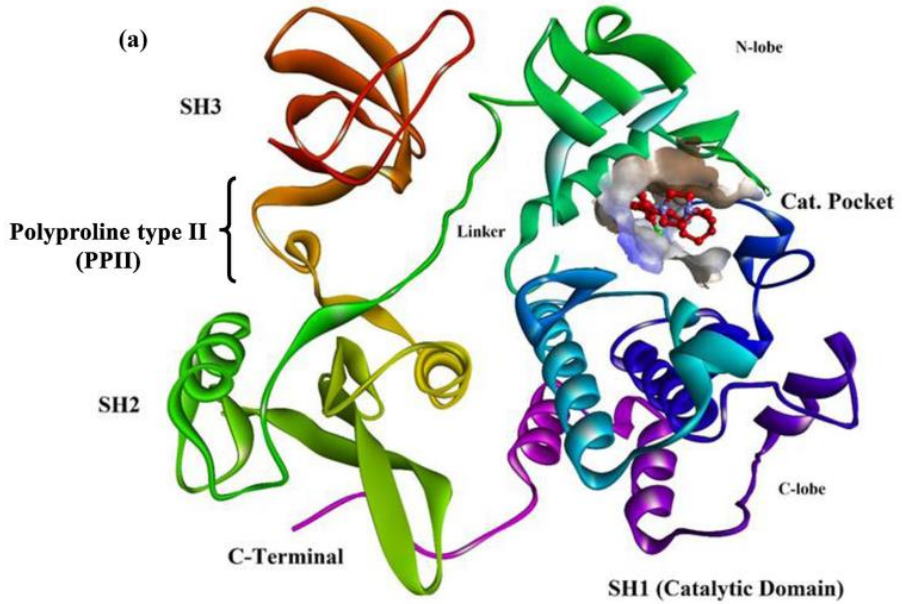
Unsupervised 3D Protein Domain Identification

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

CHLOE GUERRERO | CS584

Protein domains

- **<u>Protein domain</u> – 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.*

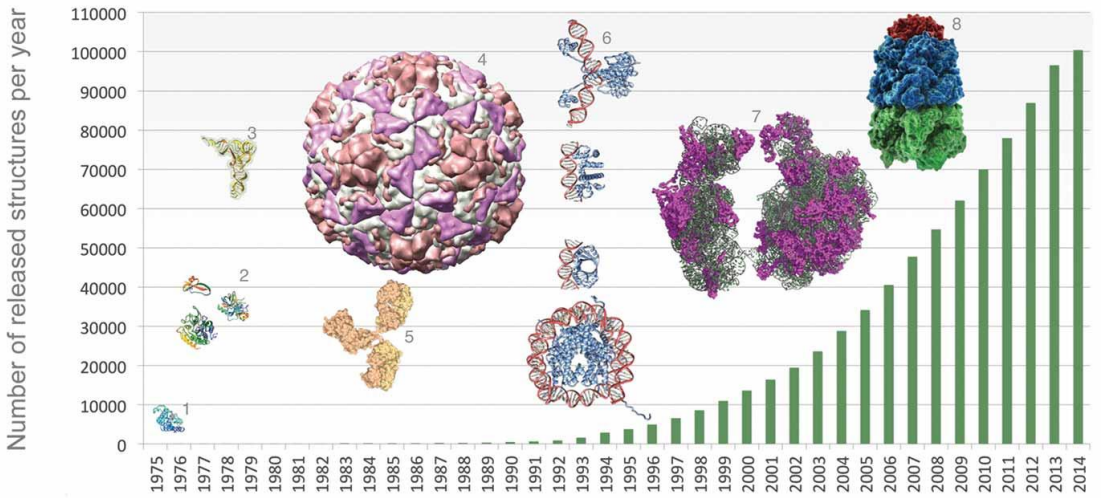
Teli et al. (2023). Chemistry & Biodiversity, 20:e202300515.

Pfam – The Sequence-Structure Gap

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

Pfam

RCSB PDB
PROTEIN DATA BANK



Xverina, J. (2014). UC San Diego Today.

