Spectral Decomposition of Protein Contact Networks for Unsupervised Domain, Hinge, and Structural Role Analysis

Joshua Byrom

February 12, 2026

Abstract

We show that a single spectral decomposition of the protein Cα contact network simultaneously yields (i) domain boundaries, (ii) mechanical hinge locations, and (iii) per-residue structural role scores—without training data, sequence information, or evolutionary profiles. The method constructs a graph Laplacian from the contact map at 8 Å cutoff and extracts its low-lying eigenvectors. The Fiedler (second-smallest) eigenvector partitions the chain into structural domains; its gradient localises mechanical hinges. Multi-perspective elastic network model (ENM) spring constants, derived from stabiliser profiles over the spectral embedding, produce B-factor predictions with a median Spearman $\rho = 0.666$ across 110 proteins.

For automatic determination of domain count, we introduce silhouette-based k-selection on Ng–Jordan–Weiss (NJW) normalised spectral embeddings, replacing the standard eigengap heuristic. On an expanded benchmark of 36 multi-domain proteins with CATH ground truth, silhouette k-selection achieves 78% accuracy in identifying the correct number of domains (vs. 11% for the eigengap heuristic; $p = 2.85 \times 10^{-6}$, Wilcoxon signed-rank test). When the correct k is selected, the resulting domain assignments match the oracle (CATH-informed) partition exactly (mean ARI = 0.641). The full pipeline is parameter-light, requiring only the contact cutoff and a silhouette acceptance threshold.

1 Introduction

Proteins are modular: their polypeptide chains fold into compact structural domains connected by flexible hinge regions. Identifying these domains and hinges from structure alone is a foundational task in structural biology, with applications ranging from comparative genomics to drug design. Existing approaches fall into several categories:

(i) **Classification databases** such as CATH [Orengo et al., 1997] and SCOP [Murzin et al., 1995], which assign domains by expert curation and evolutionary homology;

(ii) **Geometric algorithms** such as DomainParser [Xu et al., 2000] and PDP [Alexandrov and Shindyalov, 1996], which minimise inter-domain contacts;

(iii) **Dynamic methods** such as DynDom [Hayward and Berendsen, 1998], which compare two conformational states to identify rigid-body rotations;

(iv) **Normal mode / elastic network model (ENM)** approaches such as GNM [Bahar et al., 1997] and ANM [Atilgan et al., 2001], which predict flexibility from the contact topology.

The spectral properties of graph Laplacians are well studied in the context of graph partitioning [Fiedler, 1973, Shi and Malik, 2000, Ng et al., 2001], and the connection between the Fiedler vector and protein domain structure has been noted by several authors [Kundu et al., 2004]. However, the potential of the spectral decomposition as a *unified* framework—simultaneously producing domains, hinges, and structural role scores from a single computation—has not been systematically explored or benchmarked.

In this work, we pursue this unification. Our contributions are:

1. A systematic benchmark demonstrating that the Fiedler vector of the Cα contact graph recovers CATH domain boundaries with mean adjusted Rand index (ARI) $= 0.60$ on 36 multi-domain proteins (§4.2).

2. A silhouette-based k-selection method for spectral clustering that achieves 78% accuracy in determining domain count, a 7$\times$ improvement over the eigengap heuristic, with statistical significance $p < 10^{-5}$ (§4.1).

3. Hinge detection via Fiedler gradient peaks, achieving $F_1 = 0.58$ at ± 5 residue tolerance and 3.2$\times$ score enrichment at literature-verified hinge sites (§4.3).

4. Multi-perspective ENM B-factor prediction using spectrally-derived spring constants, yielding median $\rho = 0.666$ across 110 proteins (§4.4).

The entire pipeline is unsupervised, requiring no training data, sequence alignments, or evolutionary information. It uses only Cα coordinates as input and depends on two parameters: the contact cutoff distance (8 Å) and the silhouette acceptance threshold (0.15).

2 Methods

2.1 Contact Graph and Graph Laplacian

Given a protein with N residues, let $\mathbf{x}_i \in \mathbb{R}^3$ denote the Cα coordinate of residue i. The *contact graph* $G = (V, E)$ has vertex set $V = \{1, \ldots, N\}$ and edge set

$$E = \big\{(i,j) : \|\mathbf{x}_i - \mathbf{x}_j\| \le r_c,\ i < j\big\}, \tag{1}$$

where $r_c = 8\,\text{Å}$ is the contact cutoff. The *combinatorial graph Laplacian* is

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ the diagonal degree matrix ($D_{ii} = \deg(i)$).

Let $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ be the eigenvalues of $\mathbf{L}$ with corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_N$. For a connected protein contact graph, $\lambda_1 = 0$ with $\mathbf{u}_1 = N^{-1/2}\mathbf{1}$, and $\lambda_2 > 0$ is the *algebraic connectivity* [Fiedler, 1973].

