Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BioGraphX-Sol: A Structure-Free, Fully Interpretable Protein Solubility Predictor Exceeds AlphaFold3-Based Methods and Reveals Nine Universal Cross-Species Determinants

Abubakar Saeed

abubakarsaeed915@gmail.com

Research Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11150377/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

BioGraphX-Sol: Sequence-Derived Physicochemical Graph Encoding Outperforms Structure-Based Methods for Interpretable Protein Solubility Prediction

Abubakar Saeed*[1]

[1]Independent Researcher, Chiniot, Pakistan

*Email: abubakarsaeed915@gmail.com

Abstract

Protein solubility is a critical determinant of recombinant expression yield, yet current state-of-the-art predictors often rely on either AlphaFold3-predicted structures or large ensembles of protein language models, limiting accessibility and interpretability. We present BioGraphX-Sol, a sequence-only predictor that constructs biophysically grounded constraint graphs directly from amino acid sequences to extract 200 physicochemical features spanning graph topology, IDP and polymer physics, aggregation-hotspot geometry, and novel surface- and terminal-conditioned descriptors, without requiring 3D coordinates, multiple sequence alignments, or pretrained models. On the standard eSOL benchmark, the XGB+ESM post-hoc ensemble essentially matches ProtSATT in predictive performance ($R^2 = 0.5451$ vs. $R^2 = 0.5450$), despite training only 1.44M neural parameters and requiring no structure prediction. The physics-only XGBoost configuration ($R^2 = 0.499$, ROC–AUC $= 0.873$) exceeds FGNNSol ($R^2 = 0.487$, ROC–AUC $= 0.867$), which combines AlphaFold3-derived 3D coordinates with ESM-C embeddings and structural descriptors across roughly 1,831 input dimensions, showing that explicit 3D structure prediction is unnecessary when the relevant physicochemical information is correctly encoded. A post-hoc α sweep quantifies the physicochemical-to-evolutionary information balance at approximately 59%/41%, providing, to our knowledge, the first quantitative decomposition of information sources for protein solubility prediction. In zero-shot transfer to *S. cerevisiae*, the physics-only XGBoost achieves the highest ROC–AUC (0.8524) and PR–AUC (0.9246) among all evaluated methods. Permutation-importance analysis identifies nine universal solubility determinants exhibiting consistent directional effects across *E. coli* and *S. cerevisiae* (9/9 agreement). The physics-only configuration requires no GPU at inference and runs efficiently on standard CPU hardware, supporting the Green AI principle that compact, interpretable representations grounded in domain knowledge can provide a viable alternative to parameter scaling.

Keywords

protein solubility prediction, physicochemical graph encoding, interpretable machine learning, cross-species generalization, Green AI

Introduction

Protein solubility is a fundamental physicochemical property governing success across industrial biotechnology, biopharmaceutical development, and basic research. High solubility promotes proper folding and recovery during bioprocessing, while insolubility drives inclusion body formation, aggregation, and immunogenicity in therapeutic formulations [1, 2]. Beyond bioprocessing, aberrant protein aggregation underlies a broad class of human disorders, including amyloidoses, neurodegenerative diseases [3, 4], and cataract formation [5], linking protein solubility directly to cellular proteostasis and disease pathogenesis. In recombinant Escherichia coli expression systems, the dominant workhorse of protein production, a large fraction of overexpressed proteins misfold into insoluble aggregates, creating a screening bottleneck that experimental means alone cannot resolve. Turbidimetry, dynamic light scattering, and aggregation-based assays are low-throughput, require purified material, and cannot scale to the volumes of candidate sequences

generated by modern genomics and protein design pipelines [6, 7]. Computational solubility prediction from sequence alone would enable rapid screening before experimental commitment, substantially cutting development time and reagent use.

Early computational predictors established that amino acid composition carries measurable solubility signal, exploiting the statistical enrichment of charged residues and depletion of hydrophobic stretches in soluble proteins. Methods such as SOLpro [8] and CamSol [9] offered proteome-scale applicability but lacked the representational capacity to model residue interactions that are spatially proximal in the folded structure yet distant in primary sequence. The availability of large benchmarks, particularly the eSOL dataset of 3,173 E. coli proteins with continuous solubility measurements from PURE cell-free expression [10], enabled more expressive models. SoluProt [11] advanced the field with random forest classifiers over physicochemical sequence descriptors, while DeepSol [12] introduced convolutional neural networks over one-hot and physicochemical encodings, demonstrating the value of local sequence context.

The field's subsequent shift toward structure-informed prediction raised performance substantially. GraphSol [13] showed that graph convolutional networks operating on residue contact maps outperform sequence-based approaches, establishing spatial proximity as a key inductive bias. GATSol [14] extended this by integrating graph attention over AlphaFold2-predicted structures with protein language model embeddings. The most recent structure-dependent method, FGNNSol [15], combines ESM-C embeddings with AlphaFold3-predicted structural coordinates in a hybrid graph model, representing the current frontier of structure-aware solubility prediction. Despite strong empirical performance, structure-dependent methods carry three fundamental liabilities. First, they introduce a computational bottleneck: AlphaFold3 prediction demands substantial GPU memory and minutes per protein, ruling out real-time proteome-scale screening [16]. Second, they propagate structural prediction uncertainty through every network layer, a concern that is sharpest for low-homology and intrinsically disordered proteins, precisely those most in need of expression optimization, since AlphaFold produces structured loop representations for disordered regions that are physically inaccurate [17]. Third, their predictions cannot be attributed to named biophysical features, limiting their usefulness for mechanistic reasoning and sequence design.

A parallel paradigm, protein language models (PLMs), encodes evolutionary co-occurrence patterns from hundreds of millions of protein sequences without structural coordinates. PLM-Sol [18], MTPSol [19], and RP3Net [20] progressively showed that frozen or fine-tuned PLM embeddings provide competitive solubility prediction without the structure-prediction overhead. The current state of the art, ProtSATT [21], fuses three complementary PLMs, UniRep, ESM-2, and ProtT5, totalling roughly 1.87 billion backbone parameters. PLM-based methods sidestep the structure-prediction dependency, but introduce their own limitations: substantial computational infrastructure for training and inference, evolutionary biases that may not transfer across organisms, and the same interpretability deficit as structure-based methods, predictive signal distributed across billions of learned parameters. Both paradigms embody the "Red AI" trend of trading efficiency and interpretability for marginal accuracy gains [22], together forming the dominant but limiting approach to this problem.

We propose that these limitations share a common root: neither paradigm has systematically encoded the biophysical mechanisms that govern solubility from first principles. Protein aggregation and insolubility are driven by identifiable, quantifiable physicochemical forces, the density and spatial distribution of hydrophobic and aromatic sticker residues driving phase separation [23], the segregation of oppositely charged residues into complementary electrostatic patches [24], the number and spatial arrangement of independent aggregation-prone segments, and the completeness of folding topology captured by C-terminal long-range contact order. These properties are biologically universal, recoverable from sequence alone, and directly actionable as protein engineering targets. A predictor that encodes them explicitly and deterministically should need fewer parameters, generalize across species, and yield attributions that connect prediction to mechanism. The central premise of this work is that solubility prediction does not need more information, it needs the right information, correctly encoded.

This work addresses six research questions: (i) Can sequence-derived physicochemical constraint graphs, without 3D structure prediction or multiple PLMs, match or exceed current state-of-the-art solubility predictors on the standard eSOL benchmark? (ii) Does explicit 3D structure prediction add predictive value beyond what a deterministic, sequence-derived structural proxy already recovers? (iii) What fraction of the predictive signal for protein solubility comes from physicochemical principles versus evolutionary information, and does this balance hold across evaluation settings? (iv) Does the model's learned allocation of physicochemical versus evolutionary information shift systematically under cross-species distribution shift, and is that shift consistent with known solubility biology? (v) Do physicochemical features encoding species-independent biophysical properties generalize better to a held-out species than evolutionary embeddings calibrated on a single-organism training corpus? and (vi) Which individual physico-

chemical features are universal solubility determinants across species, and what mechanism explains generalization failure for features that do not transfer?

To address these questions, we present BioGraphX-Sol, a protein solubility predictor that builds biophysically grounded constraint graphs directly from amino acid sequences, without 3D coordinates, multiple sequence alignments, or pretrained models, and extracts 200 physicochemical features spanning graph topology, IDP and polymer physics, aggregation hotspot geometry, surface- and terminal-conditioned variants, and global physicochemical scalars. Building on the structural proxy paradigm of BioGraphX [25] and BioGraphX-RNA [26], BioGraphX-Sol introduces novel features: five aggregation-hotspot-geometry descriptors novel to any biological prediction domain, and surface- and terminal-conditioned physicochemical variants novel to solubility prediction. Notably, the original BioGraphX constraint-graph encoding was already validated, without modification, on the E. coli eSOL solubility benchmark, achieving a recall of 0.726, comparable to the structure-dependent fGNNSol (0.734), providing early evidence that sequence-derived physicochemical graph features capture structural information relevant to solubility. A post-hoc ensemble combining the physicochemical encoding with a frozen ESM-2 backbone provides a quantitative decomposition of physicochemical versus evolutionary information contributions.

We make four specific contributions:

(i) Performance. The XGB+ESM post-hoc ensemble matches ProtSATT on the standard eSOL benchmark ($R^2 = 0.5451$ vs. 0.5450) while requiring no structure prediction, no language model fine-tuning, and a fraction of the parameters, 1.44M. Physicochemical features alone already exceed FGNNSol, a method that relies on AlphaFold3-predicted structure, showing that explicit 3D structure prediction offers no advantage over a deterministic, sequence-derived structural proxy.

(ii) Mechanistic interpretability. A post-hoc alpha sweep quantifies the physicochemical-to-evolutionary information balance for solubility at roughly 59%/41%, the first such quantification for this task, and situates it within a cross-task pattern in which physically defined biological tasks lean toward physicochemical signal, while functionally annotated tasks are evolution-dominated. Leave-one-out ablation and permutation importance independently identify IDP and polymer physics features, particularly sticker–spacer patterning and charge segregation, as the most predictive category.

(iii) Cross-species generalization. In zero-shot evaluation on an external S. cerevisiae benchmark, physicochemical features outperform every baseline on both ranking metrics tested, and permutation importance identifies a small set of universal solubility determinants with consistent directional signal across three independent datasets.

(iv) Green AI and generalizability. The physics-only configuration requires no GPU at inference, no language model, and no structure prediction, running competitively on standard CPU hardware. The alpha-sweep information-decomposition methodology, applied here to solubility for the first time, generalizes to any two-stream biological prediction problem.

Through extensive benchmarking and ablation analysis, BioGraphX-Sol demonstrates that a compact, interpretable, sequence-derived encoding can outperform structure-based solubility predictors and match language-model-based ones, while requiring far fewer computational resources and generalizing across species

Methods

Datasets and Preprocessing

We used the eSOL dataset as our primary regression benchmark, E. coli proteins expressed in a PURE cell-free system, with solubility measured as the supernatant-to-total protein ratio [10]. We followed the split introduced by GATSol [14] and later adopted by ProtSATT [21]: 2019 training sequences and 660 test sequences, with the target variable continuous in (0,1).

For zero-shot cross-species evaluation, we used the external S. cerevisiae set of 108 sequences [27], following the same GATSol/ProtSATT protocol. No yeast sequences entered training, validation, or hyperparameter selection at any point; every model was applied to this set only after training on eSOL was complete.

Each protein was encoded into a 200-dimensional feature vector using the BioGraphX-Sol encoder. We also generated per-residue embeddings from ESM-2 (esm2_t36_3B_UR50D) for all proteins [28]. After removing the BOS and EOS tokens, the final hidden states were mean-pooled into a single 2560-dimensional vector per protein. Embeddings were precomputed once and reused across all experiments.

Overview of BioGraphX-Sol Architectures

BioGraphX-Sol separates physicochemical and evolutionary information into two independent streams. The pipeline consists of three stages: sequence-derived physicochemical encoding, independent training of each stream, and post-hoc fusion or learned gating.

Each protein is encoded into a 200-dimensional feature vector that captures graph-topological, polymer-physics, aggregation-geometry, surface-conditioned, and terminal-conditioned properties. Two models are trained on these handcrafted features: an XGBoost regressor that serves as the primary physicochemical stream and a graph-only multilayer perceptron (MLP). In parallel, the same protein sequence is processed by a frozen ESM-2 language model. Per-residue embeddings are mean-pooled to generate a 2560-dimensional representation, which is used to train an ESM-only neural network.

