

Deep Learning of Protein Structural Classes: Any Evidence for an 'Urfold'?

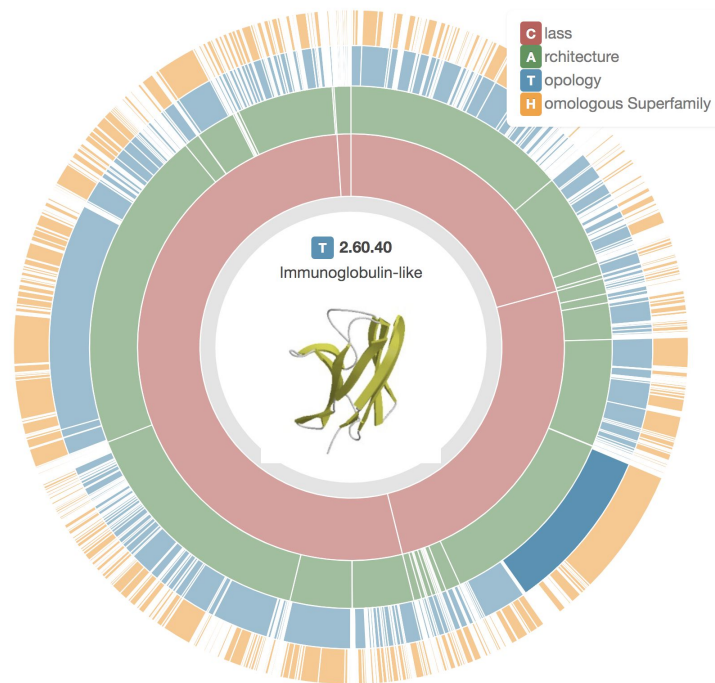
**3DSig, ISMB 2020
July 15th, 2020**

**Eli Draizen
Phil Bourne's Lab
University of Virginia**

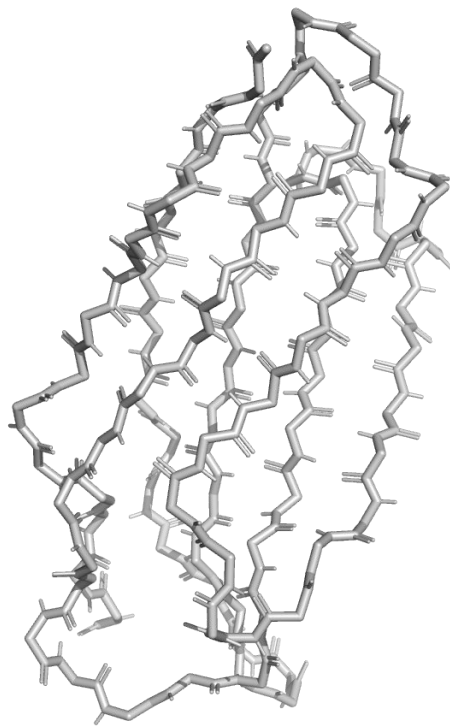
Find these slide online: [doi:10.5281/zenodo.3909755](https://doi.org/10.5281/zenodo.3909755)

Background

- Sequence -> Structure -> Function -> Evolution questions encompass a majority of (structural) bioinformatics problems
- Hierarchical classification of structure provide an important framework for understanding these relationships
 - **CATH**
 - SCOP
 - ECOD

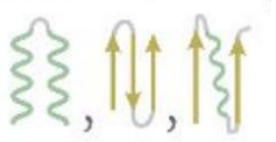


Structure Hierarchy



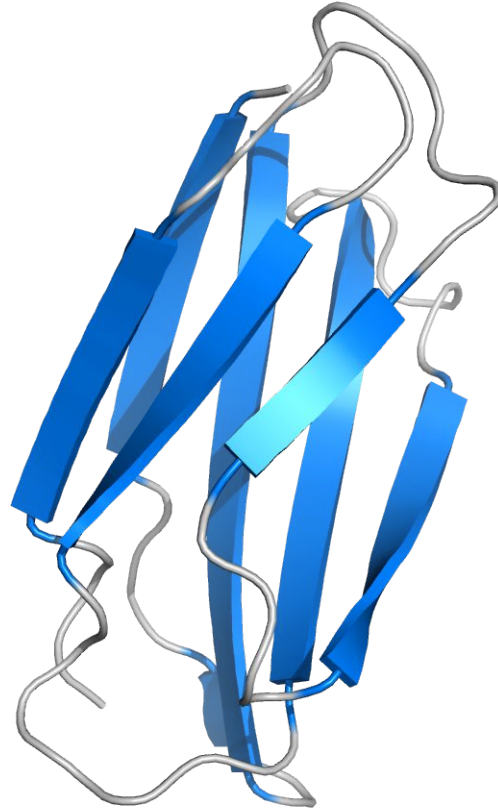
Structure Hierarchy

1. Class



Types of 2° structure
elements (SSE)

E.g. Mostly Beta (2)



Structure Hierarchy

1. Class



Types of 2° structure elements (SSE)

E.g. Mostly Beta (2)

2. Architecture



3D arrangements of SSE

E.g. Sandwich (2.60)



180°

Structure Hierarchy

1. Class



Types of 2° structure elements (SSE)

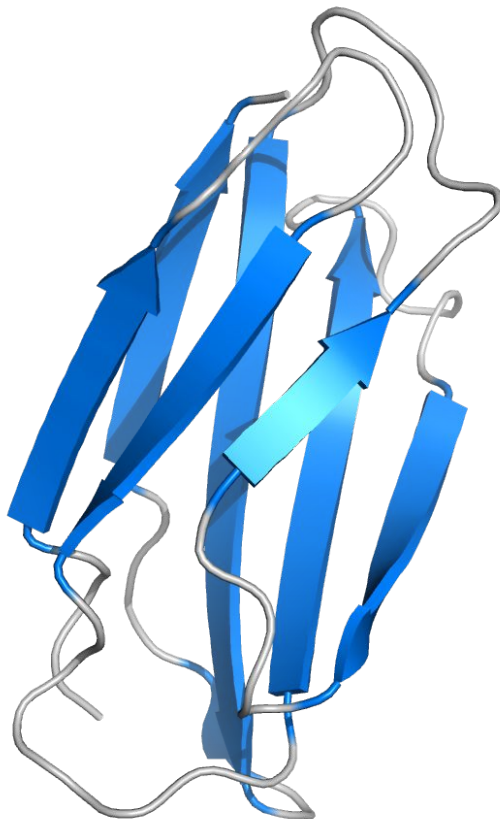
E.g. Mostly Beta (2)

2. Architecture

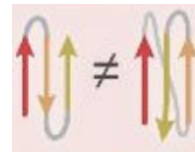


3D arrangements of SSEs

E.g. Sandwich (2.60)



3. Topology



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

E.g. Immunoglobulin-like (2.60.40)

Structure Hierarchy

1. Class



Types of 2° structure elements (SSE)

