



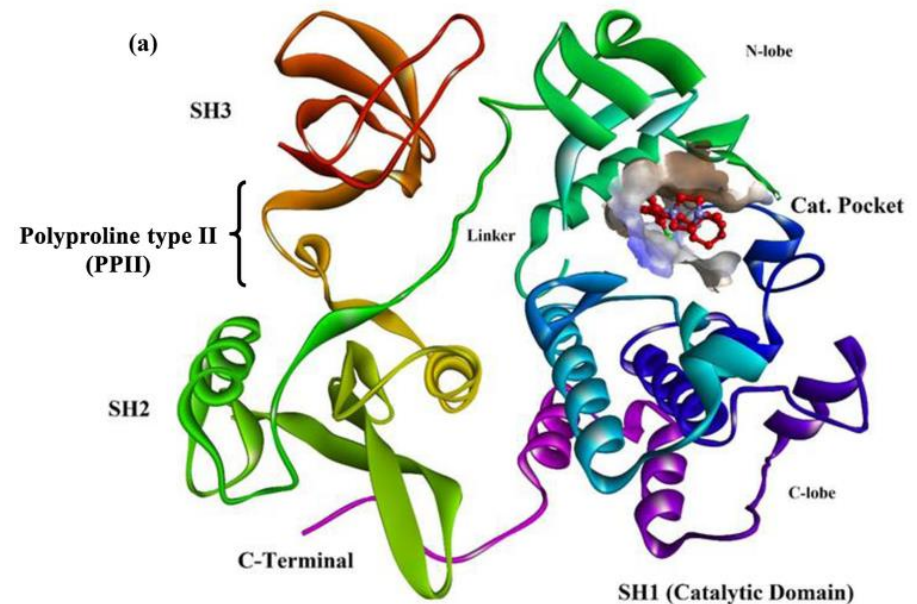
Unsupervised 3D Protein Domain Identification

*3D STRUCTURE-BASED ALGORITHM FOR PREDICTING PROTEIN DOMAINS VIA
LOUVAIN AND TWO-STAGE SPECTRAL CLUSTERING METHODS.*

CHLOE GUERRERO | CS584

Protein domains

- Protein domain – a distinct, conserved functional and/or structural unit within a protein that has a unique and well-defined three-dimensional fold



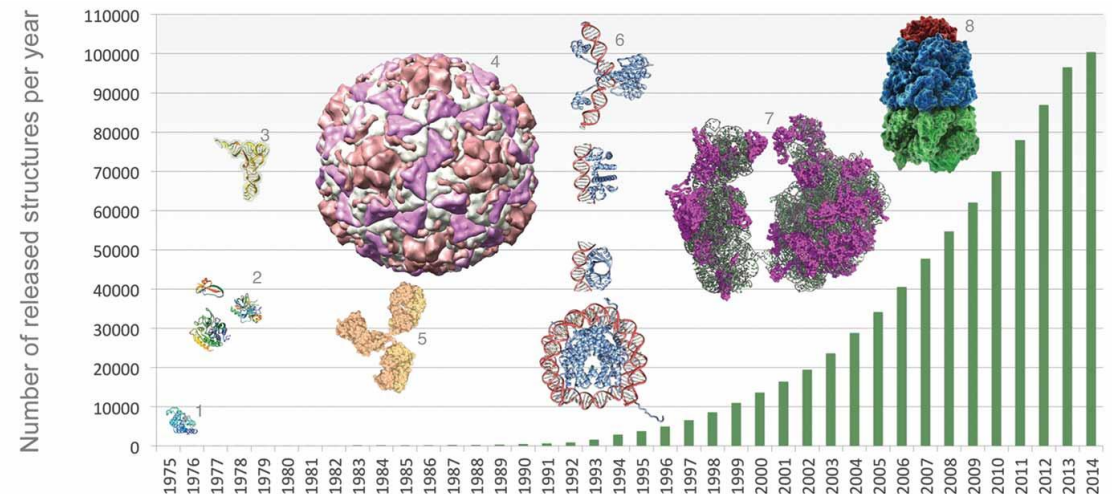
Src kinase architecture. Ribbon diagram of the multi-domain architecture of the Src tyrosine kinase.