Post-hoc ensemble models combine the physicochemical and evolutionary streams using a weighting coefficient, α, selected on the validation set. The optimal value of α provides a quantitative estimate of the relative contribution of each information source. A gated hybrid model instead learns protein-specific fusion weights, enabling analysis of how confidence is dynamically allocated between physicochemical and evolutionary signals under distribution shift, as illustrated in Figure 1.

Because all 200 features are explicitly defined and biophysically interpretable, the framework supports multiple interpretability analyses, including permutation importance, Spearman directional validation, gate attribution analysis, and α-sweep information decomposition. The ESM-2 backbone remains frozen throughout training, restricting trainable parameters to lightweight downstream models and reducing computational requirements.

This design enables BioGraphX-Sol to address not only whether accurate protein solubility prediction can be achieved, but also how much predictive signal originates from physicochemical principles versus evolutionary information, and which biophysical features generalize most effectively across evaluation settings.

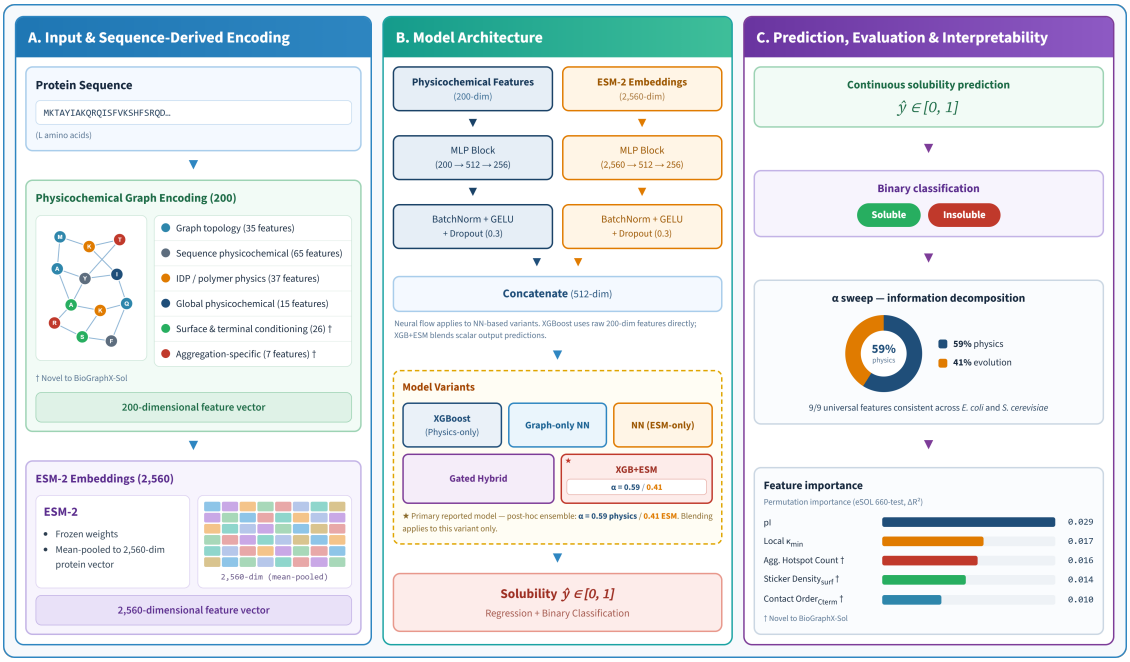


Figure 1: **BioGraphX-Sol workflow.** **(A)** Sequence-derived physicochemical graph encoding (200 features) and frozen ESM-2 embeddings (2,560 dimensions) are generated in parallel from the input protein sequence. **(B)** Six model variants are evaluated, with the primary XGB+ESM ensemble combining physicochemical and evolutionary predictions at a per-seed validation-selected α (mean $= 0.59$, SD $= 0.06$) **(C)** Solubility prediction, classification, and interpretability analyses reveal the physicochemical-to-evolutionary information balance and key determinants of protein solubility.

BioGraphX-Sol Encoding Framework

Constraint graph construction. Following the structural proxy paradigm introduced in BioGraphX [25] and extended in BioGraphX-RNA, [26], BioGraphX-Sol constructs a weighted undirected constraint graph $G(V, E)$ directly from the primary amino acid sequence, without requiring 3D coordinates, multiple sequence alignments, or pretrained models. Each residue is represented as a node, while edges connect residue pairs satisfying any of twelve biophysically grounded interaction rules: hydrophobic contacts, hydrogen bonds, salt bridges, disulfide bonds, aromatic π–π stacking, cation–π interactions, van der Waals contacts, C–H$\cdots\pi$, N–H$\cdots\pi$, carbonyl–carbonyl interactions, sulfur–π interactions, and backbone bonds. Each interaction type is associated with a calibrated strength coefficient, w_{type}, and a maximum allowable sequence separation, $d_{\max}$. Edge weights decay with sequence distance according to

$$w_{ij} = w_{\text{type}} \left(1 - 0.3 \frac{d_{ij}}{d_{\max}}\right). \tag{1}$$

where d_{ij} denotes the linear sequence distance between residues i and j. When multiple interaction types simultaneously satisfy their criteria for the same residue pair, the edge is classified as a hybrid interaction, assigned the dominant interaction type according to strength, and up-weighted using a type-specific cooperative multiplier. Backbone edges between adjacent residues are added at a fixed weight of 1.0 whenever no other interaction already exists. Interaction definitions, distance thresholds, and strength coefficients follow [25] without modification. Graph construction is fully deterministic and reproducible from sequence alone, as illustrated in Figure 1 (panel A).

Multi-scale feature extraction

From each constraint graph, BioGraphX-Sol extracts a 200-dimensional protein-level feature vector organized into six feature groups (Table 1). Per-residue quantities are aggregated to the protein level using four summary statistics: mean, standard deviation, minimum, and maximum. The resulting feature set captures complementary aspects of protein topology, physicochemical properties, intrinsic disorder, aggregation propensity, and sequence composition. Complete feature definitions and group assignments are provided in Supplementary Table S1.

Graph topology features

Graph topology features characterize the structure of the intramolecular interaction network. Node-level descriptors, including degree, weighted degree, betweenness centrality, closeness centrality, clustering coefficient, hybrid-edge ratio, surface exposure proxy, and local edge density, are aggregated using the same four summary statistics. Distributional descriptors such as the maximum and standard deviation of betweenness and weighted degree capture network heterogeneity beyond mean-level summaries and consistently rank among the most important features across species.

Sequence physicochemical features

Sequence physicochemical features describe local residue properties using sliding windows of 7, 15, and 51 residues. These features include local hydrophobicity (Kyte–Doolittle scale) [29], local charge, local helix and β-sheet propensity (Chou–Fasman scale) [30, 31], disorder and order residue fractions, hydrophobicity and charge autocorrelation at lag 1, local bulkiness, Pro–Gly density, and predicted secondary-structure fractions derived from local propensity comparisons.

IDP and polymer physics features

IDP and polymer physics features encode the biophysical grammar of intrinsically disordered proteins and liquid–liquid phase separation, transferred here to the protein-solubility domain. Fractional charge (FCR) and net charge per residue (NCPR) are computed within 21-residue sliding windows. Local_Kappa quantifies the minimum local deviation of absolute charge density from the global net charge magnitude across 15-residue windows. A high Local_Kappa_min indicates that even the least charged window retains substantial charge density, i.e., charge is uniformly distributed, whereas a low value indicates windows depleted of charge that may promote aggregation. Sticker density (Y, F, W, R), spacer density (G, S, P), and sticker dispersion are computed using the Pappu sticker–spacer framework for phase-separating polymers. Additional features include MoRF propensity scores (core, flank, and dip components)

[32], polyampholyte score, per-residue Uversky distance, multi-scale Shannon entropy across window sizes of 5, 10, 25, and 50 residues, complexity-scale variance, non-backbone contact order, and Ramachandran backbone flexibility.

Global physicochemical features

Global physicochemical features capture whole-protein scalar properties, including isoelectric point (pI), GRAVY score, a binary-approximated instability index derived from dipeptide composition, aliphatic index, aromaticity, mean amyloid propensity (WALTZ scale), maximum and mean lengths of hydrophobic stretches (A, V, I, L, M, F, W, Y), and cysteine count. Two charge-patterning descriptors are also included. Sequence charge decoration (SCD) is defined as

$$\mathrm{SCD} = \frac{1}{L} \sum_{i<j} \frac{q_i q_j}{\sqrt{j-i}}. \tag{2}$$

where $q_i \in \{+1, -1, 0\}$ represents the charge state of residue i, and L denotes sequence length. Global charge asymmetry between the N-terminal and C-terminal halves is also included.

Surface-conditioned and terminal features

Surface-conditioned and terminal features represent a primary methodological novelty of BioGraphX-Sol. Existing surface-aware solubility predictors, including SurfSol [33] and FGNNSol [15], require solvent-accessible surface area derived from experimentally determined or AlphaFold-predicted structures. In contrast, BioGraphX-Sol estimates residue surface exposure directly from graph topology:

Table 1: Composition and biological rationale of the 200 BioGraphX-Sol features. Features were selected from a larger candidate set using gain-based importance on the training set while preserving representation from all feature groups. Complete feature definitions are provided in Supplementary Table S1.

Feature Group	Count	Representative Features	Biological Rationale
Graph Topology	43	Degree, weighted degree, betweenness, clustering coefficient, hybrid-edge ratio, local edge density, surface proxy, frustration, contact order	Captures intramolecular interaction network organization as a sequence-derived proxy for folding topology and aggregation-prone regions.
Sequence Physicochemical	65	Local hydrophobicity, local charge, helix/sheet propensity, disorder fraction, hydrophobicity and charge autocorrelation, bulkiness, Pro–Gly density, multi-scale Shannon entropy	Represents local residue environments, secondary-structure tendencies, and sequence complexity across multiple scales.
IDP / Polymer Physics	37	FCR, NCPR, κ, sticker density, spacer density, sticker dispersion, polyampholyte score, MoRF propensity, Uversky distance, non-backbone contact order	Encodes charge patterning, polymer-physics behavior, and phase-separation properties associated with aggregation.
Global Physicochemical	15	pI, GRAVY, instability index, aromaticity, aliphatic index, WALTZ amyloid propensity, hydrophobic stretch statistics, SCD, charge asymmetry	Describes whole-protein stability, charge state, hydrophobicity, and aggregation propensity using global descriptors.
Surface & Terminal Conditioning	26	Surface-derived graph features and N-/C-terminal physicochemical descriptors	Captures surface-exposed aggregation determinants and terminus-specific folding and charge signals.
Aggregation & Interface	14	Aggregation hotspot count, hotspot variance, maximum hotspot score, top-5 hotspot mean, mean hotspot score; interface propensity statistics and interface-derived disorder scores	Quantifies the number, strength, and spatial distribution of aggregation-prone regions, and integrates graph-centrality and sequence-derived interface propensity to identify disordered regions and potential interaction surfaces.

$$\mathrm{Surface}_i = \sigma\left[1 - 2\left(\tilde{b}_i + \tilde{c}_i\right)\right], \tag{3}$$

where $\tilde{b}_i$ and $\tilde{c}_i$ denote the min–max normalized betweenness and closeness centralities of residue i, respectively, and σ is the logistic sigmoid function. Residues with high betweenness and low closeness are inferred to be buried, while the complement corresponds to surface-exposed regions.

Up to ten key physicochemical features are averaged over residues satisfying $\mathrm{Surface}_i > 0.5$, producing _surf variants. The same features are computed over the first and last 30 residues, producing _Nterm and _Cterm variants. Four explicit terminal descriptors are additionally included: Nterm_Charge, Cterm_Charge and Nterm_HydroFraction.

Aggregation-specific features

Aggregation-specific features characterize the spatial distribution of aggregation propensity. A per-residue aggregation score is defined as

$$\mathrm{agg}_i = \mathrm{hydro}_{\mathrm{local}}(i) + \mathrm{sheet}_{\mathrm{local}}(i) - 0.5\,\left|\mathrm{charge}_{\mathrm{local}}(i)\right|. \tag{4}$$

Five aggregation-landscape descriptors are then derived: maximum aggregation score (Agg_Max), mean of the five highest-scoring residues (Agg_Top5_Mean), overall mean (Agg_Mean), the number of residues exceeding a threshold of 0.5 (Agg_Hotspot_Count), and the spatial variance of hotspot positions (Agg_Hotspot_Variance). Unlike prior solubility predictors that summarize aggregation propensity using global averages [34], these features explicitly encode the number and spatial distribution of independent aggregation-prone regions, providing a complementary view of the protein aggregation landscape.