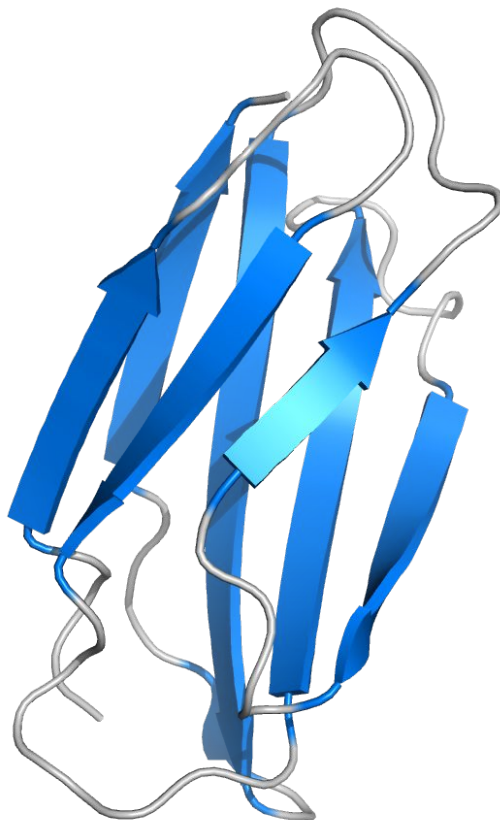
E.g. Mostly Beta (2)

2. Architecture

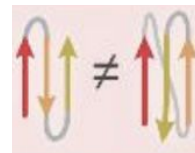


3D arrangements of SSEs

E.g. Sandwich (2.60)



3. Topology



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

E.g. Immunoglobulin-like (2.60.40)

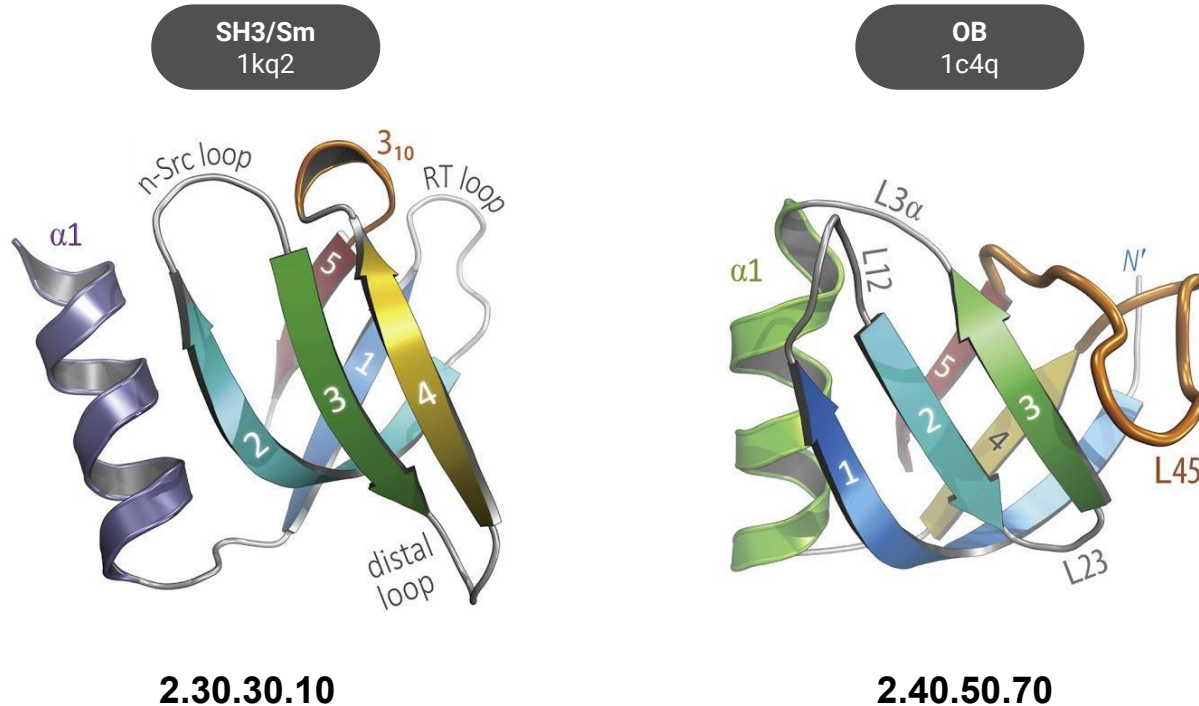
4. Homologous Superfamily

A	C	P	R	L	D	V	D	S	Q
A	C	P	R	-	E	V	D	C	N
G	C	P	R	I	E	L	D	S	H
G	C	G	K	I	E	V	E	S	D
-	C	G	K	L	E	I	E	A	T
A	C	A	R	V	D	-	D	A	Y
G	C	R	R	K	E	L	D	C	E

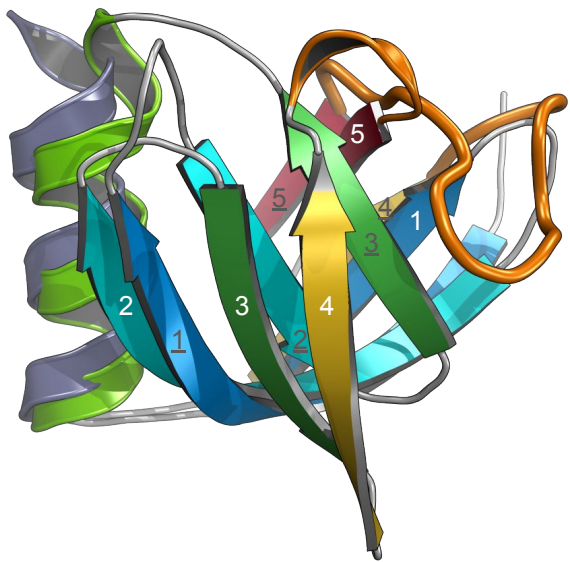
Evolutionary relationships via sequence

E.g. Immunoglobulins (2.60.40.10)

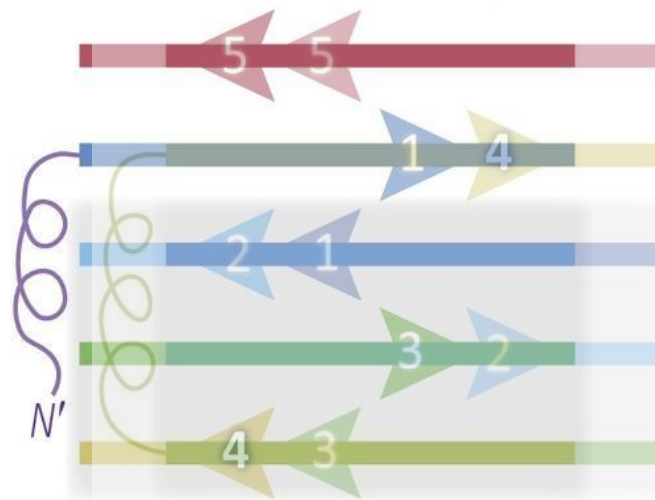
Small β -Barrels (SBBs) Exhibit *Architectural Similarity Despite Topological Variability*



