ESMRank: a Learn-to-Rank Approach for protein variant effect prediction

Riccardo Arnese, Gennaro Gambardella

COZYME

Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale
Università degli Studi di Napoli Federico II

TELETHON INSTITUTE OF GENETICS AND MEDICINE
igem

INTRODUCTION

MOTIVATION

- Predicting protein mutation effects is crucial for disease **variant interpretation** and **protein engineering**
- **MAVEs** provide large-scale data but assay heterogeneity limits cross-experiment comparability

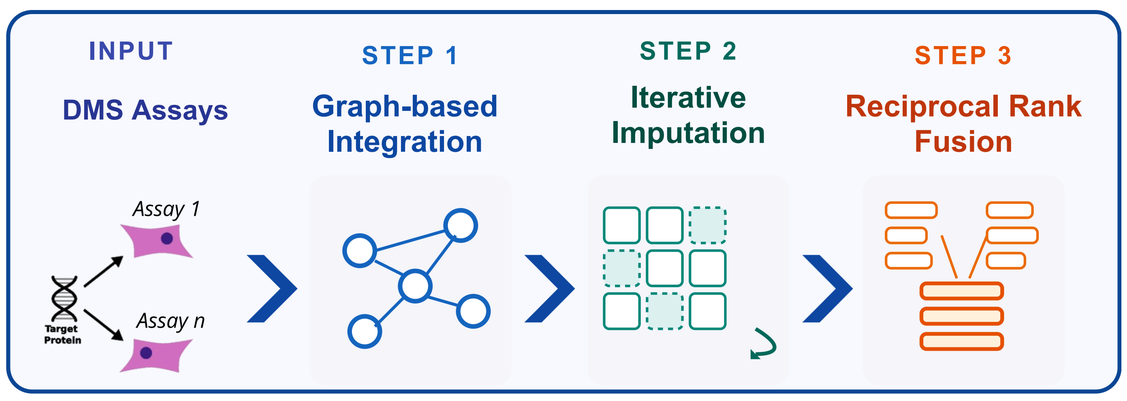
KEY INSIGHTS

- **Relative ordering** of variant effects is more reproducible across assays than absolute magnitudes
- **Partial overlap** among independent experiments encodes a **transferable ordinal signal.**

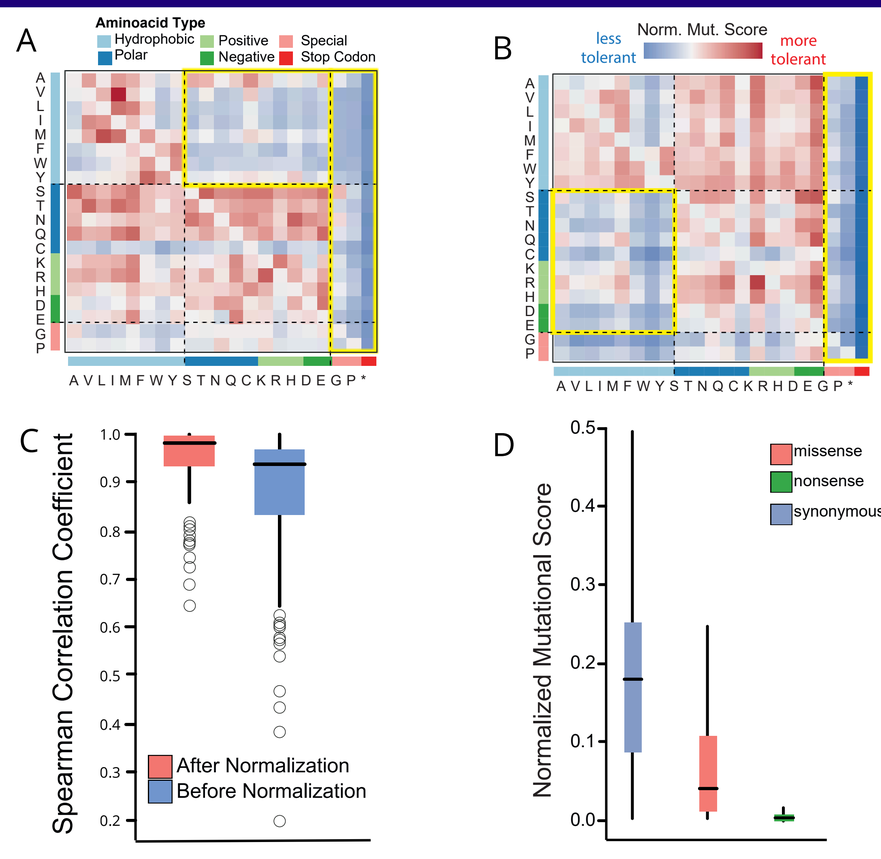
OUR APPROACH

- **Variant soundness:** overlap-aware integration via Reciprocal Rank Fusion (RRF) allows assay-agnostic mutational tolerance scoring
- **ESMRank:** learn-to-rank model combining ESM2 and biophysical descriptors, trained on ~2M variants from MAVEdb

