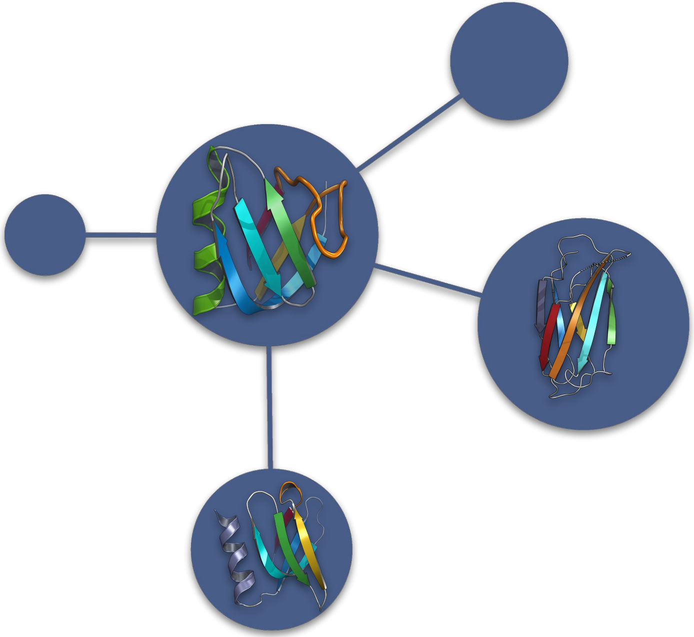


Uncovering Distant Protein Relationships with Deep Generative Models

Thesis Project Proposal
December 16th, 2021

Eli Draizen
Phil Bourne's Lab
http://bournelab.org
Biomedical Engineering
University of Virginia

Motivation

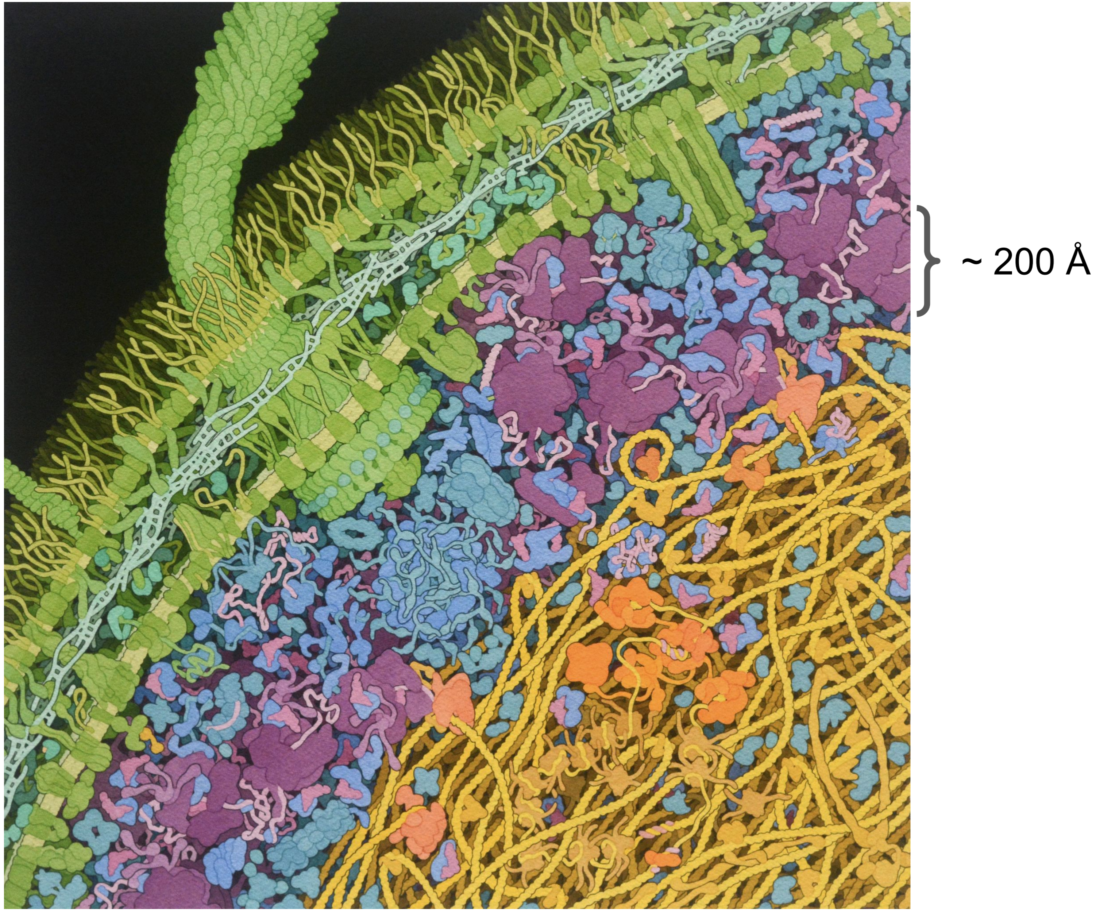


Image from David Goodsell

- Proteins mediate many biological functions and are crucial to understanding biological pathways.
- If we know their 3D structures, it will help identify:
  - Protein interactions
  - Drug targets
  - Interrelationships (evolution)
- We want to be able to **find distant relationships** between proteins to understand **protein evolution,** and **annotate functions** of a new protein if it is related to a known protein with known function

Background

- Traditionally, **hierarchical classification** systems are used to understanding these protein relationships
- **Discrete clustering** of proteins is an important **first step** of organizing the protein universe
- However, this approach breaks down for **distantly related proteins** that **don't fit into discrete bins**, highlighting the continuity of protein structure/fold space

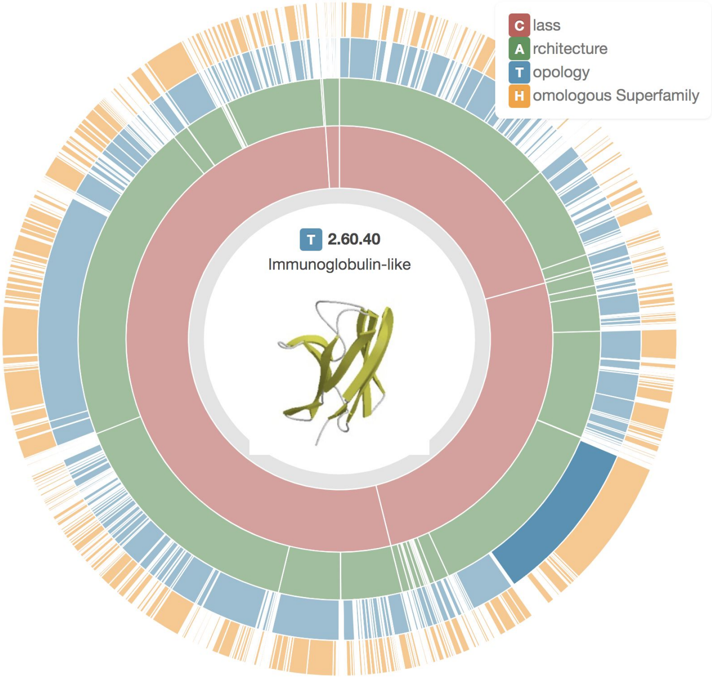
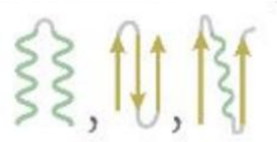


Image from cathdb.info

A Hierarchy of Structural Levels

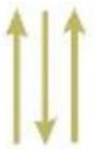
1. Class



Types of 2° structure elements (SSE)

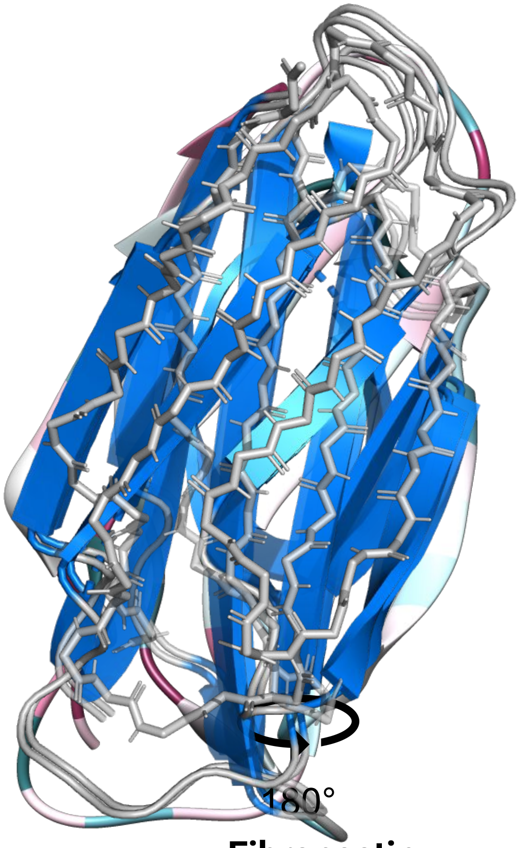
E.g. Mostly Beta (2)

2. Architecture



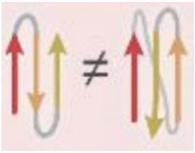
3D arrangements of SSEs

E.g. Sandwich (2.60)



Fibronectin

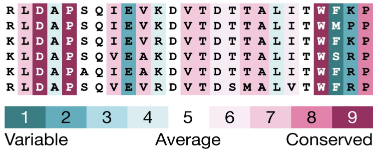
3. Topology



3D arrangement **AND** pattern of connectivities between SSEs

E.g. Immunoglobulin-like (2.60.40)

4. Homologous Superfamily



```
R L D A P S Q I E V K D V T D T T A L I T W F K P
R L D A P S Q I E V K D V T D T T A L I T W M P P
K L D A P S Q I E V R D V T D T T A L I T W F K P
K L D A P S Q V E A K D V T D T T A L I T W S R P
K L D A P A Q I E A K D V T D T T A L I T W F R P
R L D A P S Q V E V R D V T D S M A L V T W F R P
```

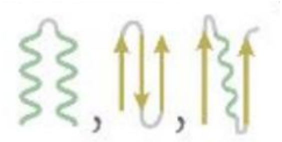
Evolutionary relationships via sequence, ≥25% sequence identity

E.g. Immunoglobulins (2.60.40.10)

PDB ID: 1TEN

A Hierarchy of Structural Levels

1. Class



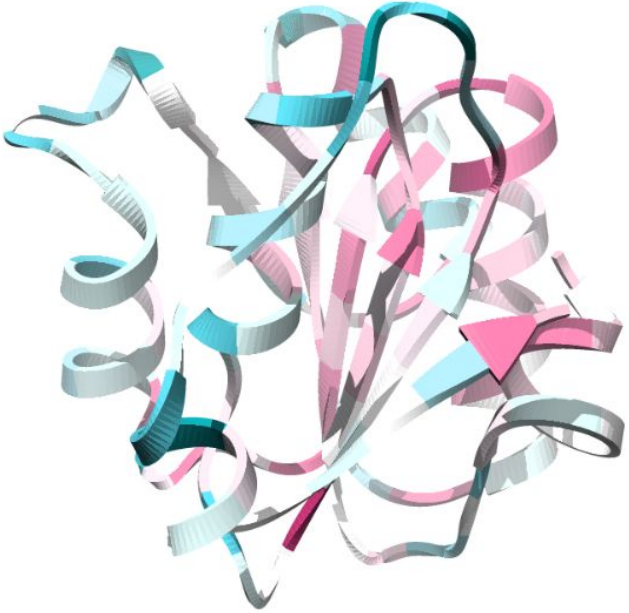
Types of 2° structure elements (SSE)

E.g. Alpha/Beta (3)

2. Architecture

3D arrangements of SSEs

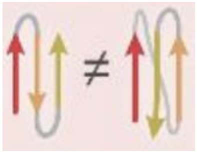
E.g. 3-Layer Sandwich (3.40)



Face 1 Face 2 Face 3

Alcohol Dehydrogenase

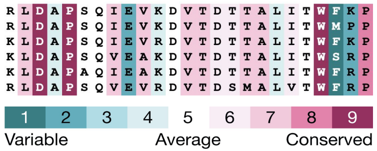
3. Topology



3D arrangement **AND** pattern of connectivities between SSEs

E.g. Rossman fold (3.40.50)

4. Homologous Superfamily



```
R L D A P S Q I E V K D V T D T T A L I T W F K P
R L D A P S Q I E V K D V T D T T A L I T W M P P
K L D A P S Q I E V R D V T D T T A L I T W F K P
K L D A P S Q V E A K D V T D T T A L I T W S R P
K L D A P A Q I E A K D V T D T T A L I T W F R P
R L D A P S Q V E V R D V T D S M A L V T W F R P
```

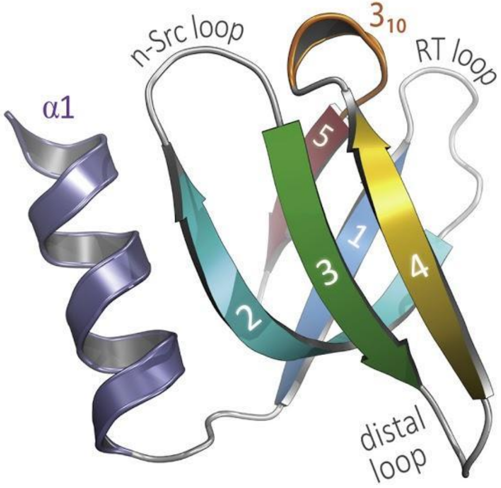
Evolutionary relationships via sequence, ≥25% sequence identity

E.g. NAD(P)-binding Rossmann-like Domain (3.40.50.720)

PDB ID: 1cdoB

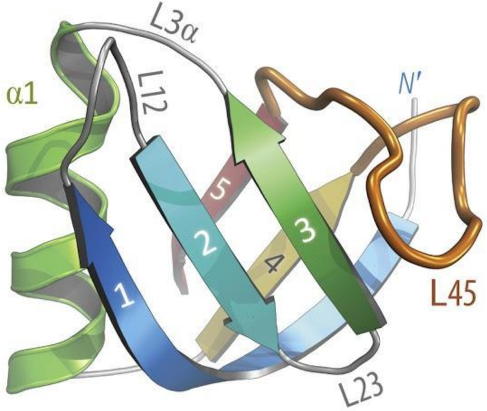
Small β-Barrels (SBBs) Exhibit *Architectural Similarity Despite Topological Variability*