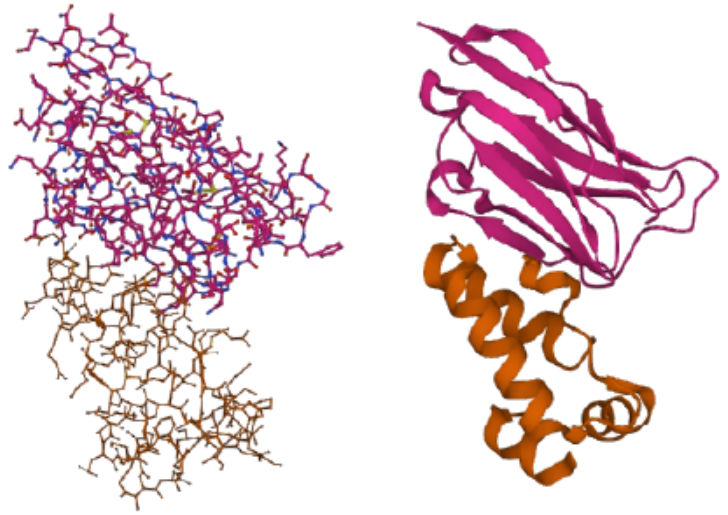
Proteins as Topographical Networks

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

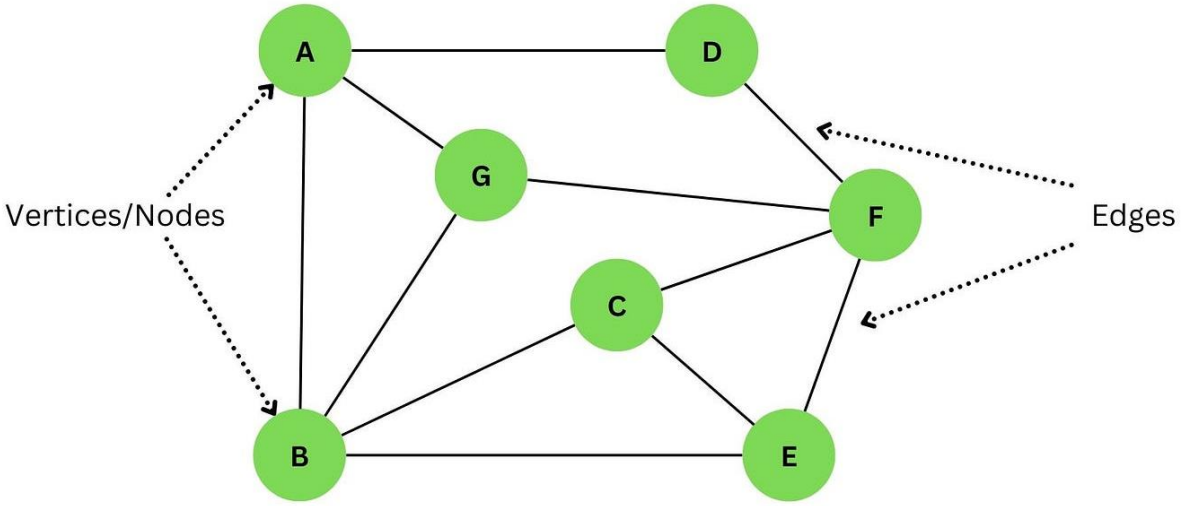
$G = (V, E)$

Nodes (V): amino acid residues

Edges (E): connections based on proximity



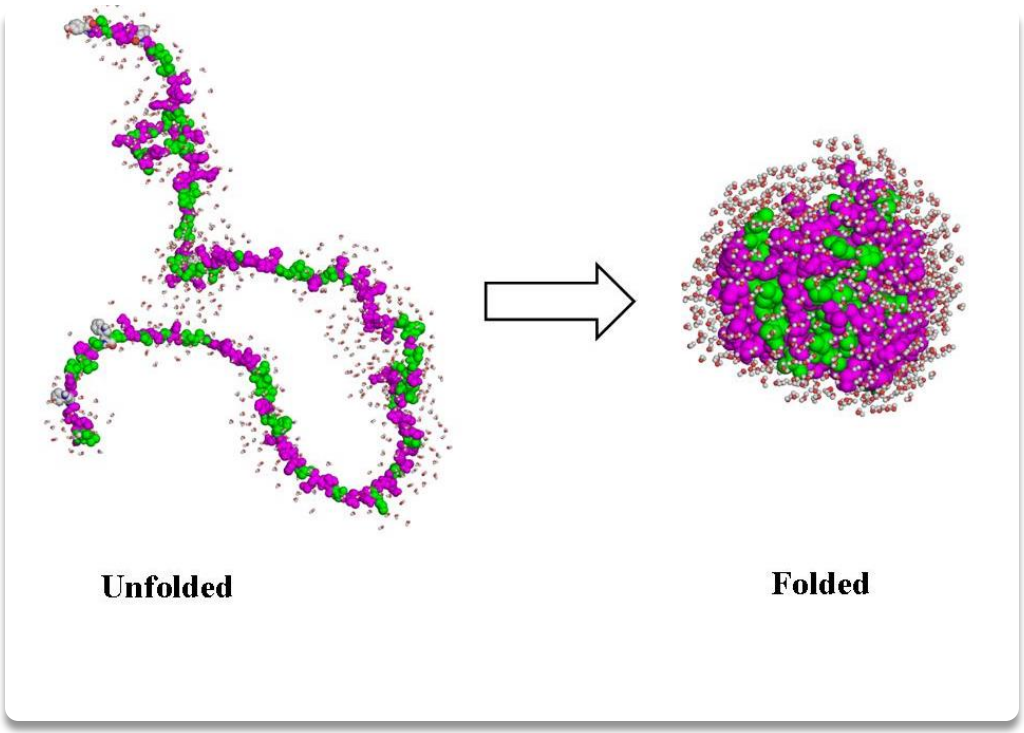
HIV-1 capsid protein C-terminal domain (146- 220) in complex with a camelid VHH.



Graph as data structure.

Lim, S. (2022). FTLOScience.
Nelsonjoseph. (2022) Medium.

Key Objective



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

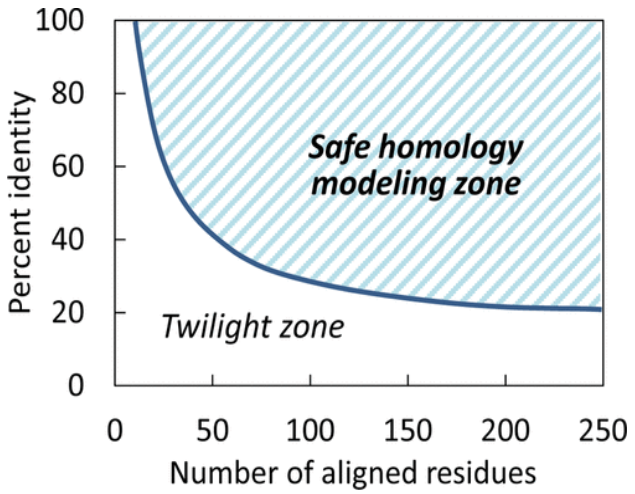
Protein Data Bank (PDB) File Format:

```
                                   Chain name
      Amino Acid                  /   Sequence Number
                   \             /   /
       Element      \           /   /      -----Coordinates-----
                     \   \     /   /       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
```

https://step1.medbullets.com/biochemistry/102094/protein-folding-and-degradation
https://lammpstube.com/2020/04/03/pdb-file-format/

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 ≤ length ≤ 2000 resid.; ≤ 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+

J. Med. Chem. 2018, 61, 1, 1–46

Feature Extraction – Distance Matrix

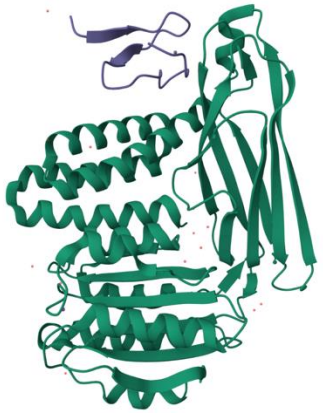
Objective: *To translate 3D spatial coordinate data into a distance matrix*

Simplification of protein chains:

