasm	0.38 ± 0.02	0.61 ± 0.01	0.62 ± 0.01	0.612 ± 0.002	$\mathbf{0.64 \pm 0.01}$
Nucleus	0.42 ± 0.02	0.66 ± 0.02	0.69 ± 0.01	$\mathbf{0.718 \pm 0.002}$	$\mathbf{0.71 \pm 0.02}$
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	0.66 ± 0.01	0.662 ± 0.006	$\mathbf{0.67 \pm 0.01}$
Mitochondrion	0.47 ± 0.02	0.73 ± 0.03	$\mathbf{0.76 \pm 0.02}$	0.754 ± 0.003	$\mathbf{0.76 \pm 0.03}$
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	$\mathbf{0.59 \pm 0.02}$
Lysosome / Vacuole	0.07 ± 0.03	0.24 ± 0.03	0.28 ± 0.04	$\mathbf{0.341 \pm 0.027}$	0.33 ± 0.04
Golgi apparatus	0.11 ± 0.04	0.36 ± 0.06	0.34 ± 0.05	0.373 ± 0.003	$\mathbf{0.43 \pm 0.07}$
Peroxisome	0.05 ± 0.02	0.48 ± 0.05	$\mathbf{0.56 \pm 0.08}$	0.524 ± 0.017	0.54 ± 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.

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:** This measures the percentage of cases where the predicted labels match the actual labels, providing a clear sense of the model's accuracy.
- **Jaccard Index:**This compares the similarity of set of predictable and real labels, and assists in checking accuracy of 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 0.64, 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 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 0.78, higher than DeepLoc 2.0 (0.73) and LocPro (0.76). Higher Jaccard (0.72) and Macro-F1 (0.69) 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 158 graph features. Importantly, BioGraphX-Net trains only 13.46M 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

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	1.23
Accuracy	0.23	0.37	0.38	**0.45**	0.34	0.39	0.39
Jaccard	0.41	0.42	0.46	0.52	0.48	0.53	**0.54**
MicroF1	0.51	0.46	0.52	0.56	0.57	**0.60**	0.59
MacroF1	0.34	0.35	0.39	0.43	0.44	**0.46**	0.45
MCC per location (↑)							
Cytoplasm	0.14	0.23	0.23	0.33	0.29	**0.36**	0.29
Nucleus	0.20	0.28	0.41	**0.45**	0.41	0.44	**0.45**
Cell membrane	0.20	0.23	0.32	0.30	0.34	**0.36**	**0.36**
Mitochondrion	0.37	0.39	0.33	0.59	**0.60**	0.56	0.51
Endoplasmic reticulum	0.12	**0.23**	0.14	0.22	0.20	0.17	0.15
Golgi apparatus	0.08	0.10	0.24	0.26	0.17	0.31	**0.32**

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.

difficult to classify due to sparse data and dynamic interactions.

Challenging Compartments: For the Golgi apparatus, the framework achieved an MCC of 0.43, surpassing both DeepLoc 2.0 (0.34) and LocPro (0.373). Similarly, the model reached an MCC of 0.54 for the Peroxisome, which is higher than the 0.52 reported for LocPro and comparable to the 0.56 achieved by DeepLoc.

Targeting Signal Identification: In mitochondrial localization, BioGraphX-Net outperforms state-of-the-art methods, achieving an MCC of 0.76, surpassing the LocPro ESM2+PROFEAT hybrid (0.754) and matching DeepLoc 2.0 (0.76).

Standardized Benchmarks: Performance on the nucleus (0.71) and lysosome/vacuole (0.33) remains highly competitive on standardized DeepLoc benchmarks, outperforming DeepLoc 2.0 and matching LocPro 2.0.

Regarding the multi-label prediction, BioGraphX-Net identified an average of 1.52 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 (0.78 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 evaluate generalization beyond the training distribution, BioGraphX-Net was benchmarked on an independent blind test set derived from the Human Protein Atlas (HPA; $N =$ 1,717). This dataset provides a particularly rigorous testing environment because all included sequences share less than 30% similarity with the SwissProt training data. This is a critical challenge, as purely sequence-based models typically experience a sharp drop in performance under these conditions. Irrespective of this difficulty, BioGraphX-Net demonstrated a high level of robustness, as a Micro-F1 score of 0.59 was achieved with a conservative average of 1.23 predicted labels per protein.

We analyzed the gating mechanism and observed that the contribution pattern is protein-specific. BioGraphX features contributed 39.2%, while ESM embeddings contributed 50.5%. We further noted compartment-specific reliance: mitochondrial proteins depended approximately 40% on BioGraphX features, which is consistent with their physicochemical properties. Conversely, Golgi apparatus proteins showed the lowest contribution from physics-based features and the strongest reliance on ESM embeddings, indicating a high degree of evolutionary conservation.

The adaptive nature of the gating ratios is further demonstrated by their distribution across proteins: 63.1% of proteins showed a balanced contribution between BioGraphX features and ESM embeddings, while only 4.3% exhibited strong reliance on physics-based features. This fine-grained, protein-specific balancing underscores the adaptive behavior of the model.

For example, nuclear proteins exhibited strong structural determinants, reflected by high physics contributions (up to 64.0%). In contrast, proteins with high ESM dependence (up to 79.8%) primarily relied on patterns of evolutionary conservation.

