GRP: A Representation-Aware Effective Geometry of Protein State Space

Chuanyang Liu

liuchuanyang13@nudt.edu.cn

National University of Defense Technology

Research Article

Keywords:

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10857002/v1

Additional Declarations: The authors declare no competing interests.

GRP: A Representation-Aware Effective Geometry of Protein-State Space

Chuanyang Liu[1,2,#]

Affiliations:

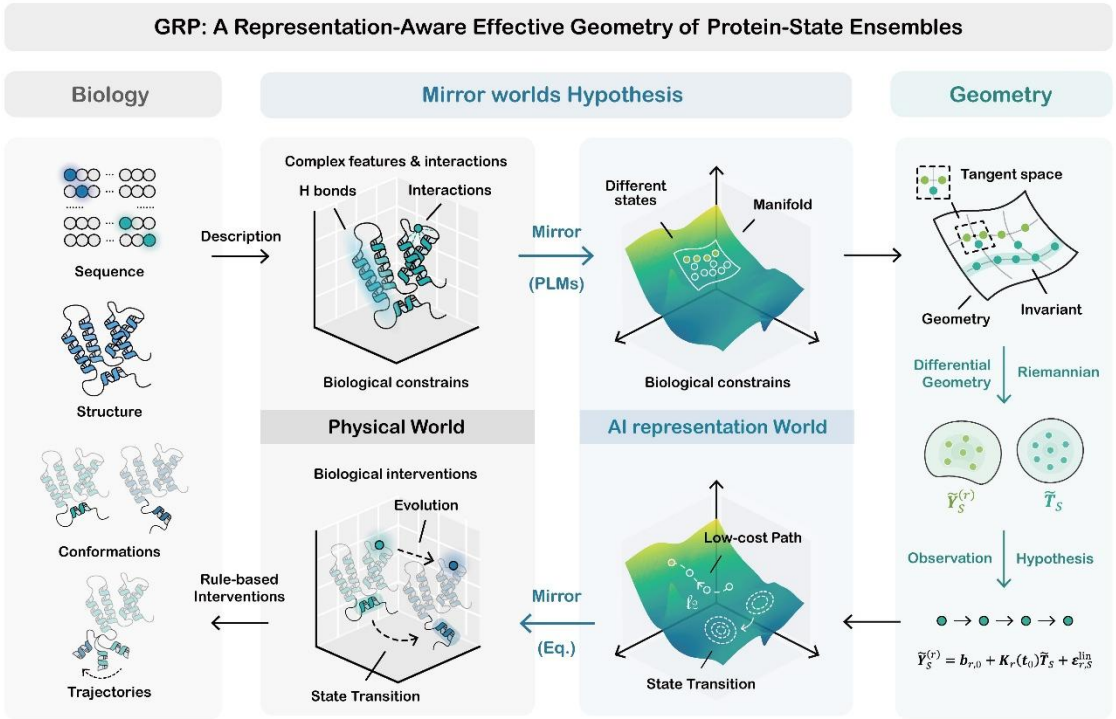
[1] College of Science, National University of Defense Technology, Changsha, Hunan, 410072, China

[2] College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, 410072, China

[#] Correspondence:

Chuanyang Liu: liuchuanyang13@nudt.edu.cn

Graphical Abstract



Abstract

Proteins are not isolated structures but multilevel state spaces shaped by sequence and context. We introduce GRP, a general-relativity-inspired framework that treats AI representations as a lossy mirror space for searching for candidate invariants - relations intended to remain stable across models, coordinates and scales - and linking them to biological constraints. GRP formulates coordinate-aware geometric, tensorial and operator relations among empirical measures, covariance geometry, perturbational responses and distributional transport, providing an effective description of protein-state space. Across vast conformations and deep-mutational-scanning measurements from eight proteins, covariance spectra were concentrated, with 3-18 modes capturing 95% of retained variance; regularized inverse-covariance costs showed weak cross-resolution correspondence and modest fitness association. Biological descriptors predicted selected geometric responses, while structured paths showed lower Gaussian transport than matched nulls (11.75 versus 22.90). GRP is a testable statistical route from AI-observed organization to biological principles across proteins, nucleic acids, complexes and cells. Code and workflows are available at https://github.com/Travis13197/GRP.