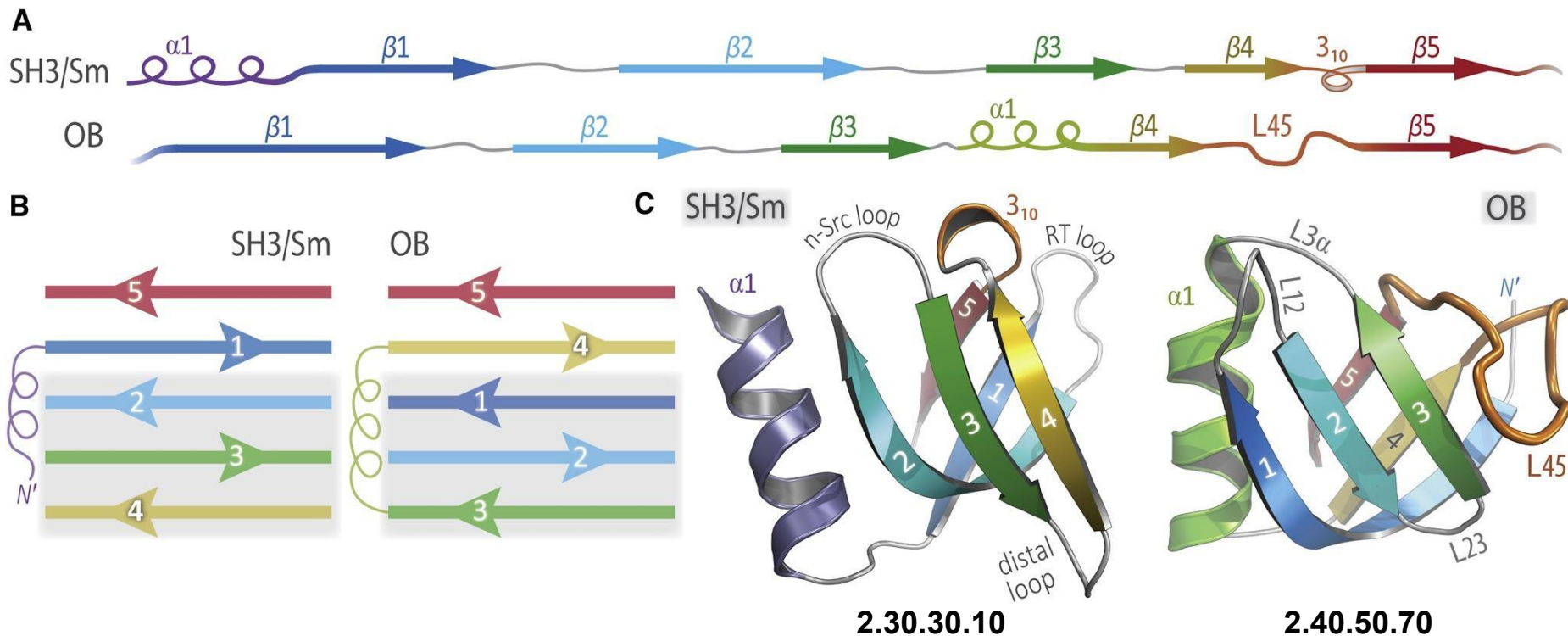
Small β -Barrels (SBBs) Exhibit *Architectural Similarity Despite Topological Variability*



RMSD: 3.7 Å



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



Possible new entity between Architecture+Topology

1. Class



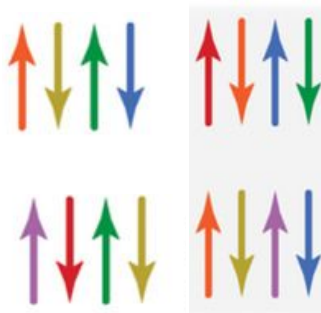
Types of 2° structure elements (SSE)

2. Architecture



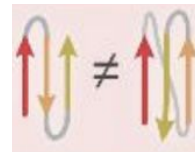
3D arrangements of SSE

3. Urfold



3D architectural similarity despite topological variability

4. Topology



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

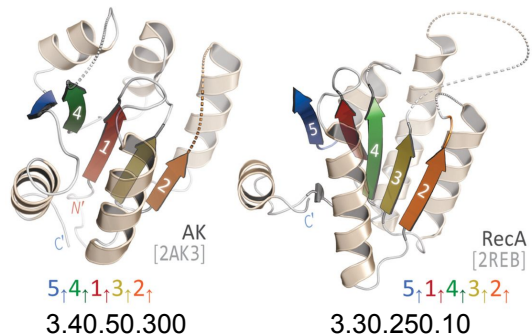
5. Homologous Superfamily



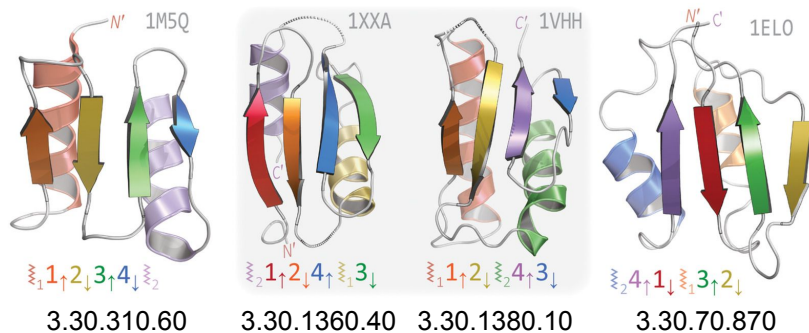
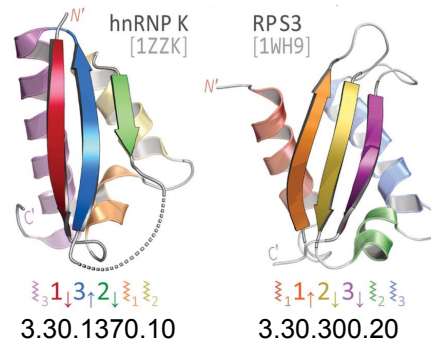
Evolutionarily relationships via sequence

Other Potential 'Urfolds'