Pfam – The Sequence-Structure Gap

- ▶ Pfam is the current status quo for evolutionary, sequence-based domain classification
 - ▶ **Sequence homology-based**
- ▶ Sequence dissimilarity does not guarantee structural or functional dissimilarity

Pfam

RCSB PDB
PROTEIN DATA BANK



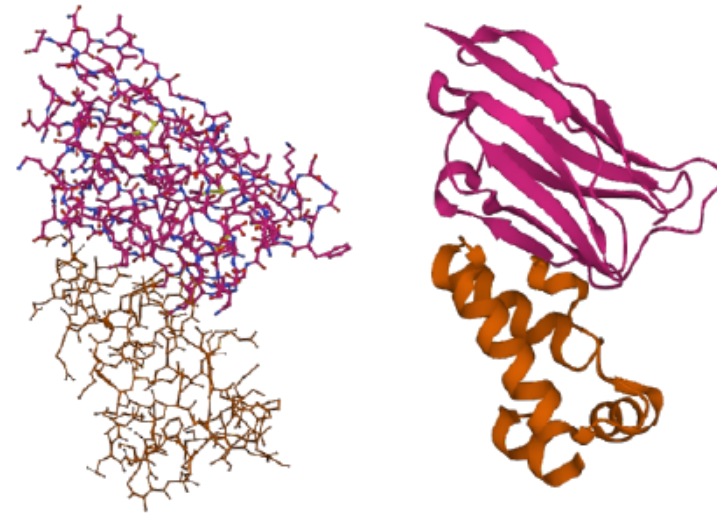
Proteins as Topographical Networks

We need to **translate** the 3D
structure of proteins into
mathematical graph.