Introduction

Proteins do not function as single static structures, but through coupled changes across sequence, structure, conformational ensembles and trajectories. Sequence constrains the accessible states, structure specifies an instantaneous configuration, an ensemble describes their probability distribution, and a trajectory records transitions through that distribution. These four levels are commonly studied separately, although together they define the biological existence of a protein. The relationship between amino-acid sequence and the native three-dimensional structure has long been understood as a central organizing problem in molecular biology[1]. Energy-landscape theories have provided a physical framework for describing protein folding and conformational transitions as motion over rugged but structured free-energy surfaces[2]. Experimental and computational studies have further shown that proteins populate heterogeneous conformational ensembles and undergo functionally relevant motions across multiple timescales[3,4]. Elastic-network and normal-mode approaches have yielded important insights into collective motions, allostery and large-scale conformational changes, but they usually begin from predefined coordinates, potentials or reduced degrees of freedom[5]. ***This static and a priori choice of objects and coordinates may be a fundamental obstacle to the deeper mathematization of biology: it describes where a system is more readily than how its space of possibilities is organized.*** A theoretical framework that unifies sequence, structure, ensemble and trajectory as related levels of one state space of biological entity is therefore needed.

We introduce GRP a general-relativity-inspired framework for the effective geometry of protein-state space as an effective description of this multilevel protein world. Rather than treating sequence, structure, ensemble or trajectory as a complete description in isolation, GRP takes their coupled state space as the primary object of analysis: sequence indexes and constrains the space, structure is a state within it, the ensemble is a probability measure over accessible states, and the trajectory is a path through those states. **The conceptual shift is from objects to distributions, from static configurations to organized possibilities, and from a science of points to a science of paths.** To make this complex biological organization mathematically tractable, AI-based molecular representations provide computational coordinates to transform discrete, heterogeneous and strongly coupled biological observations into computational objects whose covariance structure, effective metrics and path geometry can be quantified. Recent protein language models demonstrate that learned representations can encode substantial structural and evolutionary information even when trained primarily on sequence data[6-9]. This principle is not intrinsically limited to proteins; it may extend to DNA, RNA, molecular complexes, cells and other systems for which AI can construct informative state representations.

We hypothesize that ***the physical world of biomolecules and the world of AI representations form an <u>effective mirror world</u>***: biological constraints leave statistical

traces in learned representations, and the mathematical structure of those traces can generate hypotheses about the original system. This correspondence is necessarily lossy, task-dependent and scale-limited rather than an identity between representation and reality. ***AI is therefore not the biological world itself, but a scientific instrument that can make otherwise inaccessible collective structure measurable and mathematically investigable.*** Its role is not merely to predict predefined outcomes, but to expose candidate variables and relations that can subsequently be translated into physical quantities and tested experimentally.

Inspired by General Relativity, GRP adopts a methodological rather than literal analogy. General Relativity shifted attention from coordinates to the relations that remain meaningful under coordinate transformations; GRP asks the corresponding question for protein space: ***what remains valid or invariant when coordinates, models, scales or representations change? and how can they be described mathematically? What effective laws or equations characterize them?*** It neither identifies learned protein geometry with physical spacetime nor assumes that proteins obey Einstein's equations. Instead, we geometrically formulate the sequence–structure–ensemble–trajectory hierarchy and connect biological constraints to ensemble geometry, local perturbation costs and state-space paths. ***The aim is to attempt to establish a preliminary but testable effective theoretical framework of the structures that remain informative across representations.***

Results

1. Protein states define local effective geometry across protein space

An ensemble was defined as a probability measure over declared protein observations, including conformations, mutations, amino-acid substitutions, generated candidates or learned representations. Each ensemble was analyzed in its own representation space.

If X_{Sj} denotes an observation and $\Phi_{r,S}$ a representation map, then $\mathbf{x}_{Sj}^{(r)} = \Phi_{r,S}(X_{Sj})$. BioEmu generated conformational observations, whereas ProstT5, ProtT5 and ESM-C defined maps into learned spaces. Cartesian ensembles were rigidly aligned; discrete and learned ensembles were analysed through explicit representation maps.