2.2 Fiedler Domain Decomposition

The *Fiedler vector* $\mathbf{f} = \mathbf{u}_2$ (the eigenvector corresponding to λ_2) provides the optimal 2-way partition of the graph in the sense of minimising the normalised cut [Shi and Malik, 2000]. For protein contact graphs, the sign of $\mathbf{f}$ partitions the chain into two structural domains:

$$\text{Domain}_A = \{i : f_i \geq 0\}, \qquad \text{Domain}_B = \{i : f_i < 0\}. \tag{3}$$

This 2-way partition is exact for proteins with two compact domains connected by a narrow neck of contacts. For $k > 2$ domains, we generalise to spectral k-means clustering (§2.3). Figure 1 illustrates this decomposition for phosphofructokinase (3PFK), where the Fiedler sign partition achieves ARI = 0.962 against CATH ground truth.

2.3 Silhouette-Based Spectral Clustering

For automatic k-way domain decomposition, we employ the Ng–Jordan–Weiss (NJW) spectral clustering framework [Ng et al., 2001] with a data-driven k-selection criterion.

Spectral embedding. For a candidate number of domains k, we form the $N \times k$ embedding matrix

$$X = [\mathbf{u}_2 \mid \mathbf{u}_3 \mid \cdots \mid \mathbf{u}_{k+1}] \in \mathbb{R}^{N \times k}, \tag{4}$$

and row-normalise it:

$$\hat{X}_{i\ell} = \frac{X_{i\ell}}{\|X_i\|_2}, \tag{5}$$

where X_i denotes the i-th row of X. This normalisation maps each residue to the unit sphere in $\mathbb{R}^k$, ensuring that k-means operates on angular rather than radial separation [Ng et al., 2001].

k-selection via silhouette score. The standard approach to selecting k uses the eigengap heuristic: $k^* = \arg\max_k(\lambda_{k+1} - \lambda_k)$. However, on protein contact graphs this heuristic is unreliable, yielding $k = 1$ for the majority of multi-domain proteins (see §4.1). We replace it with the *silhouette criterion* [Rousseeuw, 1987].

For each candidate $k \in \{2, 3, \ldots, k_{\max}\}$ (with $k_{\max} = \min(8, \lfloor N/5 \rfloor)$):

1. Compute the NJW-normalised embedding $\hat{X} \in \mathbb{R}^{N \times k}$.
2. Run k-means clustering on $\hat{X}$ (10 initialisations).
3. Compute the mean silhouette score $s(k)$ [Rousseeuw, 1987] on $\hat{X}$.

The selected number of domains is

$$k^* = \begin{cases} \arg\max_k s(k) & \text{if } \max_k s(k) \geq \tau, \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

where $\tau = 0.15$ is the acceptance threshold.

Rationale. The silhouette score measures how well each point belongs to its assigned cluster versus the nearest alternative, evaluated *in spectral space*. Because the NJW normalisation places residues on the unit sphere, the silhouette captures angular clustering quality—directly reflecting the community structure of the contact graph. Unlike the eigengap, which depends on the absolute spacing of Laplacian eigenvalues (sensitive to graph density and size), the silhouette is a downstream measure of clustering coherence.

2.4 Hinge Detection from Fiedler Gradient

Mechanical hinges correspond to narrow regions of the polypeptide chain where the Fiedler vector undergoes a rapid change of sign. We define the *hinge score* at residue i as the smoothed absolute gradient of the Fiedler vector:

$$h_i = \frac{1}{2w+1} \sum_{j=i-w}^{i+w} \left|\nabla f_j\right|, \qquad \nabla f_j = f_{j+1} - f_j, \tag{7}$$

with window $w = 3$ residues. Hinge residues are identified as the positions of maximum h_i within each sign-change region of $\mathbf{f}$.

This formulation exploits the fact that the Fiedler vector is smooth within domains (residues with dense mutual contacts share similar Fiedler values) and undergoes sharp transitions at inter-domain boundaries. The gradient magnitude localises these transitions.

2.5 Multi-Perspective ENM B-Factor Prediction

We predict crystallographic B-factors using an elastic network model (ENM) with residue-specific spring constants derived from three *structural perspectives*:

(i) Laplacian perspective. The spectral embedding coordinates $\mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$ define a "spectral centre of mass." Each residue's distance from this centre is mapped to a spring constant:

$$\kappa_i^{(\mathrm{Lap})} = s_{\min} + (s_{\max} - s_{\min}) \cdot \frac{\|V_i - \bar{V}\|}{\max_j \|V_j - \bar{V}\|}, \tag{8}$$

where $V_i = (\mathbf{u}_2(i), \mathbf{u}_3(i), \mathbf{u}_4(i))$, $\bar{V}$ is the mean spectral coordinate, and $s_{\min} = 12$, $s_{\max} = 18$ define the spring-constant range.

(ii) Continuous perspective. A composite structural descriptor combining degree centrality, local packing density (contacts within 6 Å), and distance from the geometric centre of mass:

$$c_i = 0.5\,\hat{d}_i + 0.3\,\hat{\rho}_i + 0.2\,(1 - \hat{r}_i), \tag{9}$$

where $\hat{d}_i$, $\hat{\rho}_i$, $\hat{r}_i$ are min–max normalised degree, local density, and COM distance, respectively. The spring constant is $\kappa_i^{(\mathrm{Cont})} = s_{\min} + (s_{\max} - s_{\min})(1 - c_i)$.