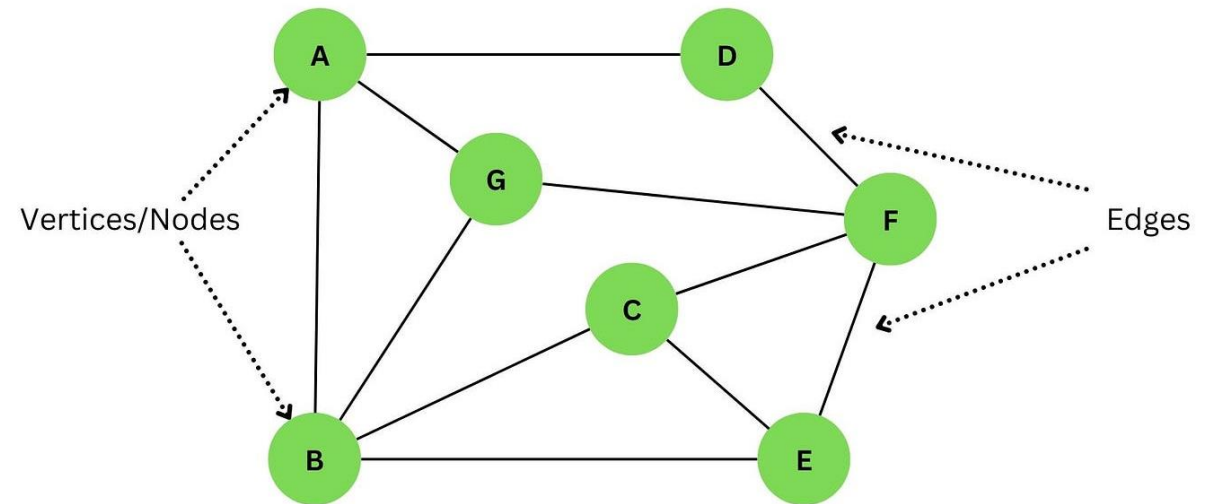
$$G = (V, E)$$

Nodes (V): amino acid residues

Edges (E): connections based on
proximity

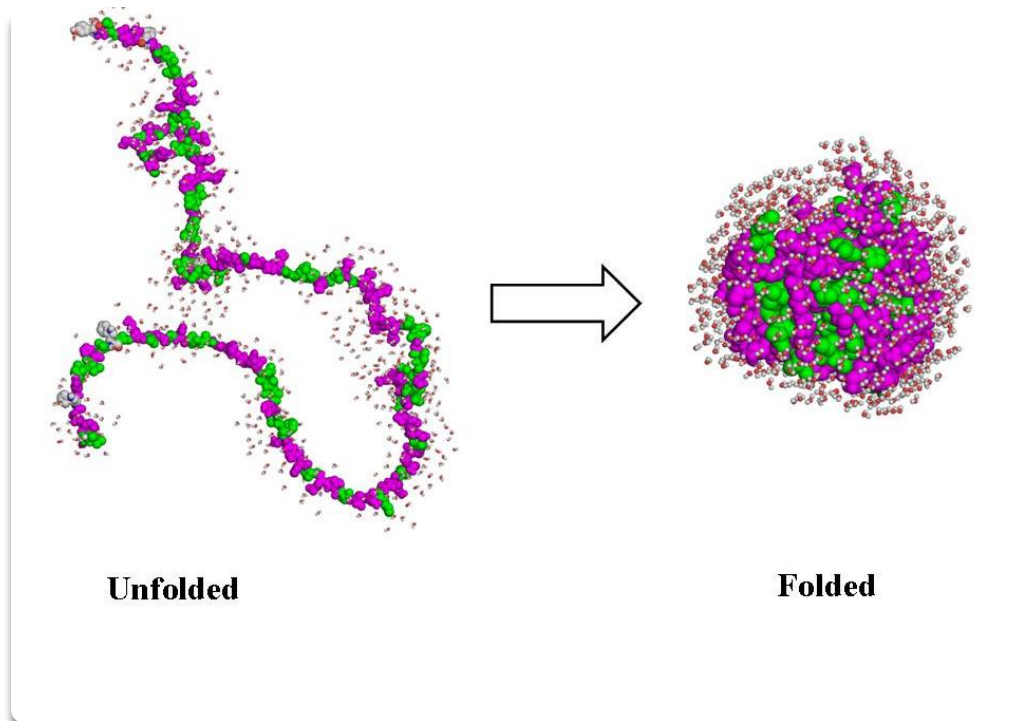


HIV-1 capsid protein C-terminal domain (146- 220) in complex with a camelid VHH.



Graph as data structure.

Key Objective



- To accurately identify protein domains using only structural topology without being informed of sequence homology

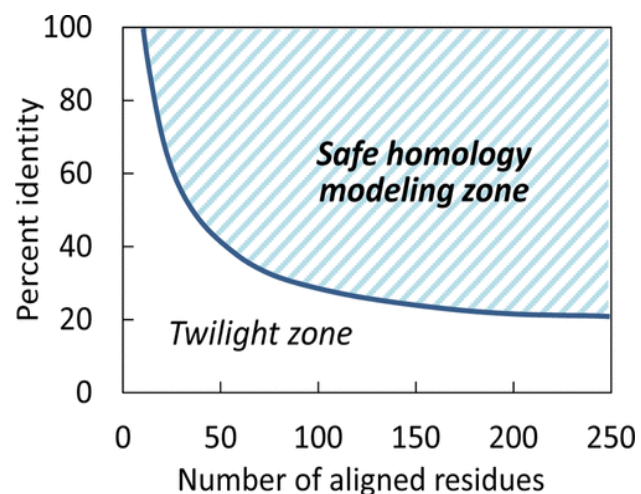
Protein Data Bank (PDB) File Format:

		Amino Acid		Chain name		Sequence Number			-----Coordinates-----		
		Element							X	Y	Z
ATOM	1	N	ASP	L	1				4.060	7.307	5.186
ATOM	2	CA	ASP	L	1				4.042	7.776	6.553
ATOM	3	C	ASP	L	1				2.668	8.426	6.644
ATOM	4	O	ASP	L	1				1.987	8.438	5.606
ATOM	5	CB	ASP	L	1				5.090	8.827	6.797
ATOM	6	CG	ASP	L	1				6.338	8.761	5.929
ATOM	7	OD1	ASP	L	1				6.576	9.758	5.241
ATOM	8	OD2	ASP	L	1				7.065	7.759	5.948

\\
Element position within amino acid

Data Acquisition & Preprocessing

- ▶ Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB)
- ▶ **Twilight zone** – range of sequence identities below 30% at which sequence homology cannot effectively identify true homologs



- ▶ **Prefiltration:** length, completeness, resolution
 - ▶ $50 \leq \text{length} \leq 2000$ resid.; $\leq 5\%$ unknown resid.; resolution ≤ 3.0 Å
- ▶ **Cluster representatives** were selected (94 protein sequences) and stratified by:
 - ▶ Length: small (50-299 res.); medium (300-799 res.); large (800-2000 res)
 - ▶ Domain Count: 1; 2; 3; 4+

Feature Extraction – Distance Matrix

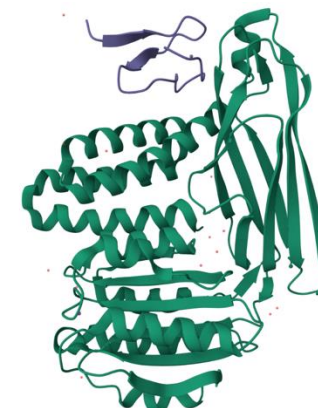
Objective: To translate 3D spatial coordinate data into a distance matrix

Simplification of protein chains:

- ▶ residue $\cong C_{\alpha}$
- ▶ pairwise Euclidian distances between C_{α} 's within a given chain

$$D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

(\mathbf{x}_i represents the C_{α} coords. of the i^{th} residue in a give chain)

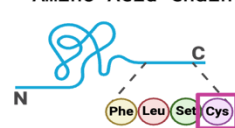


pd_b_00008s05: ArnAB complex an archaeal ortholog of the Sec23/24 core motif

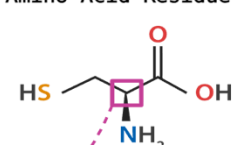
PDB File (Macromolecule)