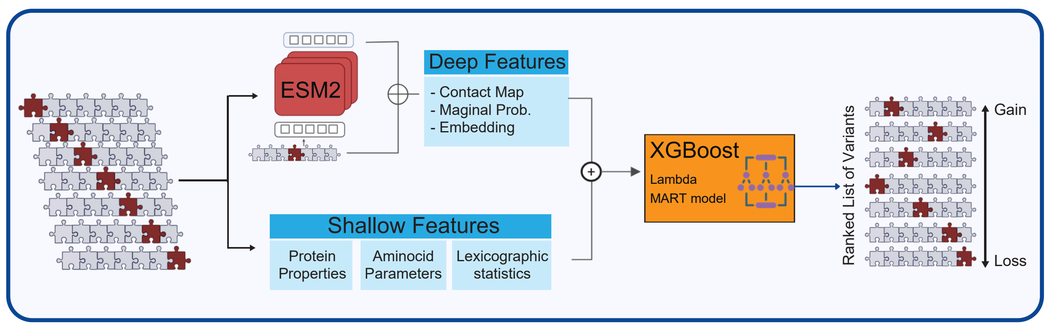
DMS DATA INTEGRATION



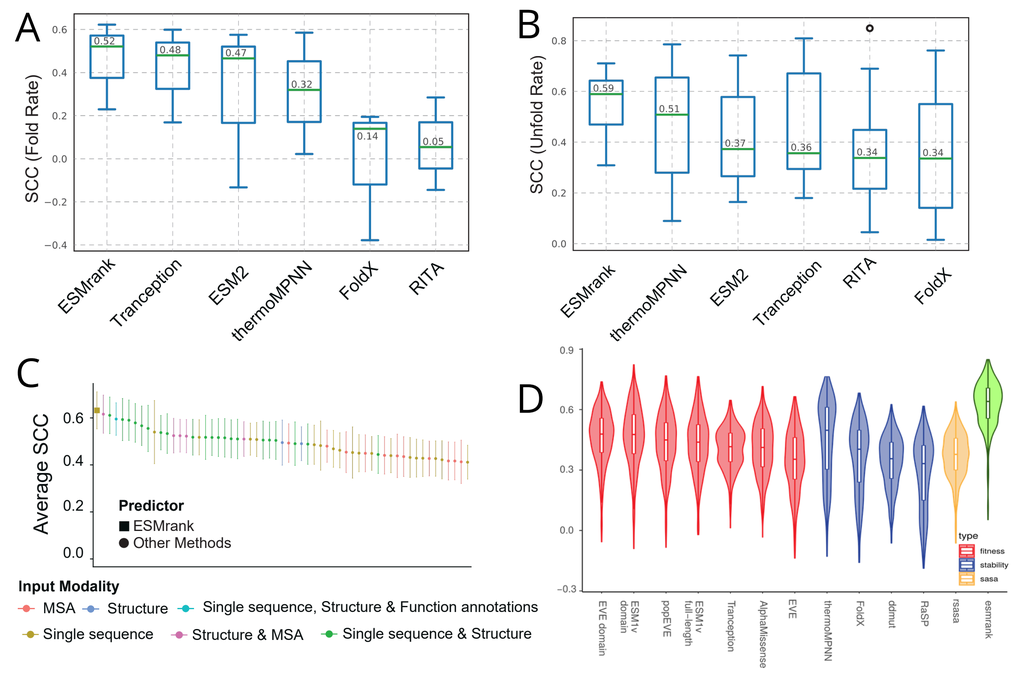
- Soundness profiles show **increased cross-experiment coherence**: normalized scores shows higher correlation with DMS assays than before integration (**C**)
- **Integrated landscape** separates synonymous, missense, and nonsense variants onto a common scale (**D**)
- **Core substitution matrix** recapitulates classical stability constraints: hydrophobic→polar/charged strongly deleterious (**A**)
- **Surface residues** show distinct electrostatic sensitivities, consistent with disruption of interaction interfaces (**B**)