Notably, our model achieved superior performance on nucleus and Golgi appraturs with MCC of 0.45 and 0.32, respectively. The results suggest that BioGraphX-Net effectively capture the physical princples which transformed across species.

Overall, this physics-first advantage, propose that explicit biophysical encoding can achieve performance comparable to that of significantly larger black box models.

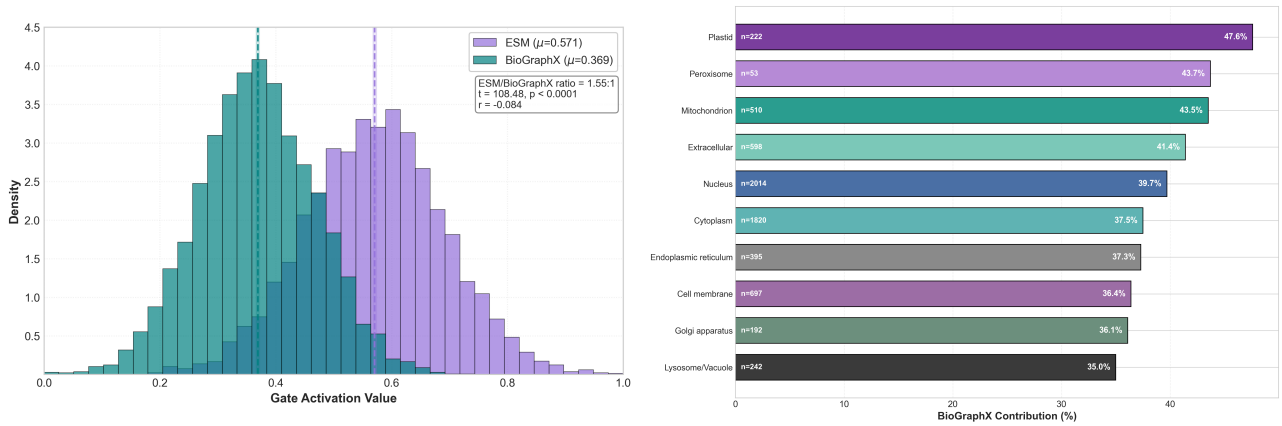


Figure 3: (A)Distribution of dynamic gating weights. The gating mechanism assigns consistently higher activation to sequence embeddings while maintaining a substantial contribution from BioGraphX features, demonstrating complementary feature fusion. (B)The hierarchy of reliance on BioGraphX features.

3.2.3 Explainability Analysis

3.2.4 Gating Mechanism: Evolutionary versus Biophysical Signal Weighting

Analysis of the gate distribution across 5,442 validation samples revealed that the BioGraphX-Net architecture employs a complex, adaptive gating mechanism that dynamically balances evolutionary and biophysical information. Overall, the model explains an average of $60.8\% \pm 9.9\%$ of the variation through ESM embeddings and $39.2\% \pm 9.5\%$ through BioGraphX physics features, corresponding to a ratio of 1.55:1. Statistical testing confirms significant differential gating between the two information sources ($t = 108.48$, $p < 0.0001$). Notably, only a weak negative correlation between the two gates was observed ($r = -0.084$), indicating largely independent modulation rather than strict competition.

The architecture further exhibits, organelle-specific patterns of biophysical feature utilization that align with known biological complexity. Subcellular compartments requiring stringent targeting validation show the highest reliance on physics-based features, including plastid, peroxisome, and mitochondrion, organelles characterized by complex import machinery and strong structural constraints. In contrast, compartments with more straightforward evolutionary signatures demonstrate reduced physics dependence, such as the cell membrane, Golgi apparatus, and lysosome/vacuole.

These compartment-adaptive gating patterns suggest that BioGraphX-Net learns to modulate the degree of biophysical validation required for accurate localization, applying stronger physics-based reasoning where evolutionary similarity alone is insufficient and relying more heavily on sequence homology where it provides a reliable signal, as shown in Figure 3.

3.2.5 SHAP Analysis

To elucidate the biophysical rules underlying BioGraphX-Net predictions, we performed a comprehensive SHapley Additive exPlanations (SHAP) analysis on a held-out validation set comprising 200 stratified protein samples. This approach quantifies the directional contribution of each of the 158 BioGraphX features across ten eukaryotic subcellular compartments, thereby revealing compartment-specific biophysical "codes." In addition to improving model interpretability, this analysis provides mechanistic insight into protein localization by identifying the biochemical rules that drive predictions for each compartment. The resulting hierarchy of physics-based feature utilization is illustrated in Figure 3B.

3.2.6 Directional Impact of Biophysical Features

The SHAP analysis showed a surprising pattern: the vast majority of features in the profile are so-called repellers, not attractors (Figure 4). This finding suggests that the implementation of a biologically plausible exclusionary logic with localization being determined by what a protein is not rather than what it is, is brought about in BioGraphX-Net. The most important feature, Profile_Cell_membrane_Residue, had strong negative SHAP values of 7 of the compartments but positive contribution for cell membrane localization. This confirms with biological reality where membrane proteins have distinctive physicochemical properties which exclude them by natural affinity from soluble compartments. Similarly, Profile_Peroxisomal_Residue showed extreme compartmentalization where SHAP values of 0.0434 were observed for peroxisome but consistently negative for every other compartment.