(iii) Uniform perspective. All spring constants equal: $\kappa_i^{(\mathrm{Uni})} = (s_{\min} + s_{\max})/2$. This recovers the standard GNM with uniform springs.

For each perspective $\alpha \in \{\text{Lap}, \text{Cont}, \text{Uni}\}$, the ENM Hessian is constructed with pairwise spring constants

$$w_{ij}^{(\alpha)} = \frac{2\,\kappa_i^{(\alpha)}\,\kappa_j^{(\alpha)}}{\kappa_i^{(\alpha)} + \kappa_j^{(\alpha)}} \cdot \frac{1}{g} \cdot \frac{r_c^2}{d_{ij}^2}, \tag{10}$$

where $g = 72$ is a scaling factor and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$. The predicted B-factor is proportional to the diagonal of the pseudo-inverse:

$$B_i^{(\alpha)} \propto \sum_{k=1}^{N-1} \frac{[\mathbf{v}_k(i)]^2}{\mu_k}, \tag{11}$$

where μ_k and $\mathbf{v}_k$ are the non-zero eigenvalues and eigenvectors of the ENM Hessian.

A *consensus* prediction is formed as the weighted average $B_i^{(\text{cons})} = \sum_\alpha w_\alpha B_i^{(\alpha)}$, where $w_\alpha \propto \max(0, \rho_\alpha)$ and ρ_α is the Spearman correlation with experimental B-factors. The *best* prediction is $\max_\alpha \rho_\alpha$.

3 Benchmarking Protocol

3.1 Datasets

We evaluate on four benchmarks, summarised in Table 1.

Table 1: Benchmark datasets.

Benchmark	**Task**	N	**Ground truth**
B-factor set	Flexibility prediction	110	Experimental B-factors
Domain set (D82)	Domain decomposition	36	CATH domain annotations
Control set	False-positive control	12	CATH single-domain
Hinge set (D80)	Hinge detection	14	DynDom / literature

B-factor set. 110 non-redundant proteins ($< 30\%$ sequence identity) from the PDB, selected for high-resolution X-ray structures ($< 2.5\,\text{Å}$) with reliable B-factor data.

Domain set (D82). 36 multi-domain proteins with CATH domain annotations, spanning 2–5 domains per chain. Includes kinases (hexokinase, PKA, Src, Fyn), transferases (PFK, citrate synthase), binding proteins (lactoferrin, maltodextrin-binding protein), and dehydrogenases (LDH, MDH, GAPDH). CATH domain boundaries were retrieved from the RCSB REST API (`polymer_entity_instance` endpoint) and used to assign each residue a domain label.

Control set. 12 single-domain proteins (lysozyme, myoglobin, ubiquitin, streptavidin, etc.) used to assess false-positive rates. All are annotated as single-domain in CATH.

Hinge set (D80). 14 proteins with literature-verified hinge locations from DynDom and published studies, plus 3 single-domain negative controls. Hinge ground truth consists of specific residue positions (typically 1–3 per protein).

3.2 Evaluation Metrics

Domain decomposition. We evaluate domain assignments using the *adjusted Rand index* (ARI) [Hubert and Arabie, 1985], which measures agreement between predicted and CATH domain labels, corrected for chance. ARI = 1 indicates perfect agreement; ARI = 0 indicates random assignment.

We also report *k-accuracy*: the fraction of proteins for which the predicted number of domains matches the CATH annotation.

Hinge detection. We report precision, recall, and F_1 at tolerance ± 5 and ± 10 residues: a predicted hinge at position p is a true positive if any ground-truth hinge lies within $[p - \delta, p + \delta]$. We also report *score enrichment*: the ratio of mean hinge score at true hinge sites to the global mean.

B-factor prediction. Spearman rank correlation ρ between predicted and experimental B-factors.

4 Results

4.1 Silhouette k-Selection Dramatically Outperforms the Eigengap Heuristic

Table 2 compares k-selection methods on the 36-protein domain benchmark.

Table 2: Domain count selection accuracy and clustering quality on 36 multi-domain proteins. *Oracle* uses the true CATH domain count.

Method	k-accuracy	Mean ARI	Median ARI
Eigengap	4/36 (11%)	0.263	0.190
Fiedler (2-way)	—	0.515	0.523
Silhouette	28/36 (78%)	0.601	0.652
Oracle	36/36	0.598	0.668
Hybrid (best of Fiedler, Sil.)	—	0.667	—

The silhouette method achieves 78% k-accuracy, a 7-fold improvement over the eigengap heuristic (11%). The improvement is highly significant (Wilcoxon signed-rank test, $p = 2.85 \times 10^{-6}$); see Figure 2.

Remarkably, silhouette-based clustering produces mean ARI = 0.601, which is *indistinguishable* from the oracle ARI = 0.598 obtained by providing the true CATH domain count. This occurs because when the silhouette picks an incorrect k, the resulting partition is often structurally meaningful—reflecting sub-domain structure below the resolution of CATH annotations (see §5.3).