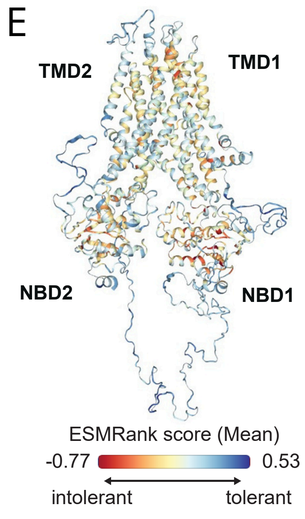
ESMRANK MODEL BENCHMARK



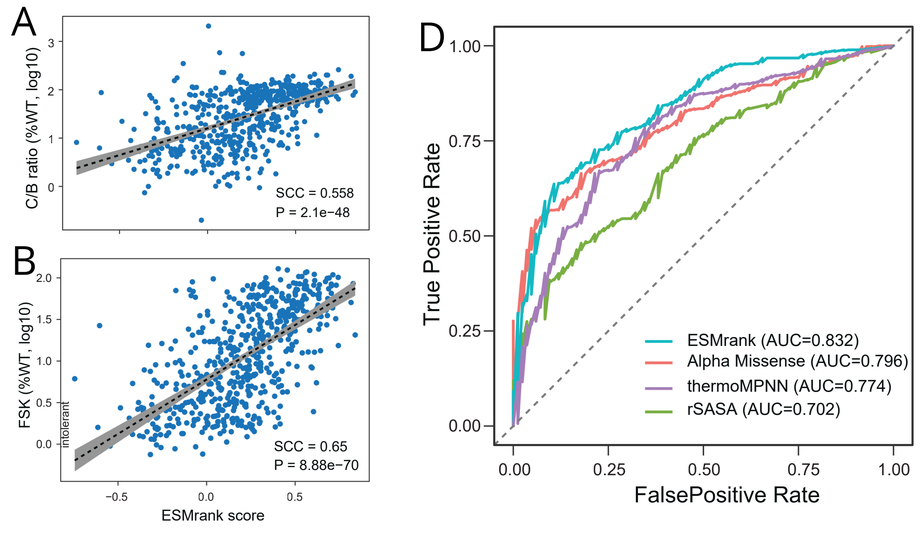
- **VariBench folding kinetics**: ESMRank outperforms all competitors on both folding (A) and unfolding (B) rate prediction (median ρ = 0.52 and 0.59).
- **ProteinGym stability leaderboard**: ESMRank ranks first among all methods (mean SCC = 0.63), including structure-based and MSA-based predictors, using sequence alone (C).
- **Human Domainome benchmark:** ESMRank achieves median ρ = 0.62, significanlly outperforming all other stability/fitness predictors.



CFTR CHARACTERIZATION



- **ESMRank scores mapped onto CFTR structure** (E). Intolerant regions (red) localize to transmembrane helices and interdomain interfaces.
- ESMRank correlates with (**A**) **maturation efficiency** (ρ = 0.56) and (**B**) **basal channel function** (ρ = 0.65) across 585 missense variants — without CFTR-specific training.
- **Classification of therapeutically responsive variants**: ESMRank achieves AUC = 0.83 (**D**), outperforming AlphaMissense, ThermoMPNN, and rSASA. Higher ESMRank scores predict better response to modulator therapy (ELX+TEZ+IVA).



CONCLUSIONS

- Overlap-aware integration of 2M+ variants across 596 proteins recovers a coherent axis of mutational constraint enriched for structural stability determinants
- ESMRank achieves state-of-the-art on stability benchmarks from sequence alone, under strict protein-level partitioning
- Without clinical supervision, predicted constraint stratifies disease mechanisms (GOF→HI) and predicts CFTR drug response (AUC = 0.83), linking sequence constraint to therapeutic tractability

REFERENCES

1. Beltran, A., Jiang, X., Shen, Y. et al. **Site-saturation mutagenesis of 500 human protein domains.** Nature 637, 885–894 (2025).
2. Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Alexander Rives et al. **Evolutionary-scale prediction of atomic-level protein structure with a language mode**l. Science Vol.379 No. 6637 (2023)
3. Notin, P., Kollasch, M., Franceschi, D., Gal, Y., & Marks, D. et al. (2023). **ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design**. In Advances in Neural Information Processing Systems (pp. 64331–64379). Curran Associates, Inc..
4. Rubin, A.F., Stone, J., Bianchi, A.H. et al. **MaveDB 2024: a curated community database with over seven million variant effects from multiplexed functional assays**. Genome Biol 26, 13 (2025).