This exclusionary logic is based on the biological concept of signal specificity in protein targeting [Akopian et al., 2013]. Different organelles use different recognition machineries, nuclear pores for nuclear

localization [Paci et al., 2021], TIM/TOM complexes for mitochondria [Wiedemann and Pfanner, 2017] and Pex5 receptors for peroxisomes [Mindthoff et al., 2016], creating natural exclusion boundaries, which BioGraphX-Net detects using negative SHAP relationships.

3.2.7 Cross-Compartment Profile Affinities in Protein Targeting

SHAP analysis reveals counterintuitive cross-compartment profile affinities, providing insights into the functional architecture and evolutionary constraints of cellular trafficking systems. Convergent profile attractions show that organelles function as integrated hubs rather than isolated destinations.

Notably, the Golgi, cytoplasm, and membrane profiles show a strong association with the endoplasmic reticulum (ER), reflecting its central role in the biosynthesis of secretory and membrane proteins [Alberts et al., 2002]. This pattern highlights the functional integration of the ER: all membrane proteins are synthesized in the ER [Rapoport, 2007], all secreted proteins are processed there [Braakman and Bulleid, 2011], and both types of proteins subsequently transit to the Golgi apparatus [Glick and Nakano, 2009, Brandizzi and Barlowe, 2013]. The ER and Golgi thus form a continuous secretory network, explaining the co-attraction of their profile features, as shown in Figure 4.

Similarly, the plastid profile's attraction to mitochondria provides computational evidence of their shared endosymbiotic ancestry, capturing conserved targeting features in transit peptides that distinguish these organelles from non-endosymbiotic compartments [McFadden, 2014, Schleiff and Becker, 2011]. Conversely, the cytoplasmic profile shows self-repulsion, indicating that cytoplasmic localization is determined more by the absence of competing signals than by the presence of specific motifs, an exclusion-based targeting logic.

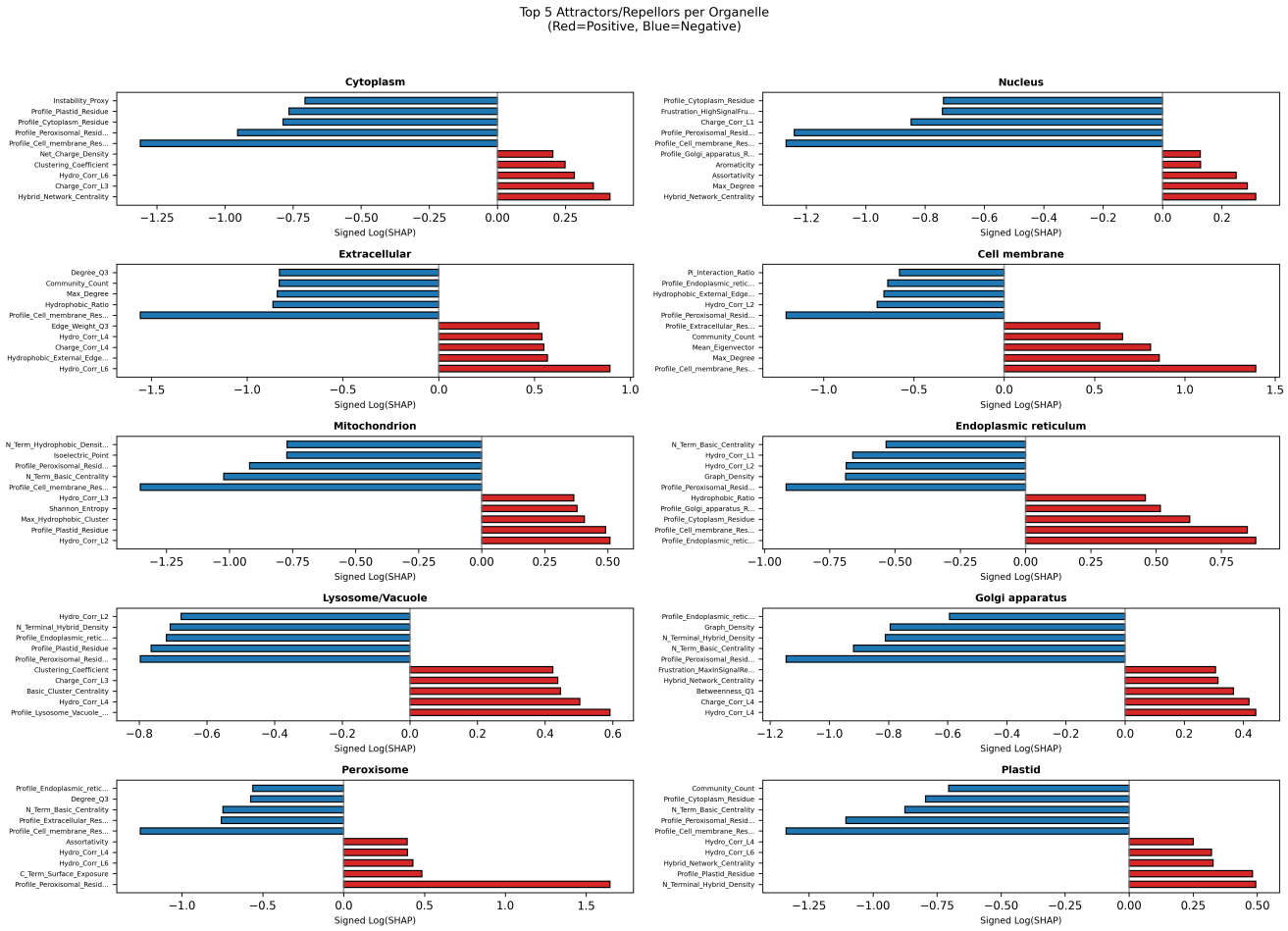


Figure 4: Top five organelle-specific features ranked by signed log-transformed feature importance values for improved visibility.