SH3/Sm
1kq2



2.30.30.100

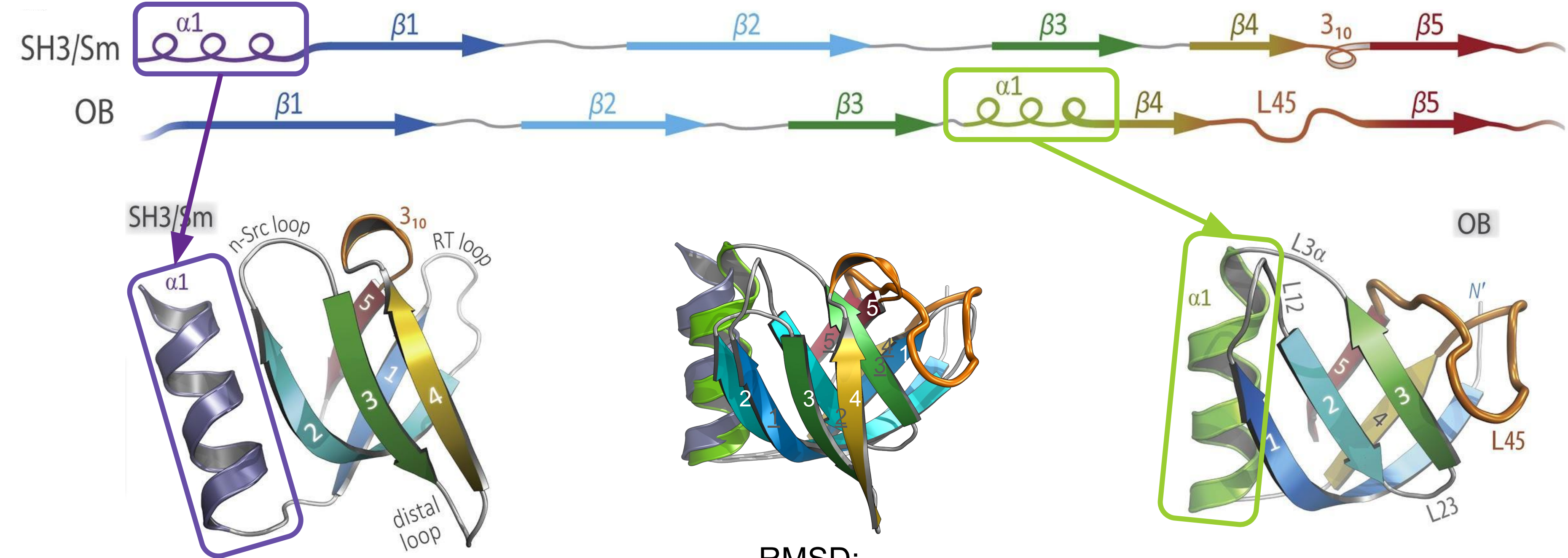
OB
1c4q



2.40.50.70

Youkharibache, Veretnik, et al. *Structure* (2019); https://doi.org/10.1016/j.str.2018.09.012

Small β-Barrels (SBBs) Exhibit *Architectural Similarity Despite Topological Variability*



2.30.30.100

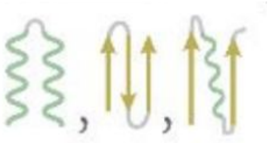
RMSD:
3.7 Å

2.40.50.70

Youkharibache, Veretnik, et al. *Structure* (2019); https://doi.org/10.1016/j.str.2018.09.012

Possible new entity between Architecture+Topology

1. Class



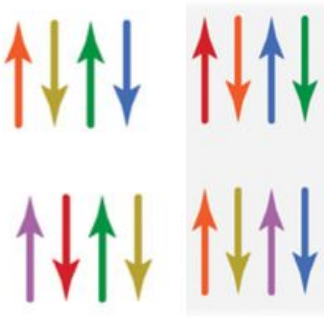
Types of 2° structure elements (SSE)

2. Architecture

3D arrangements of SSE

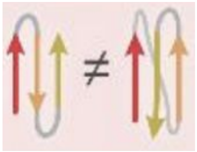
3. *Ur*-fold

Prefix meaning “proto-, primitive, original.” Origin: German.



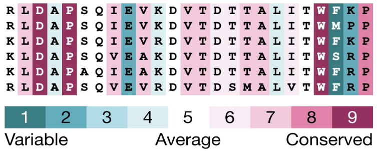
3D architectural similarity despite topological variability

4. Topology



3D arrangement **AND** pattern of connectivities between SSEs

5. Homologous Superfamily



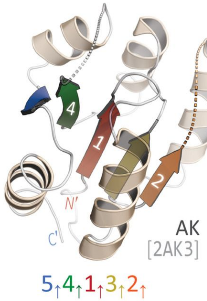
```
R L D A P S Q I E V K D V T D T T A L I T W F K P
R L D A P S Q I E V K D V T D T T A L I T W M P P
K L D A P S Q I E V R D V T D T T A L I T W F K P
K L D A P S Q V E A K D V T D T T A L I T W S R P
K L D A P A Q I E A K D V T D T T A L I T W F R P
R L D A P S Q V E V R D V T D S M A L V T W F R P
```

Evolutionarily relationships via sequence, ≥25% sequence identity

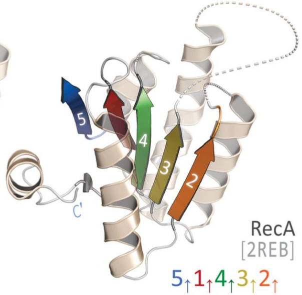
Mura, Veretnik, Bourne. *Prot Sci* (12/2019); https://doi.org/10.1002/pro.3742

Other Potential ‘Urfolds’

P-loop NTPases

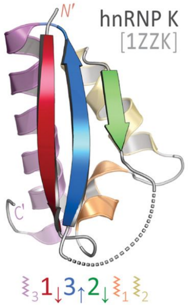


3.40.50.300

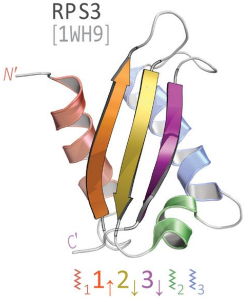


3.30.250.10

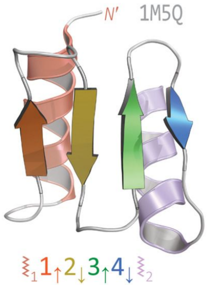
KH Domains



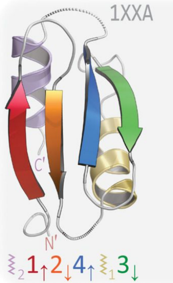
3.30.1370.10



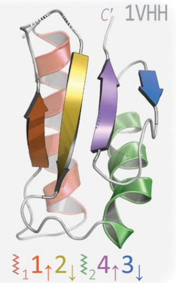
3.30.300.20



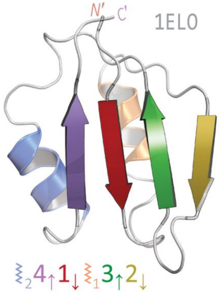
3.30.310.60



3.30.1360.40



3.30.1380.10

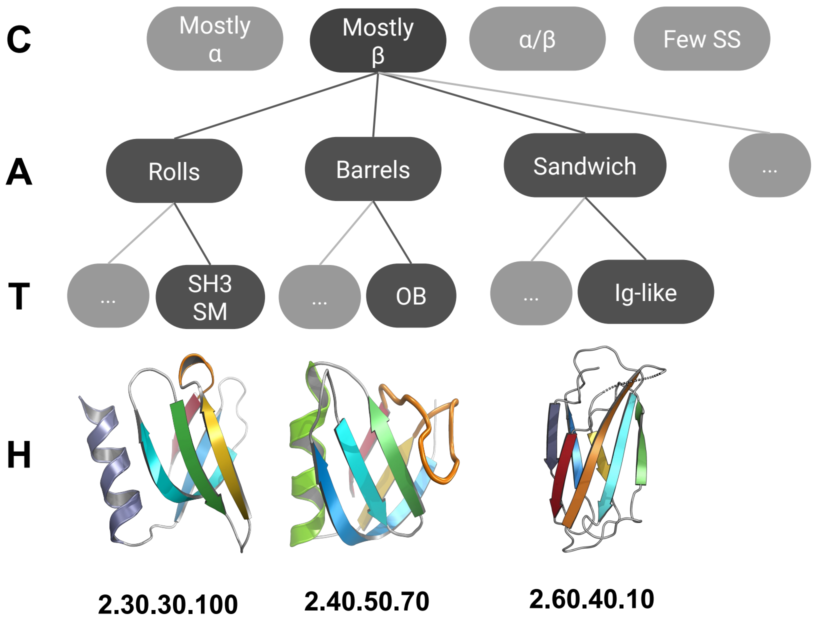


3.30.70.870

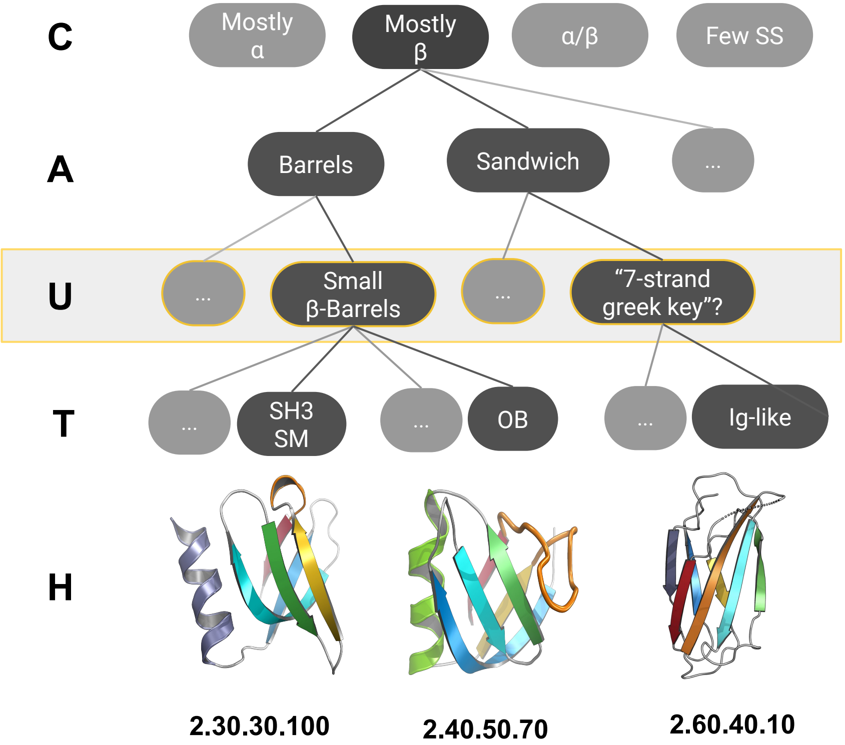
Mura, Veretnik, Bourne. *Prot Sci* (12/2019); https://doi.org/10.1002/pro.3742

A Different View of Clustering Relationships

Current Clustering



Potential Clustering



DeepUrfold

Can we learn local substructures of biophysical properties and geometry that bridge ‘gaps’ in hierarchical classification systems?

Thesis Aims

1. **Create a database of biophysical atomic-level properties, in 3D, for the known protein universe**

   Develop a scalable, reproducible workflow to prepare proteins and calculate atomic properties; intended to be shared as a community resource

2. **Build and interrogate Deep Generative Models to learn superfamily-specific geometries and properties**

   Learn the defining geometries and biophysical properties for different superfamilies, allowing us to assess the Urfold hypothesis

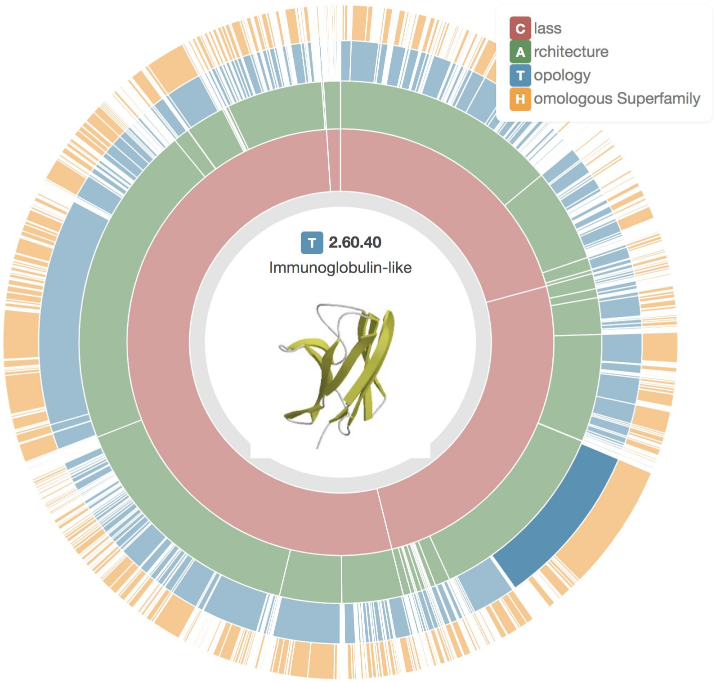
3. **Identify distant evolutionary relationships that bridge protein architectures and topologies that define an Urfold**

   Use Explainable AI techniques to understand model decisions and cluster proteins in light of the continuous nature of fold space

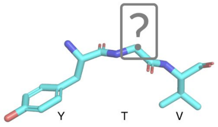
Aim 1

Create a database of biophysical atomic properties in 3D for the known protein universe

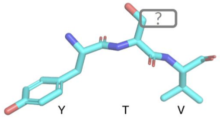
Data Engineering is the first step in machine learning



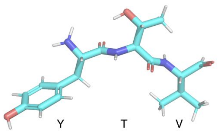
Step 1: Protein Structure Preparation



1. Add missing residues

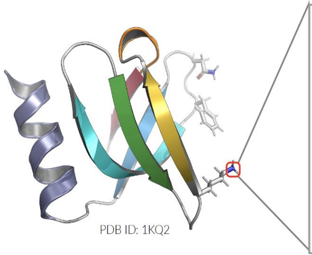


2. Add missing atoms



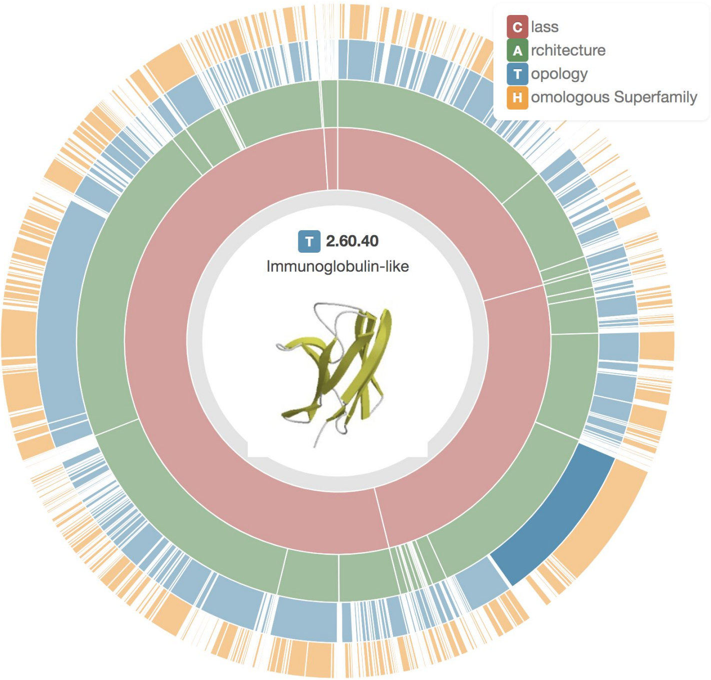
3. Add hydrogens and energy minimize structure