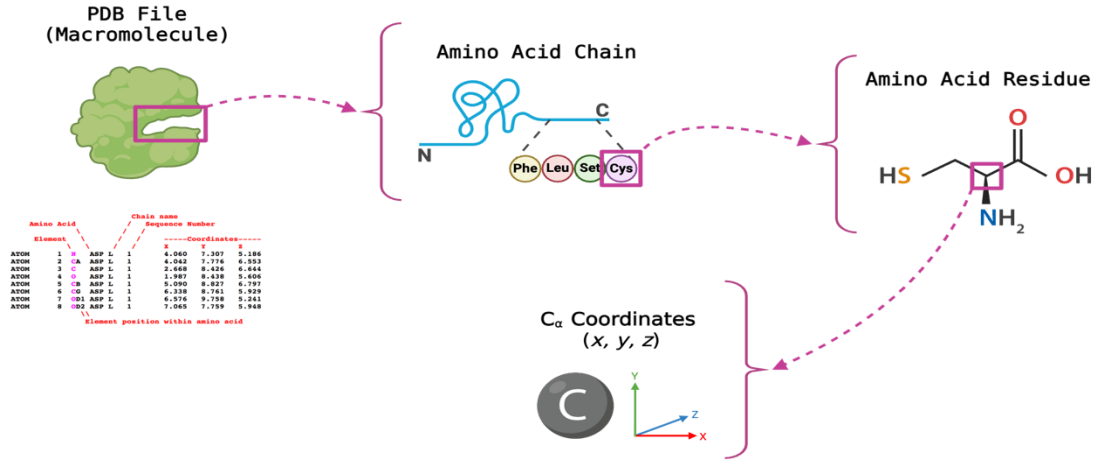
- residue ≅ C_α
- pairwise Euclidian distances between C_α's within a given chain

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

($\mathbf{x_i}$ represents the C_α coords. of the i^{th} residue in a give chain)



pdb_00008s05: *ArnAB complex an archaeal ortholog of the Sec23/24 core motif*



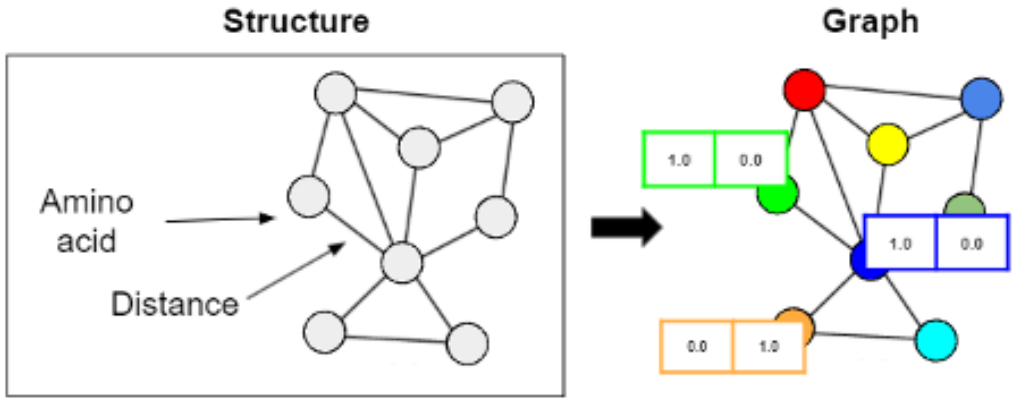
Created in https://BioRender.com

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



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

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

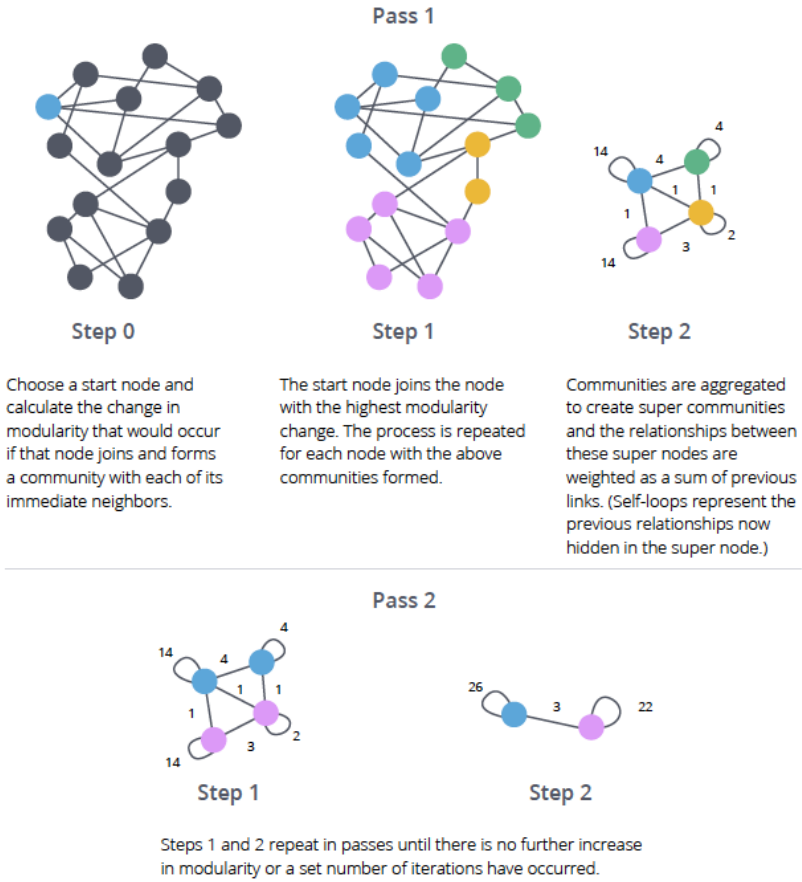
Louvain clustering maximizes *modularity scores*.