The Golgi profile's attraction to nucleus predictions likely reflects shared intrinsically disordered regions, as both Golgi-resident glycosyltransferases and nuclear transcription factors contain polar/charged-enriched disordered tails with similar amino acid compositions [Dyson and Wright, 2005, Tu and Banfield, 2010]. Frustration features provide the structural validation needed to differentiate true organelle-specific signals from compositional similarities.

Overall, these cross-compartment affinities show that BioGraphX-Net captures hierarchical targeting constraints and functional relationships within the cellular transport network, rather than relying on simple one-to-one sequence-to-compartment mappings. The model views organelles as integrated entities, emphasizing the endoplasmic reticulum (ER) as a key convergence point, the functional continuity between compartments of the secretory pathway, and the common targeting logic of organelles that have related functions. The model displays a complex understanding of cellular organization that transcends simple cytoplasmic compartmentalization.

3.2.8 Frustration Features Resolve Targeting Ambiguities

The proposed features based on frustration also present compartment-specific properties. These features can also account for the intricate import mechanisms of proteins. In nuclear localization, Frustration_HighSignalFrustration was identified as the fourth most important feature. In the endoplasmic reticulum, Frustration_CTerminal_Mean was also identified as a notable feature. In both cases, the negative sign of the features indicates that higher frustration is related to lower prediction confidence. This indicates that highly frustrated regions can create difficulties in the localization decisions of these compartments, as can be seen in Figure 4.

This also relates to the idea of frustration in protein folding, which states that interactions between residues can create conflicts in the energy of the protein sequence. [Ferreiro et al., 2018]. In the context of cellular compartments, these conflicts can relate to unclear signals that require additional regulatory mechanisms. For example, in nuclear localization, proteins with high N-terminal frustration can use chaperone-assisted transport or modification after translation to efficiently import [Christie et al., 2016]. This provides a biological rationale for why the model interprets high frustration as an alert signal rather than a direct determinant of localization.

3.2.9 Hydrophobicity Periodicity Encodes Compartment Identity

The SHAP analysis revealed that hydrophobic autocorrelation at certain sequence lags is a feature that is dependent on protein compartments. For instance, Hydro Corr L6, which is used to identify hydrophobic repeating patterns of 6 residues, was found to be the strongest positive predictor of extracellular proteins. The pattern of 3.6 residues is often seen as a helix [Rost et al., 1996]. In contrast, Hydro_Corr_L2 consistently contributed negatively to predictions for membrane, endoplasmic reticulum, and lysosomal compartments, indicating that short-range hydrophobic repetition is poorly tolerated in these targeting contexts.

3.2.10 Graph Topology Features Distinguish Structural Contexts

The relationship of Graph_Density with staying in the Golgi or endoplasmic reticulum demonstrates that the pattern of a protein's structure is important to how it behaves when it is located in those compartments. Some enzymes, such as glycosyltransferases, that are located in the Golgi have one membrane anchor and a catalytic domain that is located in the lumen [Tu and Banfield, 2010]. The number of strong interaction networks that enzymes such as this can have is restricted, which is why it is not good for graph density to be high for enzymes that are located in these compartments. The findings suggest that it is picking up on structural biases that are inherent in a protein's pattern of sequences. Also, larger hydrophobic patterns (lags of 4-6) could be associated with stable helical loops that facilitate proteins moving out of the cell and staying outside of it [Deber et al., 2001]. Shorter periodicities, on the other hand, may correspond to β-strand–like or structurally irregular regions, which appear more compatible with membrane-associated environments [Ulmschneider et al., 2010].

3.2.11 Charge Distribution Patterns as Targeting Determinants

N-terminal charge features exhibited pronounced compartment-specific effects. Notably, N_Term_Basic_Centrality was found to be very negatively correlated with localization to mitochondria, the Golgi, and plastids. This may seem strange at first because, of course, mitochondrial localization signals are well known to be amphipathic alpha helices with high basic residue content [von Heijne, 1986]. However, this makes more sense when one recalls that this is a graph-based algorithm, and it is not just the basic residue content that is important, but the centrality of those basic residues to the entire network. In mitochondrial presequences, basic residues may be distributed along the periphery of the network rather than occupying highly central positions, leading the model to down-weight this feature despite overall charge enrichment.

Charge periodicity features provided further insight into localization-specific coding. Charge_Corr_L4 contributed positively to extracellular and Golgi predictions, suggesting that regularly spaced charged residues may support interactions with components of the secretory pathway or enhance solubility during transit through these compartments [Kauko et al., 2010].

3.2.12 The Two-Stage Exclusion-Attraction Framework

Integrating the full set of SHAP patterns suggests that BioGraphX-Net operates through a two-stage decision process.