For the primary Cartesian conformational analysis, state S was represented by M_S aligned observations:

$$\hat{\mathrm{P}}_S = \frac{1}{M_S}\sum_{j=1}^{M_S}\delta_{\tilde{\mathbf{x}}_{Sj}}, \qquad \hat{\mathbf{C}}_{S,d} = \frac{1}{M_S-1}\sum_{j=1}^{M_S}\mathbf{z}_{Sj}\,\mathbf{z}_{Sj}^{\top} \qquad \text{(R1)}$$

where

$$\mathbf{z}_{Sj} = \mathbf{Q}_S^{\top}\left(\tilde{\mathbf{x}}_{Sj} - \hat{\boldsymbol{\mu}}_S\right) \qquad \text{(R2)}$$

is the centered coordinate in the retained local statistical subspace. This equation applies to the primary Cartesian conformational analysis. Discrete and learned-representation ensembles are treated through their corresponding representation maps. The covariance describes internal variation after rigid-body alignment rather than variation in the laboratory coordinate frame.

Let $I_S^{\mathrm{sum}} = \{1, \ldots, d_S^{\mathrm{sum}}\}$ denote the leading-mode spectrum used for normalized summaries. Participation ratio, normalized anisotropy and all other normalized spectral summaries use this same index set. The finite-range spectral-decay parameter α_S was estimated over the separately declared fitting interval I_S^{fit}, which may be narrower than I_S^{sum} and is not used to redefine the normalized-spectrum denominators.

Define

$$p_{Si} = \frac{\lambda_{Si}}{\sum_{k=1}^{d_S^{\mathrm{sum}}}\lambda_{Sk}}, \; i = 1, \ldots, d_S^{\mathrm{sum}}$$

We quantified covariance organization by

$$\mathrm{PR}_S = \left(\sum_{i=1}^{d_S^{\mathrm{sum}}} p_{Si}^2 \right)^{-1}, \; a_S^{\mathrm{norm}} = \frac{d_S^{\mathrm{sum}} p_{S1} - 1}{d_S^{\mathrm{sum}} - 1} \qquad \text{(R3)}$$

These quantities describe covariance organization in the declared representation and are not intrinsic manifold, topological or physical dimensions.

Rigid-body alignment revealed concentration of conformational variance in a limited number of empirical modes. Depending on the sequence and ensemble, approximately 3-18 modes accounted for 95% of the retained covariance: the corresponding ranks were 3 for G10, 11 for G30, 18 for G50 and 13 for K50. These threshold-based ranks were treated as sample-size- and representation-dependent descriptors rather than as intrinsic dimensions (***Figure 1***).

Under the prespecified chain-length models, the four residues with the strongest positive participation-ratio scaling, K, G, T and H, showed a robust increase in participation ratio with chain length (pooled log–log Pearson $r = 0.78$, $P = 7.5 \times 10^{-43}$, $n = 208$, ***Fig. 1e***). The fitted spectral-decay parameter showed a strong negative chain-length association ($\beta = -0.66$, $R^2 = 0.95$, ***Fig. 1f***) over the analyzed finite range. These coefficients describe finite-range associations within the analysed representation and sampling domain rather than universal scaling exponents. Anisotropy also decreased with chain length ($\beta = -0.343$, $R^2 = 0.142$, $P = 8.9 \times 10^{-38}$), indicating systematic reorganization of ensemble geometry rather than simple rescaling.

Residue chemistry modulated this reorganization. Effective-rank expansion was strongest for PolyG, GGGGS and PolyLys, whereas hydrophobic residues showed weaker or nonsignificant participation-ratio scaling. Spectral decay, in contrast, decreased consistently across the examined amino-acid and linker classes ($\rho < -0.83$). Within the analysed residue classes, spectral organization was more reproducible than effective-rank scaling, whose interpretation remained more composition dependent. These findings identify reproducible organization in aligned conformational ensembles under the tested sampling and representation protocols.

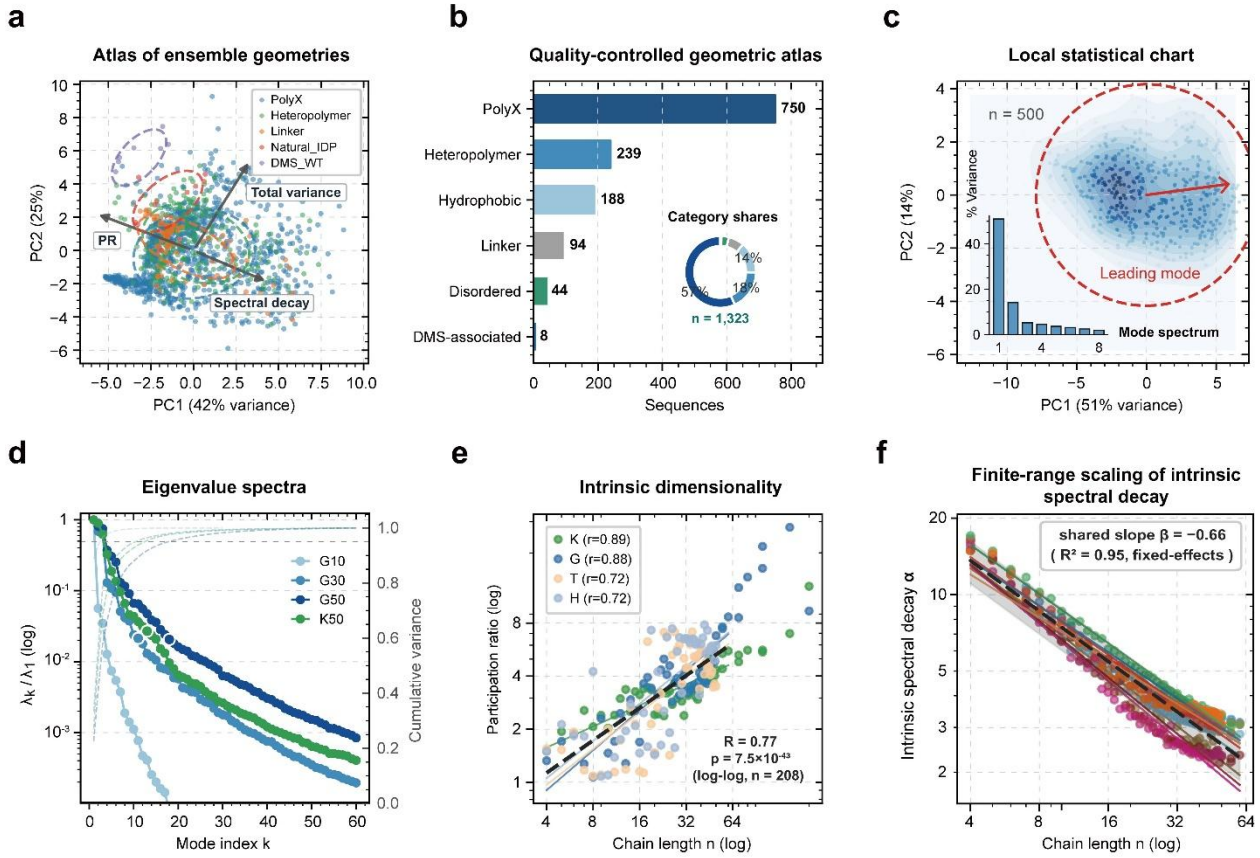


Figure 1 | Protein ensembles define local effective geometry. a, Principal-component atlas of sequence-defined conformational ensembles based on standardized intrinsic covariance and structural descriptors. Dashed ellipses denote category clouds and arrows show leading observable loadings. **b,** Composition of the quality-controlled atlas comprising 1,323 sequence-defined systems: PolyX (750), heteropolymer (239), hydrophobic-gradient L1 (188), linker (94), intrinsically disordered (44) and DMS-associated (8). **c,** Local statistical chart of the PolyG30 ensemble after rigid-body alignment, centring and covariance decomposition; the inset shows the leading covariance spectrum. **d,** Normalized eigenvalue spectra and cumulative explained-variance curves for G10, G30, G50 and K50; the 95% explained-variance ranks are 3, 11, 18 and 13, respectively. **e,** Participation-ratio scaling with chain length for the four residues with the strongest positive association (K, G, T and H). **f,** Finite-range spectral-decay scaling over the analysed chain-length interval across the 20 amino acids. Participation ratio, effective rank and spectral-decay parameters are statistical summaries of the declared representation.

2. Covariance-normalized perturbation geometry captures mutational constraint

The same sequence perturbation can produce different raw structural displacements in different proteins because each background has a distinct set of accessible fluctuations. A displacement along a highly variable direction may be biologically less exceptional than an equal Euclidean displacement along a strongly constrained direction.

We quantified perturbations relative to the regularized covariance of the unperturbed background. Using the operational Ledoit-Wolf estimator:

$$\mathbf{G}_S = \hat{\mathbf{C}}_{S,\mathrm{LW}}^{-1},\ C_{\mathrm{geo}}(\mathcal{P} \mid S) = \delta\mathbf{z}_{\mathcal{P}|S}^{\top}\mathbf{G}_S\delta\mathbf{z}_{\mathcal{P}|S} \qquad \text{(R4)}$$

This regularized Mahalanobis-type cost assigns lower weight to directions with large background variance and higher weight to directions with small regularized background variance. It measures perturbational atypicality in the declared representation; it is not a force, elastic energy, mutation free energy or Hessian-derived mechanical cost. Raw and covariance-normalized costs were compared only through the prespecified diagnostic quantities defined in Methods. No direct difference between the raw squared displacement and the covariance-normalized cost was used as an invariant biological endpoint.

We examined the diagnostic behaviour of covariance normalization using 414 controlled adjacent-chain-length perturbations across nine residue classes. Normalization increased directional consistency ($P = 9.63 \times 10^{-6}$, Cohen's $d = 0.291$, ***Fig 2b***) and reduced the coefficient of variation of perturbation costs ($P = 6.27 \times 10^{-106}$, Cohen's $d = 2.084$, ***Fig 2c***). Because whitening can mechanically alter angular similarity and relative scalar variability, these statistics were interpreted as transformation diagnostics.

Scalar geometric costs showed a weak positive correspondence between matched Cα and full-atom representations ($r = 0.219,\ P = 1.8 \times 10^{-5},\ n = 376$). Across eight proteins and 84,361 mutations, the mean protein-level association between $C_{\mathrm{geo}}$and experimental fitness was $\rho = -0.1475$, with all eight protein-level associations negative. The uncertainty of this summary was estimated by protein-level resampling rather than mutation-level resampling. Its predictive contribution was assessed against sequence-based baselines using identical mutation sets and grouped held-out evaluation. The total geometric cost was not treated as a complete fitness predictor.

Together, these analyses support a representation-dependent covariance-normalized perturbation construction. Its biological relevance was evaluated through real-mutant responses, cross-resolution scalar-cost correspondence and association with experimental fitness.

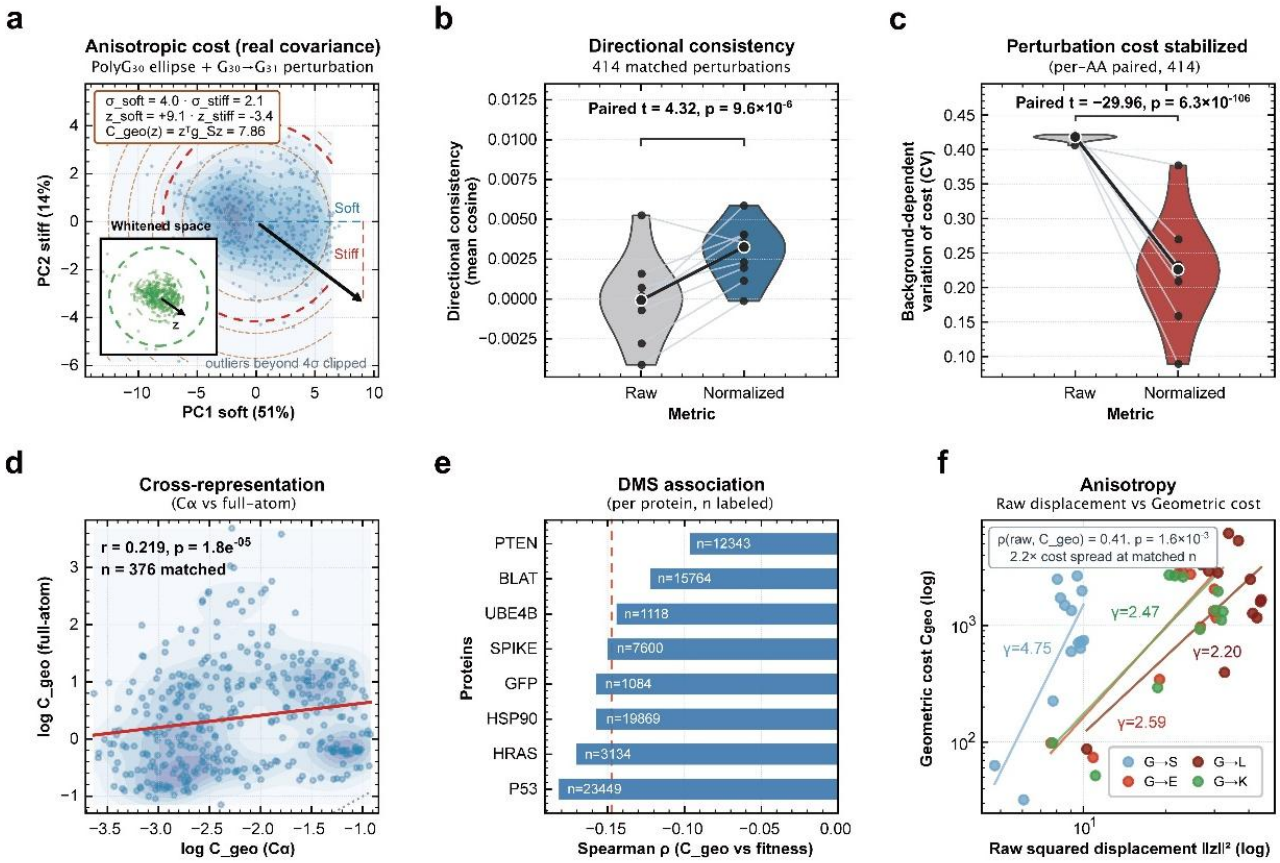


Figure 2 | Local covariance normalization makes perturbations comparable across protein backgrounds. a, Real PolyG30 fluctuation ellipse with the G30→G31 perturbation decomposed into high- and low-variance directions; the anisotropic contours represent $C_{\mathrm{geo}} = \delta\mathbf{z}^{\top}\mathbf{G}_{\mathrm{S}}\delta\mathbf{z}$, and the inset shows the corresponding normalized space. **b,** Directional consistency across 414 matched adjacent-chain-length perturbations before and after normalization ($d = 0.29$, $P = 9.6 \times 10^{-6}$, win rate 58.7%). **c,** Background-dependent coefficient of variation of perturbation cost before and after normalization (0.417 versus 0.224; Cohen's $d = 2.08$, win rate 92.0%); this was the dominant transformation effect. Panels b and c are interpreted as transformation diagnostics rather than independent biological validation; their differences correspond to the prespecified $\Delta_{\mathrm{cons}}$and Δ_{CV}, not to an additional ΔC_{geo}endpoint. **d,** Cross-representation correspondence between Cα- and full-atom-derived costs ($r = 0.219$, $P = 1.8 \times 10^{-5}$, $n = 376$). **e,** Protein-level association between C_{geo} and DMS fitness across eight proteins and 84,361 mutations (mean $\rho = -0.1475$; all eight protein-level associations negative). **f,** Anisotropy analysis showing that raw squared displacement is only a weak proxy for covariance-normalized cost across 56 G→X perturbations ($\rho = 0.41$, $P = 1.6 \times 10^{-3}$). Panels b and c are interpreted as transformation diagnostics because whitening can alter these statistics mechanically. C_{geo} is a statistical atypicality measure.

3. Biological constraints define a predictive effective-geometric field

Biological source variables were represented by a non-redundant vector containing chain length, identifiable amino-acid composition contrasts, fixed sequence-derived physicochemical descriptors and independently defined contextual variables. Predictors identical to, deterministically derived from or estimated from the same target ensemble as the response were excluded from the corresponding model. Composition coordinates were parameterized identifiably, and hydrophobicity and charge were not interpreted independently when algebraically redundant with amino-acid composition.