Novel feature contribution

BioGraphX-Sol introduces novel features spanning two levels of methodological novelty. Five features are novel to biological prediction in general and comprise the aggregation hotspot geometry block: Agg_Hotspot_Count, Agg_Hotspot_Variance, Agg_Max, Agg_Top5_Mean, and Agg_Mean. Collectively, these features encode the spatial landscape of aggregation propensity, including the number of independent aggregation-prone regions, their intensity, and their spatial dispersion across the sequence. Existing predictors typically collapse aggregation propensity into a single global score or average, whereas this representation explicitly captures the geometry of aggregation hotspots and their distribution within the protein sequence.

Additional features novel to protein solubility prediction arise from the systematic surface- and terminal-conditioning strategy introduced in BioGraphX-Sol. For each of the ten features in the solubility key feature set, the methodology can generate corresponding _surf, _Nterm, and _Cterm variants.

Unlike structure-aware solubility predictors which derive surface information from experimentally resolved structures or AlphaFold-generated models, the BioGraphX-Sol conditioning framework estimates surface exposure directly from sequence-derived graph topology. Consequently, all features can be computed without requiring three-dimensional coordinates at any stage of the prediction pipeline, preserving full sequence-only applicability while introducing a level of surface-aware representation previously unavailable to sequence-based solubility predictors.

Model Architectures

We evaluated six configurations to isolate the contributions of physicochemical and evolutionary information (Table 2).

XGBoost physics stream. A gradient-boosted tree ensemble (XGBoost) was trained on the 200-dimensional BioGraphX-Sol feature vectors. Tree-based models are invariant to feature scaling, so no normalization was applied. XGBoost serves as the primary physics-only predictor and as the physics component of the XGB+ESM post-hoc ensemble; it requires no 3D structures, sequence alignments, or language model embeddings.

Graph-only neural network. A multilayer perceptron operating on the 200-dimensional feature vector was used, with two hidden layers ($512 \rightarrow 256$ units), batch normalization, GELU activation, dropout (rate 0.3) at each layer, and a sigmoid output. This configuration isolates the contribution of network architecture from the choice of physics representation.

Table 2: Model configurations and computational requirements.

Configuration	Model size / complexity	Frozen backbone	Structure required	GPU required
XGBoost (physics stream)	∼550 trees, depth 5	—	No	No (inference)
Graph-only NN	0.236 M parameters	—	No	Training only
ESM-only NN	1.444 M parameters	ESM-2 3B	No	Embedding precompute
Gated hybrid	2.733 M parameters	ESM-2 3B	No	Embedding precompute
NN ensemble	1.680 M parameters	ESM-2 3B	No	Embedding precompute
XGB+ESM ensemble[†]	**XGBoost + 1.444 M parameters**	**ESM-2 3B**	**No**	**Embedding precompute**
Reference methods				
FGNNSol [15]	Not reported[‡]	ESM-C	Yes (AlphaFold3)	Yes
ProtSATT [21]	6.0 M parameters	UniRep + ESM-2 + ProtT5 (1.87 B)	No	Yes

[†] Primary BioGraphX-Sol model used throughout the manuscript.
[‡] FGNNSol input dimensionality is 1,831 dimensions (structural + sequence features derived from AlphaFold3 and ESM-C); trainable parameter count was not reported in the original publication.

ESM-only neural network. The same MLP architecture as the graph-only network was applied to this 2,560-dimensional input. The ESM-only NN has approximately 1.44M trainable parameters.

Gated hybrid neural network. A dual-branch network integrates both streams through an interpretable gating mechanism, following the architecture of BioGraphX-Net [25]. The physics branch uses three MLP blocks (200 $\rightarrow$ 512 $\rightarrow$ 512 $\rightarrow$ 512) to give it additional representational capacity relative to the single-block ESM branch (2,560 $\rightarrow$ 512), partially offsetting the 12.8-fold input dimensionality asymmetry between the two streams. A two-layer gate network takes the concatenated 1,024-dimensional branch representations and produces two independent scalar gate values via sigmoid activation:

$$\mathbf{g} = [g_{\text{physics}}, g_{\text{ESM}}] \in (0,1)^2. \tag{5}$$

The gated representation is defined as

$$\mathbf{h}_{\text{fused}} = \left[\mathbf{h}_{\text{physics}} \odot g_{\text{physics}} \; ; \; \mathbf{h}_{\text{ESM}} \odot g_{\text{ESM}}\right], \tag{6}$$

where $\odot$ denotes element-wise multiplication and ";" denotes vector concatenation. The fused representation is then projected to the output through a final MLP block and sigmoid head.

Joint optimization tends to systematically underweight the physics branch, a consequence of the input dimensionality asymmetry. To counteract this effect, a physics-encouragement loss was active for the first 25 training epochs:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \text{MSE}\left(\overline{g}_{\text{physics}}, 0.5\right), \tag{7}$$

where $\overline{g}_{\text{physics}}$ denotes the mean physics-branch gate activation and $\lambda = 0.2$. The gated hybrid has 2.73M trainable parameters.

Post-hoc ensembles. Post-hoc ensembles. We evaluated two post-hoc linear ensembles, each combining independently trained streams without joint optimization. The NN ensemble blends graph-only NN and ESM-only NN predictions; the XGB+ESM ensemble blends XGBoost and ESM-only NN predictions and is our primary reported configuration. For both ensembles, the blending weight α was selected independently for each of the five random seeds by

grid search over $\alpha \in \{0.00, 0.05, \ldots, 1.00\}$ (Eq. 8), maximizing validation R^2 on the 268-sequence validation subset. The per-seed validation-optimal α_s was then applied to form a seed-specific ensemble prediction:

$$\hat{y}^{(s)} = \alpha_s \cdot \hat{y}^{(s)}_{\text{physics}} + (1 - \alpha_s) \cdot \hat{y}^{(s)}_{\text{ESM}}, \tag{8}$$

where $\hat{y}^{(s)}_{\text{physics}}$ and $\hat{y}^{(s)}_{\text{ESM}}$ denote the physics and ESM predictions from seed s, and α_s is the seed-specific blending weight. Reported ensemble metrics are the mean and standard deviation across the five per-seed ensemble predictions, matching the multi-seed reporting protocol used for individual configurations. The 268-sequence validation subset comprises the first 268 rows of the 660-sequence test set; ensemble performance is subsequently reported on the full 660-sequence test set inclusive of these rows. The strictly held-out 392-sequence subset (rows 268–659), never used for any modelling decision, provides the leakage-free bootstrap confidence interval reported in Supplementary Table S6. The per-seed α_s values were also applied, without re-tuning, to the *S. cerevisiae* external benchmark. The XGB+ESM ensemble's neural component trains approximately 1.44M parameters.

Training Protocol

XGBoost. The number of estimators was determined by five-fold cross-validation within the 2,019-sequence training set. In each fold, an XGBoost model was trained with early stopping (patience of 50 rounds) on the held-out validation fold. The mean best iteration across the five folds was then used as the fixed estimator count for the final model, which was retrained on all 2,019 training sequences. The fixed hyperparameters were a learning rate of 0.01, maximum tree depth of 5, subsample ratio of 0.8, and column subsampling ratio per tree of 0.7. This procedure mirrors the ProtSATT training protocol.

Neural networks. All neural models were trained using the AdamW optimizer with a weight decay of 10^{-4}. A OneCycleLR scheduler provided linear warmup over the first 10% of training steps, followed by cosine annealing. The maximum learning rate was 5×10^{-5} for all parameters and 1×10^{-4} for the physics branch, with annealing to a final learning rate of 5×10^{-8} for the base branch and 1×10^{-7} for the physics branch. Gradient norms were clipped at 1.0, and the batch size was fixed at 16.

As with XGBoost, the stopping epoch was determined through five-fold cross-validation. For each fold, the best epoch was selected according to validation R^2. The 75th percentile of the fold-level best epochs was then used as the fixed training duration for the final model, which was retrained on all 2,019 training sequences. Training was capped at 300 epochs, with a patience of 40 epochs for cross-validation early stopping. Mean squared error (MSE) was used as the training objective for all neural configurations, with the additional physics-encouragement loss applied to the gated hybrid during the first 25 epochs.

Evaluation Protocols

Metrics. We assessed regression performance using the coefficient of determination (R^2), Pearson correlation, and root mean squared error (RMSE). For binary classification, predictions were thresholded at 0.5, and we report accuracy, precision, recall, F1 score, area under the ROC curve (AUC), and Matthews correlation coefficient (MCC). Regression metrics (R^2, Pearson, RMSE) and AUC are computed from continuous predictions clipped to $[0, 1]$; classification metrics (accuracy, precision, recall, F1, MCC) are computed from binary predictions obtained by thresholding at 0.5.
Multi-seed evaluation. All models were trained and evaluated across five independent random seeds (2024–2028). These seeds governed parameter initialization, fold assignment during cross-validation, subsampling within XGBoost, and, for the post-hoc ensembles, the per-seed selection of the blending weight α. Reported metrics are presented as mean $\pm$ sample standard deviation across the five seeds. For the ensembles, metrics are computed separately for each seed-specific ensemble prediction and then averaged, so that both model training variability and alpha-selection variability are reflected in the reported standard deviations.
Alpha selection. For the XGB+ESM post-hoc ensemble, the blending weight α was chosen by grid search over 21 values from 0.00 to 1.00 in steps of 0.05, maximizing validation R^2 on the first 268 sequences of the 660-sequence test set. We report performance on the full 660-sequence test set, inclusive of these rows, to remain directly comparable with the evaluation protocol used by ProtSATT. To confirm α selection does not inflate reported performance, the strictly held-out 392-sequence subset (rows 268–659), never used for α selection at any stage, provides a leakage-free evaluation, reported in Supplementary Table S6. The per-seed α_s values were applied to the *S. cerevisiae* external benchmark without re-tuning.

Interpretability Analysis

Permutation importance. For each of the 200 BioGraphX-Sol features, we randomly permuted the feature's values across all test proteins and recorded the resulting decrease in R^2, denoted as ΔR^2. This procedure was repeated ten times per model seed across the five XGBoost seeds, yielding 50 permutation replicates per feature. We report the mean and standard deviation of ΔR^2 across all replicates and seeds as the feature importance measure. Permutation importance was computed independently on the eSOL 660-sequence test set and the *S. cerevisiae* 108-sequence benchmark under identical methodology, enabling direct cross-species comparison. We prefer permutation importance on held-out sequences over gain-based training importance because it measures each feature's contribution to generalization directly, rather than its training-time split frequency.

Spearman directional validation. For each feature identified as important on either species, we computed its Spearman rank correlation with solubility on three independent datasets: the eSOL 2,019-sequence training set, the eSOL 660-sequence test set, and the *S. cerevisiae* 108-sequence benchmark. A feature was classified as a universal solubility determinant if it ranked among the top 20 features by permutation importance on both species and exhibited the same Spearman correlation sign across all three datasets. This criterion separates consistent biophysical signal from potential artifacts associated with the training species.

Gate analysis. Per-protein gate values $[g_{\text{physics}}, g_{\text{ESM}}]$ from the gated hybrid model were recorded for all proteins in both the eSOL test set and the *S. cerevisiae* benchmark. We computed the mean gate values and their standard deviations for each dataset to quantify dataset-level reliance on each information source. We further examined the shift in gate values between species as an indicator of adaptive behavior under distribution shift.

Alpha sweep. The validation-optimal α and the complete R^2-versus-α curve were recorded for each of the five training seeds. We report the mean optimal α and its standard deviation across seeds as an estimate of information decomposition, representing the fraction of predictive signal attributable to physicochemical versus evolutionary information. We also assessed the breadth of the performance plateau around the optimum to determine how robust the decomposition was to perturbations in α.

Graph visualization. Because the constraint graph is constructed directly and deterministically from the sequence, it can be rendered and inspected for any individual protein, showing which residues participate in each interaction type. This provides qualitative, structure-level interpretability complementary to the quantitative analyses described above. For a specific protein of interest, a researcher can visually inspect the graph to assess why a given region was identified as buried or surface-exposed by the graph-topology surface proxy.

Results

Feature Group Ablation

Feature group ablation across five random seeds (Supplementary Tables S4 and S5) reveals each group's contribution to the full 200-feature model. In forward ablation starting from constraint graph features alone ($R^2 = 0.310$), global physicochemical scalars contribute the largest single increment ($\Delta R^2 = +0.133$), followed by IDP and polymer physics descriptors ($+0.018$) and the novel aggregation-specific block ($+0.005$). Notably, constraint graph features alone, using only 35 features, already exceed DeepSol's reported performance ($R^2 = 0.239$), showing that a compact, sequence-derived graph representation can outperform a dedicated deep learning baseline with substantially fewer features.