Atom	Element	Chain name	Residue Number	Coordinates
ATOM 1	C	ARP L	1	4.060 7.307 5.186
ATOM 2	H	ARP L	1	4.042 7.276 6.153
ATOM 3	C	ARP L	1	2.668 6.426 6.444
ATOM 4	H	ARP L	1	2.897 6.426 6.666
ATOM 5	C	ARP L	1	5.090 6.827 6.797
ATOM 6	H	ARP L	1	6.139 6.761 5.929
ATOM 7	O	ARP L	1	6.576 9.758 5.241
ATOM 8	O	ARP L	1	7.065 7.759 5.948

Amino Acid Chain



Amino Acid Residue



C_{α} Coordinates (x, y, z)



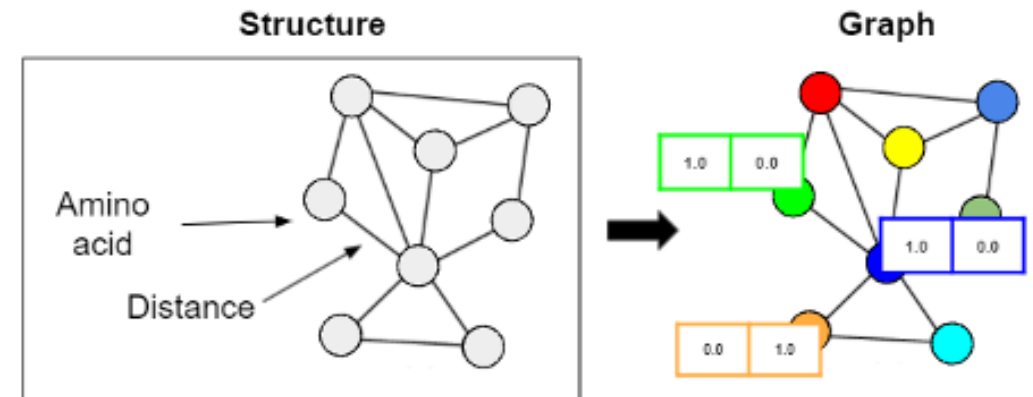
Row	V0	V1	V2	V3	V4	V5	V6	V7
0	3.8167291790175634	3.8167291790175634	6.182624970056447	6.731779135505274	11.682549043671529	14.829148311430044	17.016825188767563	20.300975692320306
1	6.182624970056447	6.182624970056447	3.828998252108025	3.828998252108025	10.14383046327181	10.14383046327181	19.323116359696964	19.323116359696964
2	6.731779135505274	6.731779135505274	3.788781638037751	0.0	6.849578319700825	10.072823254287538	13.119928748027794	15.899830326888024
3	11.682549043671529	11.682549043671529	3.788781638037751	0.0	3.788781638037751	6.742833714239418	9.39833020116882	12.541473232102582
4	14.829148311430044	14.829148311430044	10.072823254287538	10.072823254287538	6.742833714239418	6.742833714239418	6.37766579881224	6.37766579881224
5	17.016825188767563	17.016825188767563	15.899830326888024	15.899830326888024	9.39833020116882	9.39833020116882	3.787628891554267	3.787628891554267
6	20.300975692320306	20.300975692320306	19.323116359696964	19.323116359696964	12.541473232102582	12.541473232102582	6.37766579881224	6.37766579881224
7	23.592075692320306	23.592075692320306	18.3002103231432	18.3002103231432	11.9545232319507	11.9545232319507	3.787628891554267	3.787628891554267
8	26.883175692320306	26.883175692320306	17.2447682140701465	17.2447682140701465	10.30094817369698	10.30094817369698	3.787628891554267	3.787628891554267
9	29.883175692320306	29.883175692320306	16.08448695320525	16.08448695320525	9.39833020116882	9.39833020116882	3.787628891554267	3.787628891554267
10	32.883175692320306	32.883175692320306	15.000000000000001	15.000000000000001	8.498989898989899	8.498989898989899	3.787628891554267	3.787628891554267
11	35.883175692320306	35.883175692320306	14.000000000000001	14.000000000000001	7.618750000000001	7.618750000000001	3.787628891554267	3.787628891554267
12	38.883175692320306	38.883175692320306	13.000000000000001	13.000000000000001	6.754515000000001	6.754515000000001	3.787628891554267	3.787628891554267
13	41.883175692320306	41.883175692320306	12.000000000000001	12.000000000000001	5.900000000000001	5.900000000000001	3.787628891554267	3.787628891554267
14	44.883175692320306	44.883175692320306	11.000000000000001	11.000000000000001	5.050000000000001	5.050000000000001	3.787628891554267	3.787628891554267
15	47.883175692320306	47.883175692320306	10.000000000000001	10.000000000000001	4.200000000000001	4.200000000000001	3.787628891554267	3.787628891554267
16	50.883175692320306	50.883175692320306	9.000000000000001	9.000000000000001	3.350000000000001	3.350000000000001	3.787628891554267	3.787628891554267
17	53.883175692320306	53.883175692320306	8.000000000000001	8.000000000000001	2.500000000000001	2.500000000000001	3.787628891554267	3.787628891554267
18	56.883175692320306	56.883175692320306	7.000000000000001	7.000000000000001	1.650000000000001	1.650000000000001	3.787628891554267	3.787628891554267
19	59.883175692320306	59.883175692320306	6.000000000000001	6.000000000000001	0.800000000000001	0.800000000000001	3.787628891554267	3.787628891554267
20	62.883175692320306	62.883175692320306	5.000000000000001	5.000000000000001	0.0	0.0	3.787628891554267	3.787628891554267
21	65.883175692320306	65.883175692320306	4.000000000000001	4.000000000000001	0.0	0.0	3.787628891554267	3.787628891554267
22	68.883175692320306	68.883175692320306	3.000000000000001	3.000000000000001	0.0	0.0	3.787628891554267	3.787628891554267
23	71.883175692320306	71.883175692320306	2.000000000000001	2.000000000000001	0.0	0.0	3.787628891554267	3.787628891554267
24	74.883175692320306	74.883175692320306	1.000000000000001	1.000000000000001	0.0	0.0	3.787628891554267	3.787628891554267
25	77.883175692320306	77.883175692320306	0.0	0.0	0.0	0.0	3.787628891554267	3.787628891554267
26	80.883175692320306	80.883175692320306	0.0	0.0	0.0	0.0	3.787628891554267	3.787628891554267