P-loop NTPases

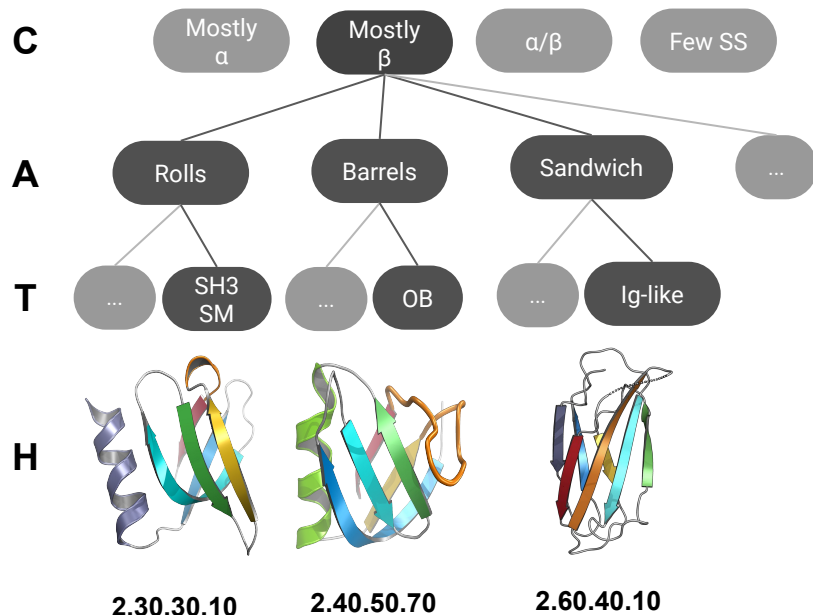


KH Domains

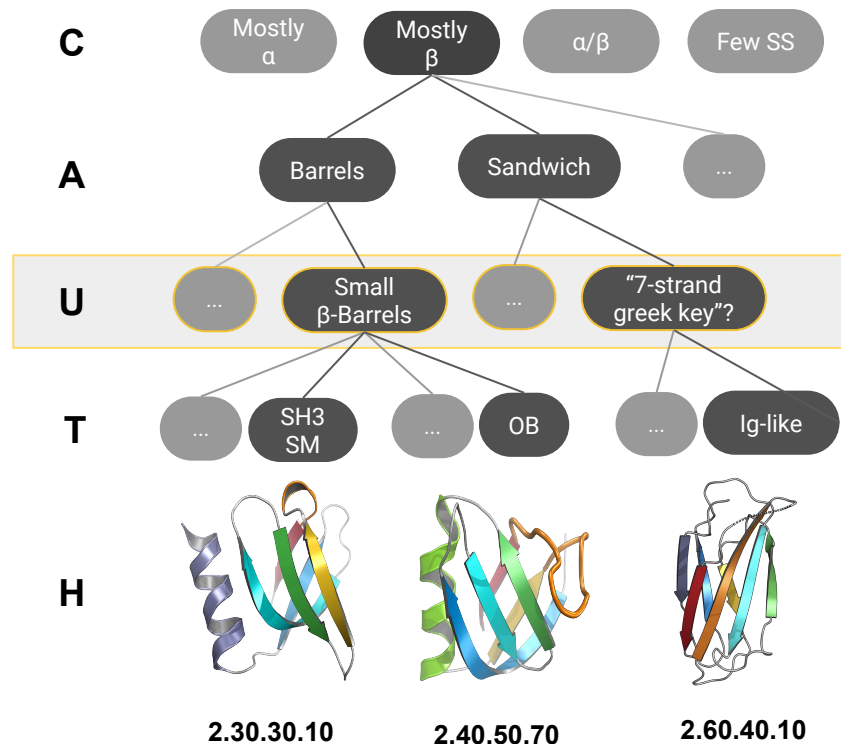


A Different View of Clustering Relationships

Current Clustering



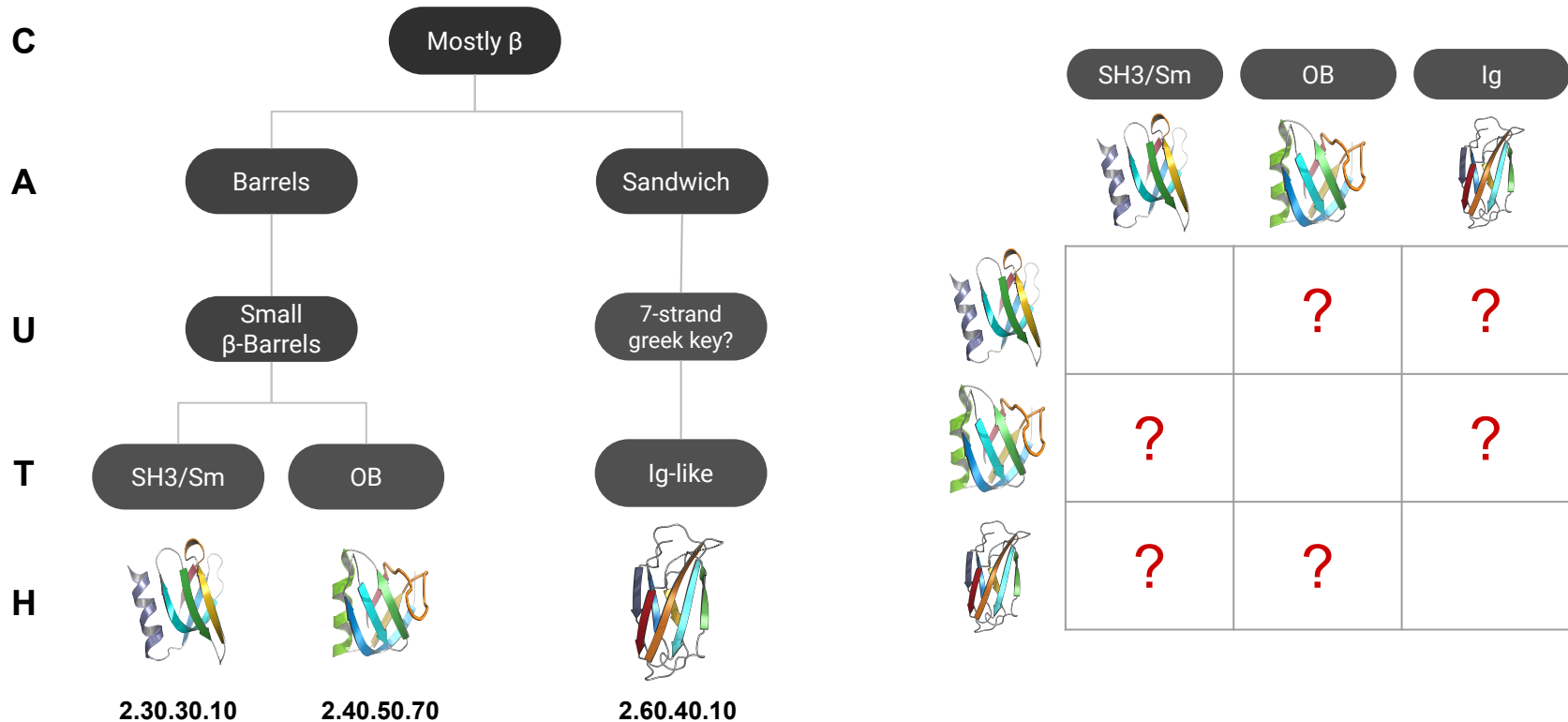
Potential Clustering



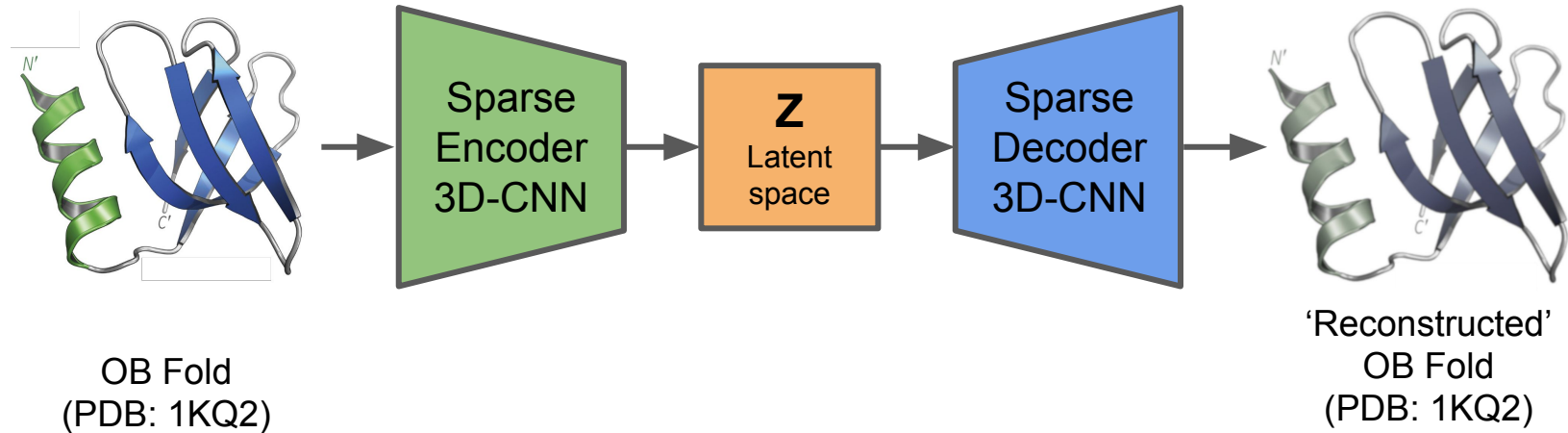
DeepUrfold

Can we learn local substructures of biophysical properties and geometry that bridge 'gaps' in hierarchical classification systems such as CATH, e.g. the Urfold ?

Create Similarity Metric for Hierarchical Clustering

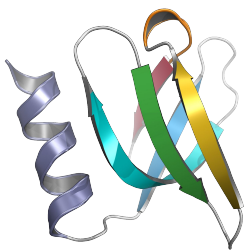


Overall Model: Reconstruct CATH domain structures for one homologous superfamily

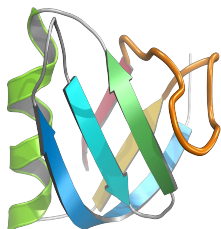


Objective: Create New Similarity Metric

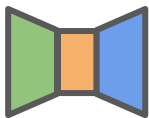
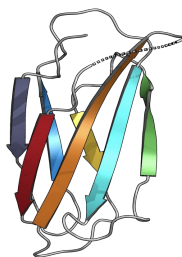