Leave-one-out ablation, which evaluates each group's marginal contribution within the complete model, identifies a consistent ranking: IDP and polymer physics features matter most ($\Delta R^2 = -0.010$, negative in 5/5 seeds), followed by sequence physicochemical statistics (-0.007, 5/5), the novel aggregation-specific block (-0.005, 5/5), and constraint-graph-derived features (-0.003, 5/5). Global physicochemical scalars appear largely redundant given the richer feature groups ($\Delta R^2 = +0.004$, positive in 4/5 seeds), suggesting their information is already captured by the windowed and graph-derived descriptors. Surface and terminal conditioning shows only a borderline within-species contribution (-0.001, 3/5 seeds), its primary value emerges instead through the cross-species transfer analysis.

Performance on the eSOL Benchmark

Table 3 summarizes the performance of all BioGraphX-Sol configurations alongside published baselines on the eSOL 660-sequence test set (Figure 2). For the primary XGB+ESM ensemble, each of the five seeds used its own validation-

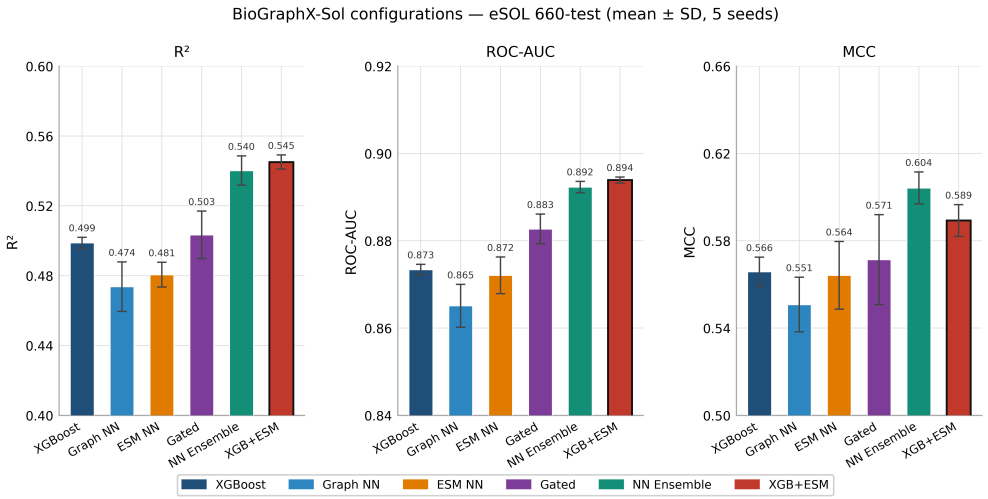


Figure 2: **Comparison of six BioGraphX-Sol configurations on the eSOL test set** ($n = 660$). Bars and error bars denote mean $\pm$ SD across five seeds for **(a)** R^2, **(b)** ROC-AUC, and **(c)** MCC. The primary XGB+ESM ensemble is highlighted in black.

optimal α_s (mean 0.59, SD = 0.06), metrics were computed separately for each per-seed ensemble and are reported as mean $\pm$ SD across the five seeds. Baseline values are sourced from ProtSATT [21] (Table 5), where all methods were evaluated on the identical 2,019/660 partition.

The XGB+ESM ensemble achieves a seed-averaged $R^2 = 0.5451$, AUC $= 0.8939$, and MCC $= 0.5893$ (Figure 3; scatter plot and confusion matrix for the eSOL test set), with bootstrap 95% confidence intervals of R^2 [0.5026, 0.5919], AUC [0.8728, 0.9181], and PR-AUC [0.8489, 0.9069]. This essentially ties ProtSATT on R^2 (0.5451 vs. 0.5450, $\Delta = +0.0001$) and is slightly behind ProtSATT [21] on AUC (0.8939 vs. 0.8962, $\Delta = -0.0023$). On

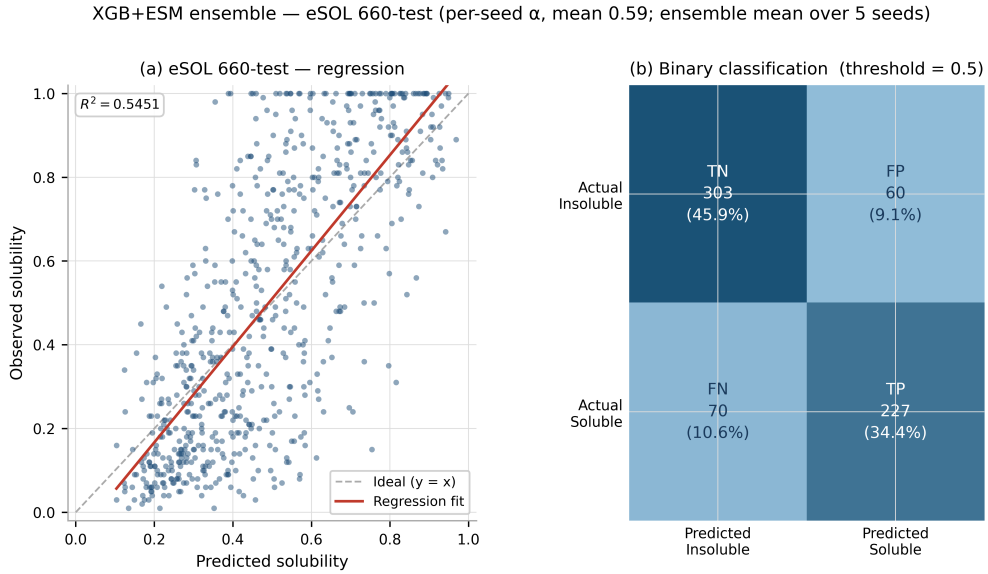


Figure 3: **Performance of the BioGraphX-Sol XGB+ESM ensemble on the eSOL test set** ($n = 660$). (a) Predicted versus observed solubility values showing regression performance ($R^2 = 0.5451$). The red line denotes the fitted regression line and the grey dashed line indicates the ideal $y = x$ relationship. (b) Confusion matrix obtained after thresholding predictions at 0.5.

Table 3: Performance comparison of protein solubility prediction methods on the eSOL dataset.

Method	R^2	ACC	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC
DeepSol [12]	0.2394	0.6924	0.7678	0.4402	0.5596	0.3809	0.7835	0.7520
SoluProt [11]	0.4601	0.7757	0.7580	0.7269	0.7421	0.5442	0.8636	0.8305
MPEPE [35]	0.0121	0.5545	0.4984	0.5529	0.5242	0.1080	0.5660	0.5153
DeepSoluE [36]	0.3913	0.7651	0.7591	0.7003	0.7285	0.5235	0.8414	0.8181
GraphSol [13]	0.3718	0.7621	**0.8333**	0.5892	0.6903	0.5263	0.8594	0.8459
HybridGCN [37]	0.3846	0.7500	0.7820	0.6161	0.6892	0.4946	0.8469	0.8363
GATSol [14]	0.4630	0.7818	0.7909	0.7003	0.7429	0.5577	0.8663	0.8502
PLM_Sol [18]	0.4836	0.7863	0.7484	0.7912	0.7692	0.5714	0.8640	0.8417
MTPSol [19]	0.4755	0.7864	0.7794	0.7235	0.7504	0.5653	0.8567	0.8401
RP3Net [20]	0.5213	0.7985	0.7993	0.7374	0.7671	0.5915	0.8921	0.8699
FGNNSol [15]	0.4867	0.7818	0.7492	0.7744	0.7616	0.5608	0.8669	0.8517
ProtSATT [21]	0.5450	**0.8121**	0.7746	**0.8215**	**0.7974**	**0.6235**	**0.8962**	0.8758
BioGraphX-Sol (XGBoost)	0.4988±0.0032	0.7855±0.0033	0.7681±0.0035	0.7495±0.0069	0.7587±0.0043	0.5658±0.0067	0.8734±0.0012	0.8566±0.0009
BioGraphX-Sol (Graph NN)	0.4737±0.0142	0.7776±0.0071	0.7667±0.0250	0.7306±0.0368	0.7470±0.0094	0.5508±0.0125	0.8651±0.0049	0.8522±0.0047
BioGraphX-Sol (ESM NN)	0.4806±0.0071	0.7839±0.0079	0.7777±0.0291	0.7327±0.0482	0.7526±0.0151	0.5642±0.0155	0.8721±0.0042	0.8485±0.0039
BioGraphX-Sol (Gated)	0.5034±0.0136	0.7876±0.0106	0.8032±0.0230	0.7024±0.0522	0.7476±0.0222	0.5714±0.0206	0.8827±0.0034	0.8652±0.0047
BioGraphX-Sol (NN Ens.)	0.5402±0.0084	0.8039±0.0038	0.7985±0.0231	0.7576±0.0396	0.7762±0.0116	0.6042±0.0073	0.8923±0.0013	0.8767±0.0025
BioGraphX-Sol (XGB+ESM)†	**0.5451±0.0041**	0.7970±0.0037	0.7854±0.0149	0.7562±0.0223	0.7701±0.0060	0.5893±0.0073	0.8939±0.0007	**0.8774±0.0014**

Note: Bold values indicate the best performance for each metric. Published baseline results are reproduced from the original studies. † Primary BioGraphX-Sol model used throughout the manuscript.

the strictly held-out 392-sequence subset, never used for α selection, the ensemble achieves $R^2 = 0.5483$ ([0.4951, 0.6062]) and AUC $= 0.8894$ ([0.8603, 0.9208]), confirming no leakage-driven inflation (Supplementary Table S6).

XGBoost alone achieves $R^2 = 0.499$ and AUC $= 0.873$, exceeding FGNNSol ($R^2 = 0.487$, AUC $= 0.867$) despite FGNNSol incorporating AlphaFold3-predicted structural features across ∼1,831 dimensions. Notably, XGBoost also exceeds the ESM-only NN ($R^2 = 0.481$, AUC $= 0.872$), a configuration with access to evolutionary patterns learned from billions of protein sequences through a 3-billion-parameter language model, demonstrating that biophysically grounded physicochemical encoding provides richer predictive signal for solubility than evolutionary embeddings alone. This finding suggests that structural information relevant to solubility prediction is recoverable from sequence alone through a biophysically grounded constraint graph, without requiring explicit 3D coordinates.

Zero-Shot Transfer to S. cerevisiae

To test whether physicochemical features capture species-independent solubility determinants, we evaluated all eSOL-trained models on the *S. cerevisiae* 108-sequence external benchmark, with no fine-tuning or re-selection of hyperparameters. The per-seed validation-optimal α_s values (mean $= 0.59$, SD $= 0.06$) were applied to the yeast data without re-tuning.

Table 4 summarizes performance across all configurations and baselines (Figure 4). Three findings emerge from this cross-species evaluation. First, the XGB+ESM ensemble achieves $R^2 = 0.424$ and ROC_AUC $= 0.846$, ranking second on R^2 behind ProtSATT (0.465). Notably, the standalone XGBoost configuration achieves the highest ROC-AUC among all evaluated methods (0.8524), exceeding ProtSATT (0.8325) by a margin of $+0.020$. Because ROC_AUC is threshold-independent, it provides a more reliable measure of ranking discrimination than threshold-dependent metrics in this small-sample setting ($n = 108$). Second, and most strikingly, BioGraphX-Sol (XGBoost) achieves PR_AUC $= 0.925$, the highest precision-recall AUC of any evaluated method on S. cerevisiae, exceeding ProtSATT (PR_AUC $= 0.902$) by $+0.023$. This indicates that physics-only constraint graph encoding achieves superior precision-recall discrimination across all decision thresholds on yeast proteins, even though ProtSATT still achieves higher R^2 and MCC at the default threshold. Third, the physics-only XGBoost configuration ($R^2 = 0.378$, ROC_AUC $= 0.852$) exceeds FGNNSol ($R^2 = 0.339$, ROC_AUC $= 0.828$) despite FGNNSol's AlphaFold3-derived structural features. XGBoost ($R^2 = 0.378 \pm 0.008$) substantially exceeds the ESM-only NN ($R^2 = 0.263 \pm 0.048$) on *S. cerevisiae* by $+0.115$ R^2 and $+0.111$ ROC-AUC (0.852 vs. 0.741). This gap is markedly larger than on eSOL ($+0.018$ R^2, $+0.001$ ROC-AUC), consistent with the hypothesis that ESM-2 embeddings encode *E. coli*-specific evolutionary patterns that transfer poorly to *S. cerevisiae*, whereas physicochemical features encoding species-independent biophysical properties maintain their advantage and widen it under cross-species distribution shift. The gated hybrid model provides convergent evidence: mean gate values on *S. cerevisiae* shift slightly toward physics ($g_{\text{physics}} = 0.531 \pm 0.012$) relative to eSOL ($g_{\text{physics}} = 0.510 \pm 0.009$), a directionally consistent $+4.3\%$ increase, though modest, with the physics/ESM ratio remaining roughly constant ($0.694 \rightarrow 0.699$). That sequence-derived physicochemical features outperform predicted 3D structures on a zero-shot cross-species evaluation lends

