Variant Effect Predictors Show Systematic Bias in Intrinsically Disordered Regions

Norbert Deutsch, Zsuzsanna Dosztányi

MTA-ELTE Momentum Bioinformatics Research Group

Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary

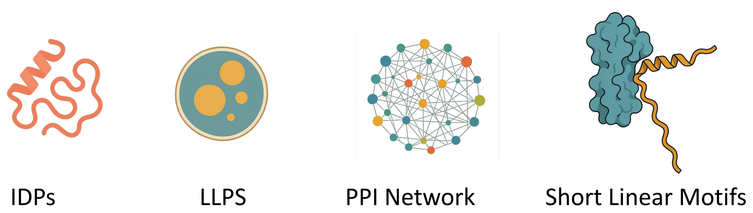
Introduction

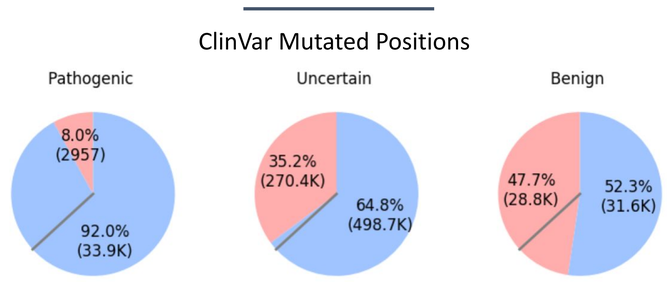
The Missing Piece: VEP benchmarking is structurally incomplete. It heavily prioritizes folded domains, largely ignoring Intrinsically Disordered Regions (IDRs) that comprise half the human proteome.[1,2]

Distinct Biology: IDRs drive critical cellular functions via Short Linear Motifs (SLiMs) and Liquid-Liquid Phase Separation (LLPS). Their sequence constraints are highly diverse, ranging from strict local conservation to broader molecular grammar.[3]

The Clinical Bias: Clinical databases are severely skewed. IDR mutations are rarely annotated as pathogenic; instead, they are overwhelmingly classified as VUS or benign, creating a massive diagnostic blindspot.[2,4,5]

The Experimental Challenge: To bypass clinical limits, the field increasingly relies on MAVEs. However, it remains unexplored whether these foundational experimental datasets inherently share the same structural bias against IDRs.[6]





Objectives

Assess Structural Bias: Evaluate VEP performance in IDRs to trace predictive gaps back to the severe lack of IDRs in clinical and experimental data.

Understand Predictive Success: Determine if VEP accuracy in IDRs relies on the successful recognition of local functional contexts (e.g., SLiMs).

Characterize Mutational Landscapes: Analyze MAVE data to contrast mutation distributions in folded vs. disordered regions, quantifying localized functional clustering.