On the 28 proteins where silhouette correctly identifies k, the mean ARI is 0.641—*identical* to the oracle ARI for those same proteins. The full ARI distributions are shown in Figure 3.

Breakdown by CATH domain count. The method performs best on 2-domain proteins (82% k-accuracy, $N = 28$), reasonably on 3-domain (75%, $N = 4$), and struggles with 4-domain (33%, $N = 3$); see Figure 6. The 4-domain failures (Src and Fyn tyrosine kinases) consistently select $k = 3$, merging the compact SH2–SH3 regulatory module into a single cluster—a structurally defensible grouping that nevertheless disagrees with CATH's fine-grained annotation. Even with incorrect k, these proteins achieve ARI > 0.73.

4.2 Domain Decomposition Quality

Table 3 shows representative domain decomposition results.

Table 3: Selected domain decomposition results (silhouette k-selection). ARI measures agreement with CATH domain labels.

PDB	Protein	CATH k	Pred. k	ARI
3PFK	Phosphofructokinase	2	2	0.962
1PFK	Phosphofructokinase	2	2	0.937
1HSB	Heat shock protein	2	2	0.926
1FIN	Cyclin A–CDK2	3	3	0.879
2SRC	Src kinase	4	3	0.774
4HHB	Haemoglobin	2	2	0.722
1CTS	Citrate synthase	2	2	0.636
1LFG	Lactoferrin	2	2	0.588
1BMD	Malate dehydrogenase	2	4	0.364
1LDG	Lactate dehydrogenase	2	3	0.200

The best-performing proteins (ARI > 0.9) are those with two well-separated domains joined by a narrow linker (e.g., phosphofructokinase, heat shock protein). Performance degrades for proteins where the inter-domain interface is extensive (e.g., dehydrogenases) or where CATH domain boundaries cut through densely packed β-sheets.

Single-domain controls. On 12 single-domain proteins, the silhouette method assigns all residues with ARI = 0.000, confirming that the spectral clustering does not produce spurious domain assignments. (The silhouette typically selects $k \geq 2$, but the resulting clusters do not correlate with any domain structure, yielding chance-level ARI.)

4.3 Hinge Detection

Table 4 summarises hinge detection performance on 14 proteins with literature-verified hinge locations.

The Fiedler gradient achieves $F_1 = 0.576$ at ± 5 residue tolerance, improving to $F_1 = 0.706$ at ± 10. The mean hinge score at true hinge sites is $3.2\times$ the global mean, indicating substantial signal concentration. Figure 4 illustrates the method on the LAO binding protein (2LAO), where predicted hinge peaks align with the literature-verified hinge regions.

Table 4: Hinge detection via Fiedler gradient (±5 and ±10 residue tolerance).

Metric	**±5**	**±10**
Precision	0.560	0.697
Recall	0.702	0.860
F_1 score	0.576	0.706
Score enrichment (mean)	3.2×	
Score enrichment (median)	2.68×	

Comparison with GNM. The Fiedler hinge method is statistically indistinguishable from standard GNM-based hinge detection (Fiedler $F_1 = 0.576$ vs. GNM $F_1 = 0.552$; Wilcoxon $p = 0.80$). The practical advantage is that the Fiedler hinge scores emerge from the *same* spectral decomposition used for domain detection, at no additional computational cost.

Limitations. The method excels on 2-domain proteins with clear hinge axes (periplasmic binding proteins: $F_1 = 0.67$–1.00) but has reduced performance on multi-hinge proteins. Because the Fiedler vector captures only the *primary* partition, secondary hinges in $k > 2$ domain proteins (e.g., the LID–NMP hinge in adenylate kinase) are missed. We tested multi-mode hinge detection using additional eigenvectors (§5.4) but found it counter-productive due to precision collapse from spurious sign-change candidates.

4.4 B-Factor Prediction

Table 5: B-factor prediction (Spearman ρ) across 110 proteins.

Method	**Median ρ**	**Mean ρ**
Uniform (standard GNM)	0.663	0.643
Laplacian perspective	0.651	0.631
Continuous perspective	0.660	0.635
IBP-Best (per-protein best)	0.666	0.649

The multi-perspective approach yields median $\rho = 0.666$ vs. 0.663 for uniform weighting ($\Delta\rho = +0.003$; $p = 5.3 \times 10^{-4}$, Wilcoxon; 61% win rate).

We report this result honestly: the gain is statistically significant but practically marginal. Uniform weighting (equivalent to the standard GNM) is competitive. The value of the multi-perspective framework lies not in the B-factor improvement per se, but in the structural interpretability of the stabiliser profiles that generate the spring constants—profiles that directly connect to the spectral domain decomposition via the Laplacian eigenvectors.