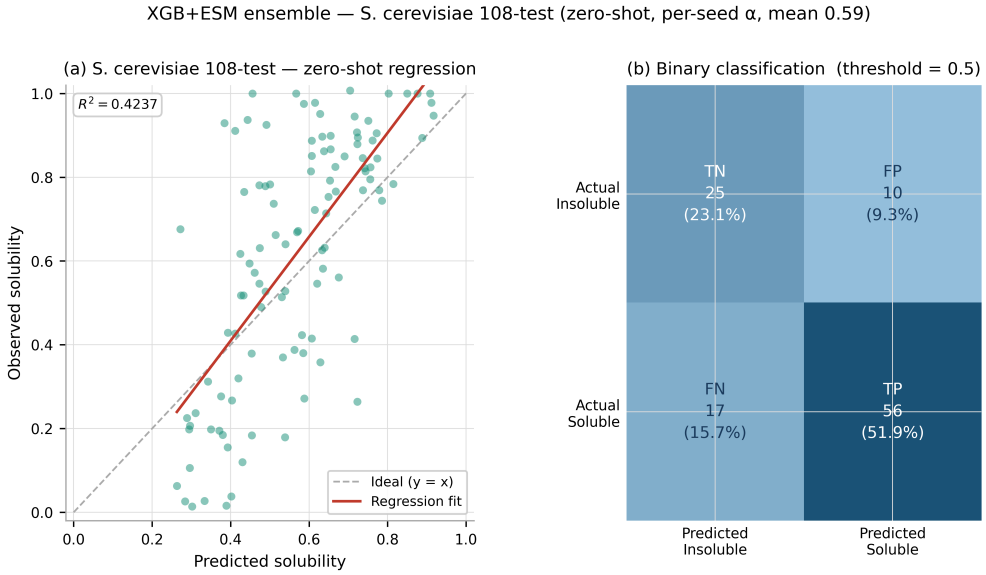


Figure 4: **Zero-shot evaluation of the BioGraphX-Sol XGB+ESM ensemble on the *S. cerevisiae* test set** ($n = 108$). Predictions are averaged over five per-seed ensembles, each formed with its own eSOL-validation-selected α (mean $= 0.59$, SD $= 0.06$). **(a)** Predicted versus observed solubility ($R^2 = 0.424$). **(b)** Confusion matrix at a threshold of 0.5.

further support to the structural proxy paradigm: the relevant structural information is recoverable from sequence, and the simplicity and determinism of constraint graph encoding avoids the error propagation that AlphaFold predictions carry on a novel proteome.

ProtSATT's advantage over BioGraphX-Sol on *S. cerevisiae* is driven primarily by recall (0.904 vs. 0.770) and MCC (0.610 vs. 0.483), reflecting stronger identification of soluble proteins at the default threshold. This likely stems from the three-PLM fusion architecture's access to broader evolutionary representations, UniRep, ESM-2, and ProtT5 collectively span diverse pretraining objectives and sequence corpora, providing redundant evolutionary signal that partially compensates for species shift in a way a single ESM-2 backbone cannot.

Table 4: Performance comparison of protein solubility prediction methods on the independent *S. cerevisiae* external test set.

Method	R^2	**ACC**	**Precision**	**Recall**	**F1**	**MCC**	**ROC-AUC**	**PR-AUC**
DeepSol [12]	0.0829	0.5648	**0.9333**	0.3835	0.5437	0.3411	0.7503	0.8708
SoluProt [11]	0.3840	0.7592	0.8852	0.7397	0.8059	0.5095	0.8504	0.9127
MPEPE [35]	0.1070	0.5462	0.6714	0.6438	0.6573	0.0130	0.5033	0.6747
DeepSoluE [36]	0.2852	0.6667	0.7936	0.6849	0.7352	0.2976	0.7733	0.8656
GraphSol [13]	0.2618	0.7037	0.8867	0.6438	0.7460	0.4422	0.8273	0.8945
HybridGCN [37]	0.2619	0.6944	0.8448	0.6712	0.7480	0.3886	0.7741	0.8661
GATSol [14]	0.3054	0.6759	0.8958	0.5890	0.7107	0.4202	0.8501	0.9227
PLM_Sol [18]	0.3569	0.7685	0.8429	0.8082	0.8252	0.4841	0.8184	0.9000
MTPSol [19]	0.2975	0.6759	0.8800	0.6027	0.7154	0.4048	0.8239	0.9051
RP3Net [20]	0.2874	0.6759	0.8519	0.6301	0.7244	0.3759	0.8043	0.8856
FGNNSol [15]	0.3387	0.6944	0.8846	0.6301	0.7360	0.4297	0.8278	0.9080
ProtSATT [21]	**0.4652**	**0.8333**	0.8571	**0.9041**	**0.8800**	**0.6102**	0.8325	0.9016
BioGraphX-Sol (XGBoost)	0.3775±0.0079	0.7593±0.0101	0.8879±0.0106	0.7370±0.0134	0.8053±0.0090	0.5120±0.0211	**0.8524±0.0024**	**0.9246±0.0022**
BioGraphX-Sol (Graph NN)	0.3243±0.0287	0.7222±0.0211	0.8412±0.0247	0.7288±0.0564	0.7790±0.0245	0.4190±0.0326	0.7995±0.0092	0.8963±0.0016
BioGraphX-Sol (ESM NN)	0.2631±0.0480	0.7315±0.0321	0.8008±0.0045	0.8027±0.0674	0.8002±0.0330	0.3917±0.0442	0.7414±0.0167	0.8425±0.0132
BioGraphX-Sol (Gated)	0.3463±0.0208	0.7537±0.0329	0.8226±0.0119	0.8110±0.0633	0.8153±0.0326	0.4468±0.0498	0.7833±0.0059	0.8673±0.0057
BioGraphX-Sol (NN Ens.)	0.3965±0.0305	0.7426±0.0159	0.8196±0.0108	0.7945±0.0347	0.8063±0.0165	0.4236±0.0245	0.8059±0.0119	0.8947±0.0057
BioGraphX-Sol (XGB+ESM)†	0.4237±0.0239	0.7574±0.0277	0.8568±0.0072	0.7699±0.0485	0.8102±0.0268	0.4828±0.0423	0.8456±0.0094	0.9163±0.0059

Note: Bold values indicate the best performance for each metric. Published baseline results are reproduced from the original studies. † Primary BioGraphX-Sol model used throughout the manuscript.

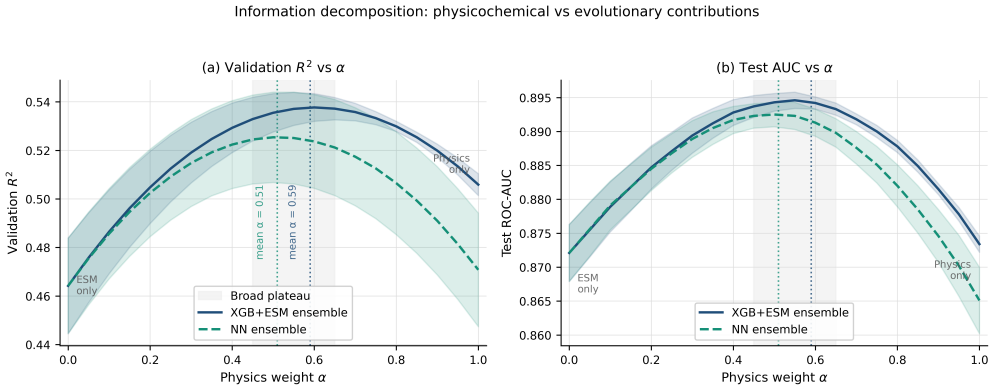


Figure 5: **Information decomposition across the blending weight** α ($\alpha = 1$: physics only; $\alpha = 0$: ESM only). **(a)** Validation R^2 and **(b)** test ROC-AUC for the XGB+ESM (solid) and NN (dashed) ensembles. Lines show means across five seeds; shaded bands indicate $\pm$SD. Dotted lines mark the mean optimal α . The broad plateau indicates robust, approximately balanced contributions with a modest physicochemical advantage.

Information Decomposition

The post-hoc α sweep directly decomposes physicochemical versus evolutionary contributions to solubility prediction. Figure 5 plots R^2 as a function of α, where $\alpha = 1.0$ corresponds to physics-only prediction and $\alpha = 0.0$ to ESM-only prediction. Both the validation and test curves are unimodal with a broad plateau, with mean per-seed $\alpha = 0.59$ for the XGB+ESM ensemble. The NN ensemble, which replaces XGBoost with a graph-only neural network as the physics stream, peaks at $\alpha = 0.50$. Both configurations converge on the same qualitative finding: cell-free protein solubility requires roughly equal contributions from physicochemical and evolutionary information, with the physics fraction ranging from 50% to 55% depending on the physics representation used.

Comparing this alpha sweep with results from the BioGraphX framework applied to intrinsic disorder prediction a companion study (manuscript under review) (Table 5) reveals a cross-task pattern. Physically defined tasks, where the target is a direct biophysical measurement (cell-free solubility) or structurally determined (Disorder-PDB, which restricts negatives to PDB-observed residues), require balanced contributions from both information sources, with physics fractions of 49–55%. By contrast, functionally annotated disorder (Disorder-NOX, where negatives include all residues lacking experimental disorder annotation) is evolution-dominated: 87% ESM versus only 13% physics. This is consistent with the hypothesis that physically defined biological properties carry substantial physicochemical signal recoverable from sequence alone, while functionally annotated properties are shaped primarily by evolutionary selection, a signal that language models capture through co-occurrence statistics across billions of homologous sequences. The alpha sweep plateau is remarkably broad: R^2 differs by only 0.001 between $\alpha = 0.45$ and $\alpha = 0.60$

Table 5: Relative contributions of physics-derived BioGraphX features and evolutionary language-model embeddings in the final ensemble models.

Task	Target Definition	Biophysics (%)	Evolution (%)	Source
eSOL Solubility (XGB+ESM)	Cell-free PURE system	59	41	This work
eSOL Solubility (NN Ensemble)	Cell-free PURE system	50	50	This work
Disorder-PDB	PDB-restricted negatives	49	51	BioGraphX-IDP
Disorder-NOX	Functional annotation negatives	13	87	BioGraphX-IDP

(Supplementary Table S2, S3), indicating that the information decomposition is robust to perturbations in the blending weight rather than an artifact of a sharp peak.

The gated hybrid model offers a complementary, per-protein estimate of modality reliance. On the eSOL test set, the mean gate values are $g_{\text{physics}} = 0.510 \pm 0.009$ and $g_{\text{ESM}} = 0.735 \pm 0.019$. Because these sigmoid-activated gates are not normalized to sum to one, their ratio (0.694) is a relative, not absolute, measure of reliance. The physics gate value (0.510) sits modestly below the alpha-sweep-derived estimate (0.55), consistent with the gate slightly underestimating

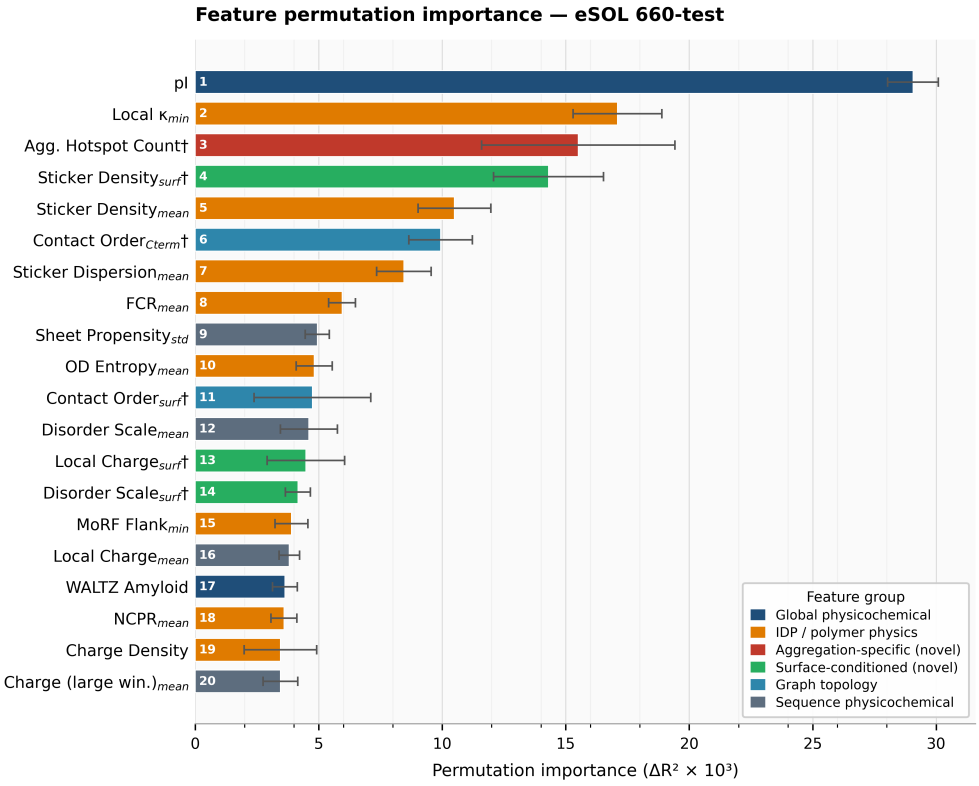


Figure 6: **Top-20 permutation importance features on the eSOL test set** ($n = 660$). Bar lengths indicate mean ΔR^2 after feature permutation, averaged over 50 replicates and five seeds; error bars show $\pm$ SD. Colors denote feature categories.

the physics contribution, likely a consequence of the 12.8-fold input dimensionality asymmetry between the ESM branch (2,560 dimensions) and the physics branch (200 dimensions), which gives the joint optimizer more gradient signal from the ESM stream.