Feature Extraction – Graph Construction

Objective: To translate a distance matrix into a network graph (via NetworkX)

$G = (V, E)$

- ▶ Edges → **adaptive k -Nearest Neighbors** (k -NN) algorithm
 - ▶ k -NN introduces sparsity into graph; noise reduction ($k = 10$)
 - ▶ adaptive to enforce connectivity



Inverse-Distance Edge Weighting:
$$w_{ij} = \frac{1}{1 + D_{ij}}$$

<https://curj.caltech.edu/2022/06/22/identifying-optimal-proteins-by-their-structure-using-graph-neural-networks/>

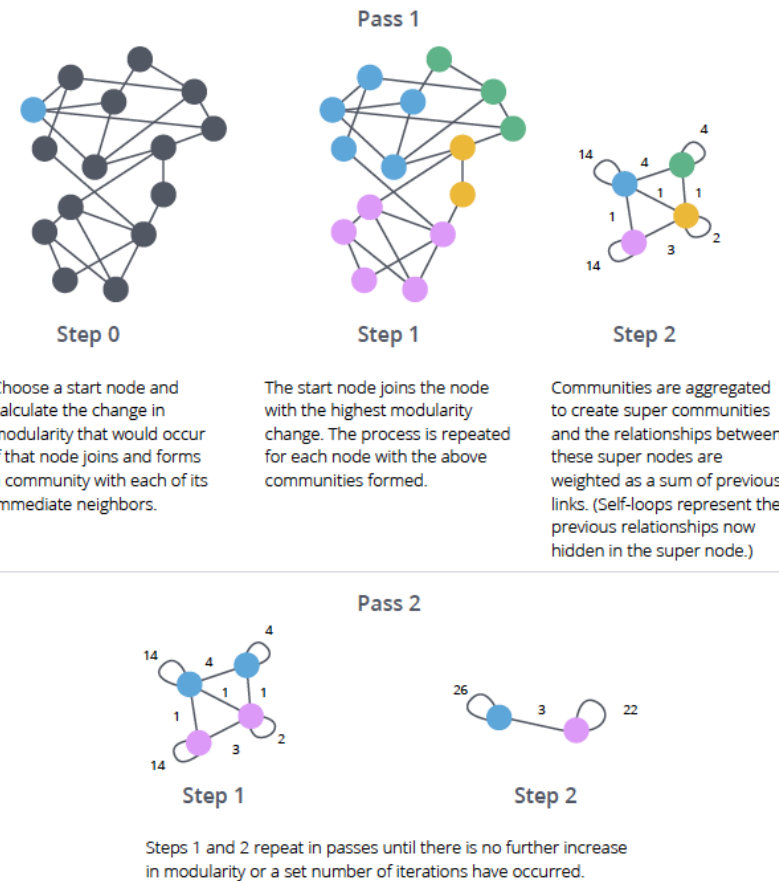
Louvain clustering maximizes *modularity scores*.

Modularity (Q) – measure of the quality of a partition

- connectedness of nodes *within* a community vs connectedness *between* communities

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{w_i w_j}{2m} \right] \delta(c_i, c_j)$$

- A_{ij} = edge weight between nodes i and j
- w_i = weighted degree of residue i (sum of all edge weights connected to residue i)
- c_i = assigned community of node i
- m = sum of all edge weights in the graph



Needham, M. (2019) neo4j.

Generating Clustering Predictions

(1) Louvain Community Detection (Unsupervised)

(python-louvain 0.16)

Two-stage spectral clustering addresses the “unknown k ” problem.