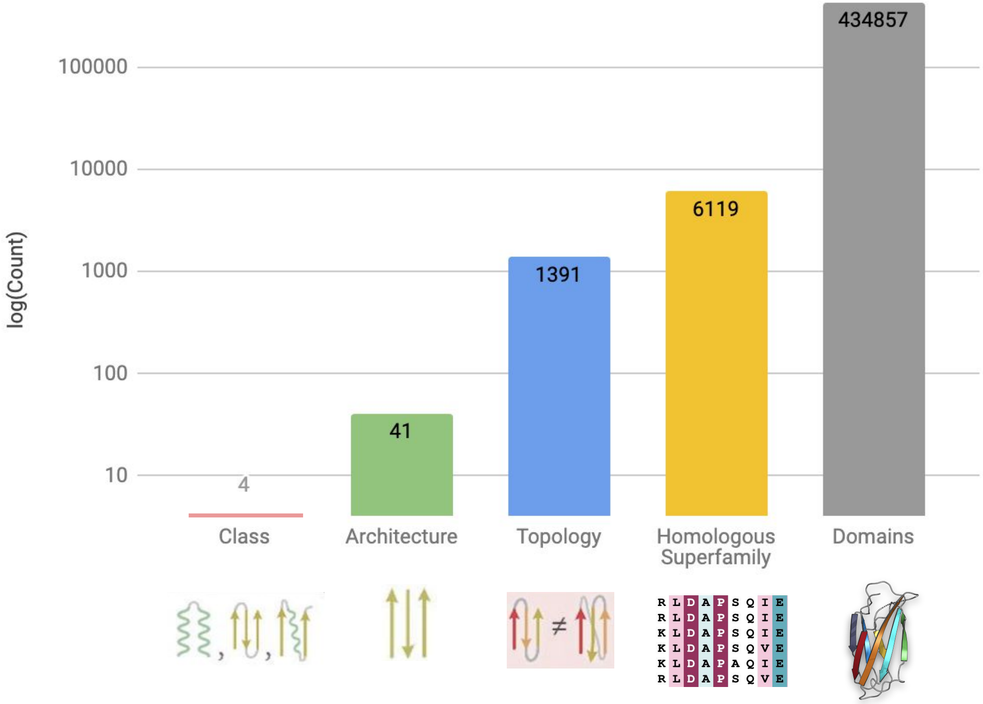
Step 2: Calculate Biophysical Properties



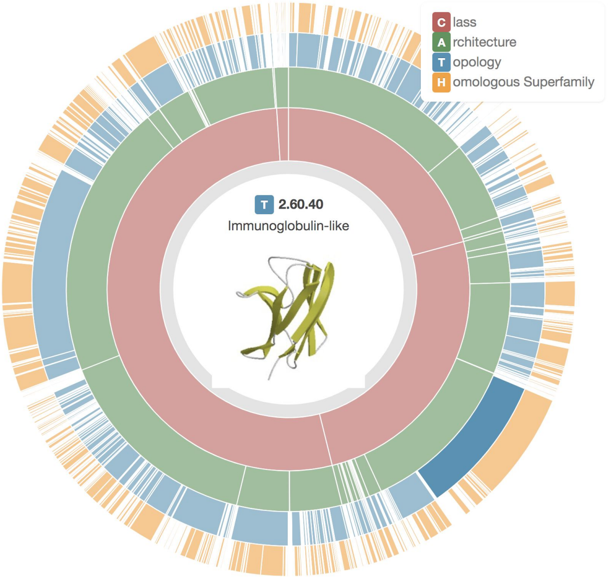
- Atom Type
- Partial Charge + Electrostatics
- Hydrophobicity
- Secondary Structure
- Evolutionary Conservation

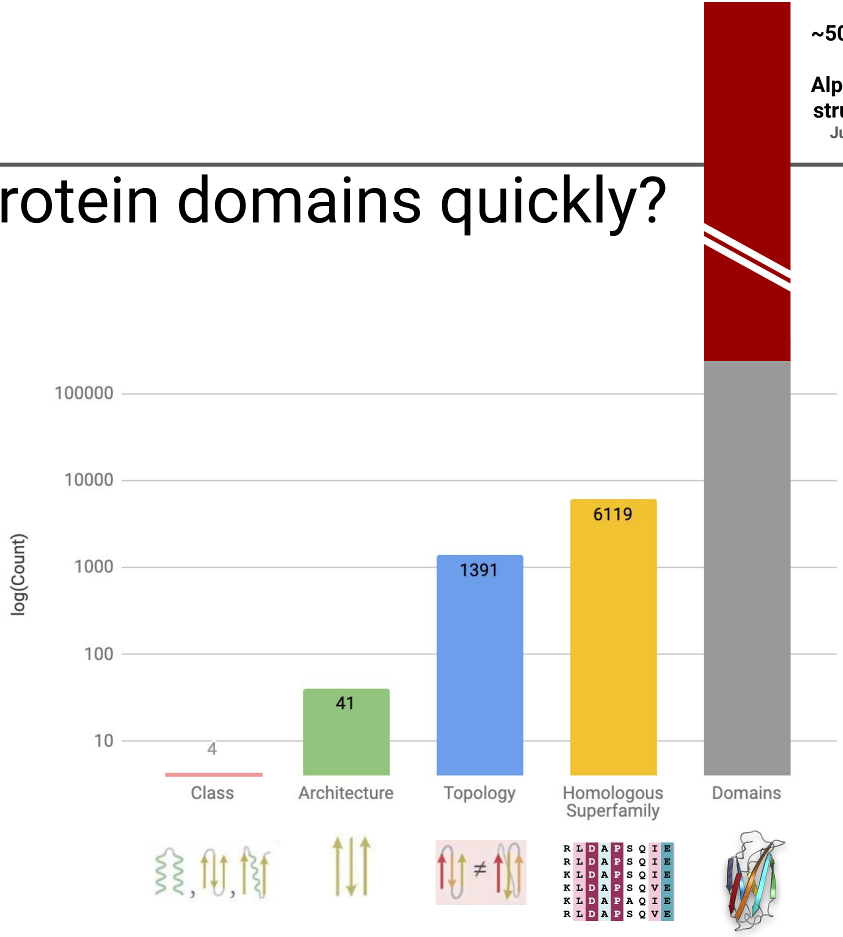
How to process ~500K protein domains quickly?





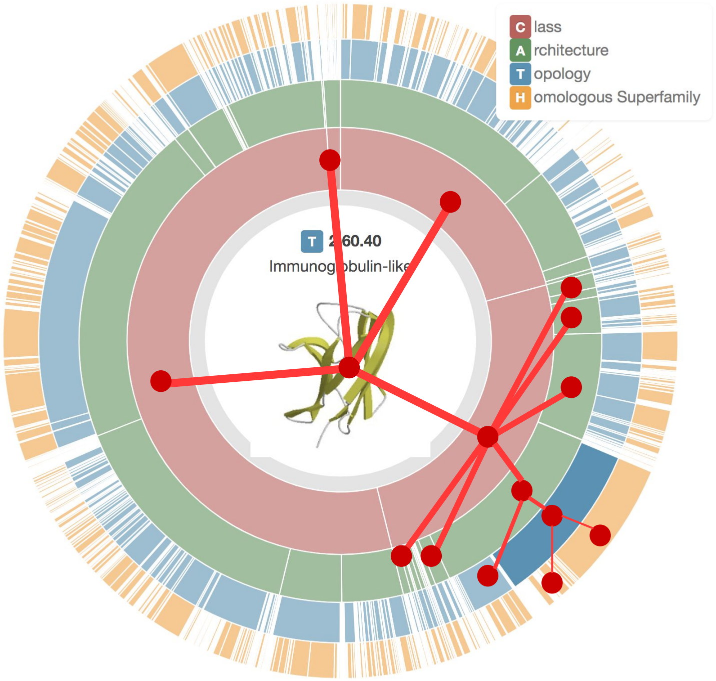
How to process ~500K protein domains quickly?





~50 Million New AlphaFold2 structures!
July 2021

Massively Parallel Workflows with TOIL are used to process the CATH Hierarchy in the cloud and HPCs



TOIL

Massively Parallel Workflows
(and fire-breathing dragon slugs)

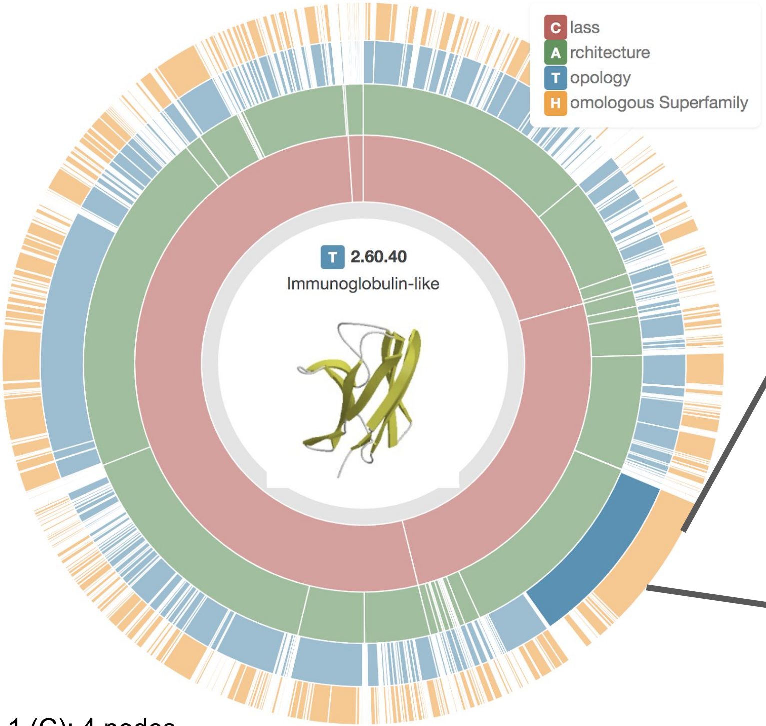
AWS

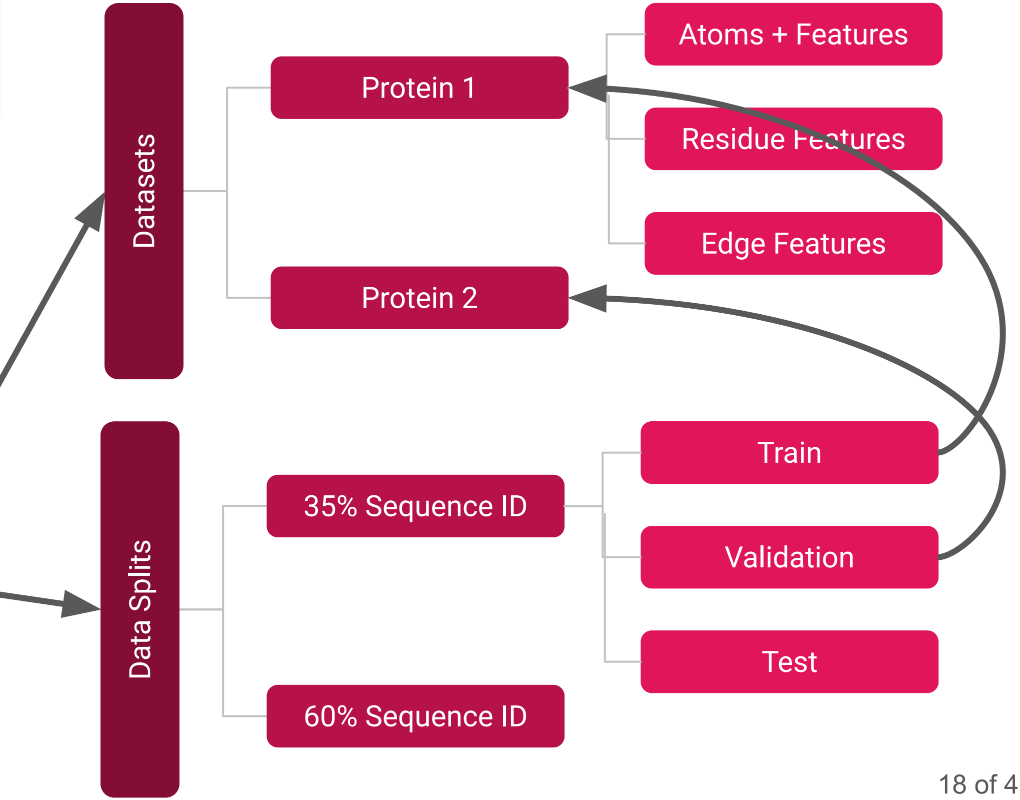
Kubernetes

slurm
workload manager

Speed-up is ~1 week instead of 3+ months

Hierarchical Data Format (HDF) files can chunk and compress the CATH hierarchy in a scalable way