SH3/Sm



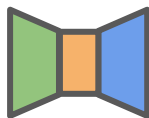
OB



Ig



SH3/Sm DNN Model



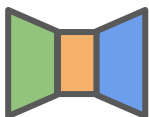
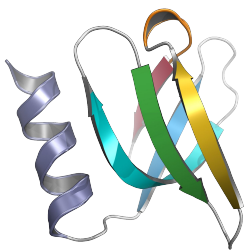
OB DNN Model



Ig DNN Model

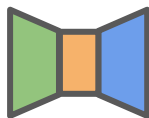
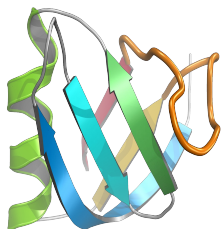
Objective: Create New Similarity Metric

SH3/Sm



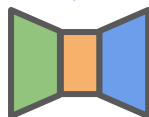
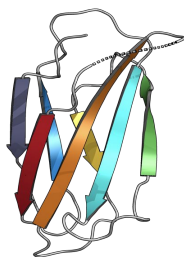
SH3/Sm DNN Model

OB



OB DNN Model

Ig

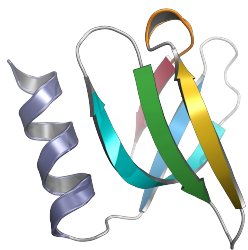


Ig DNN Model

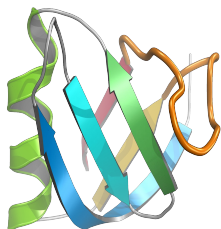
1. **Train** one model for each superfamily

Objective: Create New Similarity Metric

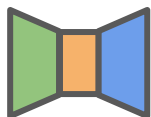
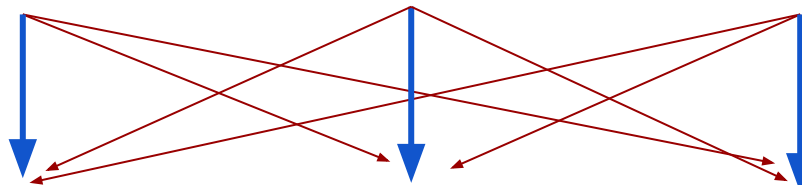
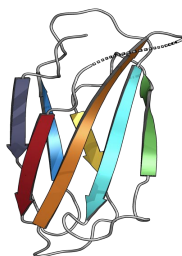
SH3/Sm



OB



Ig



SH3/Sm DNN Model



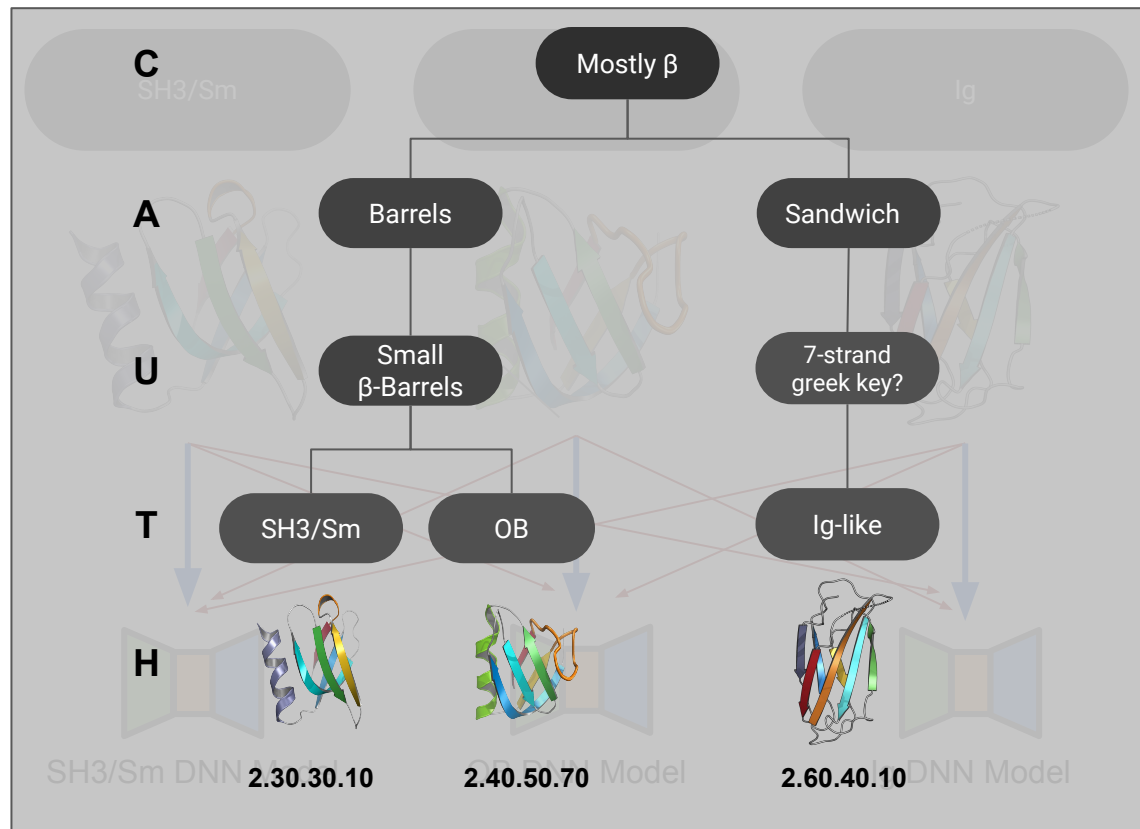
OB DNN Model



Ig DNN Model

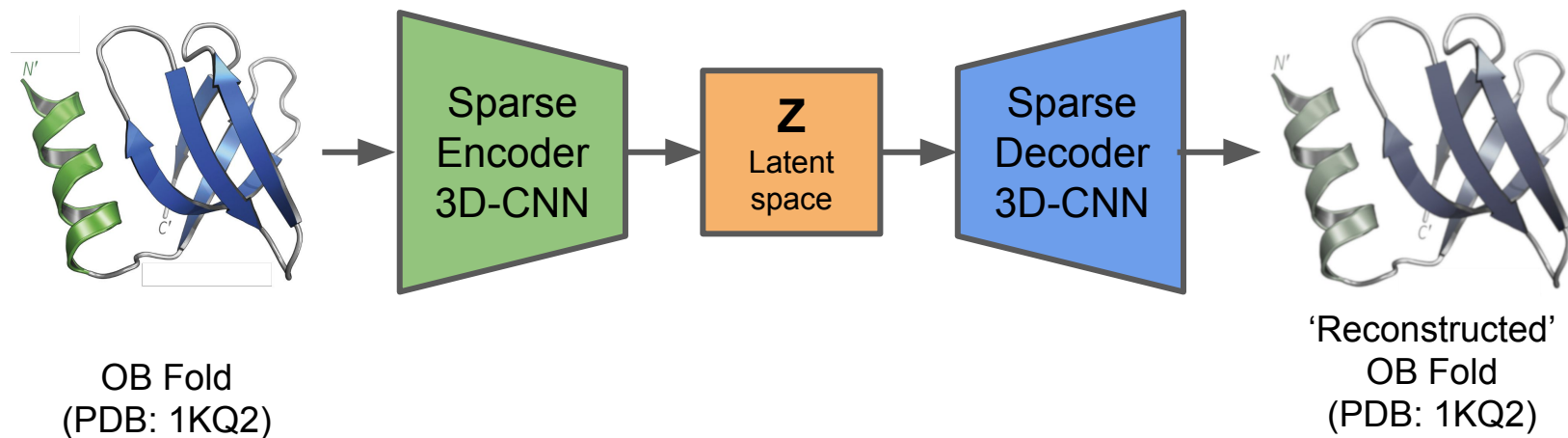
1. **Train** one model for each superfamily
2. **Subject** superfamily representatives to *all other* superfamily models

Objective: Create New Similarity Metric

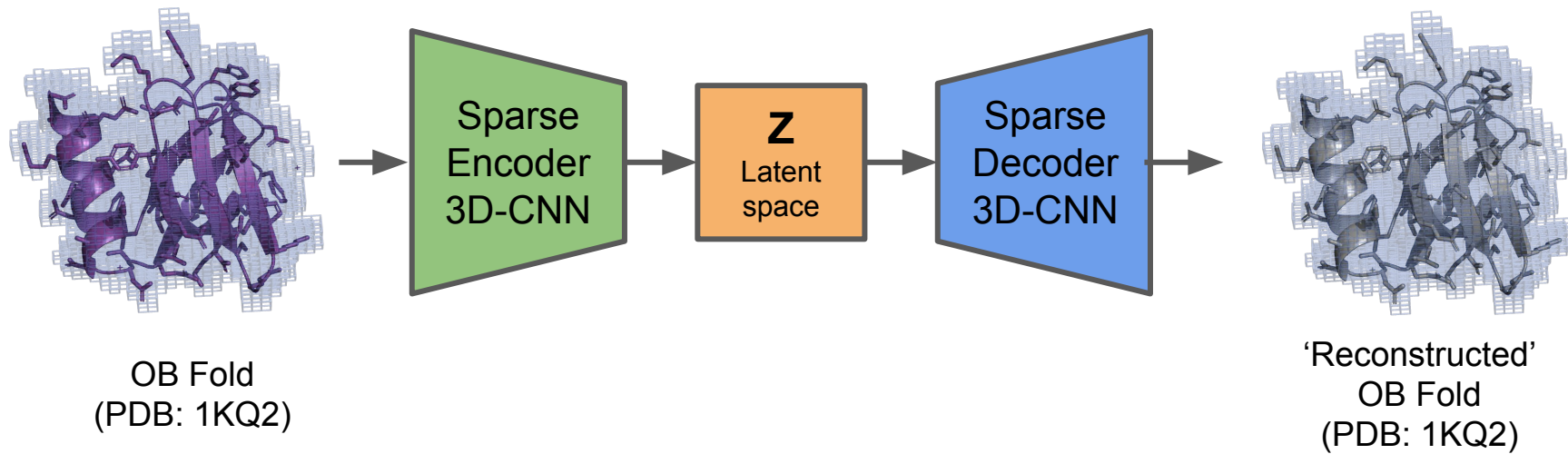


1. **Train** one model for each superfamily
2. **Subject** superfamily representatives to *all other* superfamily models
3. Create a hierarchical clustering with DNN scores: MSE, AUC

