

LAB REPORT

SmartSoluPred: A novel two-stage graph-language modeling framework for intrinsic and mutation-induced protein solubility

Suvid Bordia, Eric Chen, Rhys Llewellyn-Jones, Derek Peng, Arjun Subramanian, and Albert Wu ■ The Pingry School, Basking Ridge, New Jersey, United States of America

Reviewed on 9 May 2026; Accepted on 22 June 2026; Published on 26 October 2026

Protein solubility is a major bottleneck in biologic drug development, particularly for subcutaneous delivery, where poor solubility causes precipitation, inflammation, and dosing limitations. Computational prediction of solubility could reduce the cost and time of experimental screening, but remains difficult because solubility depends on complex interactions between amino acid sequence, structure, and local aggregation. While molecular dynamics simulations can model these interactions directly, they are computationally expensive, motivating the development of faster machine learning approaches. We present SmartSoluPred, a two-step machine learning pipeline: (1) predicting intrinsic protein solubility directly from sequence and structure (continuous percentage values), and (2) modeling mutation-induced changes (solubilizing or desolubilizing) in solubility to guide rational protein engineering. For intrinsic solubility prediction, we construct AlphaFold-derived residue graphs from the eSOL dataset, incorporating Aggrescan3D aggregation features, and 640-dimensional pre-trained ESM-2 (Evolutionary Scale Modeling) model, and fine-tuned them using a Feed Forward Network on the eSOL dataset. Our graph neural network (GNN) is based on the DeepMind graph network, a 3-fold update function in each layer to update global, node, and edge features. Fine-tuning ESM-2 achieves our best intrinsic solubility predictions ($R^2=0.479$). To our knowledge, this work is the first to use edge, node, and global embeddings to accurately predict protein solubility. For mutation-induced solubility changes, we fine-tune protein language embeddings to predict the effect of mutations on solubility using the SoluProtMutDB dataset. We fine-tune ESM-2 (35 M Parameters) using a Siamese architecture with a shared encoder for wild-type and mutant sequences, comparing embeddings through a binary classifier with focal loss to address the desolubilizing-skewed class imbalance. We benchmark against the current state-of-the-art Siamese graph convolutional neural networks on the same-protein evaluation using 5-fold CV. Our 1D-sequence-only model achieves 34.6% mean informedness, outperforming the current IEConv 3D-based state-of-the-art (33.2%) while eliminating dependence on structural data, and achieves 14.6% informedness for cross-protein generalization. However, the low cross-protein specificity of 26.5% indicates the model is best suited as a triage tool to prioritize candidates for further experimental validation, rather than as a standalone filter. These results demonstrate how our model matches or exceeds graph neural networks at mutation solubility prediction, while avoiding the computational and accessibility constraints of 3D structural data. Together, our work advances scalable, structure-aware, and mutation-sensitive solubility modeling, bridging 1D sequence embeddings and 3D structural representations for rational protein engineering.

Keywords: Protein solubility engineering, graph neural network, protein, ESM-2 protein language model, Siamese neural network, mutation effect prediction



The authors were mentored by Marie-Pierre Jolly*, The Pingry School, Basking Ridge, New Jersey, United States of America. Please direct correspondence to: Marie-Pierre Jolly* (mjolly@pingry.org). This is an Open Access article, which was copyrighted by the authors and published by BioTreks in 2026. It is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

PROTEIN solubility is a central constraint for biologic drugs because therapeutic proteins must remain soluble during expression, purification, storage, formulation, and administration. This challenge becomes especially important for subcutaneous delivery, where limited injection volumes often require high protein concentrations; under these conditions, poor solubility and aggregation can restrict dose, complicate formulation, and increase the risk of immunogenicity or other adverse immune responses (Badkar et al., 2021; Pham & Meng, 2020). As a result, accurately predicting protein solubility before extensive experimental screening could improve the selection and engineering of biologic candidates, reducing both development time and experimental cost.

Current computational solubility prediction remains difficult because solubility depends on multiple interacting factors, including amino acid composition, surface charge, hydrophobic exposure, structural packing, aggregation-prone regions, and the local context of each residue. Large experimental datasets, such as eSOL (Niwa et al., 2009) have enabled supervised modeling by measuring solubility across thousands of *Escherichia coli* proteins in a chaperone-free cell-free expression system (Niwa et al., 2009). Sequence-based tools such as SoluProt have demonstrated that machine learning can help prioritize soluble proteins. Still, solubility prediction remains a challenging task because sequence-level features alone may not fully capture long-range structural contacts or residue environments that influence aggregation and precipitation (Hon et al., 2021).

Recent advances in protein representation learning enable the integration of sequence, structure, and graph-based information to address these limitations. AlphaFold enables high-accuracy prediction of protein 3D structure from amino acid sequence, allowing proteins to be represented as residue graphs in which nodes correspond to residues and edges encode spatial relationships (Jumper et al., 2021). Aggrescan3D provides structure-based aggregation and solubility-related features that capture how aggregation-prone regions appear in a three-dimensional conformational context (Kuriata et al., 2019). In parallel, protein language models (PLM) such as ESM-2 learn rich residue-level embeddings from large-scale protein sequence data, capturing evolutionary and structural signals without requiring manual feature engineering (Lin et al., 2023). A global-aware graph network framework is particularly well-suited to combining these information sources because it can jointly update node, edge, and global protein representations, preserving both local residue context and whole-protein properties (Battaglia et al., 2018).

The first objective of this project is therefore to predict intrinsic protein solubility by combining AlphaFold-derived residue graphs, Aggrescan3D aggregation features, inter-residue distance embeddings, and ESM-2 sequence embeddings. This model is designed to test whether explicitly integrating edge-level geometry, residue-level biochemical context, and global protein information can improve solubility prediction beyond models that rely on only sequence or simplified structural descriptors. By training and evaluating this architecture on the eSOL dataset, the experiment provides a controlled benchmark for assessing whether structure-aware graph learning can better capture the determinants of baseline protein solubility.

A complementary challenge is predicting how mutations change solubility. This is essential for rational protein engineering because a candidate protein that is initially poorly soluble may become more developable through targeted sequence changes, while otherwise beneficial mutations may unintentionally reduce solubility. SoluProtMutDB has created an important foundation for this task by curating approximately 33,000 solubility-change measurements across about 17,000 protein variants and 103 proteins (Velecký et al., 2022). Recent mutation-focused models, including DeepMutSol and SoluProtMut, have used protein language embeddings, predicted structures, graph convolutional networks (GCN), and Siamese architectures to compare wild-type and mutant proteins (Velecký et al., 2025; Wang et al., 2024). However, models that depend heavily on 3D structural preprocessing may be less convenient for rapid, large-scale mutational scanning.

Thus, the second objective of this project is to model mutation-induced solubility changes using a Siamese ESM-2 architecture that compares wild-type and mutant sequences through a shared encoder. This design directly focuses the model on the representational difference introduced by mutation while avoiding the computational and accessibility constraints associated with 3D structure generation. Together, the intrinsic solubility model and the mutation-effect model form a two-step computational pipeline for rational protein engineering: first identifying solubility-relevant structural and sequence features, then predicting which mutations are likely to improve or impair solubility. That said, the mutation-effect model is intended to complement rather than replace experimental validation, as its current low specificity limits its use as a standalone filter. This approach supports the broader goal of making biologic drug development more scalable by enabling earlier, faster, and more accurate prioritization of soluble protein candidates (Figure 1).