Stage 1: Exclusion via Profile Scores. In the first stage, strong negative SHAP contributions from profile-based features act as rapid exclusion signals, enabling the model to discard unlikely compartments early in the decision pathway. For example, a high membrane profile score effectively suppresses cytoplasmic, nuclear, and extracellular predictions, thereby narrowing the search space before finer distinctions are made.

Stage 2: Attraction via Specialized Physics. In the second stage, the remaining candidate compartments are resolved using organelle-specific combinations of graph topology, hydrophobicity periodicity, charge organization, and frustration-related features. These physics-informed signals provide the discriminative power required to separate closely related localization outcomes. This hierarchical strategy closely parallels biological targeting mechanisms, in which broad sorting signals are followed by receptor-mediated or context-dependent recognition steps [Söllner et al., 1993].

3.2.13 Evolutionary Insights from SHAP Patterns

Several SHAP patterns also point to underlying evolutionary signals. The strong negative contribution of Profile_Plastid_Residue to cytoplasmic predictions likely reflects the evolutionary separation between plastid-targeting signals, which originate from cyanobacterial endosymbionts, and the default localization landscape of the eukaryotic cytoplasm [McFadden, 2014]. In a similar vein, the shared negative effect of N_Term_Basic_Centrality on both mitochondrial and plastid predictions suggests common constraints faced by import systems that evolved from endosymbiotic ancestors.

3.2.14 Context-Dependent Switch Features Enable Precise Discrimination

SHAP analysis revealed that BioGraphX-Net leverages a set of context-dependent "switch" features that allow it to finely distinguish between evolutionarily convergent targeting signals. Rather than acting as simple attractors, these features function bidirectionally: they promote localization to compatible compartments while actively suppressing incompatible destinations, as shown in Figure 5.

This phenomenon is nicely illustrated in the isoelectric point (pI). This measures the variability in polarity based on cellular compartments. For example, extracellular proteins tend to have more acidic properties, whereas peroxisomal proteins have more basic properties. These are no arbitrary properties; they reflect the biochemistry of these compartments. The periodicity in hydrophobicity also behaves in a similar manner. For example, the hydro cor l6 periodicity feature strongly correlates with extracellular proteins, which makes sense since they often require helices for secretion. However, they also strongly correlate in with membrane proteins, which

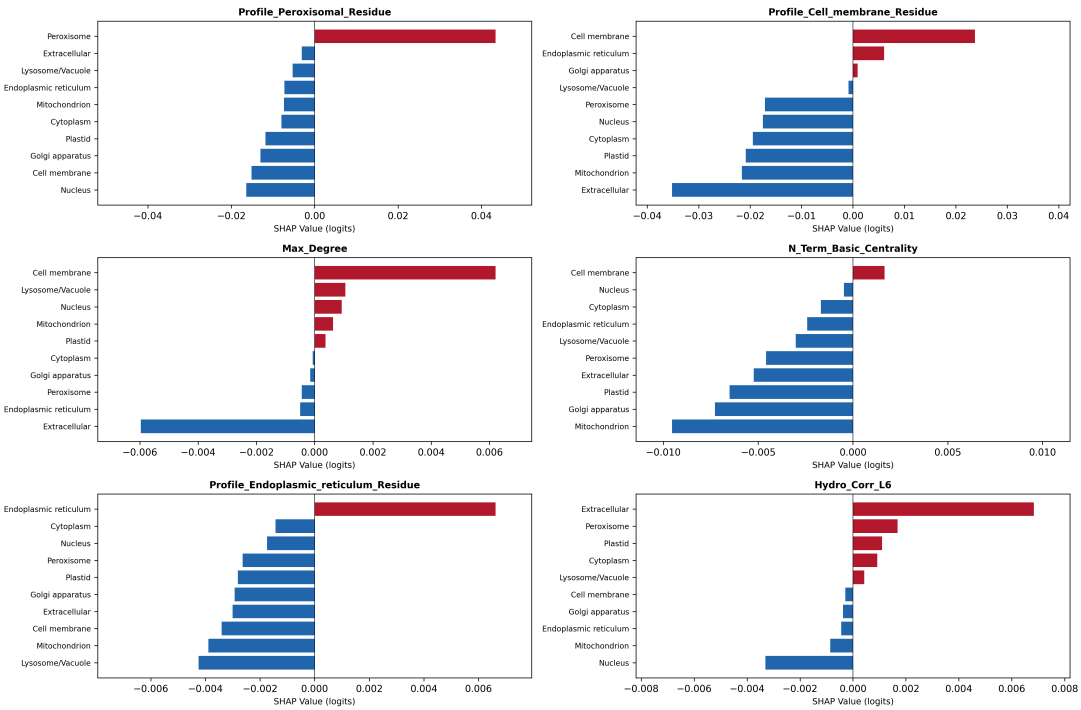


Figure 5: Bar chart highlighting six switch features that exert opposite SHAP impacts in different organelles.