Protein Representation

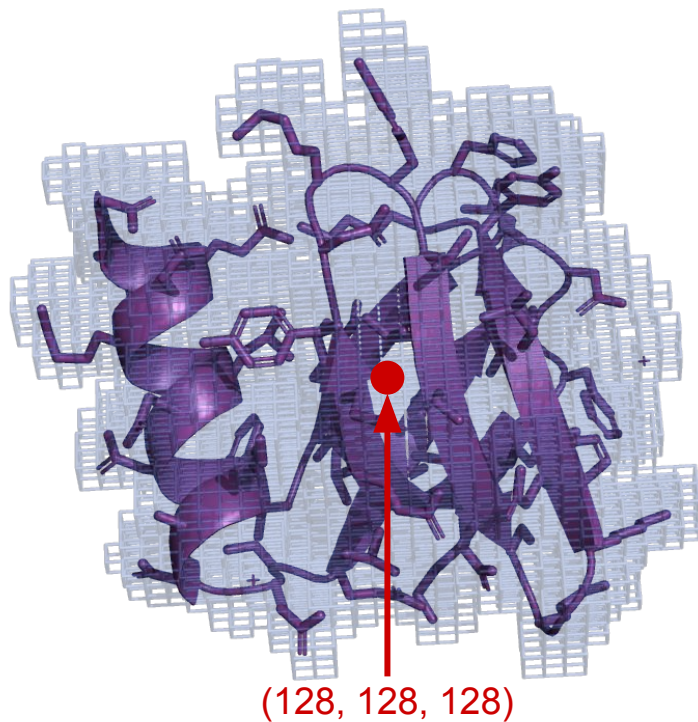


Protein Representation

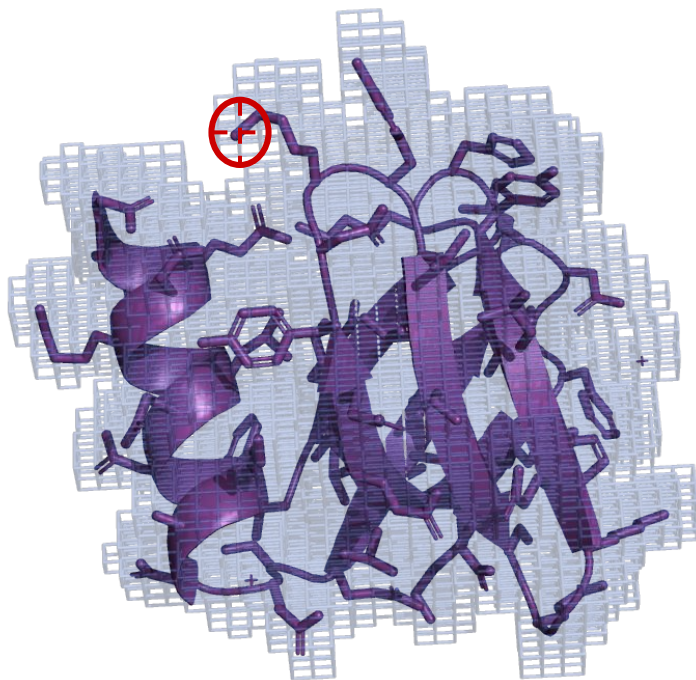


Representation: 3D Image, Voxel Space

- Protein centered in 256^3 Å volume

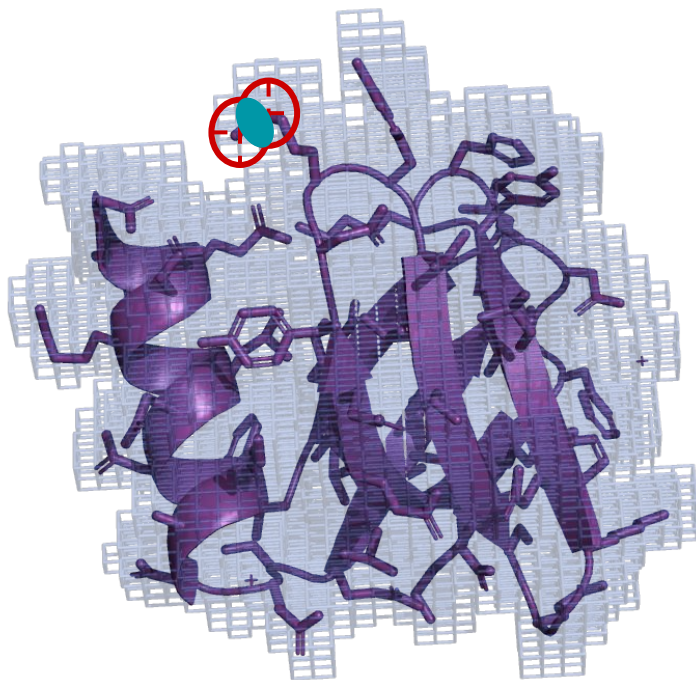


Representation: 3D Image, Voxel Space



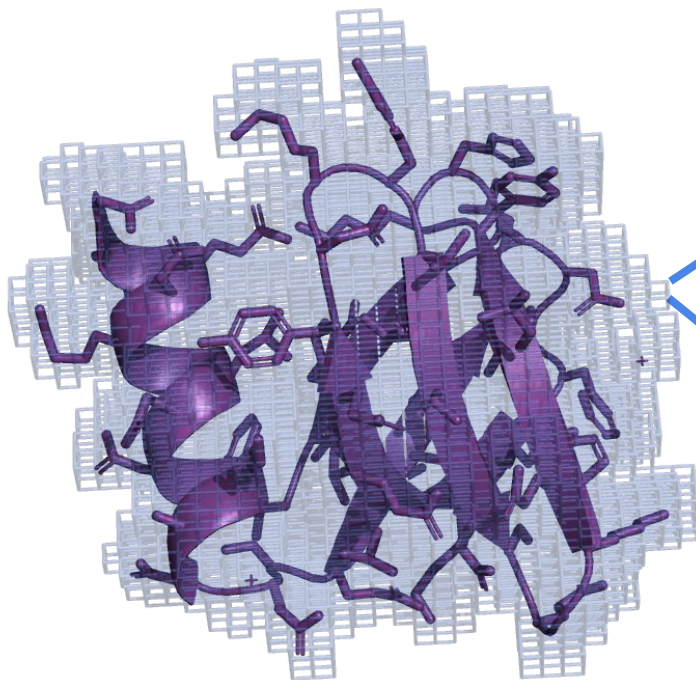
- Protein centered in 256^3 Å volume
- Van der Waals Spheres around each atom are discretized to fit 1^3 Å voxels using a KDTree
 - No need to annotate all voxels because most volume is sparse
 - Each voxel within an atomic sphere receives same set of features

Representation: 3D Image, Voxel Space



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

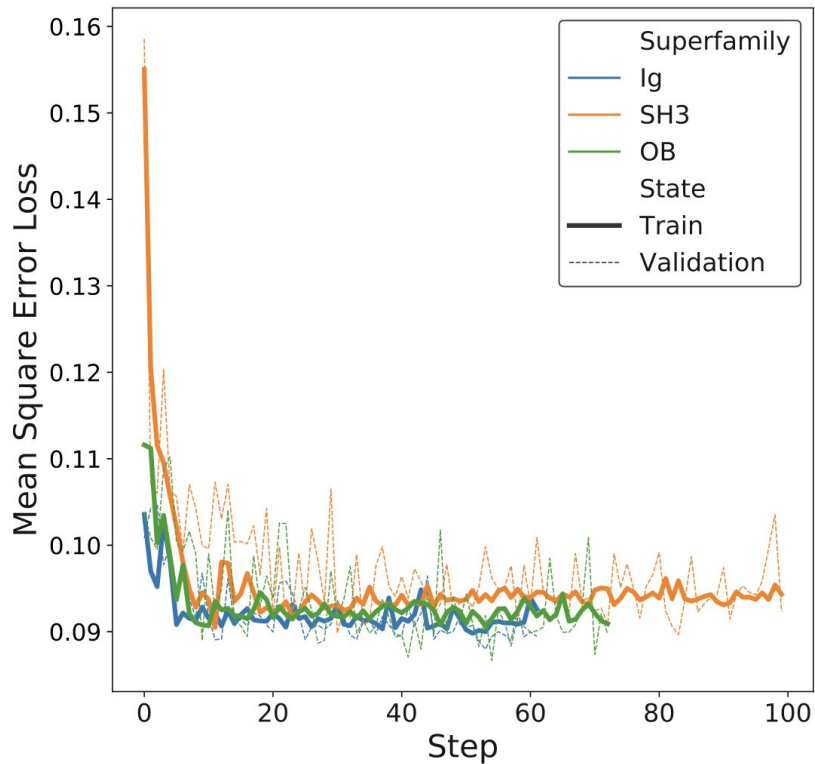
Representation: Atom-based Physicochemical Features