Results

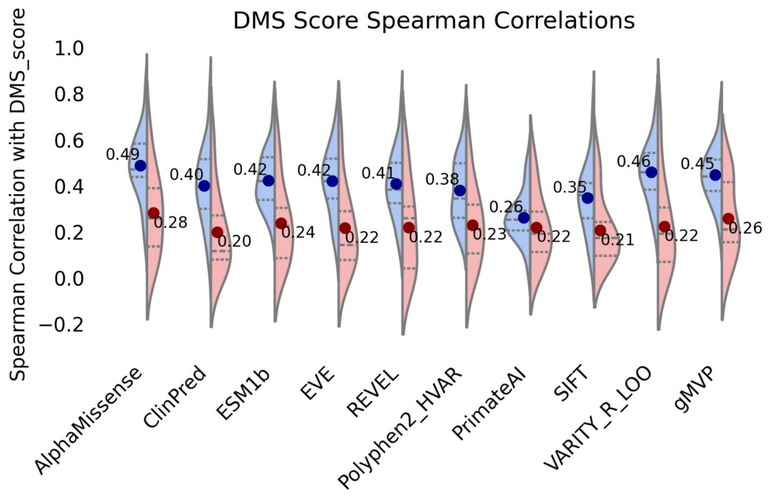
1. Structural & Dataset-Specific Biases

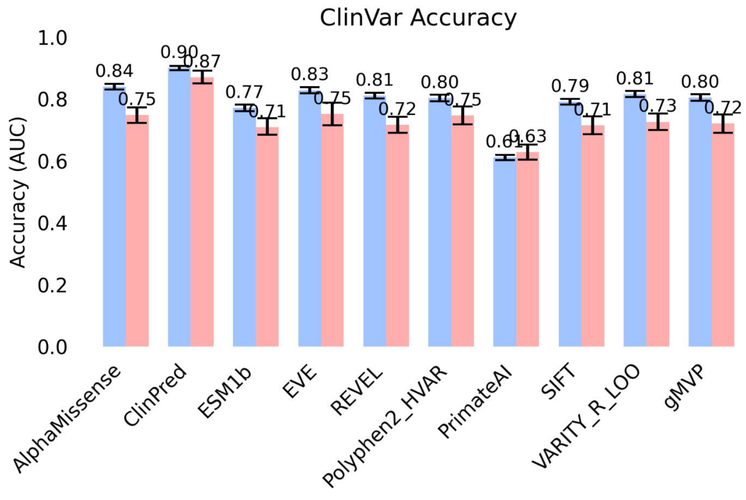
Global Underperformance

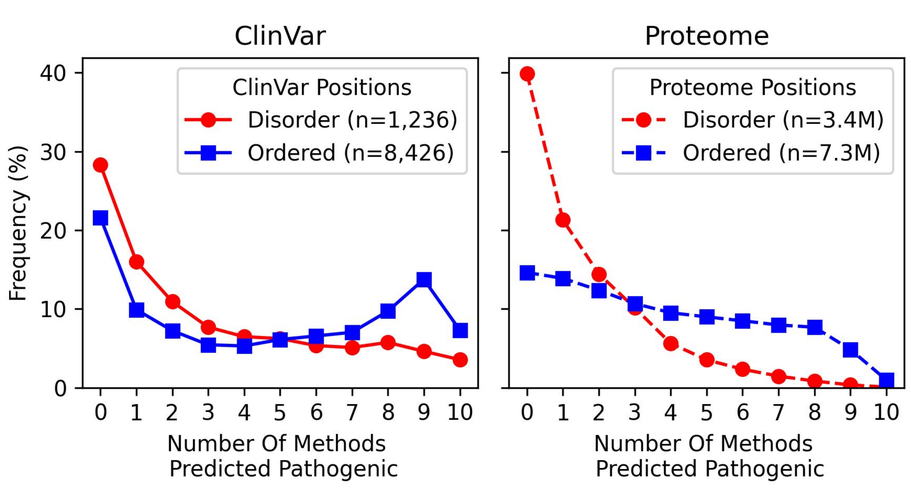
VEPs show significantly reduced accuracy in IDRs compared to ordered regions, a bias evident across both ClinVar classifications and deep mutational scanning (DMS) correlations.

The Overfitting Problem

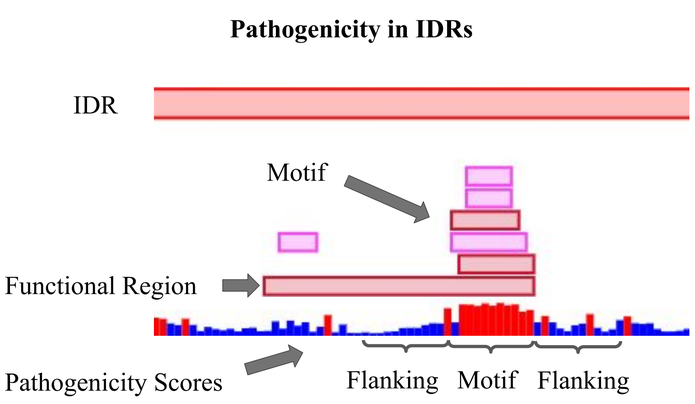
High inter-predictor correlation on known ClinVar variants collapses when applied proteome-wide to IDRs. Current VEPs overfit to folded-domain-heavy data, poorly modeling IDR functionality.







2. The Functional Signal in the "Noise"



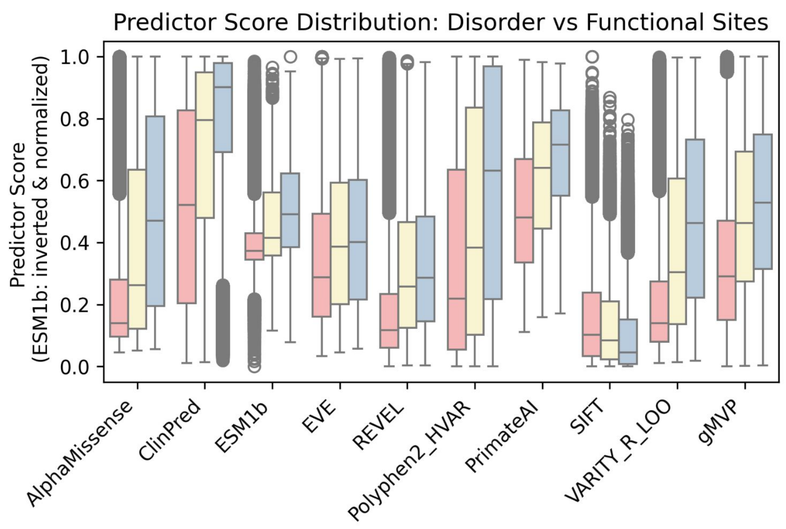
- *Pathogenic variants delineate functional sites in IDRs, with enrichment in SLiMs*
- *Predictive success in IDRs relies on capturing local functional signals.*
- *Integrating IDR-specific annotations is essential for building unbiased predictors.*

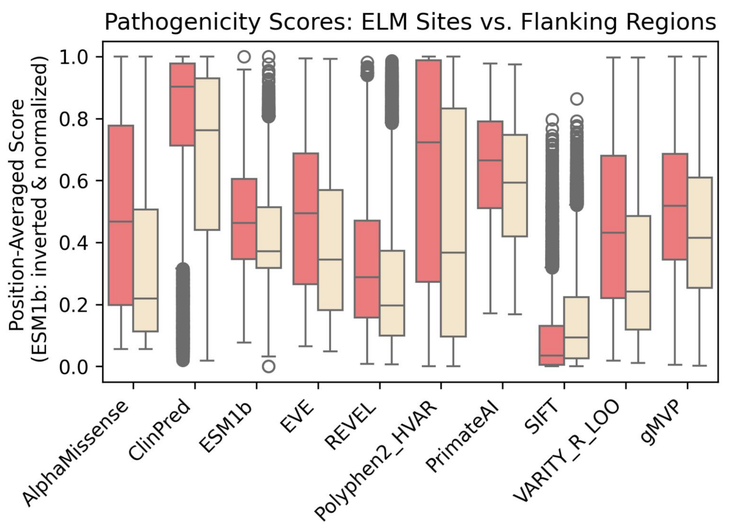
Local Context Matters

Global metrics mask critical performance differences. Top-tier models succeed in IDRs by distinguishing functional sites from the non-functional background, relying on local evolutionary constraints.

Capturing Motifs

Functional motifs in IDRs are conserved as isolated evolutionary islands. Superior VEPs effectively detect this localized constraint, producing sharp, *"island-like pathogenicity"* patterns in the sequence.





3. MAVEs Reveal Structural Bias and Pathogenic Clustering

Experimental Structural Bias

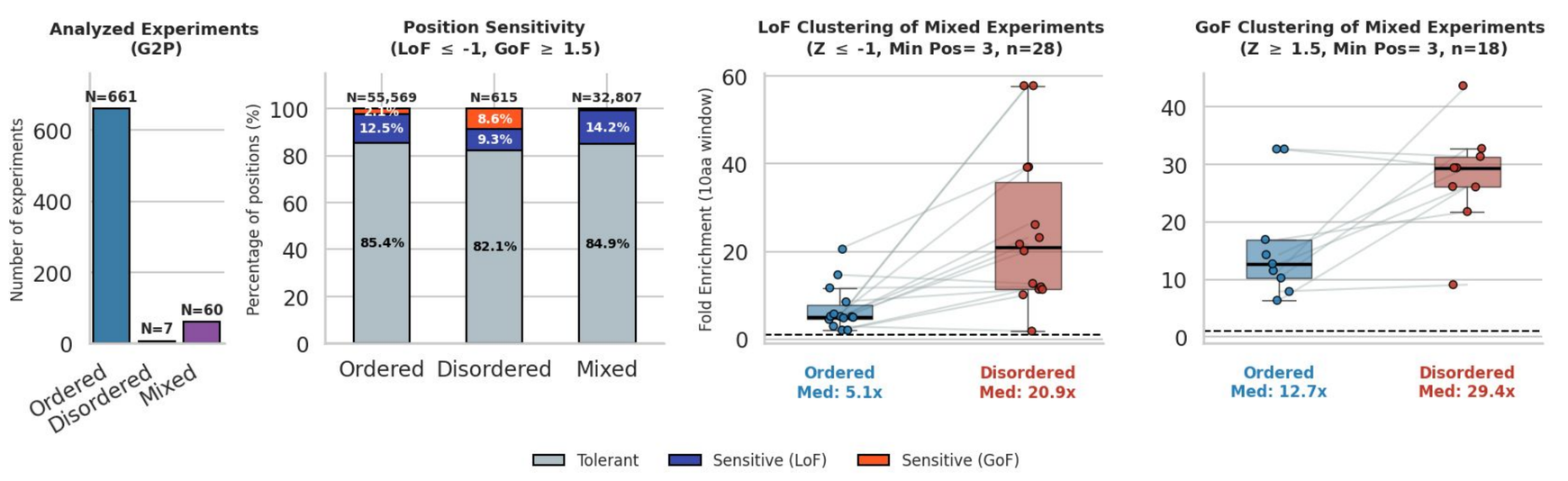
MAVE experiments predominantly target structured domains. Filtering for robust IDR coverage leaves a severely limited dataset, exposing a massive experimental blindspot.

Preferential Enrichment of GoF in IDRs

Within the available experimental landscape, intrinsically disordered regions exhibit a higher relative proportion of Gain-of-Function mutations compared to ordered domains.

Localized Clustering

Pathogenic mutations in IDRs are not uniformly distributed. They tightly concentrate into narrow, localized functional hotspots (e.g., SLiMs), directly mirroring the predictive peak signals captured by top-tier VEPs.



Conclusions

Functional Signals Drive Accuracy: Predictive success in IDRs relies on strong functional descriptions, enabling top-tier VEPs to capture localized pathogenic elements (e.g., SLiMs).

Need for IDR-Specific Assessment: To overcome the systematic bias toward structured domains, future benchmarking must explicitly evaluate predictors within IDR-specific functional contexts.

References

1. Fawzy & Marsh, *PLoS Comput Biol* **2025** – Assessing variant effect predictors and disease mechanisms in intrinsically disordered proteins.
2. Holehouse & Kragelund, *Nat Rev Mol Cell Biol* **2024** – Functional roles of intrinsically disordered regions.
3. Deutsch et al., *iScience* **2026** – Pathogenic variations illuminate functional constraints in intrinsically disordered proteins.
4. Rastogi et al., *Hum Genet* **2025** – Benchmarking missense variant effect predictors on clinical data.
5. Deutsch et al., *Protein Sci* **2023** – DisCanVis: visualization of structural and functional annotations in disordered proteins.
6. Notin et al., *bioRxiv* **2023** – ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction.

Acknowledgements

- HORIZON WIDERA 2023
  - Grant 101160233 "**IDP2Biomed**"
- HORIZON-MSCA-2023-SE
  - Grant 101182949 "IDPfun2"
- EKÖP-KDP-24 University Excellence Scholarship
  - Grant KDP-24-II-ELTE-75

This poster is based upon work from COST Action **ML4NGP**, CA21160, supported by COST (European Cooperation in Science and Technology)