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)`

Generating Clustering Predictions

(2) Two-Stage Spectral Clustering (Unsupervised)

(scikit-learn 1.3.0)

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* → Similarity matrix *A*

$$A_{ij} = \exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right)$$

(σ = scaling parameter; median of all nonzero distances in distance matrix D)

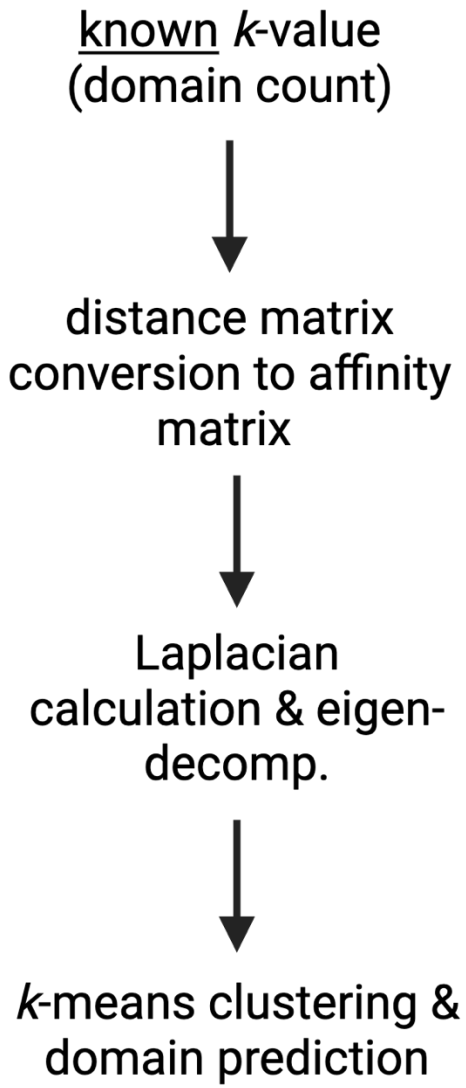
- Normalized Graph Laplacian calculation

$$L = I - D^{-1/2} A D^{-1/2}$$

(D: 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 $\{v_1 \ldots v_k\}$, $U = [v_1, v_2, \ldots, v_k]$
  - Each i^{th} row of U represents the transformed coordinates of residue i
- K-means clustering → labels clusters in the new space with predicted domain ID