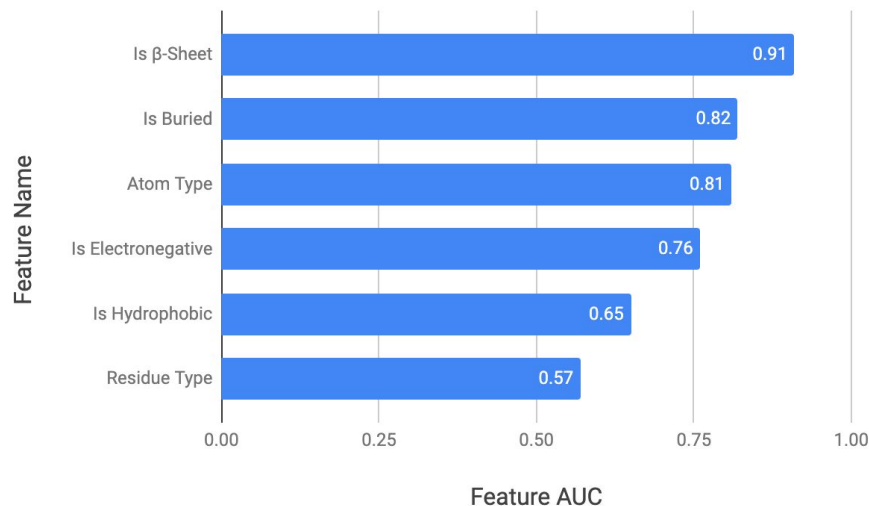
Feature Type	# of Boolean Features
Atom Type	9
Residue Type	20
SS	2
Accessibility	4
Is Hydrophobic	1
Is Positively Charged	1
Is Electronegative	1

Training Ig, SH3 & OB autoencoder models

- Learning rate of 0.01
- Batch size of 16
- Adam optimizer



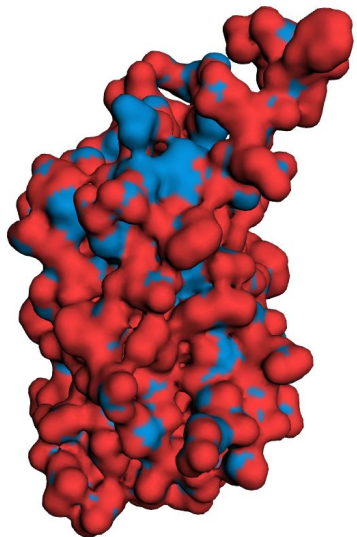
Feature Selection of **Ig** model



- We calculated the AUC for each feature in each voxel, averaging for all samples
- Secondary structure, atom type, and electrostatic potential are all more important features
- Residue type is not as important

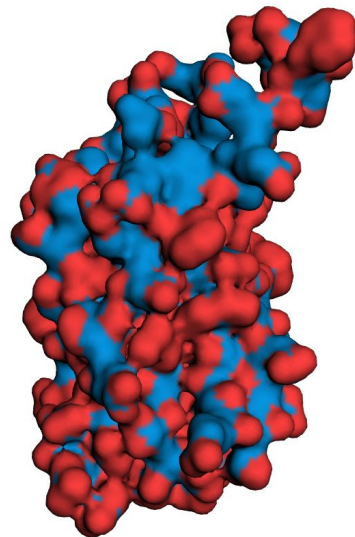
Feature Reconstruction: Electrostatic Potential (Epot)

Original







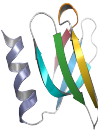

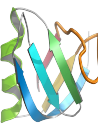


Reconstruction

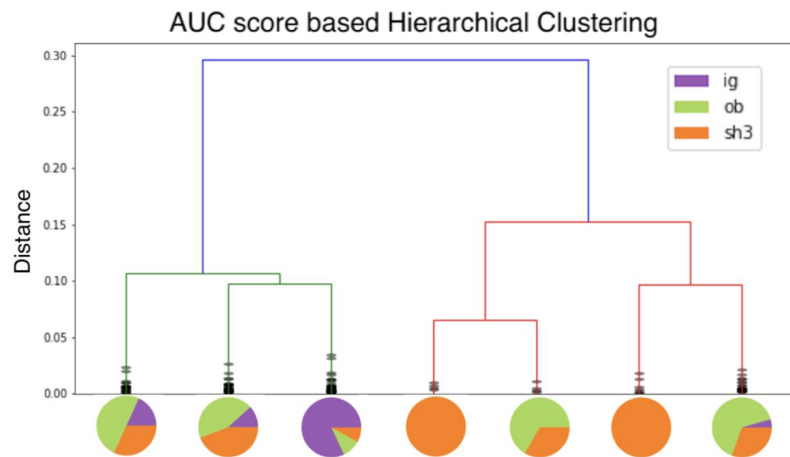
MSE = 0.0535



$E_{\text{pot}} < 0$
 $E_{\text{pot}} \geq 0$

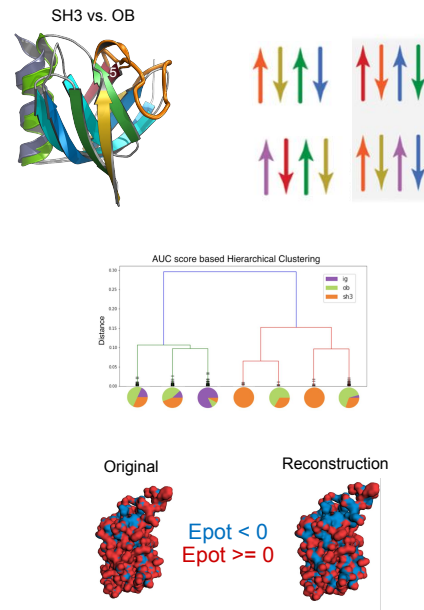
Create hierarchy based on our new distance metrics

Structures NN Models	SH3/Sm	OB	Ig
			
 	0.879 0.098	0.878 0.095	0.875 0.095
 	0.874 0.100	0.885 0.098	0.882 0.098
 	0.795 0.255	0.798 0.260	0.814 0.264
	AUC	MSE	



Conclusions

- A potential entity called the ‘Urfold’ may exist between Architecture and Topology to represent 3D architectural similarity despite topological variability
- We developed a new method to find relationships between superfamilies using 3D-CNNs
- 3D-CNNs can learn discriminatory biophysical properties and geometries for different superfamilies
- We plan to use this framework to scan the protein universe to find other Urfolds



Acknowledgements

Bourne Lab

Phil Bourne
Cam Mura
Zheng Zhao
Stella Veretnik
Menuka Jaiswal
Saad Saleem
Kwon Yonghyeon
Abby Newbury
Niraja Bohidar



Funding: Presidential
Fellowship in Data
Science program

Thanks!

If you have any questions, please come to the Q & A portion:

Wednesday, July 15th @ 10:40 AM - 11:00 AM EDT

Otherwise contact us through email, twitter, etc:

- Eli Draizen - ed4bu@virginia.edu; <http://edraizen.github.com>
- Cam Mura - cmura@virginia.edu
- Phil Bourne - peb6a@virginia.edu

Find these slide online: doi:10.5281/zenodo.3909755