Permutation Feature Importance on eSOL

The importance profile is strongly hierarchical (Figure 6). pI dominates by a wide margin ($\Delta R^2 = 0.0291 \pm 0.0010$), roughly 70% larger than the second-ranked feature. A second tier, Local_Kappa_min, Agg_Hotspot_Count, Sticker_Density_surf, Sticker_Density_mean, Contact_Order_Cterm, Sticker_Dispersion_mean, and FCR_mean, is followed by a long tail of modest contributors (ranks 9–20, $\Delta R^2 = 0.003$–0.005).

The top eight features collectively encode four distinct mechanisms of solubility: charge state (pI), charge patterning, aggregation nucleation geometry (Agg_Hotspot_Count), and sticker–spacer phase-separation thermodynamics.

pI as the dominant determinant. Proteins with pI near physiological pH (7.4) carry minimal net charge, reducing the electrostatic repulsion that maintains colloidal suspension, extreme pI (basic > 9 or acidic < 5) instead drives electrostatic aggregation or ionic-strength-dependent insolubility. A Spearman correlation of $\rho = -0.319$ confirms the association (Table 6): higher-pI proteins are, on average, less soluble in the E. coli PURE cell-free system [38–40].

Charge patterning over magnitude. Local_Kappa_min captures the minimum local charge segregation across 15-residue windows. A high $\kappa_{\min}$ indicates that even the least charged 15-residue window retains substantial charge density, i.e., charge is uniformly distributed and no low-charge, aggregation-prone patches exist, which is favorable for solubility. A low $\kappa_{\min}$ indicates windows depleted of charge, which can promote aggregation. The fact that $\kappa_{\min}$ outranks global measures such as NCPR_mean (rank 18) and Charge_Density (rank 19) confirms that local charge distribution, rather than net charge magnitude, carries the informative signal [41, 42].

Aggregation hotspot geometry. Agg_Hotspot_Count is the highest-ranked novel feature and the only aggregation-specific descriptor in the top 20. Its relatively high variability (SD/mean $= 0.25$) reflects the discrete, threshold-sensitive nature of hotspot counting. That it outranks the continuous Agg_Mean (rank 86, $\Delta R^2 = 0.00040$) shows that nucleation-site geometry, not mean aggregation intensity, is the relevant quantity [43–45].

Surface conditioning outperforms global variants. Surface-conditioned features consistently outrank their global counterparts: Sticker_Density_surf exceeds Sticker_Density_mean, Contact_Order_surf exceeds Contact_Order _mean (rank 11 vs. 39), and Local_Charge_surf exceeds Local_Charge_mean (rank 13 vs. 16), a 1.2–4.5$\times$ importance gain from restricting computation to surface-exposed residues. This validates the surface proxy directly: aggregation initiates at solvent-exposed interfaces, while buried residues are shielded from intermolecular contact by the folded structure [46–48].

C-terminal structural completion. Contact_Order_Cterm encodes mean non-backbone contact order over C-terminal residues; higher values reflect completed folding topology, while low values signal exposed, aggregation-prone hydrophobic segments ($\rho = +0.363$ with solubility) (Table 6) [49, 50].

Sticker dispersion. Sticker_Dispersion_mean measures spatial regularity of aromatic/arginine residues. Even spacing is associated with higher solubility ($\rho = +0.215$), consistent with the sticker–spacer model: evenly spaced stickers require simultaneous satisfaction of multiple contact geometries to nucleate a condensate, raising the entropic barrier to phase separation, whereas clustered stickers lower it [51, 52].

Ranks 9–20 and the most harmful features. The tail includes Local_Sheet_Propensity_std, OrderDisorder_Pattern _Entropy_mean, and the E. coli, calibrated Disorder_AA_Scale variants, the latter prominent here but absent from the S. cerevisiae top-20, foreshadowing the generalization-failure mechanism discussed there. MoRF_Flank_Score_std and Disorder_from_Ifprop_mean are the most harmful features when permuted, shuffling them improves performance, indicating a learned anti-correlation that does not generalize. Frustration_Nterm behaves similarly.

Six of the top-20 features come from novel groups introduced in this work Agg_Hotspot_Count, Sticker_Density_surf, Contact_Order_surf, Local_Charge_surf, Disorder_AA_Scale_surf, and Contact_Order_Cterm, confirming the novel encoding contributes genuine, substantial predictive signal beyond established descriptors.

Cross-Species Feature Analysis

To identify which physicochemical features are universal solubility determinants rather than species-specific artifacts, we computed permutation importance independently on the eSOL 660-sequence test set and the *S. cerevisiae* 108-sequence benchmark, then validated directional consistency by Spearman correlation across three independent datasets.

Nine features satisfy both criteria. All nine show consistent Spearman correlation sign across all three datasets (9/9), confirming that their association with solubility reflects genuine biophysical signal rather than a training-species artifact. Four of the nine come from the IDP and polymer physics feature group, independently identified as the most critical group by leave-one-out ablation, giving convergent validation from two orthogonal analyses.

Table 6: Nine universal solubility determinants identified across two species. Features appear in the top-20 permutation importance rankings on both the eSOL and *S. cerevisiae* test sets and exhibit consistent Spearman correlation signs across the training, eSOL test, and *S. cerevisiae* test datasets (9/9 directionally consistent).

Feature	eSOL Rank	SC Rank	ρ (Train)	ρ (eSOL)	ρ (SC)	Direction	Group
pI	1	3	-0.284	-0.319	-0.296	$\downarrow$ Solubility	Global
Local_Kappa_min	2	1	$+0.349$	$+0.415$	$+0.347$	$\uparrow$ Solubility	IDP / Polymer
Agg_Hotspot_Count	3	4	-0.381	-0.411	-0.419	$\downarrow$ Solubility	Aggregation (Novel)
Sticker_Density_surf	4	2	-0.321	-0.236	-0.394	$\downarrow$ Solubility	Surface (Novel)
Sticker_Density_mean	5	5	-0.320	-0.236	-0.397	$\downarrow$ Solubility	IDP / Polymer
Contact_Order_Cterm	6	11	$+0.361$	$+0.363$	$+0.421$	$\uparrow$ Solubility	Terminal (Novel)
Sticker_Dispersion_mean	7	15	$+0.283$	$+0.215$	$+0.322$	$\uparrow$ Solubility	IDP / Polymer
FCR_mean	8	6	$+0.339$	$+0.345$	$+0.265$	$\uparrow$ Solubility	IDP / Polymer
Contact_Order_surf	11	8	$+0.372$	$+0.368$	$+0.457$	$\uparrow$ Solubility	Surface (Novel)

Note: eSOL Rank and SC Rank denote permutation-importance rankings on the eSOL and *S. cerevisiae* test sets, respectively. Positive Spearman correlations indicate increasing solubility with increasing feature values, whereas negative correlations indicate the opposite. Novel features correspond to BioGraphX-Sol–specific aggregation, surface, or terminal descriptors.

The universal set is dominated by features encoding the thermodynamics of phase separation and colloidal stability. Four directly quantify sticker–spacer polymer physics: Sticker_Density_surf and Sticker_Density_mean (density of phase-separation-driving aromatic and arginine residues), Sticker_Dispersion_mean (regularity of sticker spacing, where higher dispersion resists condensate formation), and FCR_mean (fraction of charged residues promoting solvation). Two charge-patterning features complement these: pI (isoelectric point, encoding net charge at physiological pH) and Local_Kappa_min (minimum local charge segregation, encoding the absence of complementary electrostatic aggregation interfaces). The aggregation hotspot count (Agg_Hotspot_Count) directly quantifies the number of independent nucleation-prone regions. Two contact order variants, Contact_Order_Cterm and Contact_Order_surf ($\rho \approx +0.40$), capture the structural proxy contribution, where long-range interaction partners at the C-terminus and solvent-exposed surface indicate completed folding topology.

Three of the nine universal features come from novel feature groups introduced in this work: Agg_Hotspot_Count (aggregation-specific), Sticker_Density_surf (surface-conditioned), and Contact_Order_Cterm (terminal-conditioned), confirming that the novel encoding elements contribute genuine, transferable predictive signal beyond established descriptors.

Surface-conditioned features show a striking dissociation from their global counterparts. Contact_Order_surf ranks 11th on eSOL and 8th on S. cerevisiae. Sticker_Density_surf ranks 4th on eSOL and 2nd on S. cerevisiae, while Sticker_Density_mean ranks 5th on both species with nearly identical Spearman correlations across all three datasets . The permutation importance advantage of the surface-conditioned variant over the global mean is therefore not explained by a stronger linear association with solubility, both features carry equivalent marginal signal in isolation, but by reduced redundancy with other features in the full model: surface conditioning restricts computation to residues physically available for intermolecular contact, providing information orthogonal to buried-residue sticker statistics that the model exploits when all 200 features are present. This dissociation between marginal correlation and full-model importance provides empirical support for the surface proxy methodology.

Features that fail to transfer reveal the mechanism behind cross-species generalization failure. Two Disorder_AA _Scale variants, mean (eSOL rank 12) and surf (eSOL rank 14), rank within the eSOL top-20 but are absent from the S. cerevisiae top-20, while Local_mean (rank 22) and Local_std (rank 31) fall just outside the eSOL top-20 and likewise outside the S. cerevisiae top-20. Disorder_AA_Scale_Local_std shows strongly negative permutation importance on yeast (-0.0063), while Disorder_AA_Scale_Nterm is also negative but negligible (-0.00014). Spearman directional analysis, however, shows that these Disorder_AA_Scale variants retain consistent correlation signs across all three datasets (all negative with solubility), confirming that the failure reflects importance divergence rather than directional reversal: the features capture a real but species-specific signal. These scales encode proteome-level amino acid usage biases calibrated to the E. coli proteome; when applied to yeast, whose amino acid frequencies differ, they may contribute noise rather than signal. [53–55]. Similarly, Local_Sheet_Propensity_std ranks 9th on eSOL but 199th on *S. cerevisiae*, indicating that β-sheet propensity variability is informative for E. coli solubility but not for yeast.

Sticker_Density_std shows the most dramatic importance divergence of all: it ranks 43rd on eSOL but 7th ($+0.011$) on S. cerevisiae. Feature distributions are similar across species (Figure S1), ruling out a simple compositional explanation. The Spearman direction is consistently negative , so this is not a directional reversal either. The mechanistic basis of this pattern remains undetermined and warrants further investigation.

Discussion

Sequence-Derived Encoding Outperforms Structure-Based Methods

The central finding of this work is that a 200-dimensional sequence-derived physicochemical encoding, built from biophysically grounded constraint graphs, outperforms methods that incorporate AlphaFold3-predicted structures and matches a three-PLM fusion architecture. The XGBoost physics-only configuration exceeds FGNNSol on the eSOL benchmark, despite FGNNSol using 1,831 features derived from AlphaFold3-predicted coordinates and ESM-C embeddings. The result extends to the cross-species setting: on *S. cerevisiae*, XGBoost again exceeds FGNNSol. This is counterintuitive, since predicted 3D structures are widely assumed to carry richer geometric information than sequence-based encodings. Three factors explain the reversal.

First, AlphaFold3 structure predictions inherit model uncertainty. Proteins with low homology to the structural predictor's training distribution, precisely the sequences most in need of solubility screening, receive the least confident predictions. Graph convolutional networks operating on these uncertain coordinates propagate that error through every

network layer. Constraint graph encoding, by contrast, is fully deterministic: the same sequence always produces the same graph, eliminating this error propagation pathway entirely.