The Benchmarks: "Supervised" Spectral Clustering on Graph Topology



Quantitative Performance Benchmark

Louvain

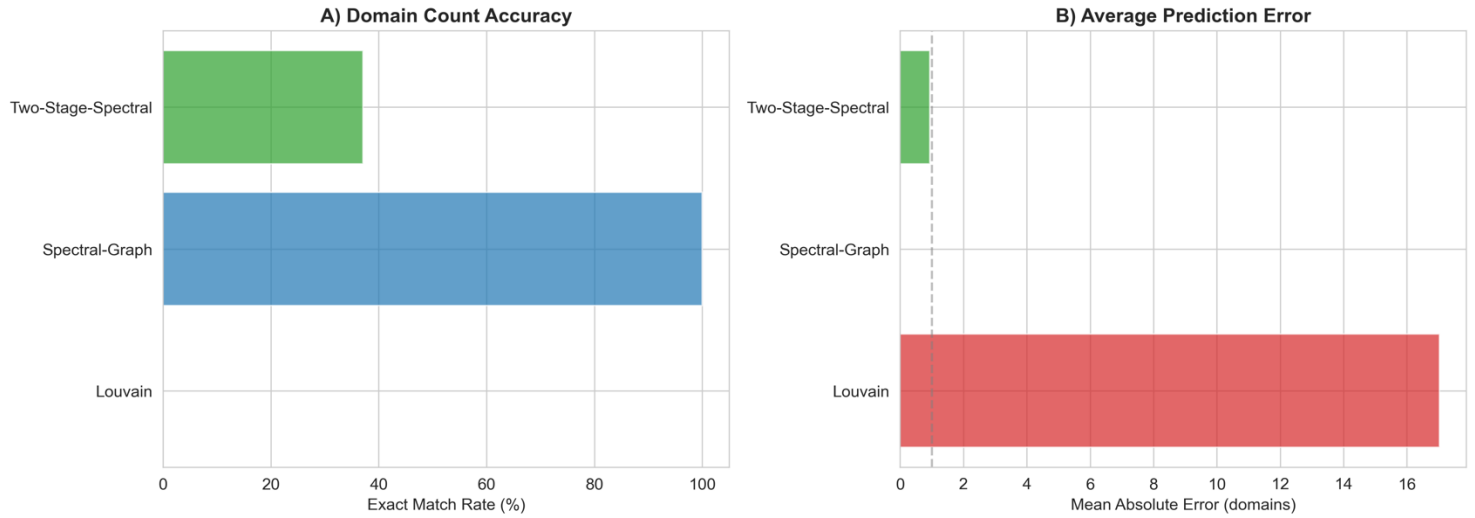
- Exact Match: 0/94
- Mean Absolute Error (MAE): 17.02