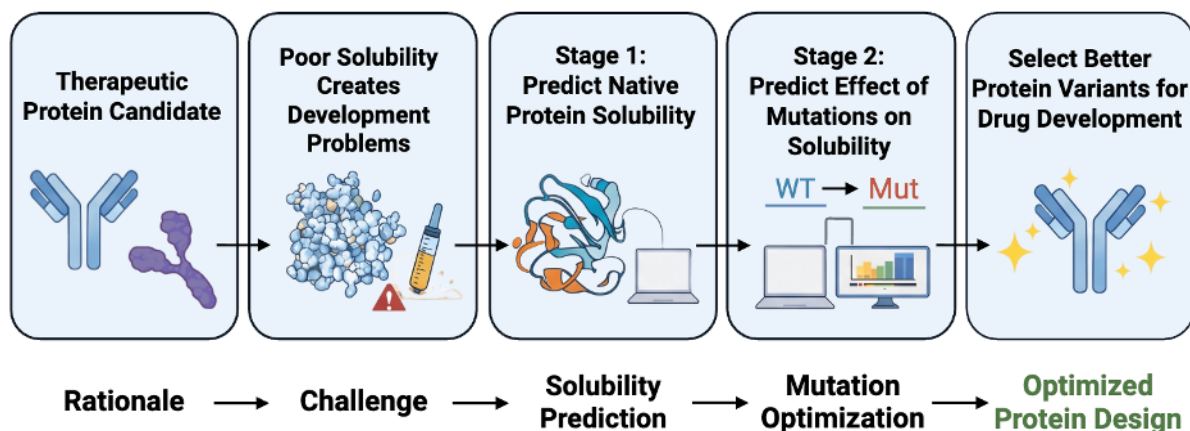


Figure 1. Generalized pipeline of SmartSoluPred framework for protein solubility modeling and drug development. Starting from a protein candidate with poor solubility, Stage 1 predicts intrinsic protein solubility from sequence and structure using a graph neural network. Stage 2 predicts whether a given point mutation increases or decreases solubility from the wild type using a Siamese ESM-2 architecture. Both stages allow for better protein engineering for drug development (Original Figure by the authors.)

Materials and methods

Study design

This study used two supervised predictive modeling workflows within the SmartSoluPred framework. The first workflow modeled intrinsic protein solubility as a protein-level regression task using eSOL soluble-expression values and AlphaFold-derived residue graphs (Niwa et al., 2009; Jumper et al., 2021). Five model configurations were compared: three GNN backbone variants (GCN, GAT, GINE), a surface-restricted GNN operating only on solvent-exposed residues, and a GNN with ESM-2 pretrained embeddings as node initialization. Each protein was represented as a graph in which residues were nodes and residue contacts were edges (with 8.0Å distance as cutoff for GCNs, GATs, GINEs, and ESM2-pretrained embeddings in a GAT, and 10.0Å distance as cutoff for the surface GNN). This graph-based approach allowed the model to use both sequence-derived features and three-dimensional structural context. Additionally, a fine-tuned ESM-2 PLM was evaluated as a sequence-only regression baseline. All configurations were trained using mean squared error loss and evaluated on a held-out test set using the coefficient of determination (R^2). Models were trained on fixed train/validation/test splits with set seeds, optimized using mean squared error, and evaluated with R^2 on

a held-out test set (Figure 2).

The second workflow modeled mutation-induced solubility change as a binary classification task using SoluProtMutDB (Velecký et al., 2022). Each example consisted of a paired wild-type and mutant sequence with an experimentally measured solubility change. Mutations with solubility change less than -0.1 were labeled desolubilizing, while all other mutations were labeled non-desolubilizing. A Siamese ESM-2 model (Jumper et al., 2021) encoded the wild-type and mutant sequences with shared weights, then classified the mutation using the wild-type embedding, mutant embedding, and their difference. Because the dataset was class-imbalanced, the classifier used balanced sampling and focal loss during training.

The two workflows were evaluated independently because they addressed different prediction problems: baseline solubility of an entire protein versus the effect of a specific mutation. Intrinsic-solubility models were assessed on a held-out eSOL test set, while mutation-effect models were evaluated using both within-protein 5-fold cross-validation on deeply mutagenized proteins and cross-protein testing on held-out proteins. Regression performance was measured with MSE, MAE, and R^2 , while mutation classification was measured with sensitivity, specificity, F1 score, and informedness, defined as sensitivity + specificity - 1. We decided to use informedness (Youden's J) because

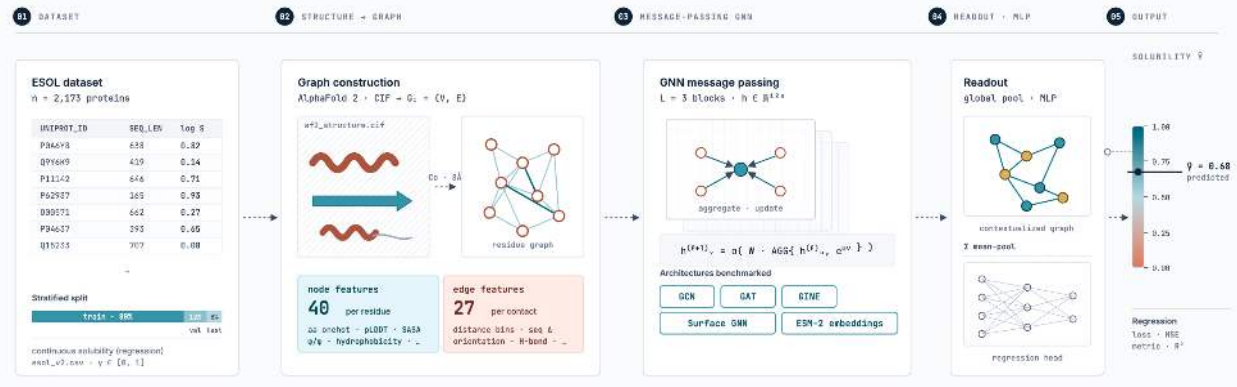


Figure 2. Pipeline of intrinsic protein solubility prediction. (1) Sequences and continuous solubility targets $y \in [0, 1]$ are drawn from the eSOL regression dataset and stratified into 80% train / 12% validation / 8% test splits. (2) For each protein, the AlphaFold 2 predicted structure is parsed from CIF and converted into a residue-level graph $G_i = (V, E)$ using a Ca-Ca contact threshold of 8 Å, yielding 40 node features per residue (amino-acid one-hot, pLDDT, SASA, dihedral angles, hydrophobicity, ...) and 27 edge features per contact (distance bins, sequence separation, relative orientation, hydrogen-bond indicators, ...). (3) Three message-passing blocks update node embeddings, $h_v^{l+1} = \sigma(W \cdot \text{AGG}_i^l h(v^l), e^{uv})$; we benchmark GCN, graph attention networks (GAT), GINE, a Surface GNN variant that operates on the molecular surface mesh, and an ESM-2-embeddings variant that initializes node features with pretrained protein-language-model embeddings. (4) A mean-pool readout collapses the contextualized graph to a fixed-size embedding which is passed through a 3-layer multilayer perceptron (MLP) regression head. (5) The model emits a continuous per-protein solubility prediction $\hat{y} \in [0, 1]$, optimized with mean-squared error (MSE) and evaluated by the coefficient of determination (R^2) (Original Figure by the authors.)

the mutation dataset is class-imbalanced, and accuracy would be misleadingly high for a model that simply predicts the majority class. Informedness treats both classes symmetrically and equals zero for a random or majority-class classifier, making it a more honest measure of true predictive discrimination. This two-pronged design tested whether SmartSoluPred could both rank proteins by predicted soluble expression and identify mutations likely to preserve or reduce solubility.

Datasets

Intrinsic solubility dataset

Intrinsic solubility modeling used the eSOL dataset, which reports soluble expression measurements for *E. coli* proteins generated in a chaperone-free cell-free expression system (Niwa et al., 2009). The raw eSOL table contained 4,132 protein entries, of which 3,173 had non-missing solubility measurements. Entries were retained if they had a valid protein identifier, a mapped AlphaFold structure, and a corresponding solubility value. Solubility was modeled as a continuous target using the reported soluble expression percentage.