Second, the structural information relevant to solubility is primarily local, surface packing, contact topology, and the spatial distribution of aggregation-prone residues, rather than requiring atomic-resolution global geometry. The constraint graph's distance-decay edge weighting (Equation 1) captures these local properties through biophysically calibrated rules, providing a coarse but sufficient structural proxy.

Third, FGNNSol explicitly acknowledges that its structural feature dimensions cannot be attributed, owing to the nonlinearity of graph aggregation, limiting its interpretability to sixteen global physicochemical descriptors. BioGraphX-Sol's 200 named features are each individually interpretable, and permutation importance on held-out sequences from two species identifies exactly which features drive predictions, a qualitative difference in the kind of evidence available for mechanistic reasoning, not just a difference in accuracy.

This finding is consistent with the earlier observation that the same constraint-graph encoding, developed for subcellular localization, transferred to the eSOL solubility benchmark without modification and achieved recall comparable to the structure-dependent fGNNSol[15, 25]. BioGraphX-Sol extends that preliminary cross-task validation into a systematic, multi-seed, cross-species benchmark with full information decomposition.

The same pattern extends to a different biological domain. In the companion BioGraphX-RNA framework, the graph-only model outperformed both the RiNALMo language model and an RNAfold partition-function graph for miRNA localization (macro-AUROC 0.9396 ± 0.0045 vs. 0.9139 ± 0.0138) [26], despite being constructed from primary sequence alone without a learned language model or a structure-prediction module. Across two biological domains and two distinct tasks, explicitly encoded biophysical constraints have thus proven competitive with, and in specific settings superior to, both sequence-based language models and structure-based predictors. This convergence supports the view that deterministic, biophysically grounded encodings capture information that large learned representations do not fully recover.

The XGB+ESM ensemble matches ProtSATT on the R^2. The contribution lies in matching ProtSATT's regression performance while requiring fundamentally different, more accessible computational resources: no structure prediction, no language model fine-tuning, and approximately 1.44M neural parameters, versus ProtSATT's 6.0M downstream parameters.

Complementary Inductive Biases in Cross-Modal Ensembles

The post-hoc ensemble design reveals the mechanism through which combining physics and evolution improves solubility prediction. The XGB+ESM ensemble exceeds both XGBoost alone and the ESM-only NN alone, a gain that requires the two streams to make partially decorrelated errors. This decorrelation stems from fundamental differences in the two approaches' inductive biases: XGBoost operates on 200 explicitly defined physicochemical features through axis-aligned decision boundaries that capture threshold effects and pairwise feature interactions, well-suited to the sharp phase boundaries of aggregation processes, whereas ESM-2 embeddings encode distributed evolutionary co-occurrence patterns learned from billions of protein sequences, capturing functional constraints invisible to physicochemical rules alone.

Notably, replacing the graph-only neural network with XGBoost as the physics stream improves ensemble performance. This finding is specific to this task and representation scale: for protein-level regression on 200 aggregated physicochemical features, tree-based models offer a more complementary inductive bias to a neural evolutionary encoding than a neural physics model does. The mechanism is straightforward: XGBoost and a neural network trained on the same 200 features make more similar errors than XGBoost and a neural network trained on ESM embeddings, which reduces the diversity bonus the ensemble can extract.

That an independently trained, post-hoc blended ensemble outperforms the jointly trained gated hybrid across every metric has a straightforward interpretation: the physics-to-evolution balance for this task is a property of the dataset, not of the individual protein. A single global weight $\alpha = 0.59$ captures this balance effectively, whereas a learned gate network adds capacity the problem does not require and that joint optimization cannot reliably direct. This parallels the finding in BioGraphX-IDP a companion study (manuscript under review), where a fixed-weight ensemble outperformed learned gating on both CAID-3 disorder benchmarks, suggesting that a global, dataset-level information balance may be a general feature of biological prediction tasks where the two modalities carry complementary, rather than redundant, signal.

Two Independent Analyses Converge on IDP/Polymer Physics as the Dominant Signal

Two orthogonal analyses converge on the same conclusion regarding the physicochemical determinants of protein solubility. At the feature-category level, leave-one-out ablation identifies the IDP/polymer physics group as the most critical contributor to predictive performance, with its removal producing the largest decrease in model accuracy ($\Delta R^2 = -0.010 \pm 0.003$ across five seeds; all 5/5 seeds showed a performance decline). At the individual-feature level, permutation importance analysis on the held-out eSOL test set yields the same conclusion: six of the top 20 most important features belong to the IDP/polymer physics category, including four of the top eight ranked features. The remaining top-20 positions comprise global physicochemical descriptors (pI, rank 1; Mean_WALTZ, rank 17), novel surface- and terminal-conditioned features (Sticker_Density_surf, rank 4; Local_Charge_surf, rank 13; Disorder_AA_Scale_surf, rank 14), graph-topology features (Contact_Order_Cterm, rank 6; Contact_Order_surf, rank 11), and sequence physicochemical descriptors (Local_Sheet_Propensity_std, rank 9; Disorder_AA_Scale_mean, rank 12; Local_Charge_mean, rank 16; Local_Charge_LargeWin_mean, rank 20). This convergence, together with the sticker–spacer and charge-patterning mechanisms identified above, is consistent with the emerging view that phase-separation thermodynamics and charge patterning are universal determinants of protein solubility [56, 57].

Graph-topology features account for two of the top-20 positions, Contact_Order_Cterm (rank 6) and Contact_Order_surf (rank 11), both of which quantify long-range intramolecular contact density and act as sequence-derived proxies for folding completeness.

Surface Conditioning Enables Cross-Species Transfer

The dissociation between surface-conditioned and global feature variants points to the mechanism through which physicochemical features transfer across species. Contact_Order_surf ranks among the top 11 features on both eSOL and S. cerevisiae; Contact_Order_mean ranks 39th on eSOL and 25th on S. cerevisiae. Sticker_Density_surf ranks 4th and 2nd across the two species. This pattern fits the physical basis of protein aggregation: aggregation initiates at solvent-exposed interfaces, where hydrophobic and aromatic residues make intermolecular contact. Buried-core residues, however aggregation-prone in isolation, are shielded by the protein's folded structure. Surface chemistry is governed by universal solvation thermodynamics, the energetics of water displacement, hydrophobic burial, and electrostatic shielding, while buried-core composition instead reflects species-specific folding preferences and amino acid usage biases shaped by evolutionary history.

Existing surface-aware methods (SurfSol, FGNNSol) derive solvent accessibility from experimentally resolved or AlphaFold-predicted structures, which introduces both a computational dependency and a species-specific bias from the structural predictor's training distribution. BioGraphX-Sol's graph-centrality surface proxy requires only the primary sequence and is fully deterministic, making it applicable at proteome scale without any infrastructure constraints. Empirical validation of this proxy comes not from a direct comparison with DSSP-computed solvent accessibility, which remains a direction for future work, but from a functional test: surface-conditioned features are universal determinants of solubility across two organisms, while the same features computed over the whole protein are not. If the proxy did not capture genuine surface-relevant information, this dissociation would not appear.

Gate Values Under Distribution Shift

The gated hybrid provides a per-protein, learned estimate of modality reliance that complements the post-hoc α sweep. On the eSOL test set, the physics gate averages 0.510 ± 0.009; on *S. cerevisiae*, it rises slightly to 0.531 ± 0.012, a $+4.3\%$ relative shift. The ESM gate shows a proportional increase ($0.735 \rightarrow 0.760$), so the physics/ESM ratio remains nearly constant ($0.694 \rightarrow 0.699$). The direction of this shift is consistent with the hypothesis that the model relies somewhat more on species-agnostic physicochemical features when evolutionary embeddings encounter out-of-distribution sequences, but the magnitude is modest.

The α sweep and gate analysis converge on the same qualitative finding, physicochemical features contribute at least half of the predictive signal for solubility and do not diminish in importance under species shift, while differing in the precise magnitude. The α sweep assigns 59% weight to physics on eSOL, whereas the gate assigns approximately 41% to physics when normalized by the combined gate sum ($0.510/1.245$). This discrepancy is consistent with the input-dimensionality asymmetry between the branches: the ESM branch's 2,560-dimensional input provides the optimizer with more gradient signal per parameter than the physics branch's 200-dimensional input, potentially biasing the learned gate toward the ESM stream. Because the post-hoc α sweep is independent of these architectural dynamics, it provides the more reliable estimate of the relative predictive contribution of the two modalities.

Practical Deployment and Computational Accessibility

The XGBoost physics-only configuration requires neither GPU acceleration at inference nor a backbone language model or structure prediction. It runs on a standard CPU, produces predictions in milliseconds per protein, and achieves $R^2 = 0.499$ on eSOL, exceeding FGNNSol ($R^2 = 0.487$), which depends on AlphaFold3 structure prediction at a computational cost of minutes per protein and tens of gigabytes of GPU memory. For research groups without high-performance computing infrastructure, including resource-constrained settings where recombinant protein production research may be particularly relevant, this configuration offers competitive predictive performance with minimal deployment overhead.

The XGB+ESM ensemble requires precomputing ESM-2 embeddings once, requiring approximately 1 hour for the full eSOL benchmark on a single T4 GPU, after which downstream training and inference operate on precomputed features at negligible marginal cost. The neural component of the ensemble contains only approximately 1.44 M parameters, compared with ProtSATT's 6.0 M downstream parameters, while FGNNSol additionally incurs the computational overhead of structure prediction. The gated hybrid configuration, with 2.73 M parameters, remains substantially smaller than ProtSATT's downstream model. Moreover, every BioGraphX-Sol configuration uses a single frozen ESM-2 backbone, whereas ProtSATT relies on three protein language models.

These efficiency characteristics align with the Green AI principle that domain knowledge encoded into compact, interpretable features can provide a viable alternative to parameter scaling for structured biological prediction tasks. The same BioGraphX encoding framework, adapted only through its interaction rules and aggregation strategy, has now achieved competitive performance across four distinct biological tasks: protein localization, RNA localization, intrinsic disorder prediction, and protein solubility, without requiring task-specific architecture design beyond lightweight downstream classifiers.

Limitations

Several limitations should be considered when interpreting these results.

First, we have not validated BioGraphX-Sol experimentally: all evaluations rely on computational benchmarks built from previously published PURE cell-free expression data. Predicted solubility scores have not guided actual construct design, and whether predicted rankings correspond to experimentally observed yields in a live synthesis campaign remains untested, a necessary step before recommending the framework for direct integration into protein engineering workflows.

Second, training relies exclusively on E. coli cell-free expression data, which differs substantially from whole-cell expression involving chaperone machinery, co-translational folding assistance, and organism-specific proteostasis networks. The learned feature weights and the information-decomposition estimate (59% physics) reflect this particular context and may not generalize to mammalian or insect-cell systems, where different aggregation mechanisms and glycosylation patterns apply. The S. cerevisiae benchmark offers partial cross-organism evidence, but it too is cell-free; in vivo eukaryotic expression remains untested.

Third, the surface proxy, which estimates residue exposure from graph betweenness and closeness centrality, has not been benchmarked against DSSP-computed solvent accessibility on a structure-annotated subset. Empirical support instead comes from a functional test: surface-conditioned features are universal determinants across two species, while their global counterparts are not. Whether the proxy identifies the same residues DSSP would, or achieves cross-species transfer through a different mechanism, remains open.

Future Directions

Three extensions follow naturally from this work.

First, extending the framework to per-residue solubility attribution would directly support mutation-guidance applications. The constraint-graph encoding already generates residue-level features; only the final protein-level aggregation step discards positional information. Retaining residue-level outputs would allow the model to identify sequence segments that drive insolubility, translating predictions into actionable engineering recommendations, such as charge-balancing substitutions in low-κ regions, aromatic-to-glycine substitutions within sticker-dense segments, or C-terminal truncations when Contact_Order_Cterm is low. Because this extension requires only a modification of the output head rather than a redesign of the encoding framework, it represents a relatively low-cost enhancement with substantial practical value.

Second, the α-sweep methodology can be extended systematically beyond solubility prediction. As a general instrument for quantifying the physicochemical-to-evolutionary information balance in two-stream biological prediction systems, it could be applied to thermostability, aggregation kinetics, binding affinity, and phase-separation propensity, tasks for which this balance has never been quantitatively measured. The current cross-task comparison contains only four data points. Expanding the analysis to ten or more biological prediction tasks would help determine whether the pattern observed here reflects a general organizing principle of biological information rather than a phenomenon specific to disorder prediction and solubility.