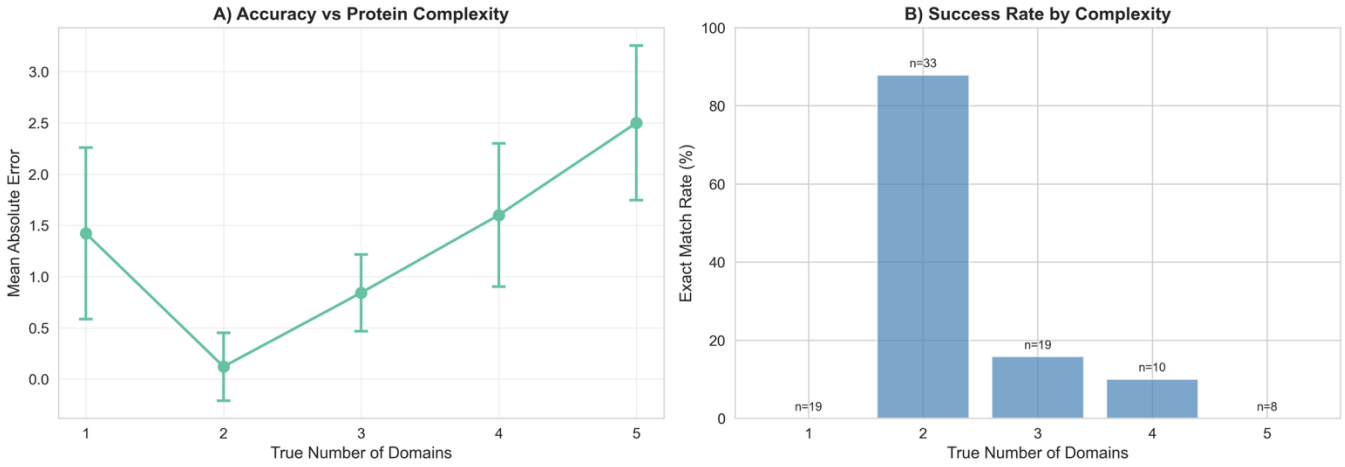
Two-Stage Spectral

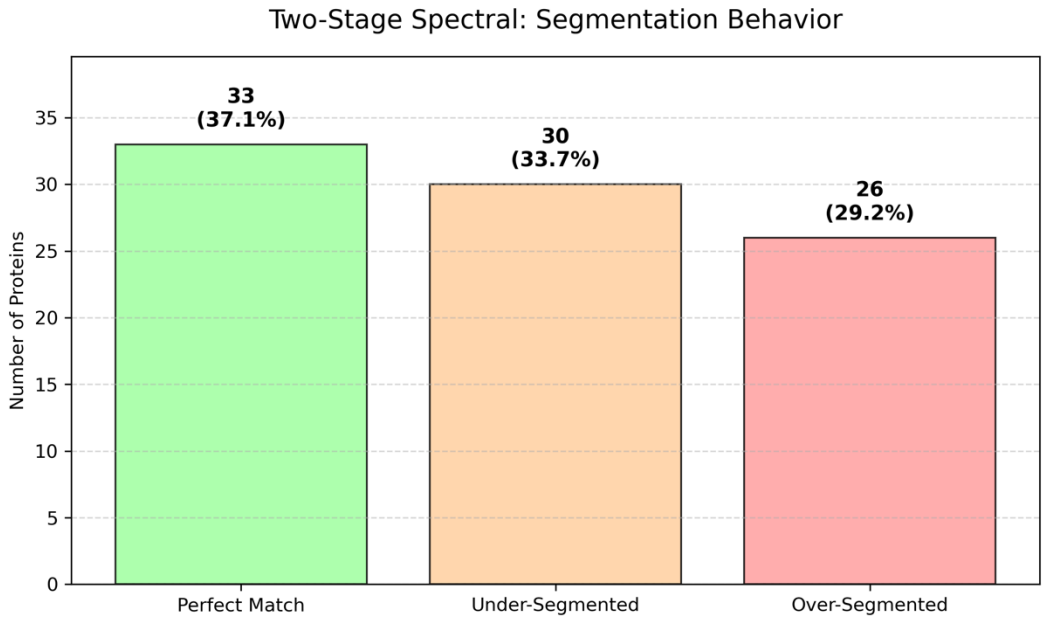
- Exact Match: 33/94
- Mean Absolute Error (MAE): 0.93

"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

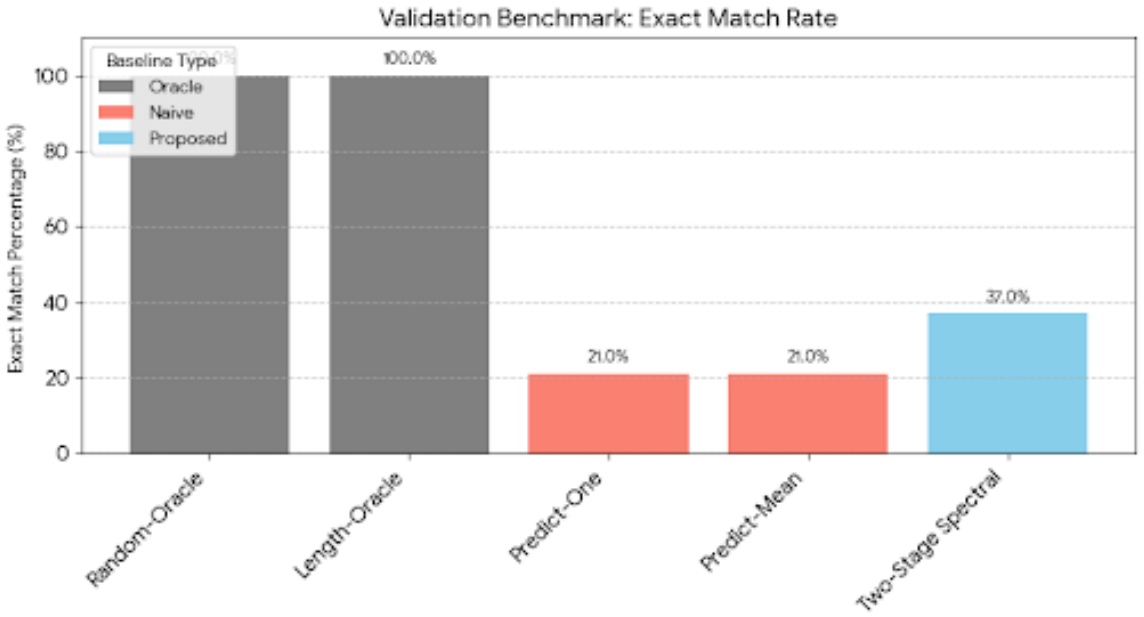
- mean_error: 1.0
- exact_match_rate: 0.429

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://lammpstube.com/2020/04/03/pdb-file-format/
8. https://pubs.acs.org/doi/10.1021/acs.jmedchem.6b01453