AlphaFold-predicted structures were collected for eSOL proteins and stored in CIF/PDB format, with CIF files used as the primary input for structure parsing.

The structure directory contained 3,142 AlphaFold-derived models. During graph construction, structures were matched to eSOL solubility values using the cleaned metadata table and Aggrescan3D filename mapping. After structure parsing and metadata validation, 3,134 proteins yielded valid residue graphs with usable labels. These graphs were divided into training, validation, and test sets using a fixed random seed and an 80:12:8 split, resulting in 2,507 training proteins, 376 validation proteins, and 251 test proteins.

Mutation-effect dataset

Mutation-effect modeling used SoluProtMutDB, a curated database of experimentally measured mutation effects on protein solubility (Velecký et al., 2022). The processed mutation table contained 32,992 variant records, each with a wild-type sequence, mutant sequence, mutation annotation, categorical solubility-effect label, and quantitative solubility-change value. Records with complete wild-type sequence, mutant sequence, and quantitative solubility-change values were retained, yielding 32,408 examples for modeling.

This task was formulated as a binary classification. Variants with a reported solubility change of less than -0.1 were labeled desolubilizing, corresponding to a decrease in solubility greater than 10 percent-

age points and variants with a solubility change of -0.1 or greater were labeled non-desolubilizing. This threshold matches the classification cutoff used in the original deep mutational scanning (DMS) experiments and adopted by SoluProtMutDB (Klesmith et al., 2017; Velecký et al., 2022). The final dataset comprised 20,643 desolubilizing variants and 11,765 non-desolubilizing variants. Stratified data splits were used to preserve class proportions across partitions.

Structural preprocessing and graph construction

Structure parsing and residue-level feature extraction

AlphaFold structures were parsed at residual resolution from CIF files. For each standard amino acid residue, the parser extracted residue identity, chain identifier, residue index, C α coordinate, side-chain center, and model confidence. Predicted local distance difference test (pLDDT) values were read from the C α B-factor field and used as residue-level confidence features. Secondary structure annotations were assigned from helix and sheet records when available; all remaining residues were labeled coil.

Solvent-accessible surface area (SASA) was computed with the Shrake–Rupley algorithm implemented in Biopython. Absolute solvent-accessible surface area was retained, and relative solvent-accessible surface area was calculated by normalizing each residue value to the amino acid-specific maximum accessible surface area. For the Surface GNN experiments, residues were then filtered according to relative solvent-accessible surface area. Amino acids with relative SASA values greater than or equal to 0.25 were classified as surface-exposed and retained for graph construction, whereas residues below this threshold were excluded from the surface-only graph. Then, on this graph, the GCN, GAT, and GINE are run. This filtering emphasized solvent-exposed residues because they are more directly involved in protein-solvent interactions and are therefore expected to have greater biological relevance to solubility than buried or structurally embedded residues. Backbone orientation was represented by the normalized vector from the backbone nitrogen atom to the carbonyl carbon atom. Side-chain orientation was represented by the normalized vector from the C α atom to the C β atom when available, or by the vector from the C α atom to the calculated side-chain center for residues lacking a C β atom.

Aggrescan3D feature mapping

Per-residue Aggrescan3D scores were mapped to each protein structure to represent local aggregation propensity in structural context (Kuriata et al., 2019). For each residue, the graph representation retained the raw Aggrescan3D score, a local window average, and a binary hotspot indicator for positive aggregation scores. Protein-level Aggrescan3D summary statistics, including minimum, maximum, mean, and total score, were stored as graph-level metadata.

Residue graph construction

Each protein was represented as a residue-level graph in which nodes corresponded to amino acid residues. Edges were added between residues whose C α atoms were within 8 Å, and sequence-adjacent residues were connected to preserve peptide-chain continuity. Graphs were encoded with directed edges, with paired reverse edges used to represent bidirectional residue relationships.

Each node was encoded with a 40-dimensional feature vector comprising a 20-dimensional amino acid one-hot representation, normalized sequence position, normalized pLDDT, three Aggrescan3D-derived features, Kyte–Doolittle hydrophobicity, polarity, charge, absolute and relative solvent-accessible surface area, one-hot secondary structure state, backbone direction vector, side-chain direction vector, and a disorder proxy derived from pLDDT.

Each edge was encoded with a 27-dimensional feature vector describing pairwise geometry and residue context. Edge features included C α distance, side-chain center distance, normalized sequence separation, C α direction vector, backbone orientation relationship, side-chain alignment terms, edge confidence, flexibility proxy, solvent-accessibility overlap, secondary-structure match indicators, peptide adjacency, and physicochemical interaction indicators. The interaction indicators encoded approximate disulfide, ionic, salt-bridge-like, hydrogen-bond-like, van der Waals-like, polar, and hydrophobic contacts, together with hydrophobicity and charge products.

Graph neural network (GNN) modeling of intrinsic solubility

Model architecture

Node features were first projected to a learned hidden representation, after which a configurable message-passing backbone was applied. Three architectures were evaluated: GCN, graph attention networks (GAT), and graph isomorphism networks with edge attributes

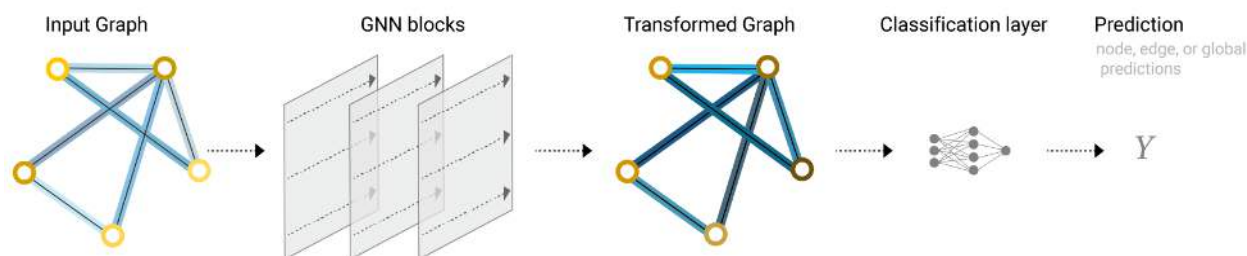


Figure 3. Schematic representing GNN creation and prediction. An input graph encoding node and edge features is passed through GNN blocks that iteratively update these representations. The transformed graph is passed through classification or regression for global predictions (Adapted from Sanchez-Lengeling et al., 2021, under CC BY 4.0).

(GINE). For architectures that used edge attributes directly, edge features were embedded with an MLP (Figure 3).

Each message-passing layer was followed by graph normalization, rectified linear unit activation, and dropout. Residual connections were used when tensor dimensions were compatible. Edge dropout was applied during training. Node embeddings were aggregated into a protein-level representation using global mean pooling and global max pooling, with attention pooling evaluated in selected configurations. The pooled representation was passed to an MLP that predicted the continuous eSOL solubility value.

Training and evaluation

Graph models were trained as supervised regression models using mean squared error loss and AdamW optimization with weight decay. The standard training configuration used a batch size of 32 for direct model training, with larger batch sizes used during grid search when hardware permitted. Models were trained for up to 100 epochs with early stopping based on validation mean squared error. Mixed-precision training was used when GPU acceleration was available.

Hyperparameter tuning varied the graph architecture, hidden dimension, number of message-passing layers, and learning rate. GCN models were evaluated with three to five layers, hidden dimensions from 64 to 512, and learning rates of 1×10^{-3} or 1×10^{-4} . GAT and GINE variants were evaluated under comparable settings. Performance was assessed on the held-out test set using mean squared error, mean absolute error, and the coefficient of determination (R^2). The validation set was used for model selection, and the test set was not used during training or tuning. Additional ablation experiments were conducted by masking predefined groups of node or edge features while preserving input dimensionality.