Stage 1: Domain Count Estimation

- ▶ Iterative testing of a range of k -values to maximize silhouette score

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- ▶ $a(i) \rightarrow$ mean distance to other nodes within same community
- ▶ $b(i) \rightarrow$ mean distance to other nodes within nearest neighboring community

The k -value for which $S(i)$ is maximized is retained.

**Notable Mentions: Eigengap & Consensus methods*

Stage 2: Spectral Clustering & Manifold Learning

- ▶ Distance matrix $D \rightarrow$ Similarity matrix A

$$A_{ij} = \exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right) \quad (\sigma = \text{scaling parameter; median of all nonzero distances in distance matrix } D)$$

- ▶ Normalized Graph Laplacian calculation

$$L = I - D^{-1/2} A D^{-1/2} \quad (D: \text{Degree matrix (diagonal matrix of row sums).})$$

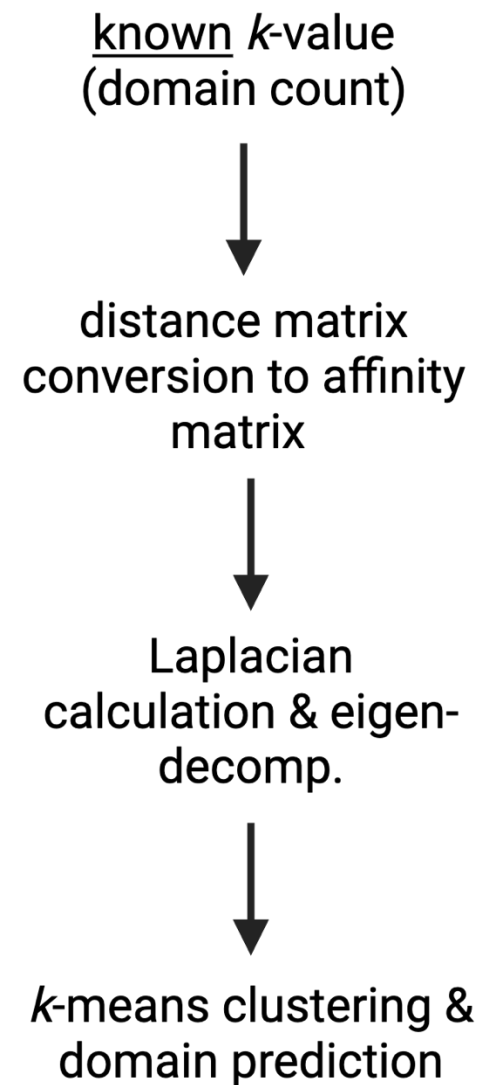
- ▶ An eigen-decomposition on L is performed, and the first k eigenvectors that correspond to the k smallest non-zero eigenvalues are retained
- ▶ Spectral embedding: Given k eigenvectors $\{v_1 \dots v_k\}$, $U = [v_1, v_2, \dots, v_k]$
 - ▶ Each i^{th} row of U represents the transformed coordinates of residue i
- ▶ K-means clustering \rightarrow labels clusters in the new space with predicted domain ID

Generating Clustering Predictions

(2) Two-Stage Spectral Clustering (Unsupervised)

(scikit-learn 1.3.0)

The Benchmarks: "Supervised" Spectral Clustering on Graph Topology



Quantitative Performance Benchmark

Louvain

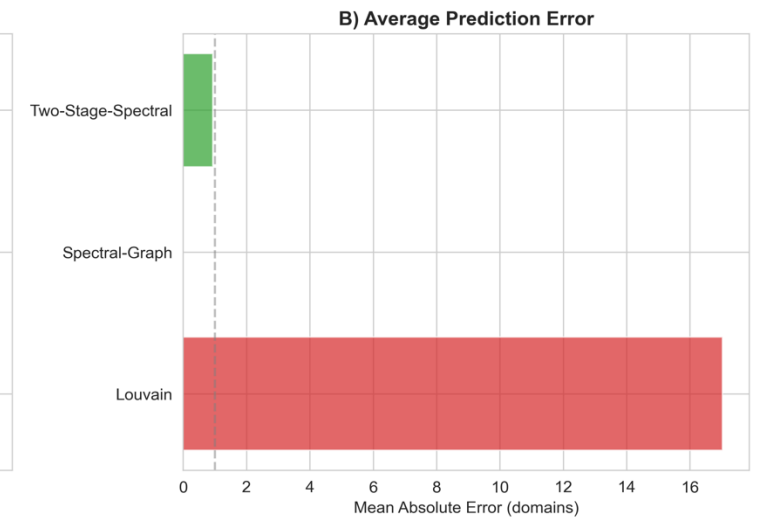
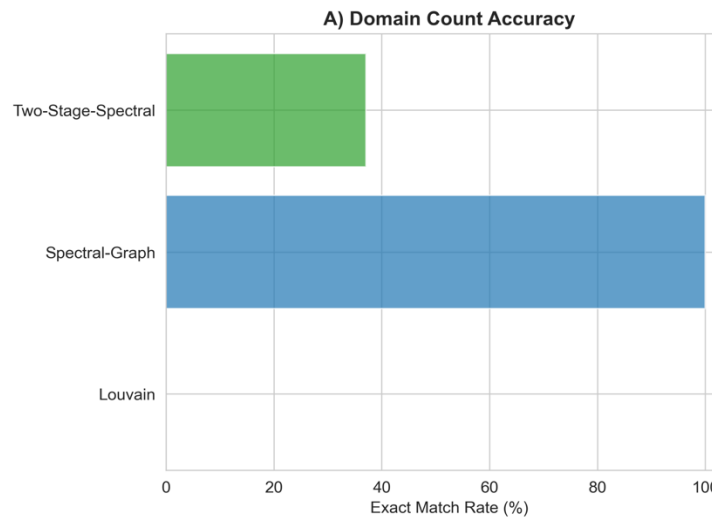
- ▶ Exact Match: 0/94
- ▶ Mean Absolute Error (MAE): 17.02