Third, evaluating BioGraphX-Sol across additional organisms and expression systems would provide a stronger test of its central mechanistic hypothesis. Whole-cell *E. coli* datasets, mammalian cell-free systems, and published yeast-expression or yeast two-hybrid solubility screens would reveal whether the physics-dominant signal (approximately 59%) persists when molecular chaperone networks and organism-specific cellular environments are present. The underlying hypothesis, that sticker–spacer thermodynamics and charge patterning are universal determinants of protein solubility, predicts that the same feature classes should remain predictive across diverse expression contexts. However, direct empirical validation is still required. The observed failure of amino-acid propensity scales to transfer cleanly to *S. cerevisiae* further suggests that broader organism coverage would help distinguish universally transferable physicochemical features from those requiring species-specific recalibration.

Conclusion

BioGraphX-Sol shows that protein solubility prediction does not need more information, it needs the right information, correctly encoded. A 200-dimensional physicochemical encoding, derived deterministically from amino acid sequence alone, without 3D coordinates or evolutionary homologs, matches current state-of-the-art performance on the standard eSOL benchmark while requiring orders of magnitude fewer computational resources. Physicochemical features alone exceed AlphaFold3-dependent methods, establishing that the structural information relevant to solubility is recoverable from sequence through a biophysically grounded constraint graph. In zero-shot transfer to S. cerevisiae, the physics-only encoding achieves the best ROC-AUC and PR-AUC of any evaluated method, and nine universal solubility determinants retain consistent directional signal across both organisms, evidence that sticker–spacer polymer physics, charge patterning, and aggregation hotspot geometry capture species-independent biophysical principles, not organism-specific statistical artifacts. The alpha sweep, applied to solubility for the first time, shows that physically defined biological tasks draw on roughly balanced contributions from physicochemical and evolutionary information, a cross-task pattern that may reflect a general property of biophysically measurable targets.

Together, these results position BioGraphX-Sol as both a competitive solubility predictor and an interpretable analytical instrument, one that identifies specific, actionable physicochemical determinants of solubility for protein engineering to target directly. The framework extends naturally to per-residue attribution, mutation guidance, and systematic information decomposition across biological prediction tasks, directions that could connect sequence-based prediction more directly to mechanistic understanding and experimental design.

Acknowledgements

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 analysed in this study were obtained from the following publicly available sources. The eSOL dataset (E. coli proteins, PURE cell-free solubility measurements) was originally reported by Niwa et al., *Proc. Natl. Acad. Sci. USA* 106(11):4201–4206, 2009 (`https://doi.org/10.1073/pnas.0811922106`). The homology-aware train/test split used in this study (2,019 training and 660 test sequences) was adopted from the GATSol benchmark and is distributed with the ProtSATT repository (`https://github.com/quietbamboo/ProtSATT`). The *S. cerevisiae* external benchmark (108 sequences) was originally reported by Hou et al., *Bioinformatics* 36(5):1445–1452, 2020 (SOLart) (`https://doi.org/10.1093/bioinformatics/btz773`) and follows the GATSol/ProtSATT protocol, distributed with the same repository. The BioGraphX-Sol source code, trained models, and precomputed feature matrices are openly available at `https://github.com/Abubakar-Saeed/BioGraphX-Sol`.

Supporting information

Supplementary materials are available.

References

[1] Mauno Vihinen. Solubility of proteins. *ADMET & DMPK*, 8(4):391–399, 2020.

[2] John W. Bye, Lucy Platts, and Robert J. Falconer. Biopharmaceutical liquid formulation: A review of the science of protein stability and solubility in aqueous environments. *Biotechnology Letters*, 36(5):869–875, 2014.

[3] J. Wang, L. Dai, and Z. Zhang. Protein aggregation in neurodegenerative diseases. *Chinese Medical Journal*, 138(21):2753–2768, 2025.

[4] Claudio Soto and Stefan Pritzkow. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nature Neuroscience*, 21:1332–1340, 2018.

[5] K. L. Moreau and J. A. King. Protein misfolding and aggregation in cataract disease and prospects for prevention. *Trends in Molecular Medicine*, 18(5):273–282, 2012.

[6] Benjamin Raynal, Pascal Lenormand, Benoît Baron, Sylvain Hoos, Philippe England, and Jean Pantel. Quality assessment and optimization of purified protein samples: Why and how? *Microbial Cell Factories*, 13:180, 2014.

[7] Max Hebditch, Manuel A. Carballo-Amador, Spyridon Charonis, Robin Curtis, and Jim Warwicker. Protein–sol: A web tool for predicting protein solubility from sequence. *Bioinformatics*, 33(19):3098–3100, 2017.

[8] Christophe N. Magnan, Alan Randall, and Pierre Baldi. Solpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, 25(17):2200–2207, 2009.

[9] Pietro Sormanni, Francesco A. Aprile, and Michele Vendruscolo. The camsol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology*, 427(2):478–490, 2015.

[10] Tatsuya Niwa, Bing Wei Ying, Kei Saito, Wataru Jin, Shoji Takada, Takuya Ueda, and Hideki Taguchi. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4201–4206, 2009.

[11] Jakub Hon, Michal Marusiak, Tomas Martinek, Ales Kunka, Jiri Zendulka, David Bednar, and Jiri Damborsky. Soluprot: prediction of soluble protein expression in Escherichia coli. *Bioinformatics*, 37(1):23–28, 2021.

[12] Sameer Khurana, Ramzi Rawi, Karan Kunji, Gwo-Yu Chuang, Halima Bensmail, and Rupa Mall. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, 2018.

[13] Jiawei Chen, Shuangjia Zheng, Hui Zhao, Yuedong Yang, et al. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *Journal of Cheminformatics*, 13(1):7, 2021.

[14] Bin Li and Dong Ming. Gatsol, an enhanced predictor of protein solubility through the synergy of 3d structure graph and large language modeling. *BMC Bioinformatics*, 25(1):204, 2024.

[15] Guanjie Song, Zhenjie Luo, Aoyun Geng, Junlin Xu, Yajie Meng, Feifei Cui, Leyi Wei, Quan Zou, and Zilong Zhang. Protein solubility prediction using fused graph convolutional networks and improved attention networks with alphafold3-derived features. *Journal of Chemical Information and Modeling*, 66(2):894–909, 01 2026.

[16] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joey Bambrick, Scott W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, others, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[17] S. Mehdiabadi, J. Chen, and B. Rost. Benchmarking alphafold2 and alphafold3 on intrinsically disordered region prediction reveals persistent structural hallucinations. *Bioinformatics*, 42(3):456–465, 2026.

[18] Xiaofeng Zhang, Xin Hu, Tianyi Zhang, Lei Yang, Chenguang Liu, Ning Xu, Hongwei Wang, and Wei Sun. Plm_sol: predicting protein solubility by benchmarking multiple protein language models with the updated Escherichia coli protein solubility dataset. *Briefings in Bioinformatics*, 25(5):bbae404, 2024.

[19] Yifan Gao, Hao Wang, Lei Zhang, Feng Liu, Jian Zhou, and Ying Gu. Mtpsol: Multimodal twin protein solubility prediction architecture based on pretrained models. *Journal of Chemical Information and Modeling*, 65(10):4878–4888, 2025.

[20] Elena Tankhilevich, Sergio Martinez Cuesta, Ian Barrett, Christina Berg, Linda Holmberg Schiavone, and Andrew R. Leach. Rp3net: a deep learning model for predicting recombinant protein production in Escherichia coli. *Bioinformatics*, 42(1):btag003, 2026.

[21] Wenqi Deng, Zhen Chen, Cheng Ji, Jing Gao, Hui Xu, and Jiansheng Huang. Protsatt: An advanced protein solubility predictor based on attention mechanism. *Journal of Chemical Information and Modeling*, 66(11):6334–6349, 2026.

[22] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, November 2020.

[23] Rohit V. Pappu, Simon R. Cohen, Faraz Dar, Mohamed Farag, and Madhav Kar. Phase transitions of associative biomacromolecules. *Chemical Reviews*, 123(14):8945–8987, 2023.

[24] Rahul K. Das and Rohit V. Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, 2013.

[25] Abubakar Saeed and Waseem Abbas. Biographx: bridging the sequence-structure gap via physicochemical graph encoding for interpretable subcellular localization prediction. *Bioinformatics Advances*, 6(1):vbag181, 01 2026.

[26] Abubakar Saeed and Waseem Abbas. Biographx-rna: a universal physicochemical graph encoding for interpretable rna subcellular localization prediction. *BMC Bioinformatics*, 27(1):1–22, 2026.

[27] Qingzhen Hou, Jean Marc Kwasigroch, Marianne Rooman, and Fabrizio Pucci. Solart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics*, 36(5):1445–1452, 03 2020.

[28] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[29] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

[30] Peter Y. Chou and Gerald D. Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2):211–222, 1974.

[31] Peter Y. Chou and Gerald D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974.

[32] Anil Mohan, Christopher J. Oldfield, Predrag Radivojac, Vladimir Vacic, Marc S. Cortese, A. Keith Dunker, and Vladimir N. Uversky. Analysis of molecular recognition features (morfs). *Journal of Molecular Biology*, 362(5):1043–1059, 2006.

[33] Y. Fang, L. Wei, B. Xing, and Z. Wei. Surfsol: A multimodal surface-based deep learning framework for protein solubility prediction. *Journal of Chemical Information and Modeling*, 66(7):3737–3746, 2026.

[34] Oriol Conchillo-Solé, Niels S. de Groot, Francesc X. Avilés, Joan Vendrell, Xavier Daura, and Salvador Ventura. Aggrescan: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, 8:65, 2007.

[35] Z. Ding, F. Guan, G. Xu, Y. Wang, Y. Yan, W. Zhang, N. Wu, B. Yao, H. Huang, T. Tuller, and J. Tian. Mpepe, a predictive approach to improve protein expression in E. coli based on deep learning. *Computational and Structural Biotechnology Journal*, 20:1142–1153, 2022.

[36] C. Wang and Q. Zou. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE. *BMC Biology*, 21(1):12, 2023.

[37] L. Chen, R. Wu, F. Zhou, H. Zhang, and J. K. Liu. HybridGCN for protein solubility prediction with adaptive weighting of multiple features. *Journal of Cheminformatics*, 15(1):118, 2023.

[38] D. L. Wilkinson and R. G. Harrison. Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/Technology*, 9(5):443–448, 1991.

[39] S. R. Trevino, J. M. Scholtz, and C. N. Pace. Amino acid contribution to protein solubility: Asp, glu, and ser contribute more favorably than the other hydrophilic amino acids in rnase sa. *Journal of Molecular Biology*, 366(2):449–460, 2007.

[40] R. M. Kramer, V. R. Shende, N. Motl, C. N. Pace, and J. M. Scholtz. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophysical Journal*, 102(8):1907–1915, 2012.

[41] L. Sawle and K. Ghosh. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of Chemical Physics*, 143(8):085101, 2015.

[42] A. S. Holehouse, R. K. Das, J. N. Ahad, M. O. Richardson, and R. V. Pappu. CIDER: Resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophysical Journal*, 112(1):16–21, 2017.

[43] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, 75:333–366, 2006.

[44] T. P. Knowles, C. A. Waudby, G. L. Devlin, S. I. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, M. E. Welland, and C. M. Dobson. An analytical solution to the kinetics of breakable filament assembly. *Science*, 326(5959):1533–1537, 2009.

[45] A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, 22(10):1302–1306, 2004.

[46] G. G. Tartaglia, A. Cavalli, R. Pellarin, and A. Caflisch. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Science*, 14(10):2723–2734, 2005.

[47] M. Belli, M. Ramazzotti, and F. Chiti. Prediction of amyloid aggregation in vivo. *EMBO Reports*, 12(7):657–663, 2011.

[48] G. G. Tartaglia and M. Vendruscolo. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews*, 37(7):1395–1401, 2008.

[49] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology*, 277(4):985–994, 1998.

[50] M. M. Gromiha and S. Selvaraj. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *Journal of Molecular Biology*, 310(1):27–32, 2001.

[51] J. Wang, J. M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, and A. A. Hyman. A molecular grammar governing the driving forces for phase separation of prion-like rna binding proteins. *Cell*, 174(3):688–699.e16, 2018.

[52] T. Mittag and R. V. Pappu. A conceptual framework for understanding phase separation and addressing open questions and challenges. *Molecular Cell*, 82(12):2201–2214, 2022.

[53] Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.

[54] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Genetics*, 42(1):38–48, 2001.

[55] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7:208, 2006.

[56] T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Molecular Cell*, 57(5):936–947, 2015.

[57] A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes, and R. A. Young. Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, 175(7):1842–1855.e16, 2018.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf