Explainable Deep Generative Models, Ancestral Fragments, and Murky Regions of the Protein Structure Universe

Eli J. Draizen, Cameron Mura, Philip E. Bourne

University of Virginia | DATA SCIENCE

edraizen@gmail.com

edraizen.github.io / bournelab.org

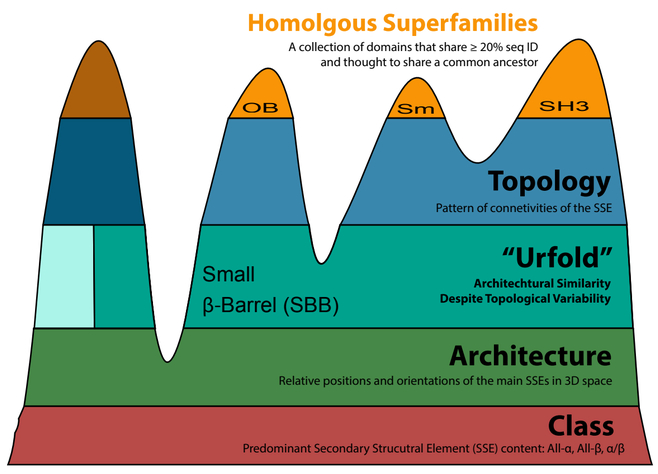
EliDraizen / BourneLabUVa

edraizen / bouralab

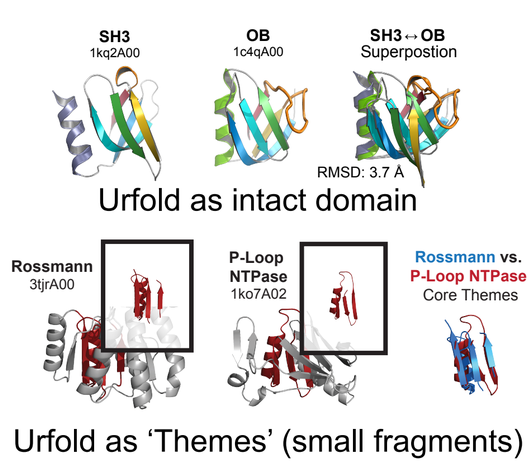
Abstract

Modern proteins did not arise abruptly, as singular events, but rather over the course of at least 3.5 billion years of evolution. Can machine learning teach us how this occurred? The molecular evolutionary processes that yielded the intricate three-dimensional (3D) structures of proteins involve duplication, recombination and mutation of genetic elements, corresponding to short peptide fragments. Identifying and elucidating these ancestral fragments is crucial to deciphering the interrelationships amongst proteins, as well as how evolution acts upon protein sequences, structures & functions. Traditionally, structural fragments have been found using sequence and 3D structural alignment, but that becomes challenging when proteins have undergone extensive permutations—allowing two proteins to share a common architecture, though their topologies may drastically differ (a phenomenon termed the *Urfold [1]*). **We have designed a new framework to identify compact, potentially-discontinuous peptide fragments by combining (i) deep generative models of protein superfamilies [2] with (ii) layer-wise relevance propagation (LRP [3]) to identify atoms of great relevance in creating the embeddings obtained via an all$_{superfamilies}$ x all$_{domains}$ analysis.** Our approach recapitulates known relationships amongst the evolutionarily ancient small β-barrels (e.g. SH3 and OB folds [4]) and P-loop–containing proteins (e.g. Rossmann and P-loop NTPases [5]), previously established via manual analysis. Because of the generality of our deep model's approach, **we anticipate that it can enable the discovery of new ancestral peptides.** In a sense, our framework uses LRP as an 'explainable AI' approach, in conjunction with a recent deep generative model of protein structure (termed *DeepUrfold*), in order to leverage decades worth of structural biology knowledge to decipher the underlying molecular bases for protein structural relationships—including those which are exceedingly remote, yet can be discovered via deep learning.

The *Urfold*: 3D Architectural Similarty Despite Topological Variability



CATH Protein Universe Organization



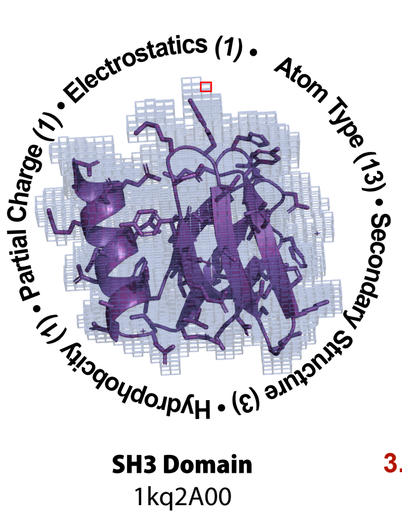
The SH3 & OB superfamilies are part of the small β-barrel urfold [4] while the Rossmann and P-loop NTPases are members of the P-Loop Binding Urfold [5]. Most members of an Urfold have similar functions (e.g. bind similar ligands). If viewed in terms of CATH [6], the Urfold would sit in between the Architecture & Topology strata.

All$_{domains}$ vs. All$_{superfamillies}$ Methodology

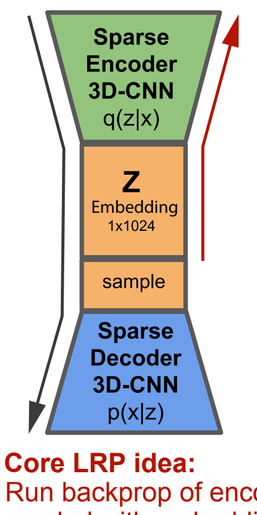
1. Energy-minimize, featurize, and voxelize 3D structures of protein domains

2. Subject domain to given VAE Model to obtain likelihood it came from superfamily VAE

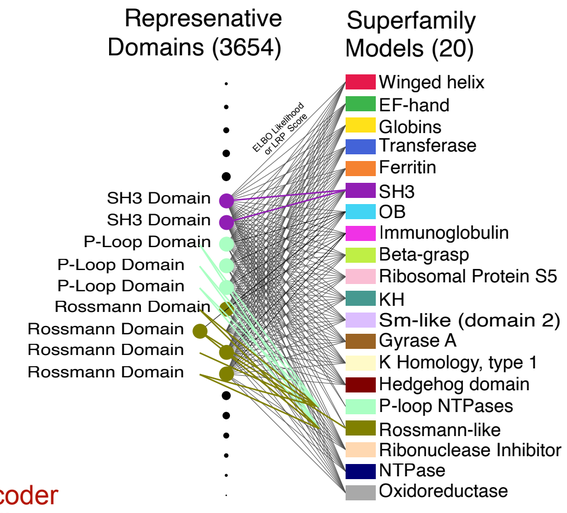
4. Cross-Model: Repeat for all representative domains through all superfamily models



SH3 Domain
1kq2A00



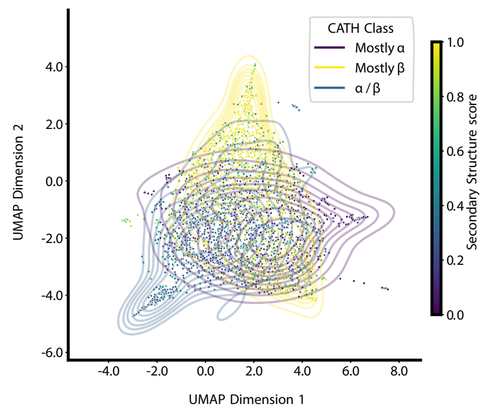
3. Core LRP idea:
Run backprop of encoder seeded with embedding



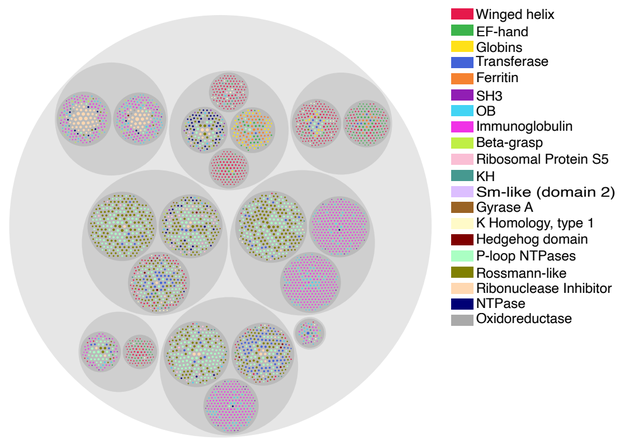
DeepUrfold Uncovers Distant Relationships

Single-Model Embeddings
e.g. SH3 domain through SH3 VAE

Cross-Model Communities
e.g. cluster likelihoods of domain X under VAE M_i and M_j for superfamilies i and j



UMAP finds SSEs as an important dimension

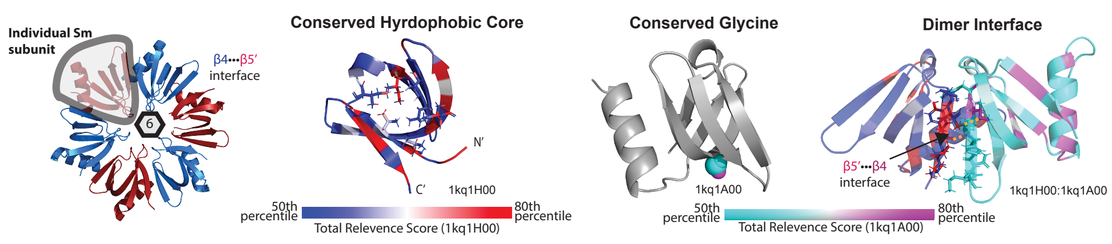


Stochastic Block Modelling (SBM) reveals intermixing between 'true' CATH superfamilies

DeepUrfold-explain Identifies Functionally Important Residues

Single-Model LRP

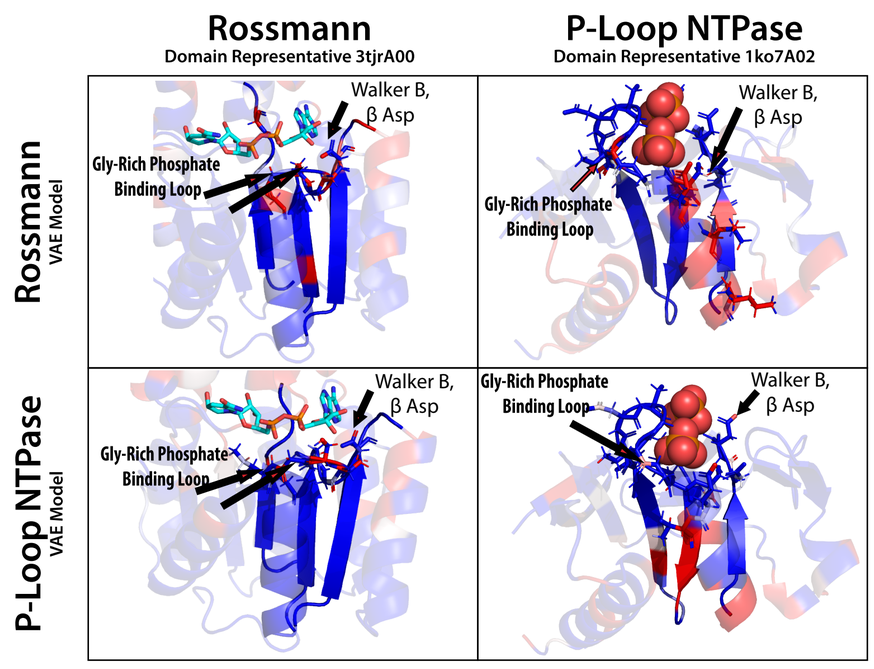
SH3 Domain Representatives subjected to SH3 VAE Model



Cross-Model LRP

Rossmann ↔ P-Loop NTPases

Representative Domains subjected to both VAE Models



LRP scores for representative domains are higher in similar 3D locations important for ligand binding, regardless of which VAE it was subjected to.

Conclusions

- **LRP correctly selects structurally important and conserved atoms** showing that the model is learning superfamily-specific features.
- Domains from the same urfold have functionally important residues in the same 3D locations, regardless of which VAE it was subject to, suggesting **the VAEs are topologically-agnostic and can find Urfolds**
- In the future, we plan to **identify more common fragments and ancestral peptides** by aligning and clustering 'important' regions from the cross-model fragments.

We thank Stella Veretnik, Luis Felipe R. Murillio and John Ready for support as well as UVA School of Data Science Presidential Fellows Program and NSF Career award MCB-1350957 for funding.

[1] Mura, Veretnik, Bourne. *Protein Science* (2019) https://doi.org/10.1002/pro.3742

[2] Draizen *et al. biorXiv* (2022) https://doi.org/10.1101/2022.07.29.501943

[3] Binder et al. *arXiv* (2016) https://doi.org/10.48550/arXiv.1604.00825

[4] Youkharibache *et al. Structure* (2019) https://doi.org/10.1016/j.str.2018.09.012

[5] Longo *et al. eLife* (2020) https://doi.org/10.7554/eLife.64415

[6] Sillitoe *et al. Nucleic Acids Research* (2021) https://doi.org/10.1093/nar/gkaa1079