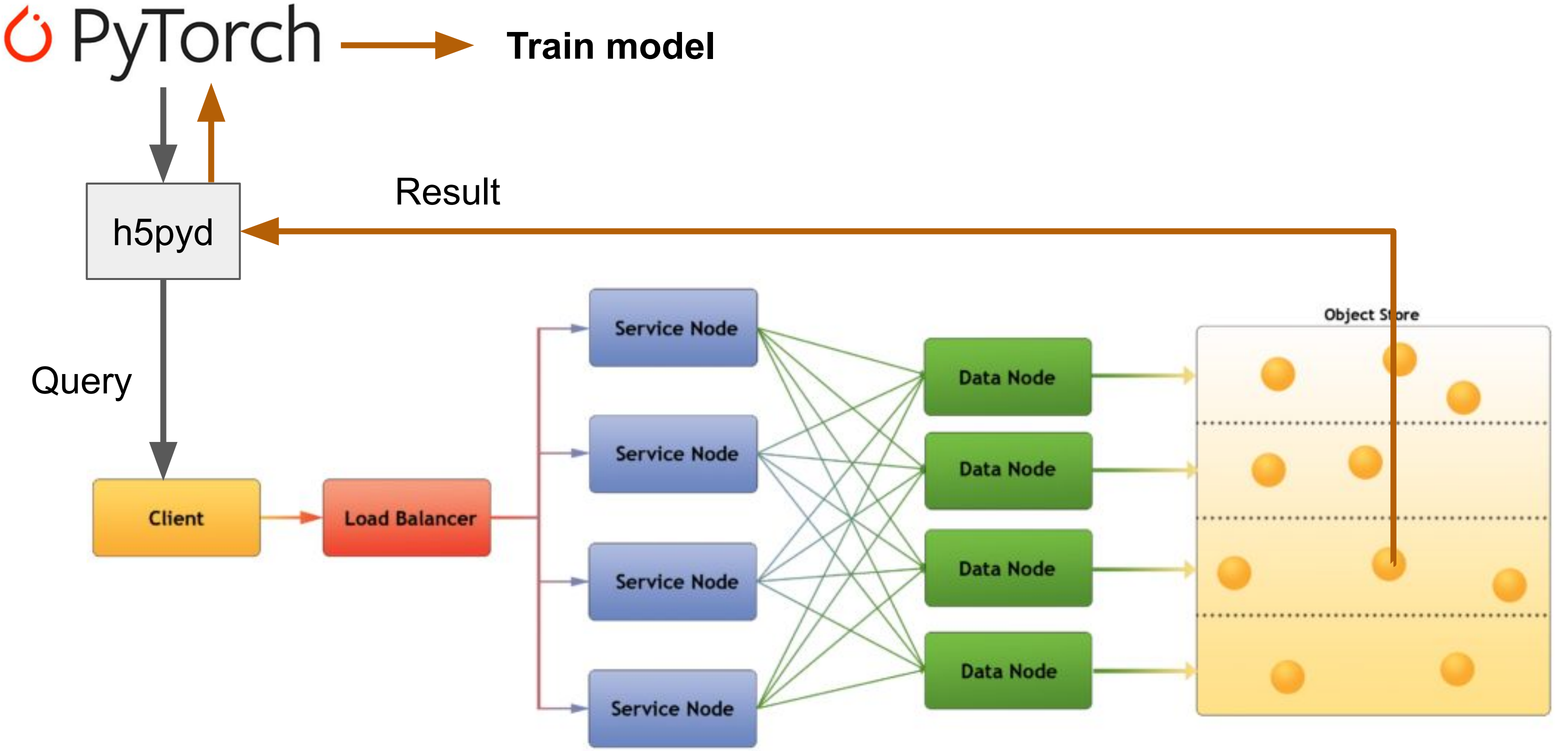




- Datasets
  - Protein 1
    - Atoms + Features
    - Residue Features
    - Edge Features
  - Protein 2
- Data Splits
  - 35% Sequence ID
    - Train
    - Validation
    - Test
  - 60% Sequence ID

Level 1 (C): 4 nodes
Level 2 (A): 41 nodes
Level 3 (T): 1391 nodes
Level 4 (H): 6119 nodes

Create a Highly Scalable Data Service (HSDS) with REST API to access biophysical properties



Process 20 Superfamilies of Interest (potential urfolds)

CATH Code	Name	# Domains	Manual Urfold
1.10.10.10	Winged helix-like DNA-binding	3444	
1.10.238.10	EF-hand	1933	
1.10.490.10	Globins	2891	
1.10.510.10	Phosphotransferase	7219	
1.20.1260.10	Ferritin	2985	
2.30.30.100	SH3 type barrels	1545	SBB
2.40.50.140	OB fold	2879	SBB
2.60.40.10	Immunoglobulin	31905	
3.10.20.30	Beta-grasp domain	520	beta-grasp (Ub)
3.30.230.10	Ribosomal Protein S5; domain 2	1274	Sm-like ribonucleoproteins
3.30.300.20	K homology (KH) domain	529	RRM/RBD(ish)
3.30.310.60	Sm-like ribonucleoprotein	28	Sm-like ribonucleoproteins
3.30.1360.40	Gyrase A; domain 2	160	Sm-like ribonucleoproteins
3.30.1370.10	K Homology domain, type 1	139	RRM/RBD(ish)
3.30.1380.10	Hedgehog domain	101	Sm-like ribonucleoproteins
3.40.50.300	P-loop NTPases	9233	P-loop NTPases
3.40.50.720	Rossmann-like Domain	11728	Rossmann-based
3.80.10.10	Ribonuclease Inhibitor	709	
3.90.79.10	NTP Pyrophosphohydrolase	850	beta-grasp (Ub)
3.90.420.10	Oxidoreductase	58	beta-grasp (Ub)

Used in previous study

Manual Searches

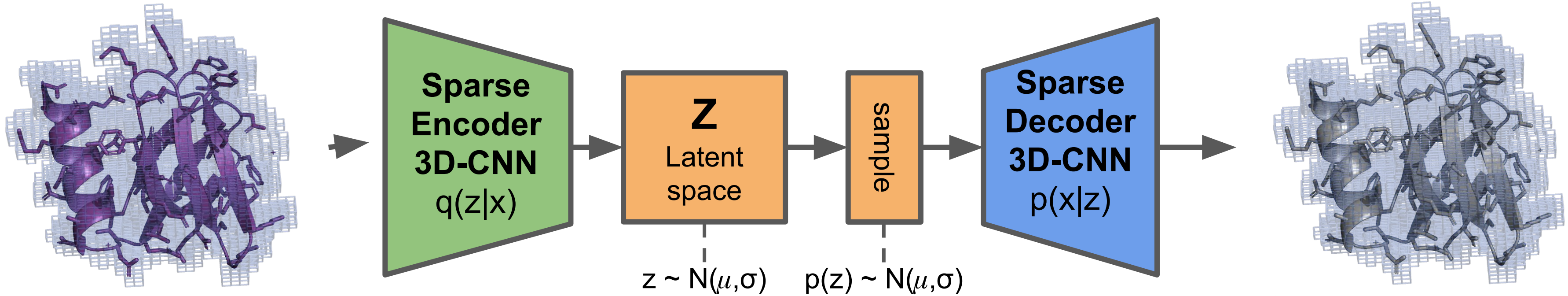
Aim 1 Progress

- Data Generation
  - ✓ Structure preparation workflow
  - ✓ Feature calculation workflow
  - ✓ Migrate to HSDS
  - ✓ Complete 20 Superfamilies of interest
  - ➢ Process all 6K superfamilies
  - ➢ Migrate to use Kubernetes as the default provisioner
- Data Access
  - ✓ Setup local HSDS instance
  - ➢ Migrate to UVA Rivanna HPC

Aim 2

Build and interrogate Deep Generative Models to learn superfamily-specific geometries and properties

Overall DeepUrfold Model: Reconstruct CATH domain structures for one homologous superfamily with Variational Autoencoders



Sparse Encoder 3D-CNN

q(z|x)

Z

Latent space

sample

Sparse Decoder 3D-CNN

p(x|z)

z ~ N(μ,σ)

p(z) ~ N(μ,σ)

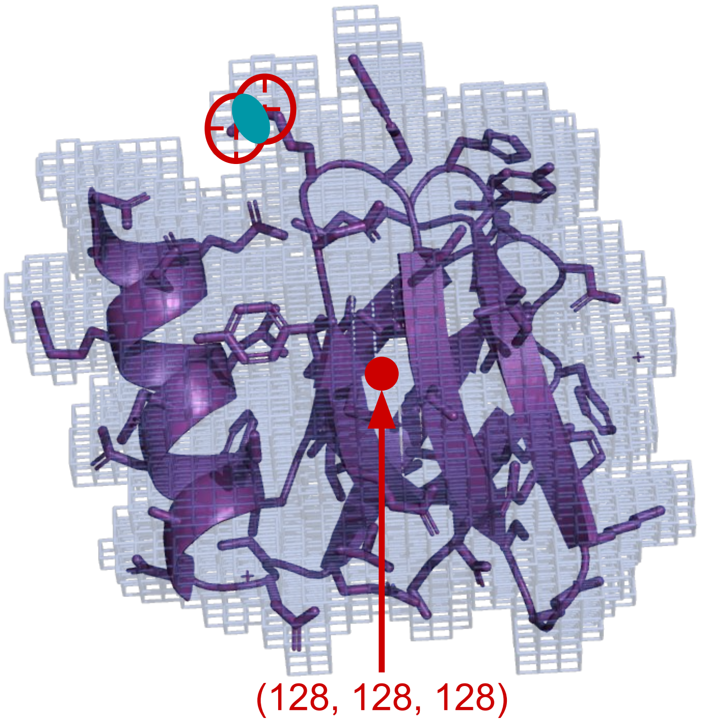
SH3 Fold
(PDB: 1KQ2)

‘Reconstructed’
SH3 Fold
(PDB: 1KQ2)

Evidence Lower Bound (ELBO):

$$\ln p(x) \geq \underbrace{E_{q(z|x)}[\ln p(x|z)]}_{\text{Reconstruction Error}} - \underbrace{D_{KL}[q(z|x)||p(z)]}_{\text{Similarity between learned distribution } (q) \text{ and true distribution } (p)}$$

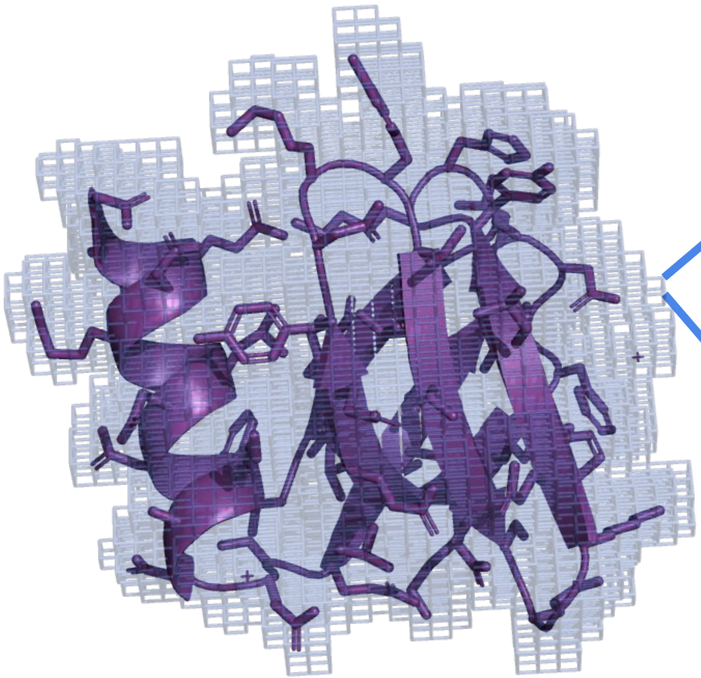
Representation: 3D Image, Voxel Space



SH3 Fold (PDB ID: 1KQ2.A)

- Protein centered in 256^3 Å^3 volume
- Van der Waals Spheres around each atom are discretized to fit 1 Å^3 voxels using a KDTree
  - No need to annotate all voxels because most volume is sparse
  - Each voxel within an atomic sphere inherits same set of features
- Covalent bonding occurs where there is overlap between voxels from different atoms
  - Bond voxels use the max between features

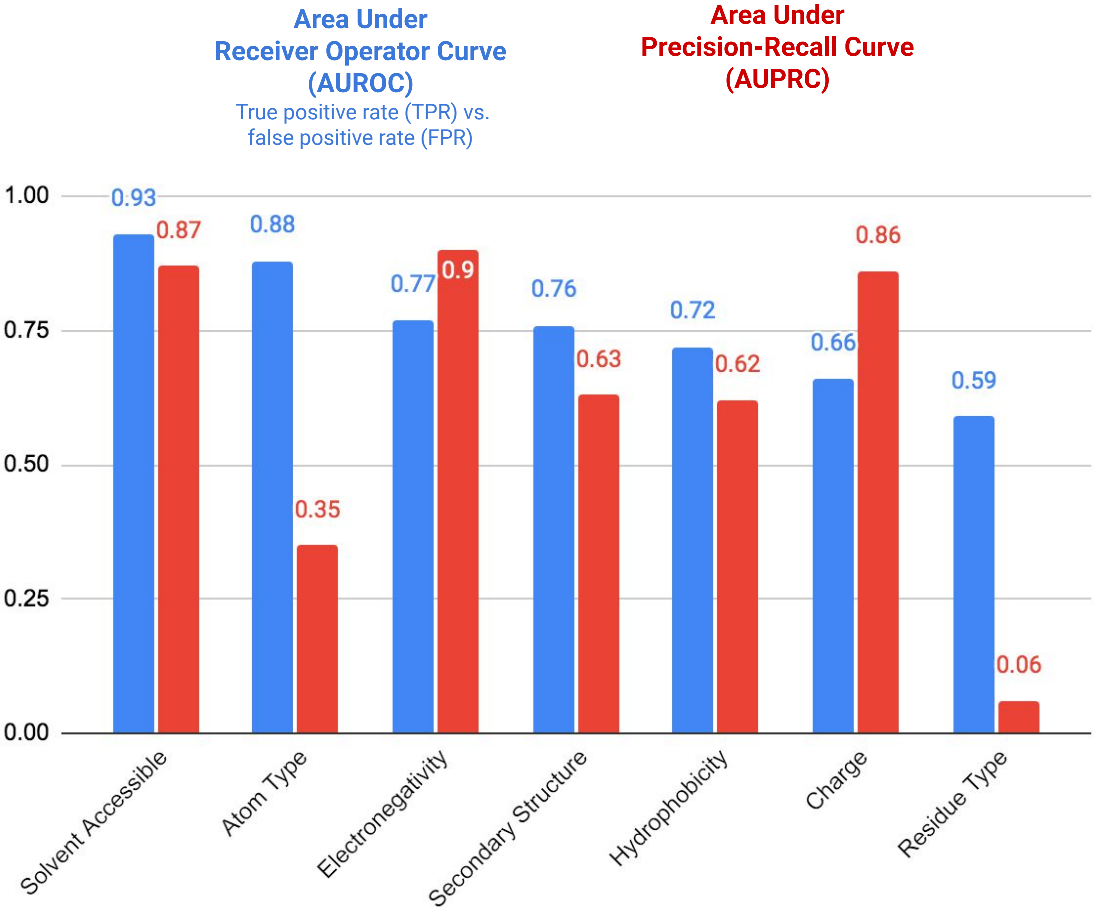
Representation: Atom-based Physicochemical Features



Feature Type	**# of Boolean Features**
Atom Type	13
Residue Type	21
Secondary Structure	3
Accessibility	1
Is Hydrophobic	1
Is Positively Charged	1
Is Electronegative	1

SH3 Fold (PDB ID: 1KQ2.A)