ESM-2 PLM fine-tuning for intrinsic solubility prediction

Model architecture

Solubility was predicted with a fine-tuned PLM designed for solubility regression. It wraps Meta's pre-trained ESM-2 (Evolutionary Scale Modeling) transformer backbone, specifically t30_150_M_UR50D, extracting residue-level representations from the ESM model. These per-residue embeddings are then aggregated into a single fixed-length sequence representation via masked mean pooling, which averages across sequence positions while ignoring padding tokens. This pooled vector is passed through a three-layer MLP regression head with ReLU activation, ultimately producing a single scalar solubility prediction per protein.

Training and evaluation

The model was trained on the eSOL dataset using mean squared error (MSE) as the loss function, comparing the predicted and actual solubility, and evaluated on a held-out test set. Training employed gradient accumulation over 64 steps with gradient norm clipping to stabilize optimization of the large ESM-2 backbone. The best-performing model checkpoint, based on validation MSE loss, was saved. Final model performance was assessed on the held-out test set by computing the coefficient of determination (R^2) between predicted and observed solubility values.

Siamese ESM-2 modeling of mutation effects

Model architecture

Mutation effects were modeled with a Siamese protein language model that used a shared ESM-2 encoder for wild-type and mutant sequences (Lin et al., 2023). The final model used the ESM-2 t12 35 M checkpoint. For each example, the wild-type and mutant

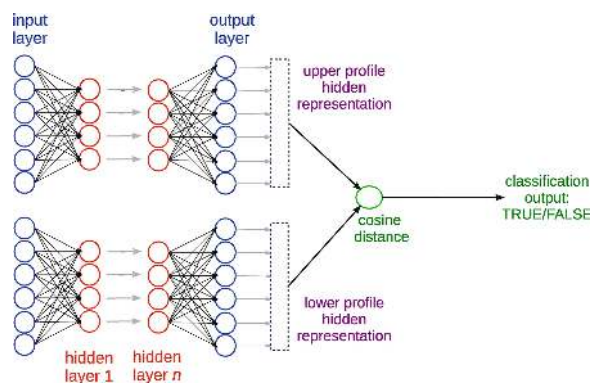


Figure 4. Architecture of the Siamese ESM-2 model for mutation-effect solubility prediction. Two twin encoders with shared weights process the wild-type and mutant sequences, producing two representations. The cosine distance is computed and passed to a classification output head (Original Figure by the authors.)

sequences were encoded independently with shared weights. Sequence-level embeddings were obtained from the model's pooled protein representation. The wild-type embedding, mutant embedding, and their difference were concatenated to form a pairwise mutation representation, which was then passed through a hidden layer with rectified linear unit activation and dropout, followed by a two-class classification head (Figure 4).

Training and evaluation

To address class imbalance, the mutation model was trained with focal loss and balanced sampling at a 1:1 class ratio during training. Hyperparameter tuning evaluated focal-loss alpha, focal-loss gamma, learning rate, and training duration. The selected configuration used alpha = 0.7, gamma = 2.0, and a learning rate of 3×10^{-5} , with early stopping.

Training used fixed random seeds and stratified data splits. Model performance was monitored using sensitivity, specificity, F1 score, and informedness, where informedness was defined as sensitivity + specificity - 1. Classification thresholds were selected by sweeping predicted probabilities on the validation data and choosing an operating point that maintained high specificity. Additional evaluation was conducted under within-protein and cross-protein settings. Within-protein evaluation used stratified cross-validation on deeply mutagenized proteins, whereas cross-protein evaluation trained on deeply mutagenized proteins and tested on held-out proteins.

Software and reproducibility

All analyses were implemented in Python. Structure parsing used Gemmi; solvent-accessibility calculations used Biopython; graph construction and graph neural network modeling used PyTorch Geometric; neural network training used PyTorch and PyTorch Lightning; and protein language model experiments used ESM-2 through the ESM and Hugging Face Transformers toolkits. Parsed graph objects were cached after conversion so that repeated training runs used identical feature tensors. Fixed random seeds were used for dataset splitting and model training, where supported by the training framework.

Results

Intrinsic solubility

The three GNN baselines varied considerably in performance. The best-performing GCN ($R^2=0.440$) is competitive with the top ESM-2 configuration, while the surface GNN ($R^2=0.320$) and embeddings GNN ($R^2=0.237$) performed substantially worse, suggesting that the choice of molecular representation is a critical factor in GNN-based solubility prediction (Figure 5).

The five ESM-2 configurations tested differ in MLP regression head depth and in the number of fine-tuned transformer layers (15, 20, 30: the maximum), though all layers are used for inference; only the last N have parameters adjusted. Increasing the number of fine-tuned layers consistently improved performance, while MLP depth showed diminishing returns; notably, extending to 4 MLP layers at 30 ESM-2 layers substantially reduced R^2 , suggesting overfitting. The best configuration (30 ESM-2 layers, 3 MLP layers, $R^2=0.479$) modestly outperformed the GCN baseline ($R^2=0.440$) and greatly outperformed all other GNNs. However, this comparison of ESM-2 fine tuning to GCNs must be taken with a grain of salt: bootstrap analysis of GCN test-set predictions ($n=251$, 2,000 resamples) yielded a 95% CI of [0.31, 0.50], indicating that the $R^2=0.04$ gap between the GCN baseline and the ESM-2 fine-tuned model is within sampling uncertainty. Due to computational constraints, multi-seed evaluation was not performed; this remains a direction for future work. All ESM-2 models greatly outperformed the surface GNN and Embedding GNN (Figure 5).

Mutation effect prediction

The within-protein generalization experiment (E1) evaluated the Siamese ESM-2 model using 5-fold cross-validation on four deeply mutagenized DMS proteins: LGK-GFP, TEM, PKS, and LGK-TAT (Figure 6). On TEM, the model achieved 44.7% informedness compared to 33.2% for the structure-based IEConv baseline, and

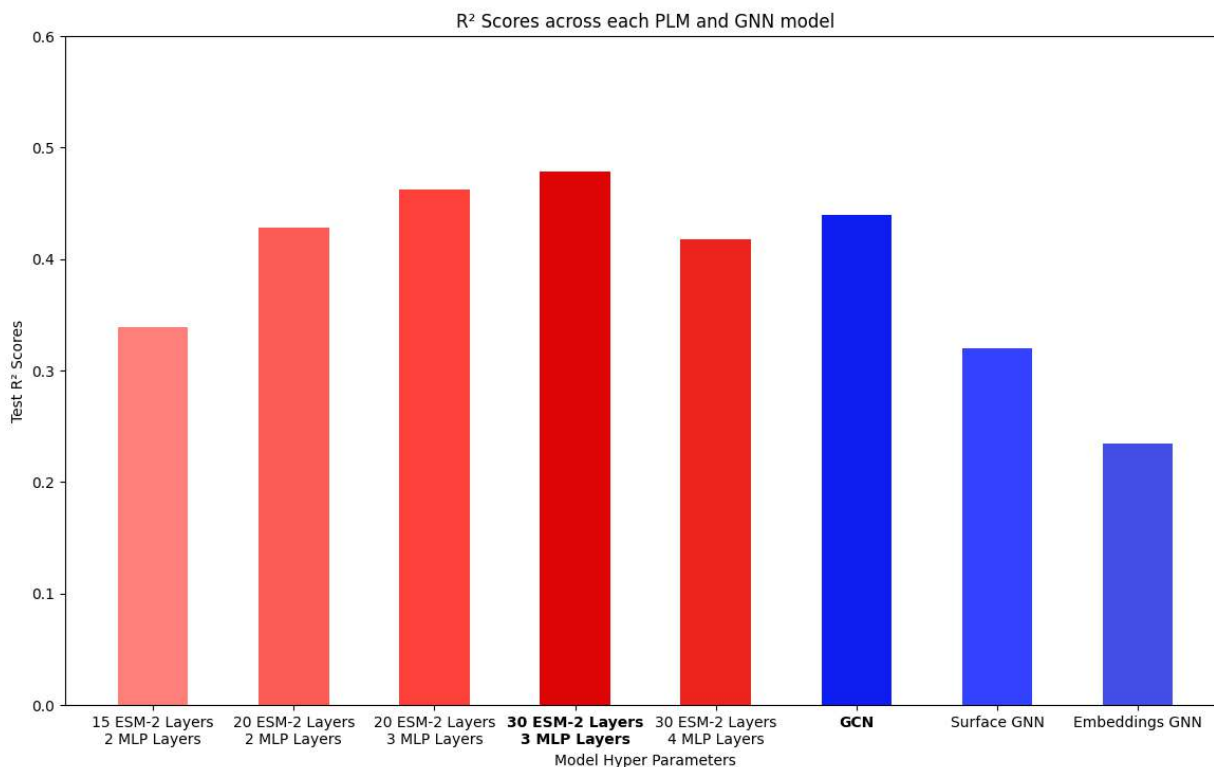


Figure 5. Test R^2 scores across ESM-2 (red) and GNN models (blue) with varying hyperparameter configurations. ESM-2 models vary in the number of fine-tuned transformer layers and downstream MLP layers. The highest performing model of each type is bolded (Original Figure by the authors).

on PKS, it achieved 36.9% versus 26.8% for IEConv, outperforming the 3D model on both. On LGK-GFP, the model scored 22.1% informedness compared to 39.6% for IEConv, representing when structural information provided a meaningful advantage. No IEConv baseline was available for LGK-TAT, where our model achieved 44.7% informedness. Averaged across all four proteins, our Siamese ESM-2 achieved a mean informedness of 34.6%, compared to 33.2% for IEConv across the three proteins where a comparison was possible. The high variance in the LGK-TAT result, reflected in the wide error bars in Figure 6, suggests that within-protein performance can be sensitive to the specific mutational landscape of individual proteins. The E1 results demonstrate that a computationally efficient, sequence-only model can match or exceed a 3D structure-dependent method on the majority of tested proteins.

In the cross-protein generalization experiment (E2), the model was trained on the four DMS proteins and evaluated on 44 held-out non-DMS proteins (Figure 7a). The Siamese ESM-2 model achieved 14.6% informedness, outperforming all five baselines, including CamSol at 10.1%, SoluProt at 10.0%, OptSolMut

at 13.0%, and the structure-based IEConv at 13.4% (Velecký et al., 2025). IEConv relies on 3D structural information yet is surpassed by our sequence-only model, indicating that the contextual representations in ESM-2 embeddings can substitute for explicit structural inputs in cross-protein generalization. The confusion matrix for the non-DMS test set (Figure 7b) provides additional detail, showing that the model correctly identified 75 of 102 true desolubilizing variants at a sensitivity of 88.2% but correctly identified non-desolubilizing variants at a specificity of 26.5%. This asymmetry indicates that the model tends to over-predict desolubilizing outcomes, likely a consequence of the 1.75 to 1 class imbalance in SoluProtMutDB that persists despite focal loss and balanced sampling during training.

The cross-protein generalization heatmap (Figure 8) displays informedness across all combinations of training and test protein sets, revealing that within-protein performance on the blue dashed diagonal is consistently the strongest, reaching up to 52.6% on TEM. Among cross-protein conditions, single-protein training is highly variable, with several off-diagonal cells showing significantly lower informedness when

training on one protein and testing on another. In contrast, training on all four DMS proteins simultaneously in the bottom row of Figure 8 yields the most consistent and highest transfer across test targets, achieving 50.8%, 25.4%, 44.6%, and 33.3% informedness on TEM, PKS, LGK-GFP, and LGK-TAT, respectively. These patterns suggest that expanding the diversity of training proteins can be more impactful than increasing in-depth on any individual protein.

a b

Discussions

Intrinsic solubility prediction discussion

The eSOL component of SmartSoluPred supports the hypothesis that intrinsic protein solubility can be learned more effectively when sequence context, predicted structure, aggregation propensity, and residue-residue geometry are modeled together. The best GNN-based model achieved an R^2 of 0.440 on the held-out test set across 3,134 validated *E. coli* protein graphs. Notably, fine-tuning the full ESM-2 transformer backbone (t30_150 M_UR50D, all 30 layers with a 3-layer MLP regression head) yielded $R^2=0.479$, modestly surpassing the best GCN ($R^2=0.440$) and substantially outperforming the surface GNN ($R^2=0.320$) and embeddings GNN ($R^2=0.237$). The consistent improvement with increasing numbers of fine-tuned layers confirms that task-specific adaptation of the pretrained language model extracts solubility-relevant signal that frozen embeddings alone do not provide, though hyperparameter sensitivity remains high. Adding a fourth MLP layer at 30 fine-tuned layers reduced performance, suggesting overfitting. These results should be interpreted in context: the eSOL label reflects a combined phenotype shaped by folding kinetics, aggregation-prone motifs, expression context, and experimental noise rather than a single physicochemical property (Niwa et al., 2009). The encoding of all of these properties may create some risk that representation-level familiarity with test sequences could occur, which would place a limit on how well these results generalize. However, this risk is relatively low as solubility is not a property ESM-2 could have implicitly learned from sequence co-occurrence patterns alone, given the complex, context-dependent nature of the eSOL phenotype and the absence of solubility labels from ESM-2's pretraining.

A central strength of the architecture is treating proteins as molecular graphs rather than flat sequence descriptors, allowing residues that are distant in primary sequence but proximal in 3D space to exchange information. The combination of AlphaFold-derived coordinates, Aggrescan3D aggregation features, multi-

channel distance embeddings, and ESM-2 residue embeddings, which are used as node features in the GNN, enables the model to represent local residue identity and broader structural context jointly (Jumper et al., 2021; Kuriata et al., 2019; Lin et al., 2023). The global-feature-aware DeepMind graph network, which jointly updates node, edge, and global protein representations at each message-passing layer (Battaglia et al., 2018), is particularly well-suited to solubility, as it learns context-dependent patterns across biological scales. A single aggregation-prone residue may be tolerated when buried, but a cluster of such residues on an exposed surface increases precipitation risk; a global readout over graph-updated embeddings can learn this distinction more directly than a model that averages residue features before prediction.

The primary practical value of the model is early-stage candidate prioritization before costly cloning, expression, and purification. Because all inputs are computable from sequence and predicted structure, the model applies to proteins lacking experimentally solved structures. Within the SmartSoluPred pipeline, the intrinsic-solubility model serves as a first-stage triage, ranking candidates by predicted soluble expression before the mutation-effect model evaluates which substitutions are likely to improve or impair solubility. The graph-language framework is also transferable: with appropriate retraining, the same architecture could model soluble expression in yeast or mammalian cells, or estimate aggregation risk under formulation conditions, by substituting new target labels while preserving the same biological inputs.