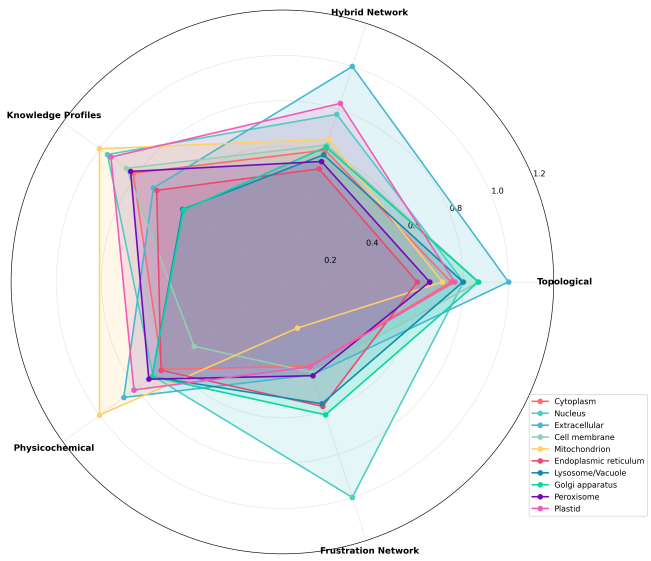


Figure 6: Radar chart summarizing the contribution of five feature categories.

makes more sense since these types of periodicities can interfere with the correct packing of transmembrane domains in a negative manner. So, it's not just learning what to favor; it's learning what to avoid.

By incorporating these conditional interactions, BioGraphX-Net addresses the phenomenon of evolutionary mimicry. Proteins can have ancestral targeting motifs, especially those of endosymbiotic origins, which can mimic true targeting motifs. These can confuse other models based purely on sequence. However, by considering these in the broader context of biochemistry, we avoid false positives. This is because the model is considering these in the same way that the cell does during quality control. The positive correlation between hydro cor l6 and extracellular localization may also suggest optimization of these features during evolution. Proteins destined for the extracellular compartment require very specific hydrophobic motifs in order to traverse the secretory pathway efficiently without aggregating. The model picks up on these requirements through the detection of long-range hydrophobic periodicity, which is in line with what is currently known in the field of secretory biology [Barlowe and Miller, 2013].

3.2.15 Organelle-Specific Biophysical "Fingerprints"

In the radar plot of feature category contributions, we find something very interesting: each compartment has a unique biophysical fingerprint (Figure 6). However, the patterns are not constant; they vary depending on the compartment. In most of the compartments, knowledge profile features are dominant. This is a logical expectation, as motif recognition based on a sequence has to be the primary localization cue. In many cases, the first question the cell asks is: does this sequence contain the right localization signal? But the story gets a little more complicated when we look at the complex localization machinery of some of the other organelles. Frustration features play a specialized role in the endoplasmic reticulum and the Golgi apparatus. This localization machinery is part of the secretory pathway, and the localization signals can be ambiguous and overlapping. In these cases, the frustration feature appears to be a disambiguating device to help the model differentiate among these very similar signals and avoid misclassifications. But perhaps the most intriguing pattern of contributions is the one we find in the nucleus. In the nucleus, we find a nice balance of contributions by the knowledge profiles and contributions by the topological and frustration features. This is a logical extension of the biological localization machinery to the nucleus. In the nucleus, localization signals are not sufficient; structural compatibility to pass through the nuclear pores is a requirement as well.

In the above discussion, we have seen how the model uses the different features for different organelles and how these patterns of feature contributions closely parallel the biological localization machinery of the different organelles. In the plasma membrane and the cytoplasm, the knowledge profile features are dominant, while in the nucleus, ER, and the Golgi apparatus, the topological and frustration features play a role as a second level of validation. This means that the model does not have a one-size-fits-all approach to localization; the approach varies depending on the biological complexity of the localization machinery of the different organelles. Physicochemical features provide moderate, consistent contributions across compartments, acting as universal biophysical validators that complement the primary knowledge-based signals. Notably, the elevated frustration contributions in secretory compartments mirror cellular quality control, where misfolded or mis-targeted proteins are actively rejected, while the balanced feature utilization in peroxisomes reflects their dual recognition system combining PTS signal detection with receptor-mediated structural compatibility.

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. 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

What makes BioGraphX-Net stand out is its capability to grasp and visualize the real complexity of protein localization. The model's use of evolutionary information and biophysical rules in a gated graph architecture allows it to retain the subtleties of the relationships between residues and domains that might otherwise go unnoticed. This allows the model to grasp the nuances of protein localization. The model does not simply look for sequence motifs; it also assesses the structural compatibility of the protein.

This makes BioGraphX-Net a reflection of a very important aspect of biological reality. Protein localization is not determined by sequence motifs; rather, it is the result of the interplay between motif detection, spatial arrangement, and biophysical considerations. These findings emphasize the importance of considering topology and biophysics in computational models. They also provide a platform for biologically informed hypothesis generation. They also highlight the capabilities of the model in guiding experiments, especially in cases of proteins whose localization cannot be determined by sequence features. What this also shows is that protein localization is not simply the detection of sequence motifs; rather, the spatial arrangement of residues also plays a role.

4.3 Resolving the Eukaryotic Secretory Pathway

SHAP analysis produced an interesting result: features of the sequence profile do not serve as simple attractors. Instead, they can serve as universal negative selection filters, which is an unexpected behavior compared to other localization models.