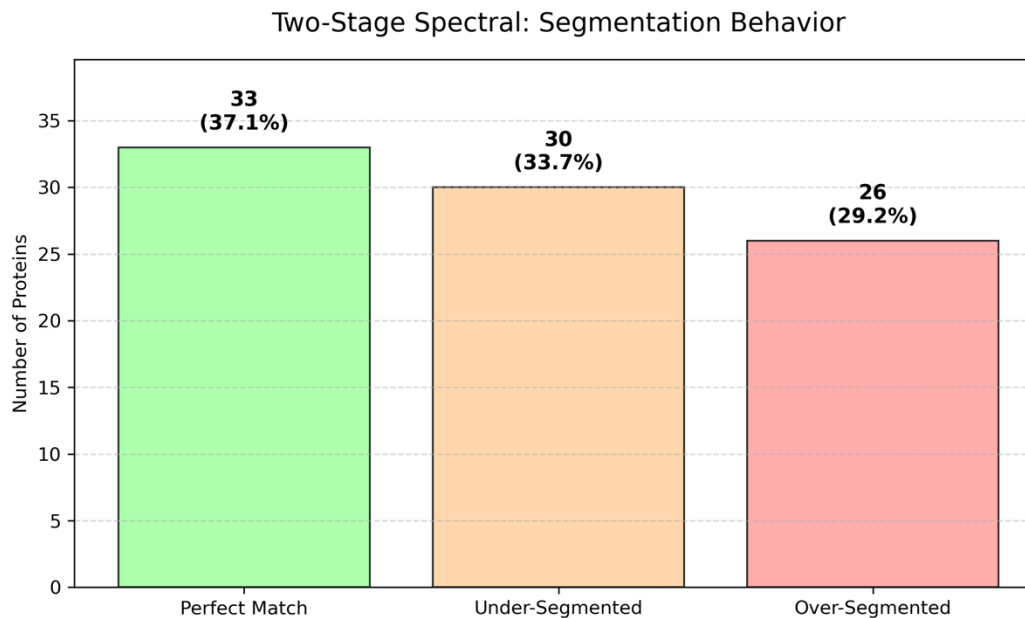
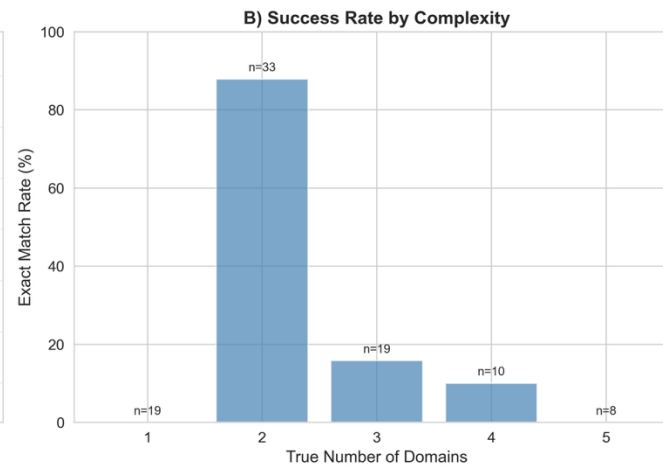
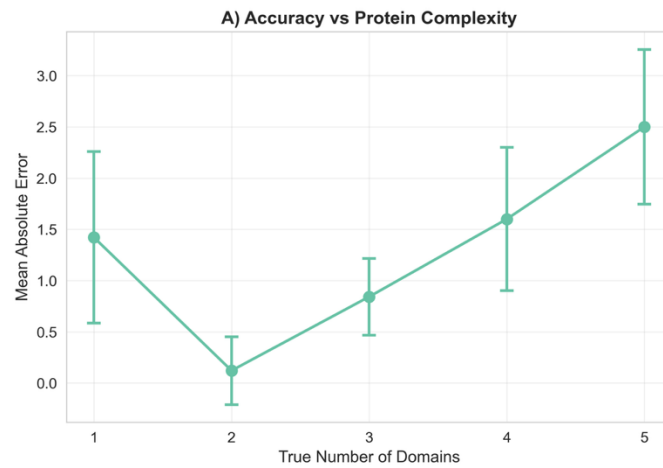
Two-Stage Spectral