The current results represent a useful predictive signal rather than a complete solution. Residual unexplained variance partly reflects the eSOL target itself: soluble expression in a chaperone-free cell-free *E. coli* system captures translation efficiency, folding pathway, and kinetic trapping together, and differs from thermodynamic solubility under purification or formulation conditions. The model also depends on static AlphaFold structures, which may miss conformational heterogeneity, intrinsic disorder, oligomerization state, and solvent-dependent rearrangements central to aggregation; pLDDT confidence features partially address this but do not represent the full conformational ensemble. Dataset composition is a further limitation: the eSOL benchmark is dominated by *E. coli* proteins from a single experimental protocol, and a random split may not fully measure transfer to therapeutic scaffolds, antibody fragments, or proteins from distant organisms. Future evaluations should include sequence-identity-clustered splits and external test sets to determine whether the model learns generalizable physicochemical determinants or patterns

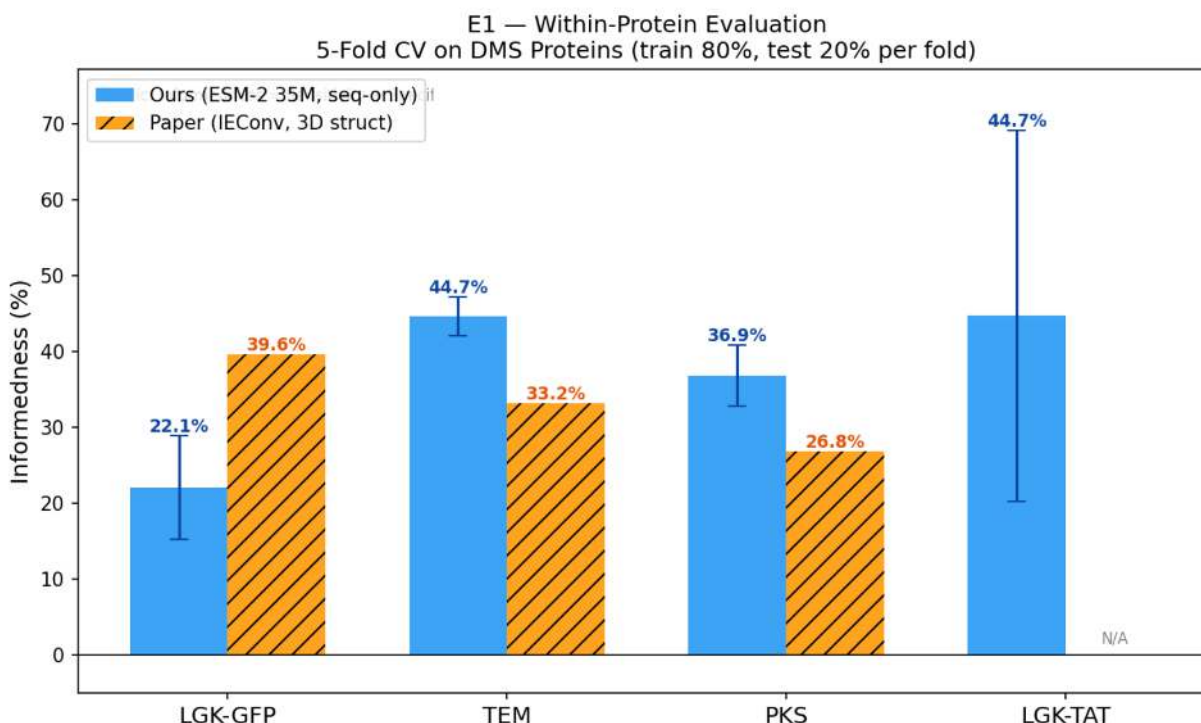


Figure 6. Within-protein evaluation of the Siamese ESM-2 model versus the IEConv baseline across four DMS proteins. Informedness (%) from 5-fold cross-validation (80/20 train/test split per fold); error bars represent ± 1 standard deviation. Our sequence-only ESM-2 model outperforms the structure-based IEConv on TEM and PKS, while IEConv leads on LGK-GFP. LGK-TAT has no IEConv baseline (Original Figure by the authors.)

specific to the eSOL proteome and assay. Finally, although the model uses biologically motivated features, attributions are not automatically causal. Residue-level saliency maps and graph perturbation analyses should be benchmarked against Aggrescan3D hotspots and mutational experiments before mechanistic conclusions are drawn.

Mutation-induced solubility prediction

Within-Protein Generalization (E1)

In the within-protein evaluation, our Siamese ESM-2 model was assessed via 5-fold cross-validation on four deeply mutagenized DMS proteins. Following standard practice, we train and test on mutations from the same protein, which may overestimate generalization to novel proteins; therefore, an additional cross-protein evaluation is administered (E2). The model matched or exceeded the 3D structure-based IEConv baseline for two of the three proteins with available comparisons, achieving 44.7% versus 33.2% on TEM and 36.9% versus 26.8% on PKS, while underperforming on LGK-GFP at 22.1% versus 39.6%. That a sequence-only model achieves competitive or superior

performance to a structure-dependent approach on the majority of tested proteins is notable, as it avoids the computational overhead of 3D structural preprocessing while remaining broadly applicable. The underperformance on LGK-GFP likely reflects the particular demands of that protein's mutational landscape. Because LGK-GFP contains a fluorescent protein fusion, mutations near the GFP interface may produce solubility changes driven more by precise three-dimensional contacts than by sequence-level patterns that ESM-2 can reliably detect, which would explain why structural information confers a meaningful advantage in that case but not on TEM or PKS. The high variance observed for LGK-TAT further suggests that within-protein performance is sensitive to the distribution of mutations across training folds and warrants further investigation.

Cross-Protein Generalization (E2)

In the cross-protein setting, the model was trained on DMS proteins and evaluated on 44 held-out non-DMS proteins. Our ESM-2 model achieved 14.6% informedness, outperforming all baselines, includ-

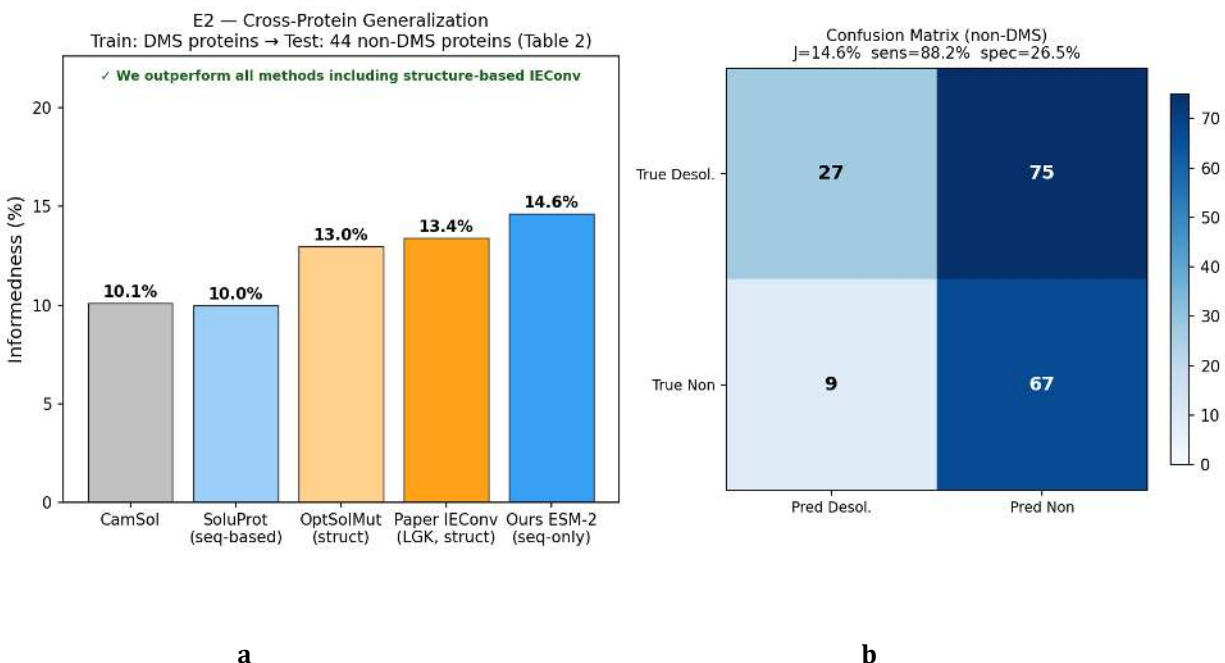


Figure 7. Cross-protein generalization performance of the Siamese ESM-2 model against existing methods. (a) Informedness J (%) across five models trained on DMS proteins and tested on 44 non-DMS proteins; our sequence-only ESM-2 model (14.6%) outperforms all baselines, including the structure-based IEConv (13.4%). (b) Confusion matrix for the ESM-2 model on non-DMS proteins (Informedness $J=14.6\%$, sensitivity=88.2%, specificity=26.5%).

ing both sequence-based methods and the structure-based IEConv at 13.4%. This result suggests that the evolutionary and biochemical patterns captured during large-scale ESM-2 pretraining generalize broadly enough across protein families to substitute for explicit structural inputs, and that the model has learned something about how mutations shift sequence context rather than simply memorizing features of the training proteins. A central limitation, however, is visible in the confusion matrix. While sensitivity is high at 88.2%, specificity is only 26.5%, meaning the model correctly flags most desolubilizing mutations but misclassifies the majority of non-desolubilizing ones. For a protein engineer screening candidate variants, this asymmetry has real consequences, as a large fraction of potentially beneficial mutations would be incorrectly rejected. That said, the low specificity means many viable variants would be wrongly rejected, so the model should complement rather than replace more rigorous downstream evaluation. This skew reflects the persistent effect of class imbalance in SoluProt-MutDB, where desolubilizing variants outnumber non-desolubilizing ones at roughly 1.75 to 1. Focal loss and balanced sampling partially corrected for this during training, but the decision boundary remains biased toward the majority class, and more aggressive interven-

tion through oversampling, threshold calibration, or cost-sensitive loss functions represents the most important direction for improving practical utility.

Overall Generalization Patterns

The protein-by-protein heatmap reveals that generalization is highly context-dependent and that training data diversity is the dominant factor in robust cross-protein transfer. Within-protein performance is consistently the strongest, as expected, since the model trains and tests on mutations from the same protein. Single-protein training, by contrast, produces highly variable transfer, with strong performance on some targets and near-chance results on others, likely because a model trained on one protein occupies a narrow region of sequence space that may not extend to evolutionarily or structurally distant targets. Training on all four DMS proteins simultaneously yields the most consistent informedness scores across test targets, underscoring the value of diverse training coverage. This has a direct implication for future dataset development, where adding new proteins to SoluProt-MutDB is likely to improve generalization more than expanding mutational coverage of proteins already well represented in the training set.

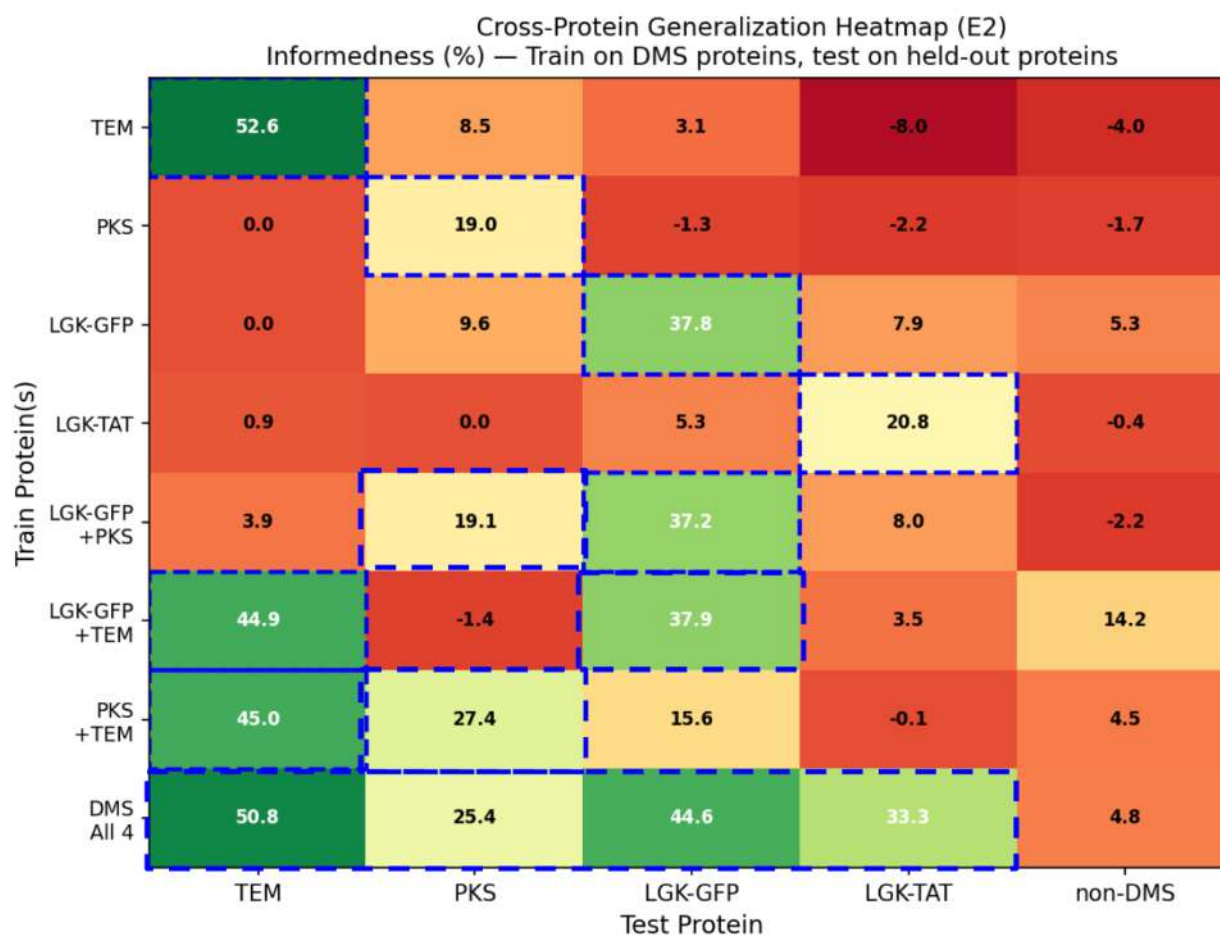


Figure 8. Generalization of the Siamese ESM-2 model measured by informedness (%). Each cell shows performance for a given training (y-axis) and test (x-axis) protein set. Blue dashed outlines indicate within-protein conditions; green and red cells indicate high and low performance, respectively. Training on all four DMS proteins (bottom row) yields the most consistent transfer. Inspired by Velecký et al. (2025) (Original Figure by the authors.)

Next steps

The next technical priority is to improve the structural resolution of the graph representation by developing hierarchical graph neural networks (HGNNs) that link atom-level and residue-level graphs within a single architecture. In the current residue-level representation, much of the fine-grained chemistry of each amino acid is compressed into residue summaries. An HGNN could reduce this loss of information by allowing atom-level message passing to capture local chemical interactions while residue-level message passing preserves the scalability needed for full-protein prediction. Atom-level edges could better represent hydrogen-bond geometry, aromatic interactions, salt bridges, side-chain packing, disulfide contacts, and solvent exposure, while the residue graph would model longer-range structural organization across the protein. This hierarchy may be especially useful for dis-

tinguishing residues that share the same amino acid identity but occupy different local chemical environments. For example, two lysines may appear identical at the sequence level but differ substantially in solubility relevance if one is solvent-exposed and forms stabilizing polar contacts while another is buried near a hydrophobic patch. By integrating atom-level chemistry with residue-level topology, HGNNs could allow SmartSoluPred to model these context-dependent effects more directly than a single-resolution graph.