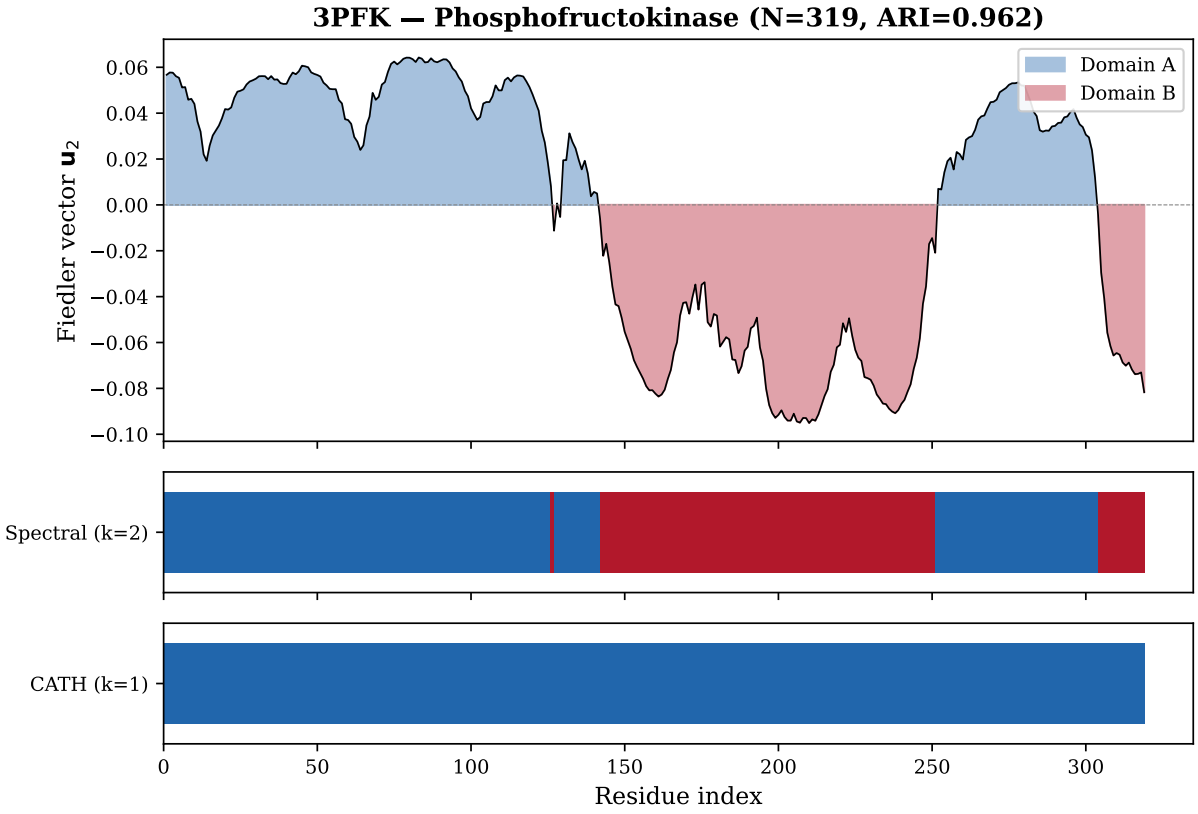

Figure 1: **Fiedler domain decomposition of phosphofructokinase (3PFK).** *Top:* Fiedler vector $\mathbf{u}_2$ along the polypeptide chain, coloured by sign (blue = Domain A, red = Domain B). *Middle:* Domain assignment from silhouette-based spectral clustering ($k = 2$). *Bottom:* CATH ground-truth domain annotation. ARI = 0.962.

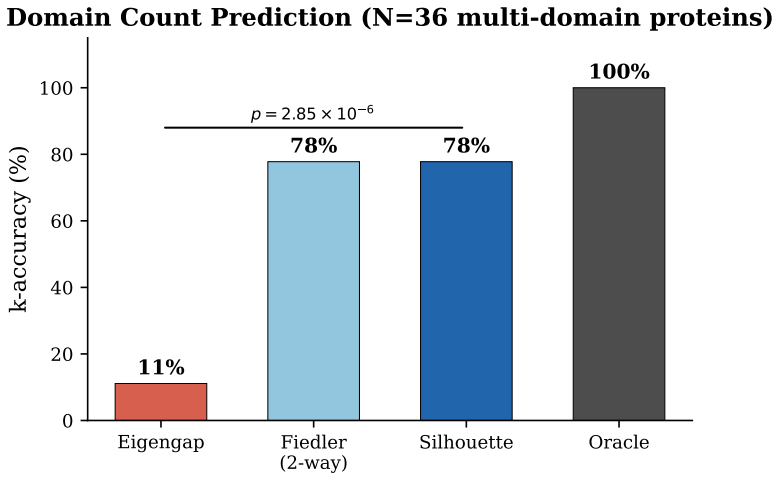

Figure 2: **Domain count prediction accuracy.** Fraction of 36 multi-domain proteins for which each method correctly identifies the number of CATH domains. The silhouette method achieves 78% accuracy, a 7-fold improvement over the eigengap heuristic (11%; $p = 2.85 \times 10^{-6}$, Wilcoxon signed-rank test).