Model Evaluation: Ig Reconstruction



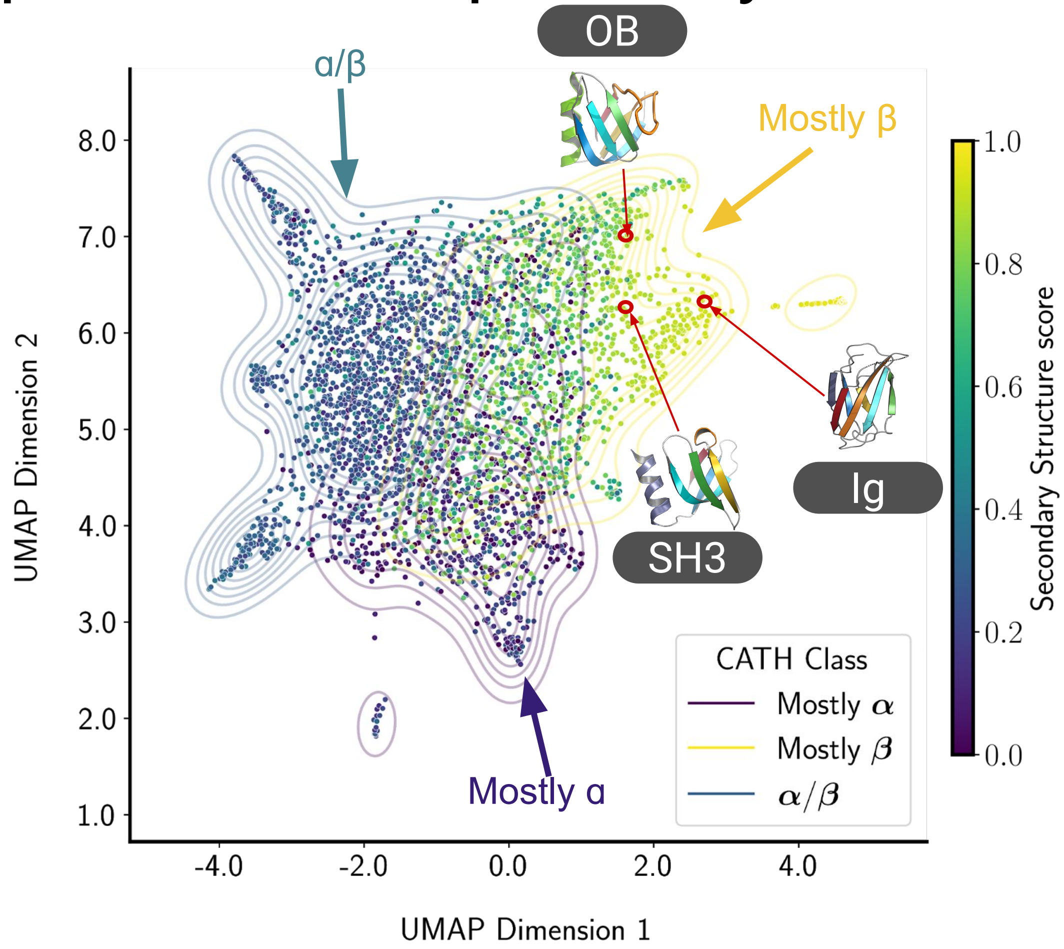
- Train **7 different Ig models** with different types of features (1 type per model)
- **Residue Type performed worst** for both metrics, so it was removed from training all other models

FPR = FP/(FP+TP)

TPR = Precision = TP(TP+FP)

Recall = TP/(TP+FN)

Superfamilies separate by CATH Class (2° structure)

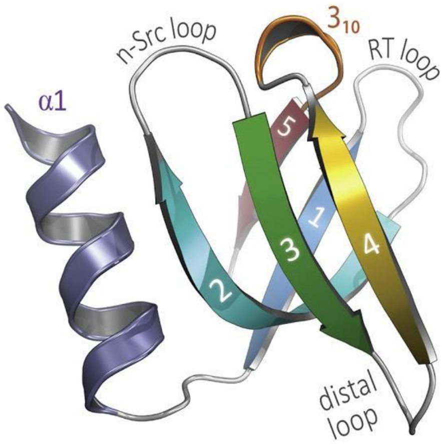


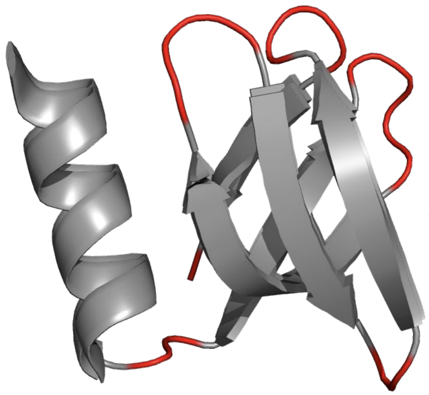
1. Trained 20 different models for each superfamily
2. Ran representatives through the models for each superfamily, saving latent space
3. **Combined all latent spaces from all different models into 1 dataset**
4. Used UMAP reduce the # of dimensions from 1024->2

$$\text{SS Score} = \frac{\#\beta \text{ atoms}}{\#\beta \text{ atoms} + \#\alpha \text{ atoms}}$$

Asses Urfold Hypothesis via Multiple Loop Permutations

SH3/Sm
1kq2

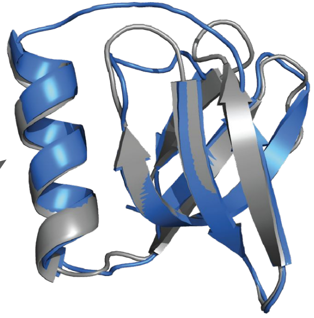


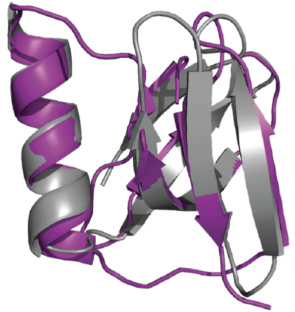


96 possible permutations!

Remodel Loops with MODELLER

3-4-α1-1-2-5
RMSD: 0.734
TM-Score: 0.61

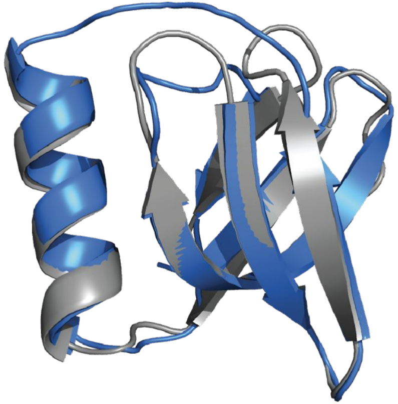




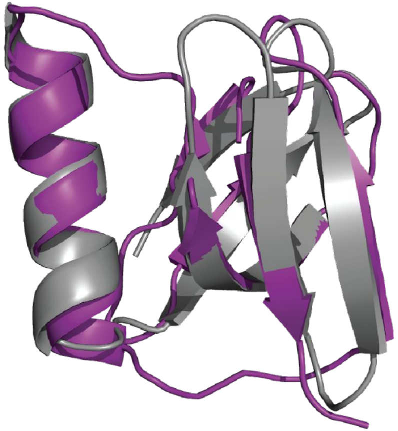
5-1-2-α1-4-3
RMSD: 1.297
TM-Score: 0.415

Likelihood ratios can be used to quantify similarities among multi-loop permuted structures

1kq2A00
SH3 (2.30.30.100)



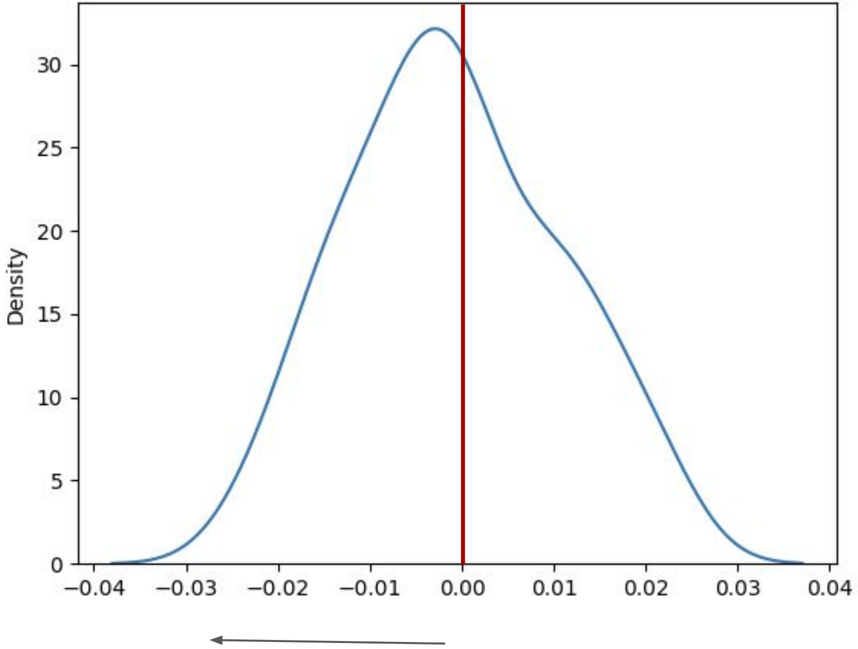
4-5-1-2-3-6
Log-odds score: -0.0228
RMSD: 0.734
TM-Score: 0.61



6-2-3-1-5-4
Log-odds score: 0.0216
RMSD: 1.297
TM-Score: 0.415

Log-odds score = log(ELBO$_{\text{Permuted}}$)-log(ELBO$_{\text{Wild-type}}$)

Log-odds score distribution



More Similar

Aim 2 Progress

✓ Train 20 SF models

✓ Visualize latent spaces

✓ Multiple loop permutations for SH3

  ➢ Complete for other SFs

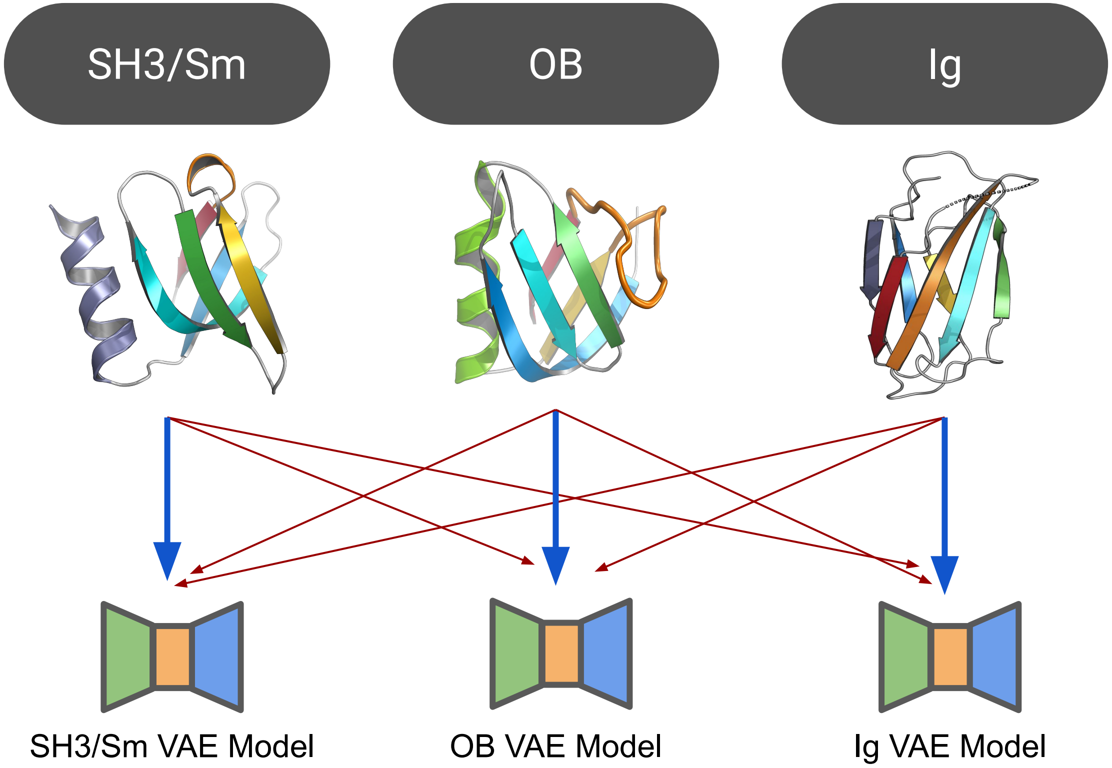
Aim 3

Identify distant evolutionary relationships that bridge protein architectures and topologies that define an Urfold

Question 3.1:

Which superfamilies might share an urfold?

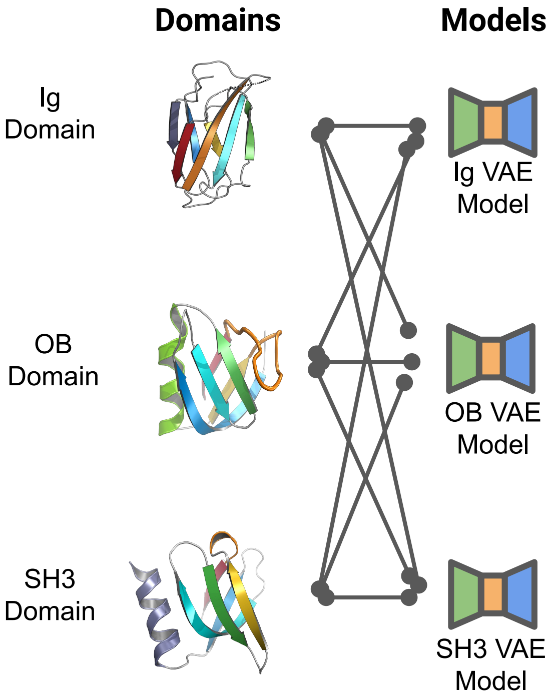
Objective: Create a New Similarity Metric

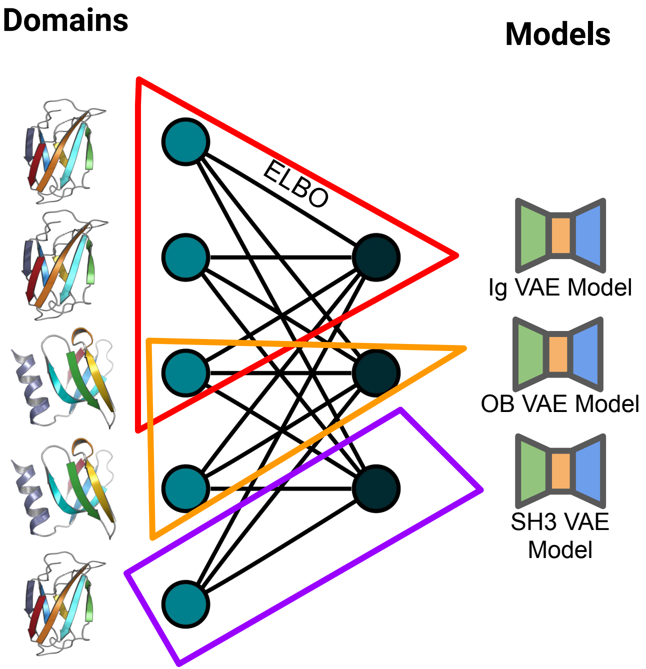


1. Train one model for each superfamily
2. Subject superfamily representatives to *all other* superfamily VAE models

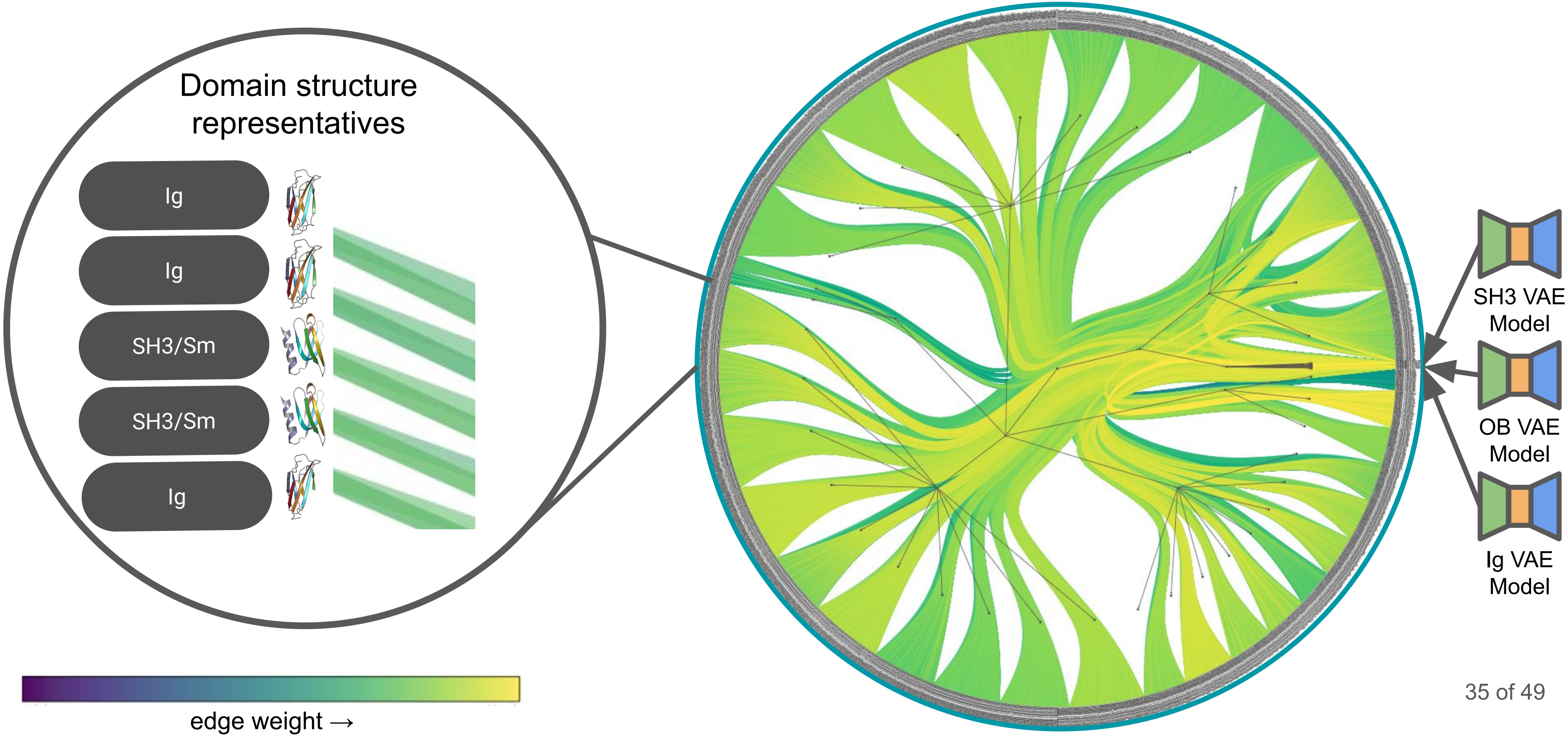
Stochastic block models (SBM) can find superfamilies that span multiple clusters

- SBMs are probabilistic graphical models that can detect **mixed-membership communities** in **fully connected bipartite graphs, with variable edge weights**





Stochastic Block Modelling finds domains from different superfamilies in the same community (potential urfolds?)



Domain structure representatives

Ig

Ig

SH3/Sm

SH3/Sm

Ig

SH3 VAE Model

OB VAE Model

Ig VAE Model

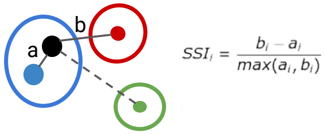
edge weight →

Measure clusters with no ground truth of Urfolds

- **Hypothesis:** Using CATH as ground truth, **“least similar”** clusterings will be stronger evidence for the Urfold
- **Compare** to known state-of-the-art protein similarity algorithms using their similarity metric in the Stochastic Block Model

Silhouette

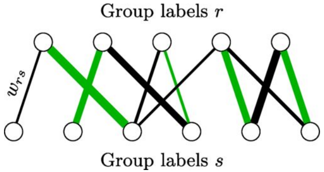
Mean intra-cluster vs mean nearest-cluster distances



-1 ≤ Silhouette ≤ 1
(wrong) (perfect)
0=overlapping

Overlap

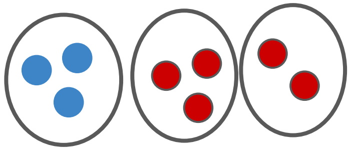
% overlap using bipartite graphs



0 ≤ Overlap ≤ 1
(wrong) (perfect)

Homogeneity

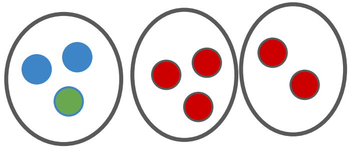
All clusters contain only members of a single class



0 ≤ Homogeneity ≤ 1
(wrong) (perfect)

Completeness

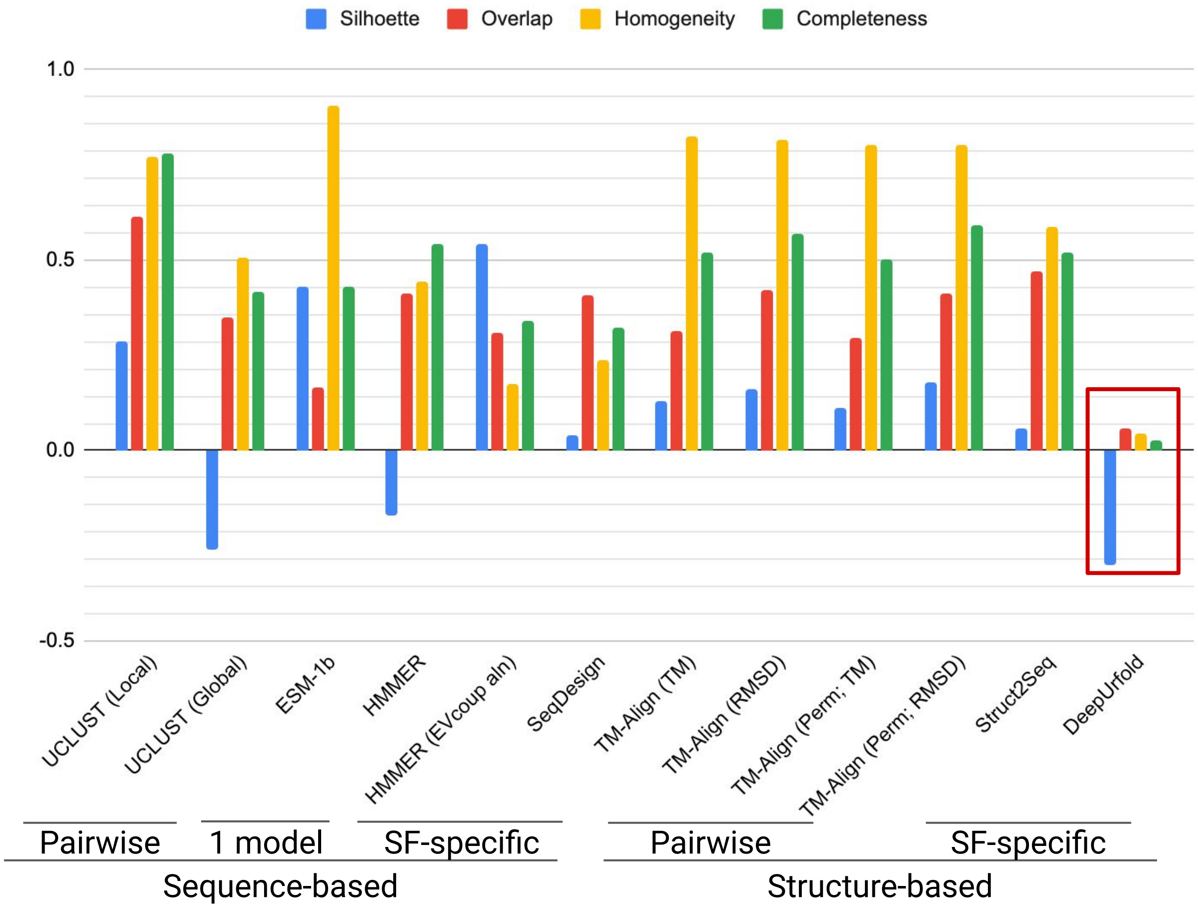
All data points of a given class are in the same cluster



0 ≤ Completeness ≤ 1
(wrong) (perfect)

Peixoto. Phys. Rev. 2021

SBM Communities vs. CATH Superfamilies



- DeepUrfold's CATH reconstructing CATH is the 'worst' – e.g. Hierarchical clustering might not be the the best view of fold space

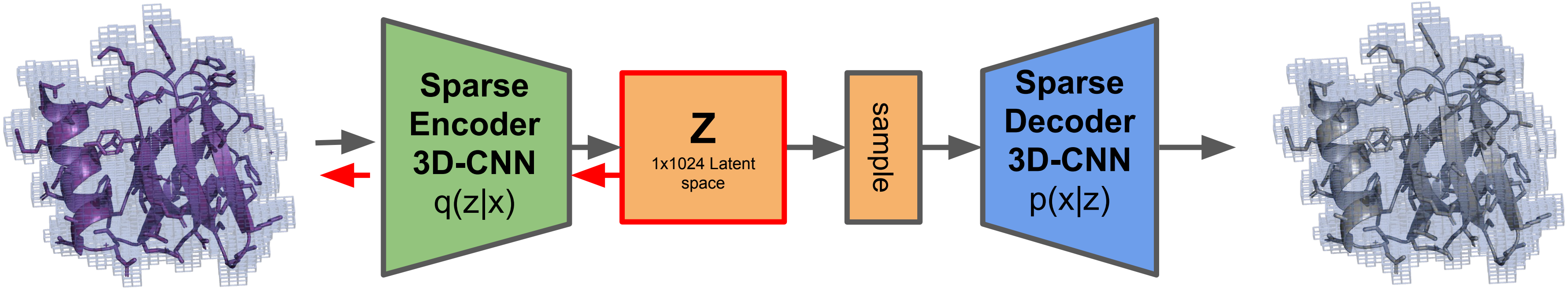
- Our van der Waals representation of each atom complete with biophysical features is so different from the others

- Our model is learning something beyond simple structural and geometric similarity, towards the realm of structure/function properties

Question 3.2:

Do particular geometric and biophysical properties contribute to an urfold?

Layerwise Relevance Propagation (LRP)



Sparse Encoder 3D-CNN q(z|x)

Z

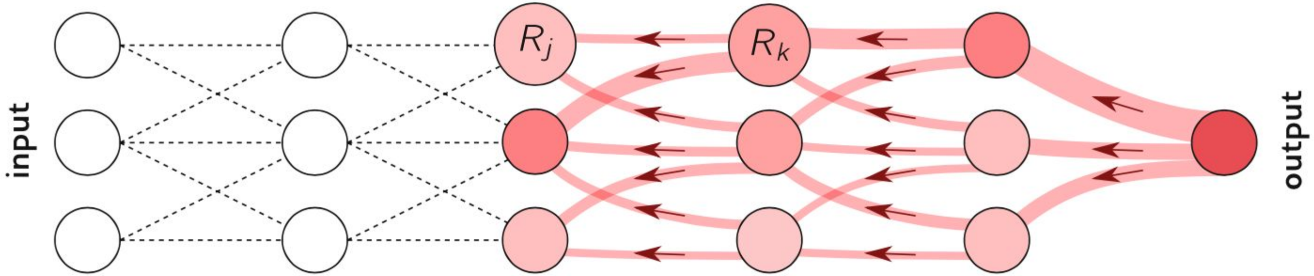
1x1024 Latent space

sample

Sparse Decoder 3D-CNN p(x|z)

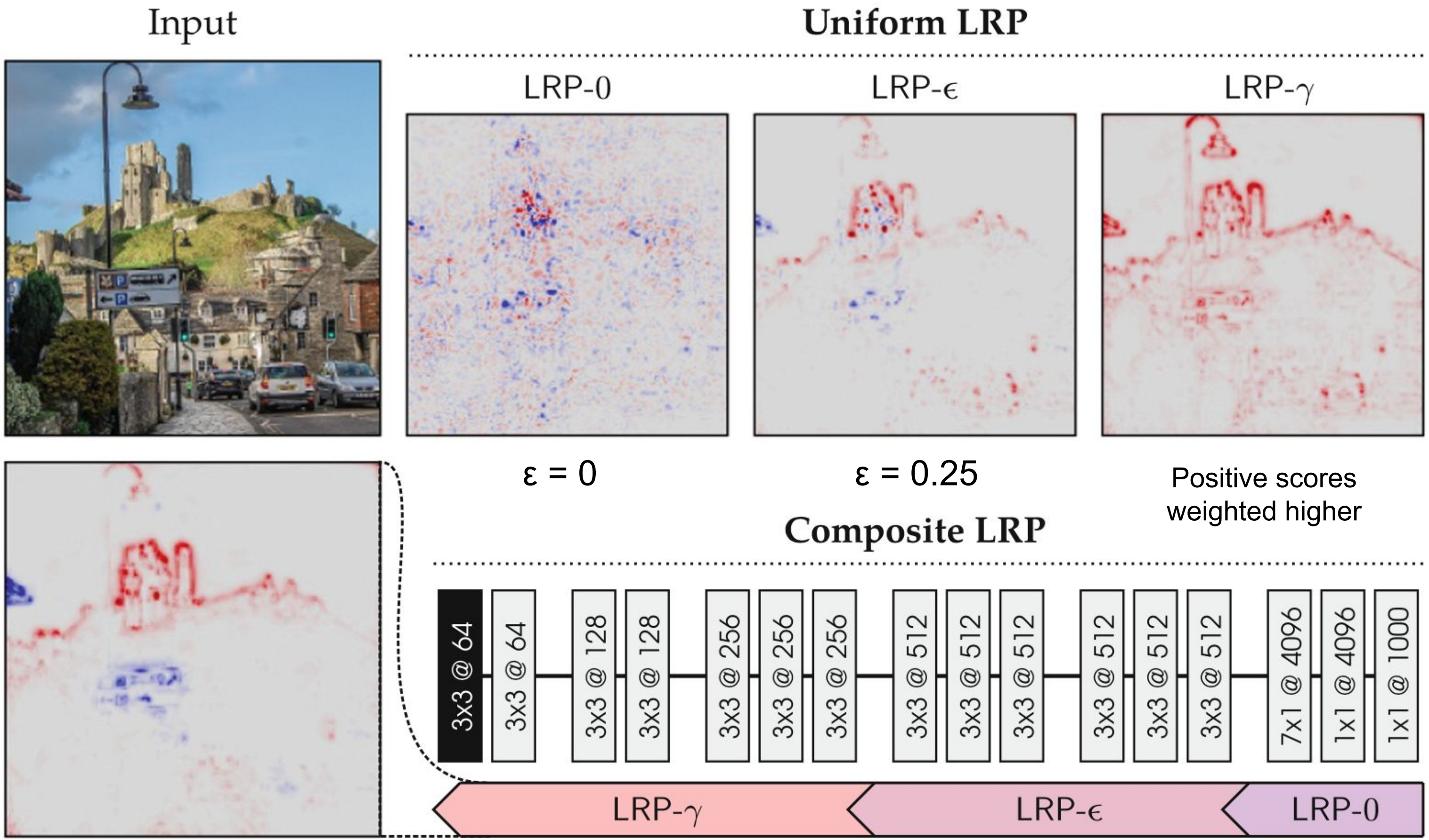
OB Fold
(PDB: 1KQ2)

‘Reconstructed’
OB Fold
(PDB: 1KQ2)



Montavon, Binder, at al. Lecture Notes in Computer Science. 2019

LRP Example For Predicting Castles



ε = 0

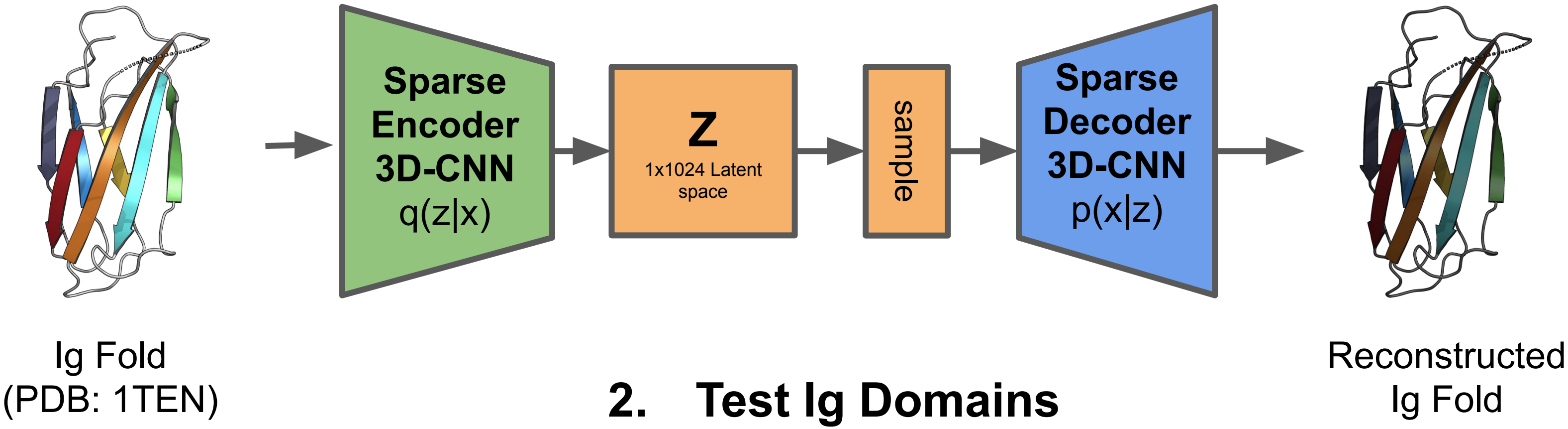
ε = 0.25

Positive scores weighted higher

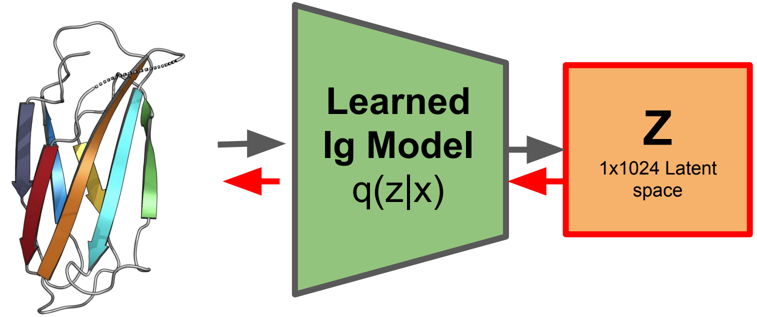
Montavon, Binder, at al. Lecture Notes in Computer Science. 2019

LRP Example: Ig structure through Ig Model

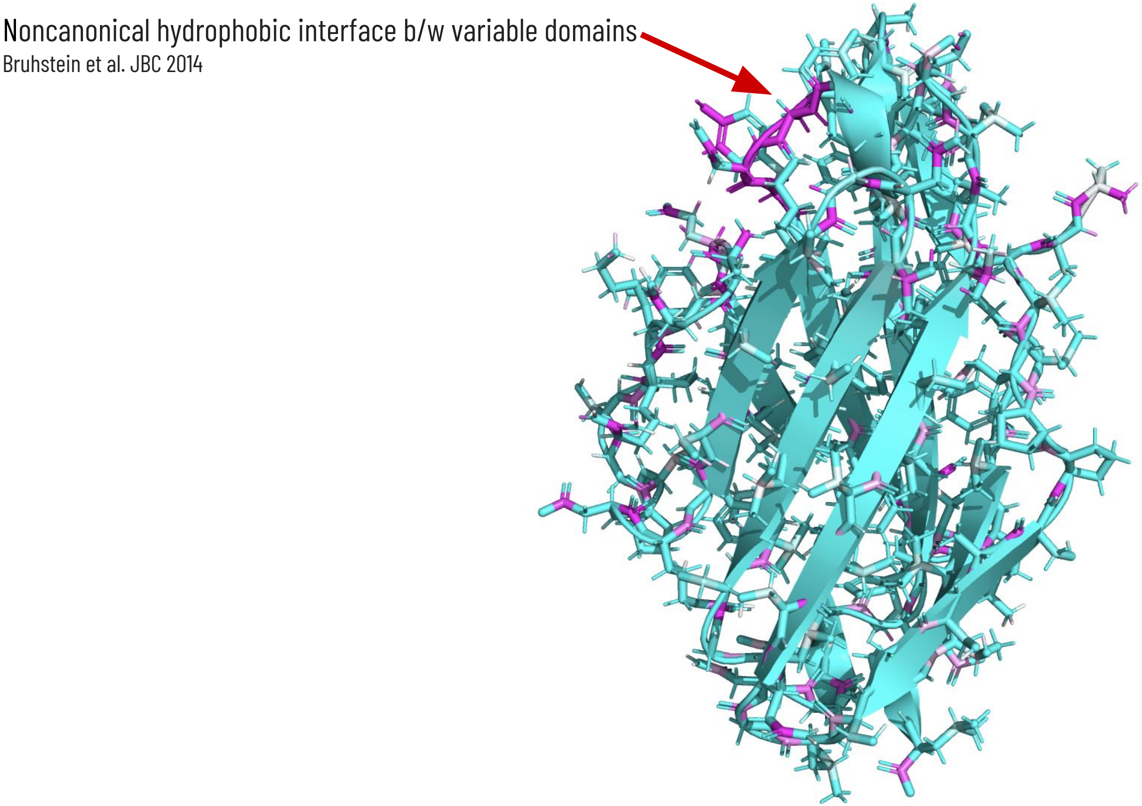
1. Train Ig Model

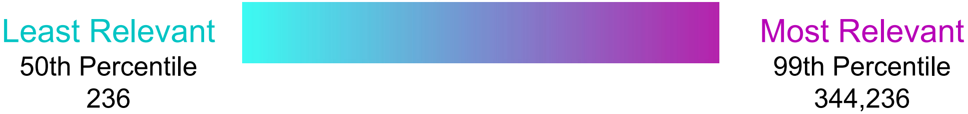


2. Test Ig Domains



LRP Uncovers Backbone Carbons as Most Relevant

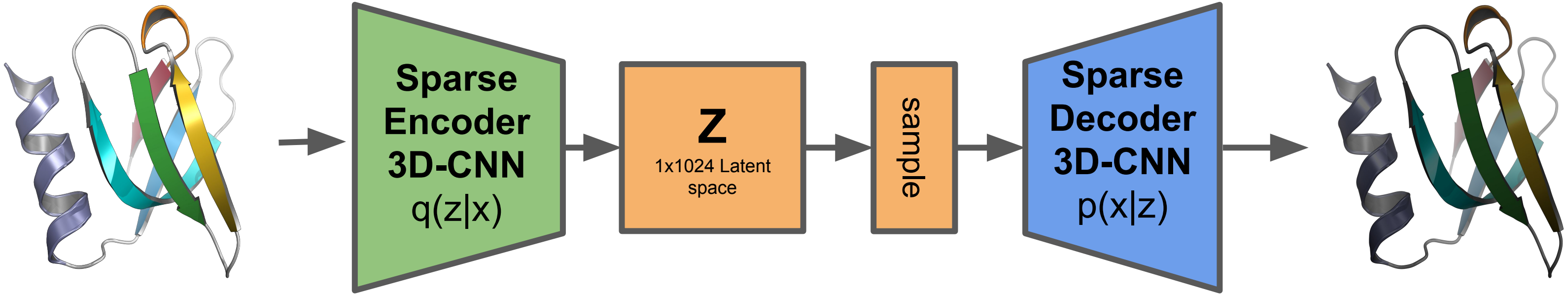

Noncanonical hydrophobic interface b/w variable domains
Bruhstein et al. JBC 2014



4unu, amyloid fiber Ig dimer

LRP Example: OB structure through SH3 Model

1. Train SH3 Model



Sparse Encoder 3D-CNN $q(z|x)$

Z

1x1024 Latent space

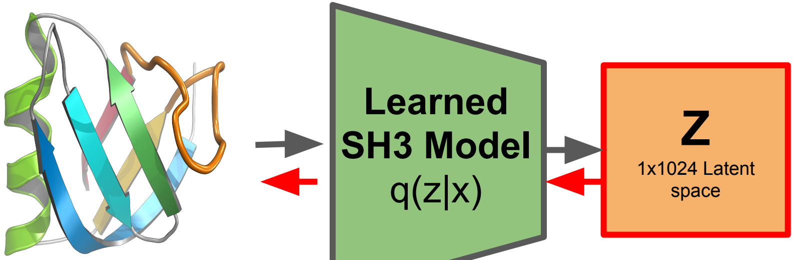
sample

Sparse Decoder 3D-CNN $p(x|z)$

SH3 Fold
(PDB: 1KQ2)

Reconstructed
SH3 Fold

2. Test OB Domains



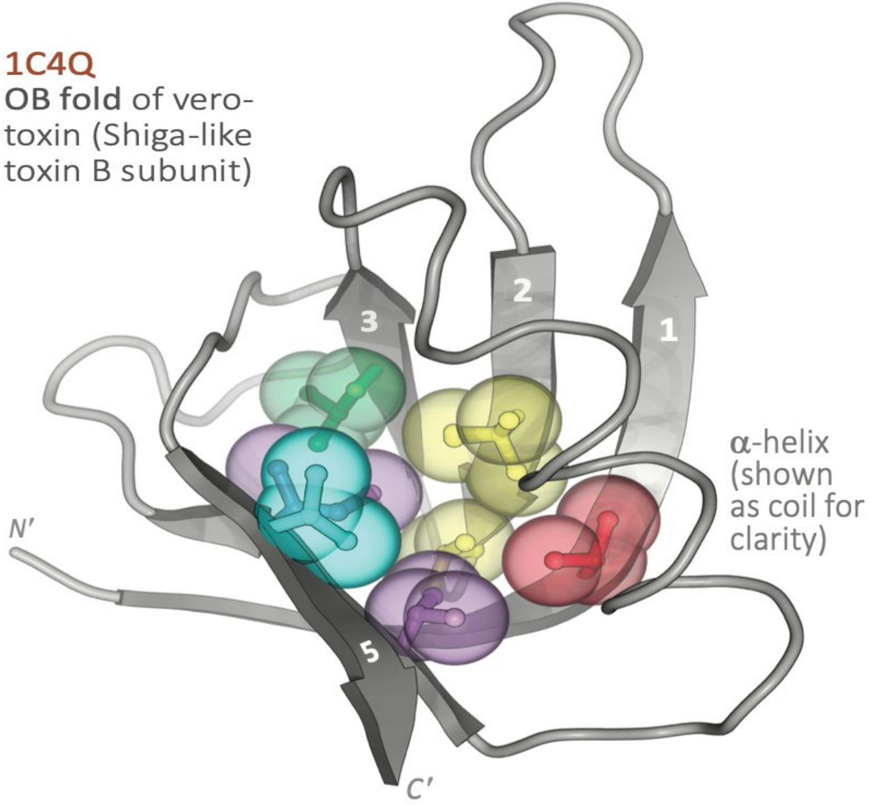
Learned SH3 Model $q(z|x)$

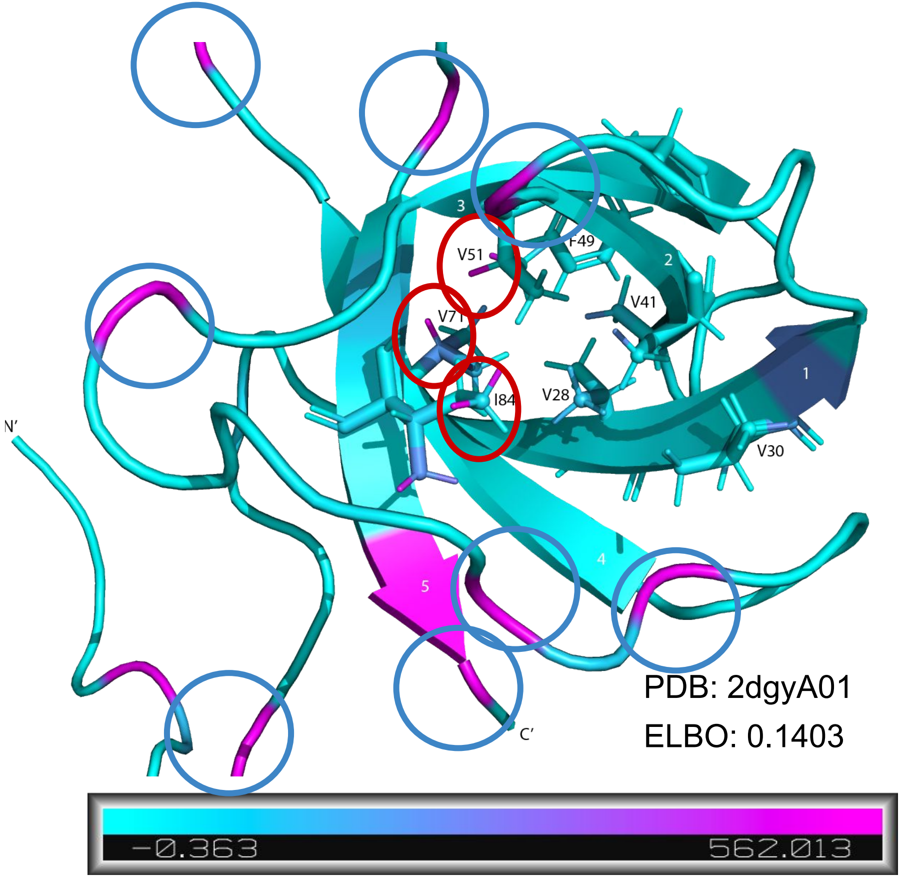
Z

1x1024 Latent space

OB Fold
(PDB: 1C4Q)

Loops and Hydrogens from Conserved hydrophobic core are uncovered by LRP


1C4Q
OB fold of vero-toxin (Shiga-like toxin B subunit)
3
2
1
5
N′
C′
α-helix (shown as coil for clarity)



PDB: 2dgyA01

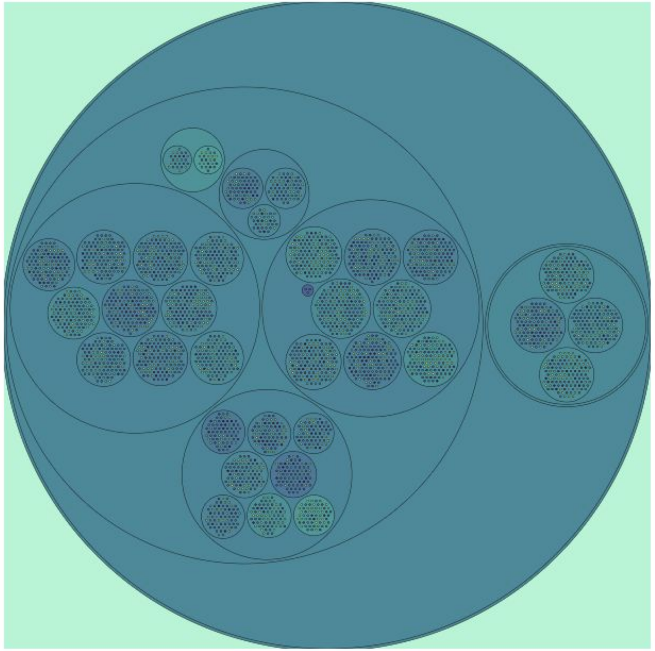
ELBO: 0.1403

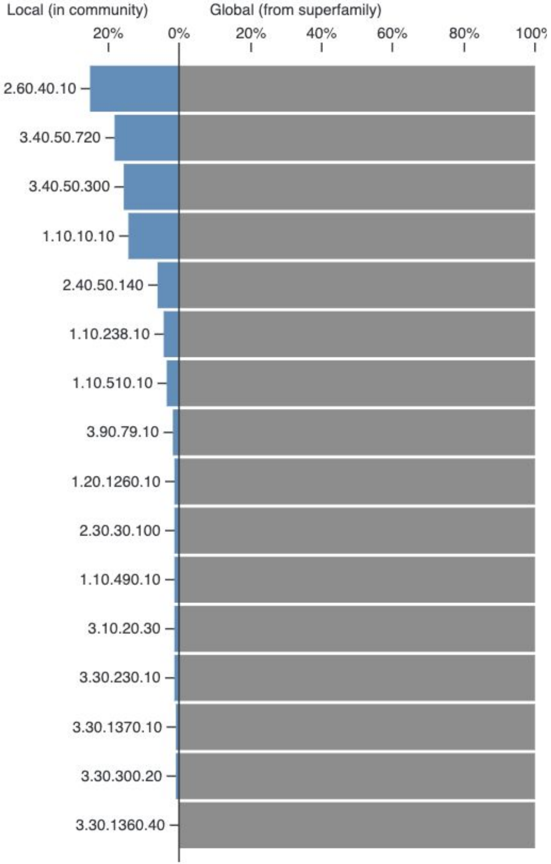
Youkharibache, Veretnik, et al. *Structure* (2019); https://doi.org/10.1016/j.str.2018.09.012

Question 3.3:

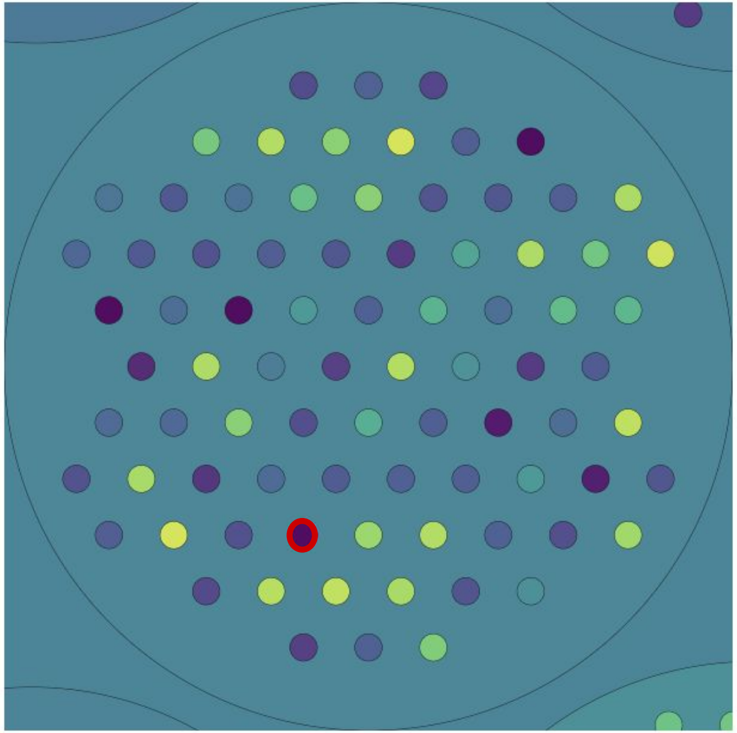
How can we define an urfold by combining SBM communities with Atomic Relevance scores from All-vs-All LRP?

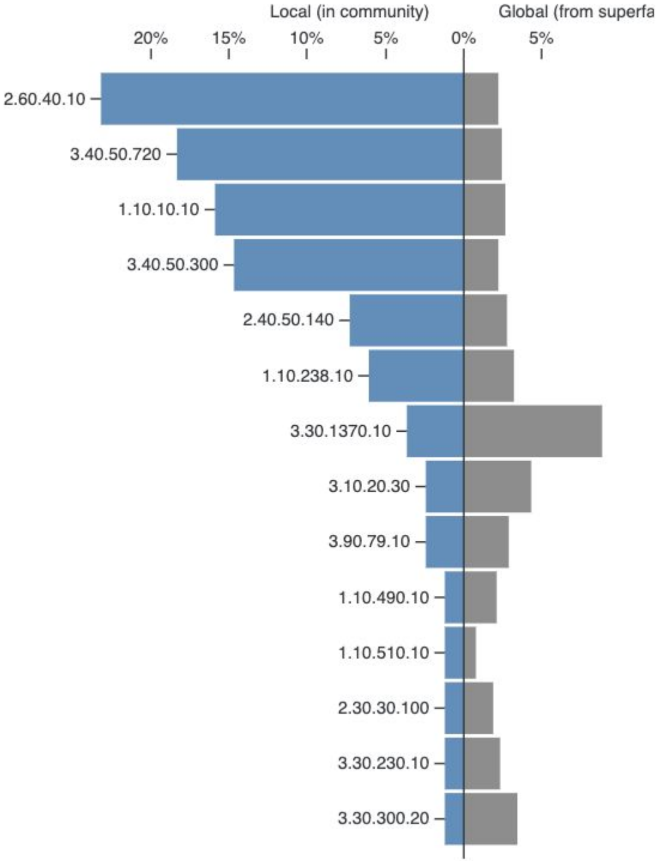
Browse SBM communities



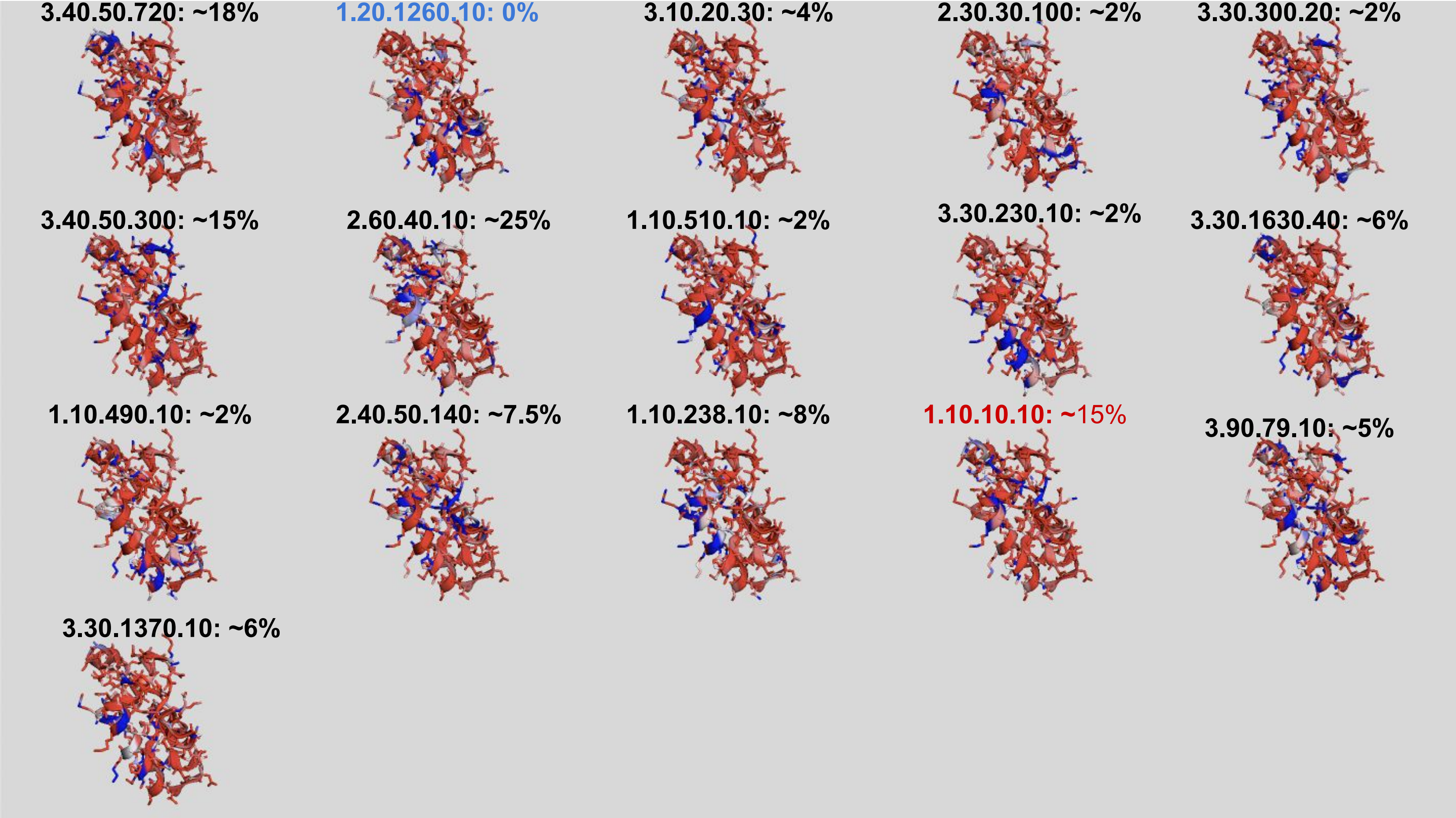


Zoom into a single community





LRP scores for one domain in community: 1a04A02



Next Steps for Aim 3

- **Question 1: Which Superfamilies may share an urfold?**
  - ✓ Complete All-vs-All for 20 superfamilies of interest
  - ✓ Detect communities with Stochastic Block Models
- **Question 2: Which geometric and biophysical properties contribute to an urfold?**
  - ✓ Run Layerwise-relevance Propagation for all domain in the All-vs-All approach
- **Question 3: How can we define an urfold?**
  - ○ Create javascript visualizations to analyze and combine SBM communities and LRP results
  - ➢ Add biophysical properties to visualizations
  - ➢ Find common structural fragments through structure alignment
  - ➢ Create a definition for an urfold by elucidating why the SBM created the communities

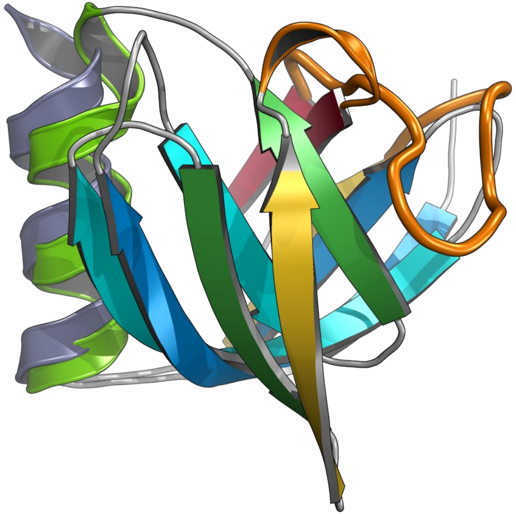
Conclusions

- **Hypothesis:** An entity called the 'Urfold' may exist as a *bona fide* level, between Architecture and Topology, to represent 3D architectural similarity despite topological variability

- Aim 1: Develop a community resource to create and **share biophysical properties and protein structures** with **Train/Test/Validation splits** to facilitate reproducible ML workflows

- Aim 2: Design and implement a novel sequence-independent, alignment-free, rotation-invariant **similarity metric of proteins** that leverages similarities in latent-spaces rather than 3D structures.

- Aim 3: A new approach to **detect clusters**, or ***communities***, of similar protein structures using Stochastic Block Models. This method takes a different approach to clustering, allowing for proteins to span multiple clusters, thereby allowing for the **continuous nature of fold space.**



Acknowledgements

Phil Bourne

Cam Mura

Stella Veretnik

Zheng Zhao

MSDS Alumni

Menuka Jaiswal

Saad Saleem

Kwon Yonghyeon

Bourne Lab Members

Lei Xie (Sabbatical Visitor)

Abby Newbury

Skylar Brodowski

Mark Bray

Niraja Bohidar

BME

UVA DATA SCIENCE

Funding: Presidential Fellowship in Data Science program

UNITED CAMPUS WORKERS
FIGHTING FOR ECONOMIC JUSTICE & DEMOCRACY - ON CAMPUS & ACROSS TENNESSEE
UCW-CWA
Communications Workers of America local 3865

UNITED CAMPUS WORKERS
UVA

Questions?