In parallel, the ESM-2 component can be improved by using sequence-aware heads, such as convolutional or attention-based readouts, that retain positional patterns across the protein rather than collapsing all residue embeddings into a single pooled vector too early. This improvement would complement HGNN development: while HGNNs would strengthen the structural branch of the model, sequence-aware

ESM-2 heads would preserve residue-order information in the language-model branch. Together, these changes could make the model more sensitive to both three-dimensional molecular interactions and sequence-level motifs associated with solubility, aggregation, and expression behavior.

For current models, more robust confidence interval testing must be done with multiple different seeds in order to truly grasp the effectiveness of each model. A full multi-seed evaluation was not done due to computational constraints, especially with the larger Protein Language Models.

The model would also benefit from broader and more rigorous training objectives. Multi-task learning could combine eSOL soluble expression with related properties such as aggregation propensity, expression yield, thermal stability, disorder, localization, and experimentally measured mutation effects. This would encourage the model to learn representations that are less dependent on a single assay. Uncertainty estimation should be added so that predictions are accompanied by confidence intervals or calibration scores. In practice, this would allow users to distinguish high-confidence rankings from cases where the model is extrapolating beyond its training distribution. Active learning could then prioritize uncertain or high-impact proteins for wet-lab testing, making each new experiment more informative.

For the mutation effect prediction component, the most immediate priority is addressing the class imbalance identified in the E2 cross-protein generalization experiments. Although focal loss and balanced sampling brought the model to 14.6% informedness, the resulting decision boundary remains skewed toward desolubilizing predictions, as reflected in the low specificity of 26.5%. The next phase of work will systematically explore alternative focal loss parameterizations and negative downsampling strategies for the minority class, with the goal of improving specificity without sacrificing the high sensitivity the model already achieves. Threshold calibration on held-out validation data represents a further avenue, since adjusting the operating point of the classifier independently of training may recover a more balanced decision boundary without requiring architectural changes. Taken together, these interventions target the single most significant gap between the current model and a practically useful screening tool.

Beyond computational improvements, we aim to validate the model experimentally, following the precedent set by the SoluProtMut authors, who tested their predictions against 64 proprietary mutations in vitro. We plan to design a small panel of targeted point mutations in a well-characterized protein such

as GFP, express wild-type and mutant variants under matched conditions in a standard bacterial expression system, and quantify soluble expression by separating soluble and insoluble fractions through sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) or fluorescence measurement. This would provide independent experimental validation of model predictions outside the SoluProtMut dataset entirely, which is important because strong benchmark performance does not guarantee that the model captures the biochemistry driving solubility changes in new proteins.

We also intend to incorporate proteins from other IRT groups at our school whose expression behavior has been difficult to characterize, bringing them into the modeling pipeline to test whether SmartSoluPred can guide productive mutations in a real engineering context. For example, the team could select proteins predicted to span high, medium, and low intrinsic solubility, express wild-type and mutant variants under matched conditions in a bacterial expression system, separate soluble and insoluble fractions, and quantify soluble expression by fluorescence or SDS-PAGE band intensity. These collaborations would expand the diversity of proteins the model is evaluated against, creating a direct feedback loop between computational predictions and wet-lab experiments. This will help determine whether the model is most useful for ranking proteins, flagging high-risk constructs, or guiding redesign.

Author contributions

SB, EC, RL, DP, AS, and AW conceived and planned the project. SB and AW designed and implemented the Siamese ESM-2 architecture for mutation-induced solubility prediction. SB led benchmarking against the SoluProtMut paper, constructed train/test splits, and stratified 5-fold cross-validation. AW formulated the original Siamese ESM-2 fine-tuning transformer architecture. SB, AW, and DP implemented different approaches to address class imbalance in SoluProtMutDB, including focal loss and direct informedness, weighted loss, and negative downsampling, respectively. EC and DP conducted exploratory data analysis of the eSOL dataset, selected the initial graph features, and experimented with different GNN architectures. DP additionally implemented NNConv architectures and the novel Surface GNN. RL and DP extracted AlphaFold structural data for eSOL proteins and created scripts to parse PDB and CIF files. DP and AS collected Aggrescan3D aggregation features and mapped proteins to PDB IDs. RL led ESM-2 embedding extraction and the incorporation of ESM-2 embeddings into

GNNs. EC integrated ESM-2 embeddings into global-aware graph neural networks. SB and DP took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Acknowledgements

We would like to thank Dr. Marie-Pierre Jolly, our research advisor, for her guidance during weekly check-ins and her insights on the project's direction. Thank you to The Pingry School and the Independent Research Team (IRT) program led by Dr. Morgan D'Ausilio for facilitating and funding this research.

References

- Badkar, A. V., Gandhi, R. B., Davis, S. P., & LaBarre, M. J. (2021). Subcutaneous delivery of high-dose/volume biologics: Current status and prospect for future advancements. *Drug Design, Development and Therapy*, 15, 159–170. <https://doi.org/10.2147/DDDT.S287323>
- Battaglia, P. W., Hamrick, J. B. C., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V., ... Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv*. <https://doi.org/10.48550/arXiv.1806.01261>
- Hon, J., Marusiak, M., Martinek, T., Kunka, A., Zengulka, J., Bednar, D., & Damborsky, J. (2021). SoluProt: Prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics*, 37(1), 23–28. <https://doi.org/10.1093/bioinformatics/btaa1102>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Klesmith, J. R., Bacik, J., Wrenbeck, E. E., Michalczyk, R., & Whitehead, T. A. (2017). Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(9), 2265–2270. <https://doi.org/10.1073/pnas.1614437114>
- Kuriata, A., Iglesias, V., Pujols, J., Kurcinski, M., Kmiecik, S., & Ventura, S. (2019). Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. *Nucleic Acids Research*, 47(W1), W300–W307. <https://doi.org/10.1093/nar/gkz321>
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), 1123–1130. <https://doi.org/10.1126/science.ade2574>
- Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., & Taguchi, H. (2009). 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. <https://doi.org/10.1073/pnas.0811922106>
- Pham, N. B., & Meng, W. S. (2020). Protein aggregation and immunogenicity of biotherapeutics. *International Journal of Pharmaceutics*, 585, 119523. <https://doi.org/10.1016/j.ijpharm.2020.119523>
- Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltchko, A. B. (2021, September 2). *A gentle introduction to graph neural networks*. *Distill*. <https://doi.org/10.23915/distill.00033>
- Song, G., Luo, Z., Geng, A., Xu, J., Meng, Y., Cui, F., Wei, L., Zou, Q., & Zhang, Z. (2026). 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. <https://doi.org/10.1021/acs.jcim.5c02262>
- Velecký, J., Faldynová, H., Hermosilla, P., Sendlerová, N., Doerr, M., Egersdorfová, S., Bornscheuer, U. T., Damborský, J., Prokop, Z., & Mazurenko, S. (2025). SoluProtMut: Siamese deep learning for predicting solubility effects of protein mutations with experimental validation. *bioRxiv* [Preprint]. <https://doi.org/10.1101/2025.09.26.676459>
- Velecký, J., Hamšíková, M., Štourač, J., Musil, M., Damborský, J., Bednář, D., & Mazurenko, S. (2022). SoluProtMutDB: A manually curated database of protein solubility changes upon mutations. *Computational and Structural Biotechnology Journal*, 20, 6339–6347. <https://doi.org/10.1016/j.csbj.2022.11.009>
- Wang, J., Chen, S., Yuan, Q., Chen, J., Li, D., Wang, L., & Yang, Y. (2024). Predicting the effects of mutations on protein solubility using graph convolution network and protein language model representation. *Journal of Computational Chemistry*, 45(8), 436–445. <https://doi.org/10.1002/jcc.27249>