SSEmb: A joint embedding of protein sequence and structure enables robust variant effect predictions

Nature Communications, 2024

LifeLU Reading Club

Amirreza Sattarzadeh - 14 August 2025
(Emir Rıza Settarzade)

SSEmb: A Joint Embedding of Protein Sequence and Structure for Robust Variant Effect Predictions

- Integrates sequence and structural information in a single model
- Addresses limitations of sequence-only or structure-only approaches
- Applications in disease variant classification & protein engineering

Background

- Small changes in amino acid sequence can impact protein structure, stability, and function
- MAVEs (Multiplexed Assays of Variant Effects) provide large-scale data
- Limitations: cost, time, incomplete variant coverage
- Computational predictors can fill the gap but often rely on one data type

Existing Approaches

- Sequence-only models (e.g., GEMME)
  - Capture evolutionary conservation
  - Limited when MSA depth is low
- Structure-only models (e.g., Rosetta)
  - Capture stability and abundance changes
  - Less accurate for activity predictions
- Need for integrated representation

SSEmb Model Overview

- Combines an MSA Transformer with a structure-based Graph Neural Network
- Structure-constrained attention focuses on spatially close residues
- Joint embedding captures complementary sequence and structure information
- End-to-end training for variant effect prediction

Model Components

- Structure-Constrained MSA Transformer
  - Initialized from pre-trained model
  - Row attention masked by 3D proximity
- GNN (Geometric Vector Perceptron)
  - Processes protein graph
  - Combines with sequence embeddings
- Masked token prediction objective

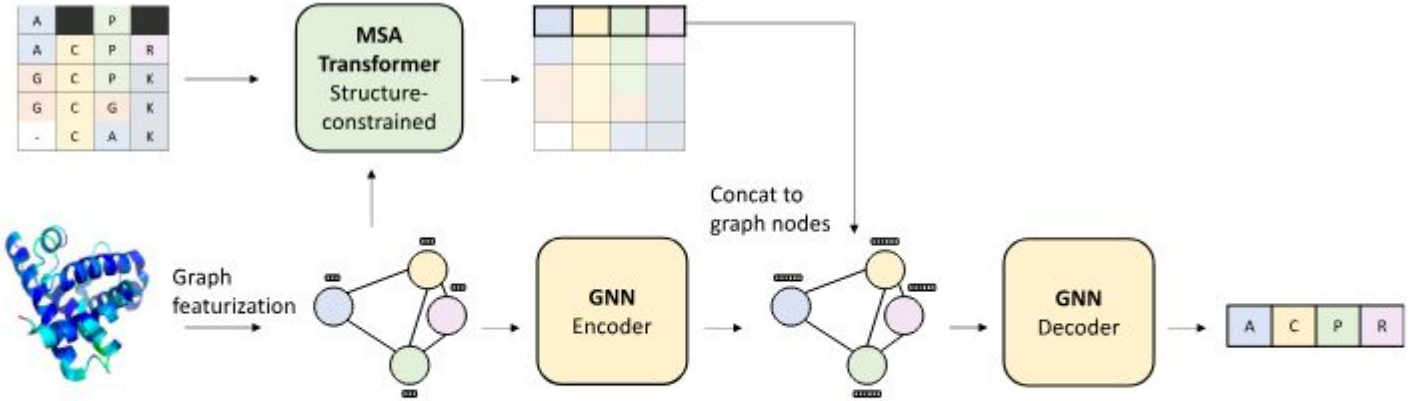

Fig. 1 | Overview of the SSEmb model and how it is trained. The model takes as input a subsampled MSA with a partially masked query sequence and a complete protein structure. The protein structure graph is used to mask (constrain) the row attention (i.e., attention across MSA columns) in the MSA Transformer. The MSA query sequence embeddings from the structure-constrained MSA Transformer are concatenated to the protein graph nodes. During training, SSEmb tries to predict the amino acid type at the masked positions. The model is optimized using the cross-entropy loss between the predicted and the true amino acid tokens at the masked positions. Variant effect prediction is made from these predictions as described in Methods.

Training & Data

- **Training set: CATH 4.2 dataset – 18,204 proteins, 40% non-redundancy**
- **MSA generation via MMSeqs2 + ColabFold**
- **BERT-style masking: 15% positions, various masking strategies**
- **Gradual unfreezing: train GNN first, then fine-tune MSA Transformer**

Validation on MAVE Datasets

- **MAVE: Multiplexed Assays of Variant Effects – large-scale functional assays**
- **Validation set: 10 datasets covering both activity & abundance assays**
- **SSEmb vs GEMME (sequence-only) and Rosetta (structure-only):**
  - **Matches GEMME in activity assays**
  - **Outperforms GEMME and Rosetta in abundance assays**
- **Mean Spearman correlation: SSEmb 0.503 > GEMME 0.484 > Rosetta 0.422**

Table 1 | Overview of SSEmb results on the MAVE validation set after model training