A good example of this is the localization of proteins to the Golgi apparatus. Initially, the sequence profile of the Golgi apparatus was found to act as a negative predictor of localization to the Golgi apparatus itself, which produced an apparent contradiction or "self-repulsion." However, upon inclusion of the frustration features, the contradiction was resolved to reveal a subtle two-tier validation mechanism. Thus, the sequence profile of proteins, particularly those with similarities arising from disordered regions of proteins found in the nucleus or cytoplasm, was not sufficient, on its own, to result in a positive prediction of localization to the Golgi apparatus. It was necessary to provide an additional level of validation through biophysical means, such as hydrophobic edge patterns or sulfur-aromatic interactions, to result in a prediction of localization to the Golgi apparatus. This means that the sequence profile was viewed with suspicion, but upon inclusion of an additional level of validation, localization was favored. Thus, the mechanism avoids false positives arising from mimicry, allowing genuine Golgi apparatus proteins to pass through structural validation while filtering out misleading similarities.

Importantly, this negative selection logic is not restricted to the Golgi. Membrane profiles are highly repulsive to cytoplasmic and nuclear predictions. Peroxisomal profiles negatively impact Golgi predictions. Finally, plastid-based predictions negatively impact cytoplasmic defaults. The model learns not only what is present but also what is absent. The incorporation of frustration features allows BioGraphX-Net to make contextual sense of the input sequence profile. Rather than being a straightforward match to a motif, predictions must be supported by biophysical data before being considered. This is analogous to quality control mechanisms within the cell, where targeting signals must be validated for structural compatibility before final decisions on localization are made. The biological implications of this approach provide an explanation for how eukaryotic cells can accurately localize proteins while being composed of similar sequences, especially when comparisons are made to other compartments that are evolutionarily similar or structurally similars. Thus, BioGraphX-Net is more than just a pattern recognition algorithm. It is an algorithm that questions, evaluates, and validates its predictions, much like the cell itself.

4.4 Biological Significance of the Physics-Gated Fusion

The architecture of BioGraphX-Net is based on the observation that high-dimensional evolutionary embeddings and low-dimensional biophysical features are asymmetric. Placing these features side by side would unfairly favor evolutionary features because of their higher dimensionality. The higher dimensionality of the evolutionary features, together with their pretrained architecture, naturally gives them a dominant advantage in the learning process.

In order to achieve symmetry between the features, the network adds more representational power to the biophysical features. The model adds an intermediate hidden projection that increases the dimensionality of the features. This allows the physics-based features to compete with the evolutionary features in a more symmetrical manner. The model adds more power in the form of an additional dimension in the hidden representation. This allows the model to have the power to accommodate complex biophysical constraints that were previously underrepresented. This architecture is more than just concatenating features. The conventional architecture of most models, in which pretrained evolutionary features, such as those of ESM-2, dominate all other features because of their higher dimensionality and pretrained architecture, is not symmetrical.

Although the power of evolutionary features cannot be de-

nied, they can also be very misleading. Sequence conservation can continue in a protein because of factors other than evolution. proteins with unique functions or localization, thus causing false positives where similarity is incorrectly assumed to reflect targeting logic.

By contrast, the biophysical features incorporated into BioGraphX allow for a direct consistency check. They force the predictions to conform to physical and structural reality, providing a grounding effect that counterbalances the evolutionary considerations. To reconcile these two forms of information, an adaptive gating mechanism is used, along with a physics-first training approach. This ensures that the biophysical information is not overshadowed by the evolutionary considerations. Instead of combining the features statically, the gating mechanism dynamically weighs the contribution of the evolutionary and biophysical signals for each protein. BioGraphX-Net thus produces a model that minimizes noise, avoids over-reliance on evolutionary considerations, and optimally balances the two forms of information. It does not privilege one view over the other but, rather, constantly balances between evolutionary memory and physical plausibility, just as the cell balances heritage and physical constraint.

4.5 Comparison with Alternative Approaches

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

The traditional Structure-Based Graph Neural Networks [Gligorijević et al., 2021] require high quality three-dimensional protein structures, which can be obtained through [Jumper et al., 2021] or through experiments using X-ray crystallography. These methods have been proven to perform well, especially for well-folded and structurally stable proteins. However, these methods are also accompanied by considerable limitations. First of all, the generation of structural predictions using AlphaFold2 for the entire proteome is 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.

4.5.2 Contact Map–Based Methods

Contact prediction approaches, such as PSICOV 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.

4.5.3 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.

4.6 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.

4.7 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 64% 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". SHAP analysis further shows that the model uses a universal negative selection strategy. Instead of acting as simple attractors, sequence profiles serve as exclusionary filters that quickly rule out unlikely compartments, while organelle-specific combinations of graph topology, hydrophobicity periodicity, and frustration features guide precise localization. This two-tier system resolves evolutionary mimicry, such as the shared secretory signatures of the Golgi and ER, while preventing mislocalization through structural validation. 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

[Akopian et al., 2013] Akopian, D., Shen, K., Zhang, X., and Shan, S.-o. (2013). Signal recognition particle: an essential protein-targeting machine. *Annual Review of Biochemistry*, 82:693–721. Epub 2013 Feb 13.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Protein sorting. In *Molecular Biology of the Cell*. Garland Science, 4 edition. Available via NCBI Bookshelf.

[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.

[Barlowe and Miller, 2013] Barlowe, C. K. and Miller, E. A. (2013). Secretory protein biogenesis and traffic in the early secretory pathway. *Genetics*, 193(2):383–410.

[Bartlett et al., 2010] Bartlett, G. J., Choudhary, A., Raines, R. T., and Woolfson, D. N. (2010). n$\rightarrow \pi^*$ interactions in proteins. *Nature Chemical Biology*, 6(8):615–620.