- ▶ Exact Match: 33/94
- ▶ Mean Absolute Error (MAE): 0.93

"Supervised" Spectral

- ▶ Exact Match: 89/94
- ▶ Mean Absolute Error (MAE): 0.00





Two-Stage Spectral Clustering & Protein Complexity

Two-Stage Spectral Clustering Hyperparameter Optimization

Grid Search Setup:

- ▶ Training set: 30 proteins (stratified)
- ▶ Test set: 64 proteins
- ▶ 240 parameter combinations tested

Training Performance

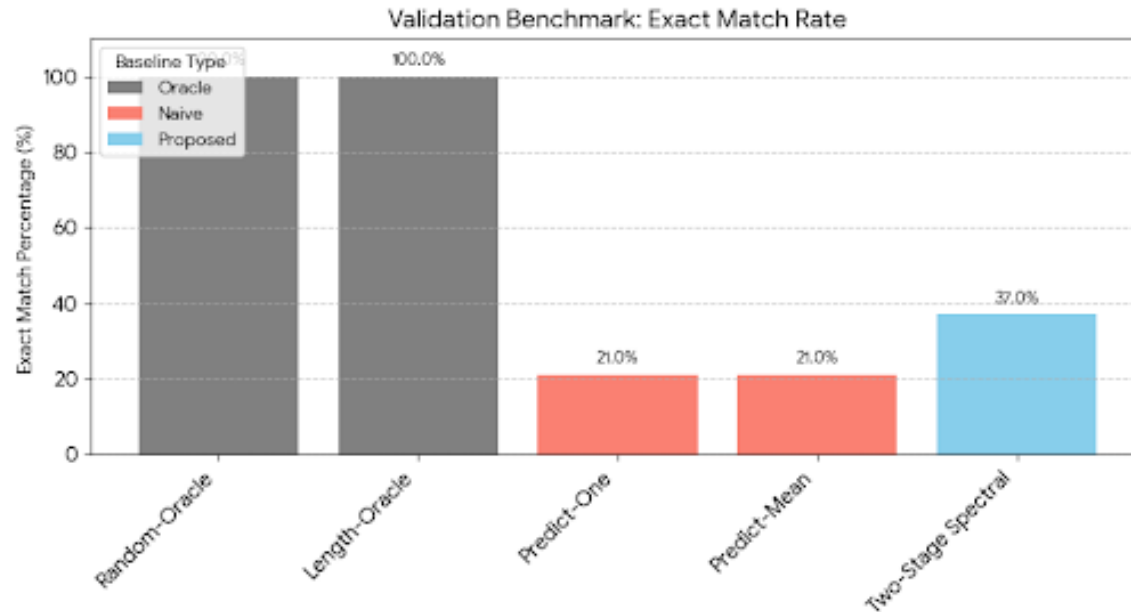
- ▶ mean_error: 1.0
- ▶ exact_match_rate: 0.429

Test Performance

- ▶ mean_error: 0.820
- ▶ exact_match_rate: 0.410

Parameter	Values	Best
Estimation Method	silhouette, eigengap, consensus	eigengap
Sigma Factor	0.5, 1.0, 1.5, 2.0, 2.5	0.5
Max Domains	6, 8, 10, 12	12
K (Graph)	5, 10, 15, 20	5

Random Control Experiments



- ▶ Oracle methods "cheat" by using true domain count; 100% match is meaningless
- ▶ Naive baselines establish the lower bound any real method must beat
- ▶ Two-Stage outperforms naive baselines by about 16% without oracle knowledge

References

1. <https://onlinelibrary.wiley.com/doi/10.1002/cbdv.202300515>
2. <https://www.sciencedirect.com/science/article/pii/S1043661815000055>
3. https://today.ucsd.edu/story/protein_data_bank_archives_its_100000th_molecule_structure
4. <https://ftloscience.com/how-to-interpret-protein-ribbon-diagrams/>
5. <https://medium.com/@nelsonjoseph123/graph-as-a-data-structure-d04db591a0e5>
6. <https://step1.medbullets.com/biochemistry/102094/protein-folding-and-degradation>
7. <https://lammptube.com/2020/04/03/pdb-file-format/>
8. <https://pubs.acs.org/doi/10.1021/acs.jmedchem.6b01453>