Protein	MAVE reference	MAVE type	Spearman $\lvert\rho_S\rvert$ (↑)		
			SSEmb	GEMME	Rosetta
NUD15	Suiter et al. 2020	Abundance	**0.584**	0.543	0.437
TPMT	Matreyek et al. 2018	Abundance	0.523	**0.529**	0.489
CP2C9	Amorosi et al. 2021	Abundance	**0.609**	0.423	0.519
P53	Kotler et al. 2018	Competitive growth	0.577	**0.655**	0.488
PABP	Melamed et al. 2013	Competitive growth	**0.595**	0.569	0.384
SUMO1	Weile et al. 2017	Competitive growth	**0.481**	0.406	0.433
RL401	Roscoe & Bolon 2014	E1 reactivity	**0.438**	0.390	0.366
PTEN	Mighell et al. 2018	Competitive growth	0.422	**0.532**	0.423
MAPK	Brenan et al. 2016	Competitive growth	0.395	**0.445**	0.307
LDLRAP1	Jiang et al. 2019	Two-hybrid assay	**0.411**	0.348	0.377
Mean	–	–	**0.503**	0.484	0.422

We use the Spearman correlation coefficient to quantify the agreement between the data generated by the MAVEs and the predictions from SSEmb, GEMME, and Rosetta. In this validation, only single-mutant variant effects were considered. The following protein structures were used in the SSEmb and Rosetta input: NUD15: 5BON_A, TPMT: 2H11_A, CP2C9: 1R90_A, P53: 4QO1_B, PABP: 1CVJ_G, SUMO1: 1WYW_B, RL401: 6NYO_E, PTEN: 1D5R_A, MAPK: 4QTA_A, LDLRAP1: 3SO6_A. The following UniProt IDs were used as input to construct multiple sequence alignments for GEMME: NUD15: Q9NV35, TPMT: P51580, CP2C9: P11712, P53: P04637, PABP: P11940, SUMO1: P63165, RL401: P0CH08, PTEN: P60484, MAPK1: P28482, LDLRAP1: Q5SW96. Some assay data points were removed during the merging of predictions in order to facilitate fair comparison between models.

[1] Bold values correspond to the best-performing model for each MAVE.

ProteinGym Benchmark

- Improved accuracy over MSA Transformer, especially for low-depth MSAs
- Comparable or better performance than other high-accuracy predictors
- Robust to predicted structures from AlphaFold

Table 2 | SSEmb performance on the originally released ProteinGym substitution benchmark compared to other variant effect prediction models grouped by UniProt ID and segmented by MSA depth

Model	Spearman ρ_S by MSA depth (↑)			
	Low	Medium	High	All
TranceptEVE L	**0.451**	**0.462**	0.502	**0.468**
GEMME	0.429	0.448	0.495	0.453
SSEmb (ours)	0.449	0.439	0.501	0.453
Tranception L	0.438	0.438	0.467	0.444
EVE (ensemble)	0.412	0.438	0.493	0.443
VESPA	0.411	0.422	**0.514**	0.438
EVE (single)	0.405	0.431	0.488	0.437
MSA Transformer (ensemble)	0.385	0.426	0.470	0.426
ESM2 (15B)	0.342	0.368	0.433	0.375
ProteinMPNN	0.189	0.151	0.237	0.175

Assays from the SSEmb validation set have been excluded from the original data set. Low: $N_{eff}/L < 1$, Medium: $N_{eff}/L < 100$, High: $N_{eff}/L > 100$[2].

[2] Bold values correspond to the best-performing model for each MSA class.