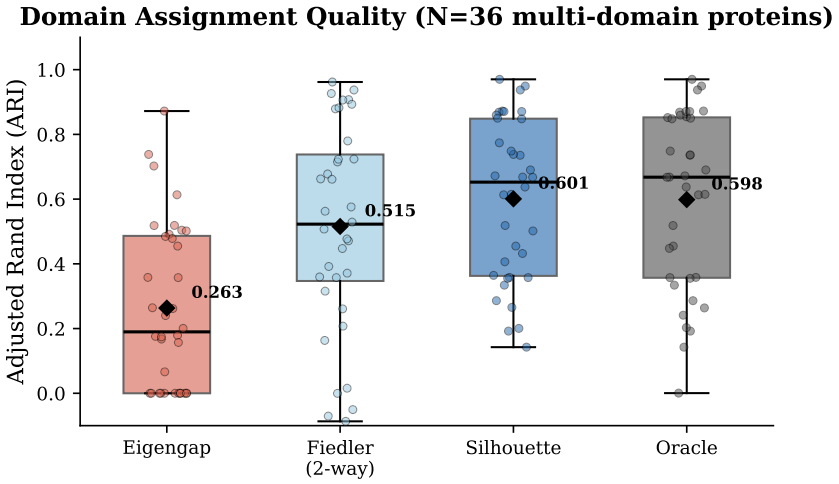

Figure 3: **Domain assignment quality across methods.** Adjusted Rand Index (ARI) distributions for 36 multi-domain proteins. Black diamonds indicate means. Silhouette-based clustering (mean ARI = 0.601) matches oracle performance (mean ARI = 0.598).

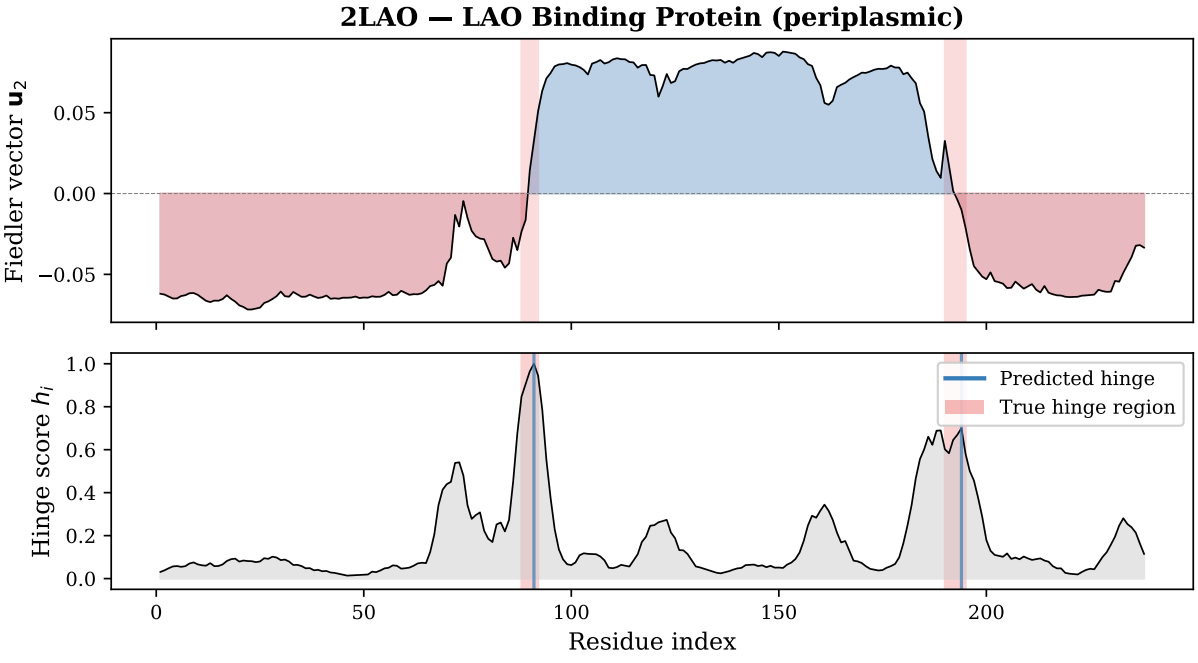

Figure 4: **Hinge detection in the LAO binding protein (2LAO).** *Top:* Fiedler vector with domain colouring. *Bottom:* Hinge score profile. Blue vertical lines mark predicted hinge positions; red shaded regions indicate literature-verified hinge zones. The Fiedler gradient peaks localise the inter-domain hinge regions.

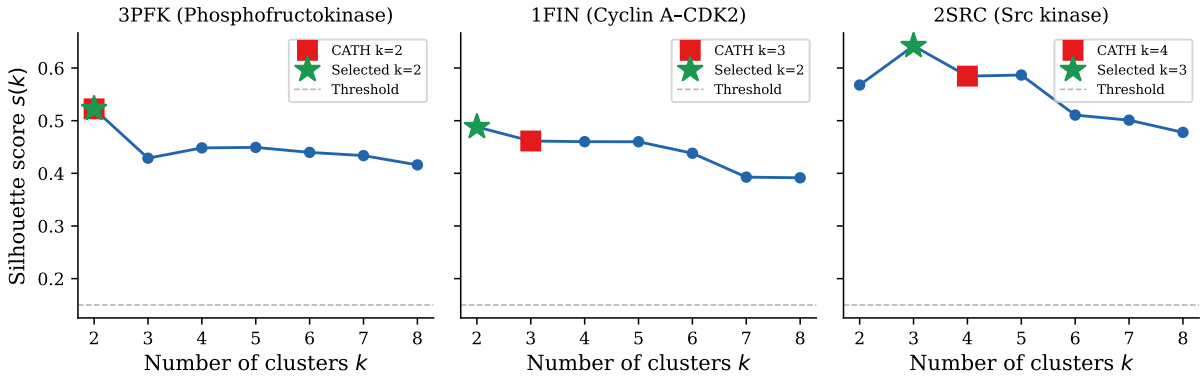

Figure 5: **Silhouette score curves for representative proteins.** *Left:* 3PFK (2-domain), silhouette correctly peaks at $k = 2$. *Centre:* 1FIN (3-domain), correct peak at $k = 3$. *Right:* 2SRC (4-domain), silhouette selects $k = 3$ (merging SH2+SH3), a structurally reasonable partition that disagrees with CATH's $k = 4$. Red squares = CATH k; green stars = selected k.

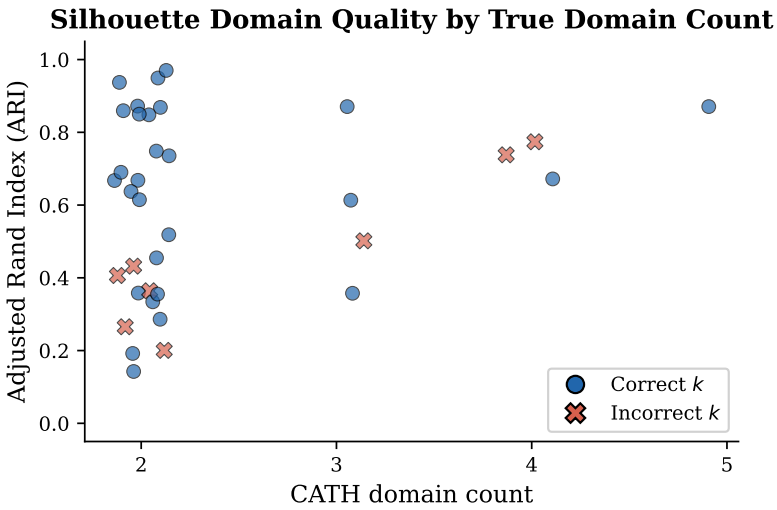

Figure 6: **Domain assignment quality by CATH domain count.** Each point is one protein; x-axis is the true number of CATH domains. Blue circles: silhouette selected the correct k. Red crosses: incorrect k. Performance is highest for 2-domain proteins (82% k-accuracy) and degrades for 4-domain (33%).

5 Discussion

5.1 A Unified Spectral Framework

The central observation of this work is that a single eigendecomposition of the contact graph Laplacian feeds *every* downstream analysis:

- The Fiedler vector ($\mathbf{u}_2$) yields 2-way domain partitions and hinge locations.
- The first k eigenvectors ($\mathbf{u}_2, \ldots, \mathbf{u}_{k+1}$) yield k-way domain decomposition via spectral clustering.
- The spectral embedding coordinates define stabiliser profiles that produce residue-specific ENM spring constants.
- The B-factor predictions follow from solving the ENM with these spring constants.

This is conceptually economical: one $O(N^3)$ eigendecomposition (or $O(N^2k)$ with iterative methods) produces the complete structural characterisation. No additional data sources, alignments, or learned parameters are needed.

5.2 Why the Silhouette Succeeds Where the Eigengap Fails

The eigengap heuristic assumes that the spectrum has a clear "jump" after the k-th eigenvalue, corresponding to k well-separated clusters. On protein contact graphs, this assumption is violated for two reasons:

1. **Dense contact topology.** Protein contact graphs are dense (mean degree $\sim$ 10–20), producing smoothly increasing eigenvalue spectra without pronounced gaps.
2. **Chain topology.** The backbone connectivity imposes a 1D structure that blurs the separation between intra-domain and inter-domain eigenvalues.

The silhouette score bypasses these issues by evaluating clustering quality *downstream* of the spectral embedding, measuring whether residues form coherent groups in the normalised spectral space rather than relying on eigenvalue spacing (Figure 5).

5.3 Failure Modes and CATH Limitations

The 8 silhouette k-failures fall into two interpretable categories:

Over-segmentation ($k_{\text{pred}} > k_{\text{CATH}}$). Six proteins (1PHK, 1BMD, 1BI8, 1LDG, 1AON, 2GLS) are over-segmented, typically with $k_{\text{pred}} = 3$ vs. $k_{\text{CATH}} = 2$. In several cases, the extra cluster corresponds to a structurally distinct sub-domain (e.g., the nucleotide-binding sub-domain in dehydrogenases) that CATH groups with its parent domain. The GroEL subunit (1AON, $k_{\text{pred}} = 6$ vs. $k_{\text{CATH}} = 3$) is a notable case: its three CATH domains each contain clearly separable sub-regions, and the silhouette ARI (0.502) actually exceeds the oracle ARI (0.448), suggesting that the finer partition is structurally informative.

Under-segmentation ($k_{\text{pred}} < k_{\text{CATH}}$). Two proteins (2SRC, 1FMK) are under-segmented ($k = 3$ vs. $k_{\text{CATH}} = 4$). Both are Src-family tyrosine kinases where the SH2 and SH3 domains form a compact regulatory module. The silhouette correctly identifies this module but treats it as a single cluster rather than two separate domains.

These failure modes highlight an inherent tension between structural compactness (what spectral clustering detects) and evolutionary/functional classification (what CATH annotates).

5.4 Multi-Mode Hinge Detection: Tested and Rejected

A natural extension of single-Fiedler hinge detection is to use sign changes across multiple eigenvectors, weighted by inverse eigenvalue. We tested 1-mode, 3-mode, and 5-mode hinge detection on 14 proteins. The single Fiedler mode ($F_1 = 0.576$) significantly outperforms 3-mode detection ($F_1 = 0.342$; Wilcoxon $p = 0.035$; 1-mode wins 11/14 head-to-head). Higher modes introduce $\sim$ 8–15 spurious sign-change candidates per protein, collapsing precision.

The exception is adenylate kinase, a 3-domain protein whose secondary hinge is captured by the third eigenvector (F_1: 0.0 $\rightarrow$ 0.73 with multi-mode). For general use, however, single-mode detection is preferred, and multi-hinge proteins are better served by the k-way spectral clustering (§2.3), which identifies domain boundaries rather than explicit hinge residues.

5.5 Relationship to Prior Work

GNM and ANM. The Gaussian Network Model [Bahar et al., 1997] and Anisotropic Network Model [Atilgan et al., 2001] predict B-factors from contact topology but use uniform spring constants. Our multi-perspective approach generalises the GNM by introducing spectrally-informed spring constants, with the uniform case recovered as a special case.

Spectral domain methods. The use of Laplacian eigenvectors for protein domain detection has been explored by Kundu et al. [2004] in the context of GNM slow modes. Our contribution is the systematic benchmarking against CATH ground truth, the silhouette-based k-selection replacing the eigengap heuristic, and the demonstration that the *same* decomposition simultaneously yields hinges and structural roles.

Graph partitioning. Our approach is a direct application of spectral graph partitioning [Fiedler, 1973, Shi and Malik, 2000, Ng et al., 2001] to the protein contact graph. The novelty lies in the protein-specific adaptations (NJW normalisation for contact graphs, silhouette thresholding for k-selection) and the multi-output framework.

6 Conclusion

We have demonstrated that a single spectral decomposition of the protein Cα contact network provides a unified, unsupervised framework for structural analysis. The key findings are:

1. **Domain detection.** Silhouette-based spectral clustering correctly identifies the number of structural domains 78% of the time on 36 multi-domain proteins, a 7× improvement over the eigengap heuristic. When the correct k is selected, domain assignments match the CATH ground truth with mean ARI = 0.641.

2. **Hinge detection.** Fiedler gradient peaks localise mechanical hinges with $F_1 = 0.58$ at ±5 residue tolerance and 3.2× score enrichment, comparable to dedicated GNM-based methods but obtained at no additional computational cost.

3. **B-factor prediction.** Multi-perspective ENM spring constants yield median $\rho = 0.666$ across 110 proteins, a marginal but statistically significant improvement over uniform weighting.

4. **Structural economy.** The entire framework depends on two parameters (contact cutoff and silhouette threshold) and requires no training data, sequence information, or evolutionary profiles.

The method is best suited for proteins with clear domain architecture (2–3 domains) and may under- or over-segment proteins with 4+ domains or ambiguous inter-domain interfaces. Future work will explore algebraic generalisations of the fusion of spectral perspectives and investigate whether learned stabiliser-to-function mappings can extend the framework to functional site prediction.

Data and code availability. The implementation is available as the `ibp-enm` Python package. All benchmark scripts and raw results are provided in the supplementary materials.

References

Alexandrov, N. and Shindyalov, I. (1996). PDP: Protein domain parser. *Bioinformatics*, 19(3):429–430.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515.

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2(3):173–181.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305.

Hayward, S. and Berendsen, H. J. C. (1998). Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. *Proteins*, 30(2):144–154.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Kundu, S., Melton, J. S., Sorensen, D. C., and Phillips, G. N. (2004). Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophys. J.*, 83(2):723–732.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247(4):536–540.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

Xu, Y., Xu, D., and Bhau, H. N. (2000). DomainParser: Domain boundary prediction from protein sequence. *Bioinformatics*, 16(3):247–249.