[Braakman and Bulleid, 2011] Braakman, I. and Bulleid, N. J. (2011). Protein folding and modification in the mammalian endoplasmic reticulum. *Annual Review of Biochemistry*, 80:71–99.

[Brandizzi and Barlowe, 2013] Brandizzi, F. and Barlowe, C. (2013). Organization of the er–golgi interface for membrane traffic control. *Nature Reviews Molecular Cell Biology*, 14(6):382–392.

[Briesemeister et al., 2010] Briesemeister, S., Rahnenfuehrer, J., and Kohlbacher, O. (2010). Yloc: an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, 38(Web Server issue):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.

[Christie et al., 2016] Christie, M., Chang, C.-W., Róna, G., Smith, K. M., Stewart, A. G., Takeda, A. A., Fontes, M. R. M., Stewart, M., Vértessy, B. G., Forwood, J. K., and Kobe, B. (2016). Structural biology and regulation of protein import into the nucleus. *Journal of Molecular Biology*, 428(10 Pt A):2060–2090.

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

[Deber et al., 2001] Deber, C. M., Wang, C., Liu, L.-P., Prior, A. S., Agrawal, S., Muskat, B. L., and Cuticchia, A. J. (2001). Tm finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Science*, 10(1):212–219.

[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.

[Dyson and Wright, 2005] Dyson, H. J. and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, 6:197–208.

[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.

[Ferreiro et al., 2018] Ferreiro, D. U., Komives, E. A., and Wolynes, P. G. (2018). Frustration, function and folding. *Current Opinion in Structural Biology*, 48:68–73.

[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.

[Gillani and Pollastri, 2024] Gillani, M. and Pollastri, G. (2024). Protein subcellular localization prediction tools. *Computational and Structural Biotechnology Journal*, 23:1796–1807.

[Glick and Nakano, 2009] Glick, B. S. and Nakano, A. (2009). Membrane traffic within the golgi apparatus. *Annual Review of Cell and Developmental Biology*, 25:113–132.

[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.

[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., 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.

[Kauko et al., 2010] Kauko, A., Hedin, L. E., Thebaud, E., Cristobal, S., Elofsson, A., and von Heijne, G. (2010). Repositioning of transmembrane alpha-helices during membrane protein folding. *Journal of Molecular Biology*, 397(1):190–201.

[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.

[Kumar and Nussinov, 1999] Kumar, S. and Nussinov, R. (1999). Salt bridge stability in monomeric proteins. *Journal of Molecular Biology*, 293(5):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.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

[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.

[Mindthoff et al., 2016] Mindthoff, S., Grunau, S., Steinfort, L. L., Girzalsky, W., Hiltunen, J. K., Erdmann, R., and Antonenkov, V. D. (2016). Peroxisomal pex11 is a pore-forming protein homologous to trpm channels. *Biochimica et Biophysica Acta*, 1863(2):271–283.

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

[Paci et al., 2021] Paci, G., Caria, J., and Lemke, E. A. (2021). Cargo transport through the nuclear pore complex at a glance. *Journal of Cell Science*, 134(2):jcs247874.

[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.

[Rapoport, 2007] Rapoport, T. A. (2007). Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*.

[Rost et al., 1996] Rost, B., Fariselli, P., and Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science*, 5(8):1704–1718.

[Schleiff and Becker, 2011] Schleiff, E. and Becker, T. (2011). Common ground for protein translocation: access control for mitochondria and chloroplasts. *Nature Reviews Molecular Cell Biology*, 12:48–59.

[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.

[Söllner et al., 1993] Söllner, T., Whiteheart, S., Brunner, M., et al. (1993). Snap receptors implicated in vesicle targeting and fusion. *Nature*, 362:318–324.

[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.

[Trachsel et al., 2015] Trachsel, M. A., Ottiger, P., Frey, H.-M., Pfaffen, C., Bihlmeier, A., Klopper, W., and Leutwyler, S. (2015). Modeling the histidine-phenylalanine interaction: The nh-π hydrogen bond of imidazole benzene. *The Journal of Physical Chemistry B*, 119(25):7778–7790.

[Tu and Banfield, 2010] Tu, L. and Banfield, D. K. (2010). Localization of golgi-resident glycosyltransferases. *Cellular and Molecular Life Sciences*, 67(1):29–41.

[Ulmschneider et al., 2010] Ulmschneider, M. B., Doux, J.-P., Killian, J. A., Smith, J. C., and Ulmschneider, J. P. (2010). Mechanism and kinetics of peptide partitioning into membranes from all-atom simulations of thermostable peptides. *Journal of the American Chemical Society*, 132(10):3452–3460.

[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.

[von Heijne, 1986] von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *The EMBO Journal*, 5(6):1335–1342.

[Wan et al., 2017] Wan, S., Mak, M.-W., and Kung, S.-Y. (2017). Fuel-mloc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. *Bioinformatics*, 33(5):749–750.

[Wiedemann and Pfanner, 2017] Wiedemann, N. and Pfanner, N. (2017). Mitochondrial machineries for protein import and assembly. *Annual Review of Biochemistry*, 86:685–714.

[Xiao et al., 2024] Xiao, H., Zou, Y., Wang, J., and Wan, S. (2024). A review for artificial intelligence based protein subcellular localization. *Biomolecules*, 14(4):409.

[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.