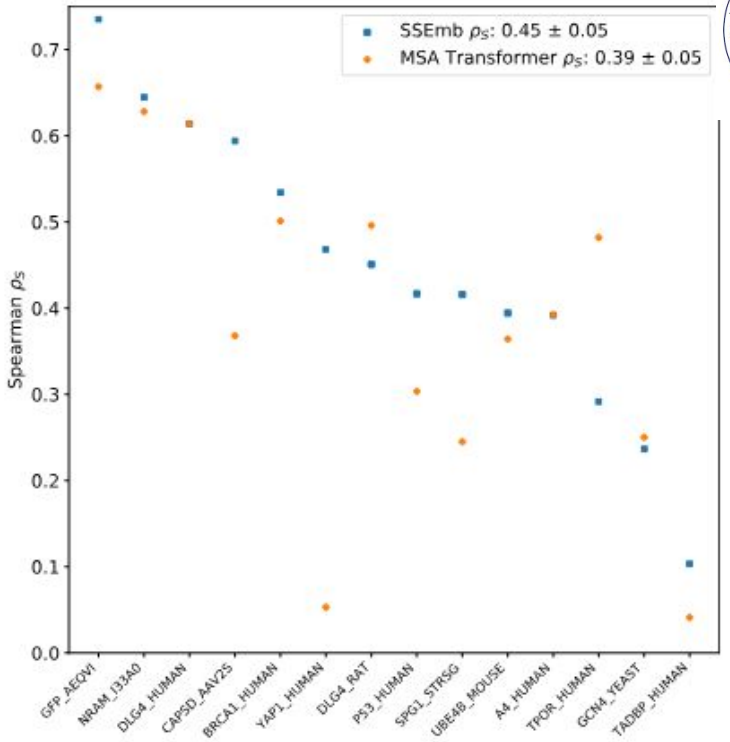

Fig. 2 | Overview of SSEmb results on the ProteinGym low-MSA ($N_{eff}/L < 1$) substitution benchmark subset grouped by UniProt ID. Spearman correlations are plotted for both SSEmb (blue) and the MSA Transformer ensemble (orange). The mean and standard error of the mean of the set of all ProteinGym Spearman correlations are presented in the legend. Assays from the SSEmb validation set have been excluded from the original data set. Source data are provided as a Source Data file.

Additional Applications

1. Zero-shot Protein Stability Prediction
   - Task: Predict ΔΔG (change in free energy) upon mutation
   - SSEmb embeddings → linear regression model
   - Spearman correlation: 0.61 (competitive with state-of-the-art)
2. Disease-causing Variant Classification
   - Dataset: ClinVar
   - Task: Classify pathogenic vs benign variants
   - SSEmb outperforms most general-purpose protein language models
3. Protein-Protein Binding Site Prediction
   - Task: Identify interface residues in protein complexes
   - SSEmb embeddings capture functional site information
   - Competitive with models trained specifically for binding site detection

Table 3 | SSEmb performance on the ProteinGym clinical substitution benchmark compared to other variant effect prediction models[67]

Model	Avg. AUC (↑)
TranceptEVE L	0.920
GEMME	0.919
EVE	0.917
SSEmb	0.893
ESM-1b	0.892

Avg. AUC is computed as the Area Under the ROC Curve averaged across genes.

Table 4 | Using the SSEmb embeddings to study protein-protein interactions

Model	PR-AUC (↑)				
	Test set (70%)	Test set (homology)	Test set (topology)	Test set (none)	Test set (all)
SSEmb downstream	0.684	0.651	0.672	0.571	0.642
Handcrafted features baseline	0.596	0.567	0.568	0.432	0.537
ScanNet	0.732	0.712	0.735	0.605	0.694

The results show the PR-AUC for our supervised downstream model compared to ScanNet and a baseline model across five different test sets. All training and test sets as well as performance metrics for ScanNet and the handcrafted-features baseline model are from[74].

Ablation Study

- Structure-based attention masking and fine-tuning reduce MSA depth sensitivity
- Removing structure information hurts low-depth MSA performance
- Even ablated SSEmb outperforms baseline MSA Transformer

Model	Ablation			Spearman ρ_S by MSA depth (↑)				
	GNN	MSA mask	Fine-tuning	Low	Medium	High	All	$\Delta_{\text{High–Low}}$
SSEmb	✓	✓	✓	**0.446**	0.440	**0.503**	0.454	+0.057
	✓	✓	×	0.435	0.445	0.498	0.453	+0.063
	✓	×	✓	0.438	0.441	0.502	0.452	+0.064
	✓	×	×	0.427	0.439	0.492	0.447	+0.065
	×	✓	✓	0.445	0.458	0.498	0.463	+0.053
	×	✓	×	0.354	0.384	0.444	0.389	+0.090
	×	×	✓	0.434	**0.464**	0.503	**0.465**	+0.069
	×	×	×	0.419	0.461	0.498	0.459	+0.079

Supplementary Table 1. Ablation study of SSEmb performance on the originally released ProteinGym substitution benchmark grouped by UniProt ID and segmented by MSA depth. Ablation of 'GNN' refers to an SSEmb version using only the MSA-processing transformer. Ablation of 'MSA mask' refers to an SSEmb version where the row attention in the MSA-processing transformer is not masked using structure information. Ablation of 'Fine-tuning' refers to an SSEmb version where the pre-trained parameters from the original MSA Transformer are kept fixed. All models were trained using early stopping as assessed by mean correlation performance on the MAVE validation set. Assays from the SSEmb validation set have been excluded from the original data set. Reported SSEmb performance differs slightly from the results presented in Table 1, Table 2, Table 3 and Fig. 2. because a fixed MSA subsample was used for validation and testing across all models in order to ensure fair comparison. Low: $N_{\text{eff}}/L < 1$, Medium: $N_{\text{eff}}/L < 100$, High: $N_{\text{eff}}/L > 100$ (***Notin et al., 2022***) [1].

Perspectives & Limitations

- Generalizable framework – can integrate alternative sequence/structure models
- Limitations:
  - Relies on both MSA and structure inputs
  - May perform poorly on intrinsically disordered proteins
  - Training data covers limited sequence-structure space

Conclusion

- SSEmb integrates sequence and structure for robust variant effect prediction
- Improves accuracy, especially for low MSA depth proteins
- Embeddings are versatile for multiple downstream tasks
- Potential for broader applications in computational biology

Teşekkürler

The End.