Because response coordinates depend on representation, the source-to-geometry relation was defined as

$$\boldsymbol{F}_r(\boldsymbol{t}) = \mathbb{E}\left[\widetilde{\boldsymbol{Y}}_S^{(r)} \mid \widetilde{\boldsymbol{T}}_S = \boldsymbol{t}\right] \qquad \text{(R5)}$$

Near a reference point $\boldsymbol{t}_0$,

$$\boldsymbol{F}_r(\boldsymbol{t}) = \boldsymbol{F}_r(\boldsymbol{t}_0) + \boldsymbol{K}_r(\boldsymbol{t}_0)(\boldsymbol{t} - \boldsymbol{t}_0) + \boldsymbol{R}_{r,2}(\boldsymbol{t}; \boldsymbol{t}_0) \qquad \text{(R6)}$$

where

$$\boldsymbol{K}_r(\boldsymbol{t}_0) = D\boldsymbol{F}_r(\boldsymbol{t}_0)$$

is the local response operator and $\boldsymbol{R}_{r,2}$ contains higher-order terms. The corresponding local affine approximation is

$$\widetilde{\boldsymbol{Y}}_S^{(r)} = \boldsymbol{b}_{r,0} + \boldsymbol{K}_r(\boldsymbol{t}_0)\widetilde{\boldsymbol{T}}_S + \boldsymbol{\varepsilon}_{r,S}^{\text{lin}} \qquad \text{(R7)}$$

with

$$\boldsymbol{b}_{r,0} = \boldsymbol{F}_r(\boldsymbol{t}_0) - \boldsymbol{K}_r(\boldsymbol{t}_0)\boldsymbol{t}_0 \qquad \text{(R8)}$$

The operator $\boldsymbol{K}_r(\boldsymbol{t}_0)$ depends on the source and response coordinates and is not a coordinate-independent physical constant.

To distinguish shared from system-specific structure, we used the representation-indexed model

$$\widetilde{\boldsymbol{Y}}_{S,c}^{(r)} = \boldsymbol{b}_{r,0,c} + \left(\boldsymbol{K}_{r,0} + \Delta\boldsymbol{K}_{r,c}\right)\widetilde{\boldsymbol{T}}_{S,c} + \boldsymbol{\varepsilon}_{r,S,c} \qquad \text{(R9)}$$

The field equation was tested through three criteria: whether biological source terms predict geometric observables, whether the response structure remains stable under resampling and representation changes, and whether a shared component transfers

across systems. For a predictive loss $\mathcal{R}$, the corresponding criterion was

$$\Delta\mathcal{R}_{\text{field}} = \mathcal{R}_{\text{test}}\left(\hat{F}_{\text{null}}\right) - \mathcal{R}_{\text{test}}\left(\hat{F}\right) > 0 \qquad \text{(R10)}$$

Biological source terms predicted a substantial and reproducible component of ensemble geometry: 73.6% of per-amino-acid responses reached $R^2 > 0.1$. The strongest per-amino-acid mean response reached $R^2 = 0.712$, whereas the strongest observable-level mean response reached $R^2 = 0.873$ (***Fig. 3b***). Across the five prespecified shared-component tests, the estimated response organization remained reproducible, although predictive performance varied across held-out systems. The result supports a partially transferable source-to-geometry component rather than universal cross-system invariance. The coupling structure was also objectively stable, with bootstrap clustering yielding an adjusted Rand index of 0.708 and all 18 leave-one-amino-acid-out analyses yielding ARI $= 1.0$ (***Fig. 3***).

These results support a transferable biological response field, but not a universal coordinate-free law. The stable object was the organization of the estimated response rather than any individual coefficient. Progressively richer sequence- and graph-based descriptors did not recover this organization under grouped held-out evaluation; the final graph model achieved $R^2 = -0.1553$. This establishes non-recovery within the tested model family, not absolute mathematical irreducibility.

The field equation further separated biological effects across distinct geometric axes. Chain length, density and effective dimensionality formed a predictable state relation ($R^2 = 0.7201$), whereas geometric distortion and anisotropy followed a different, weaker relation ($R^2 = 0.3273$). Hydrophobicity, charge and chain length therefore should not be interpreted as interchangeable drivers of a single geometric variable. Instead, their conditional responses distribute across partially distinct axes, including spatial association, compactness and other prespecified response axes.

Together, these analyses support a predictive, representation-dependent relation between biological descriptors and selected geometric responses. They do not establish a coordinate-free or causal field law.

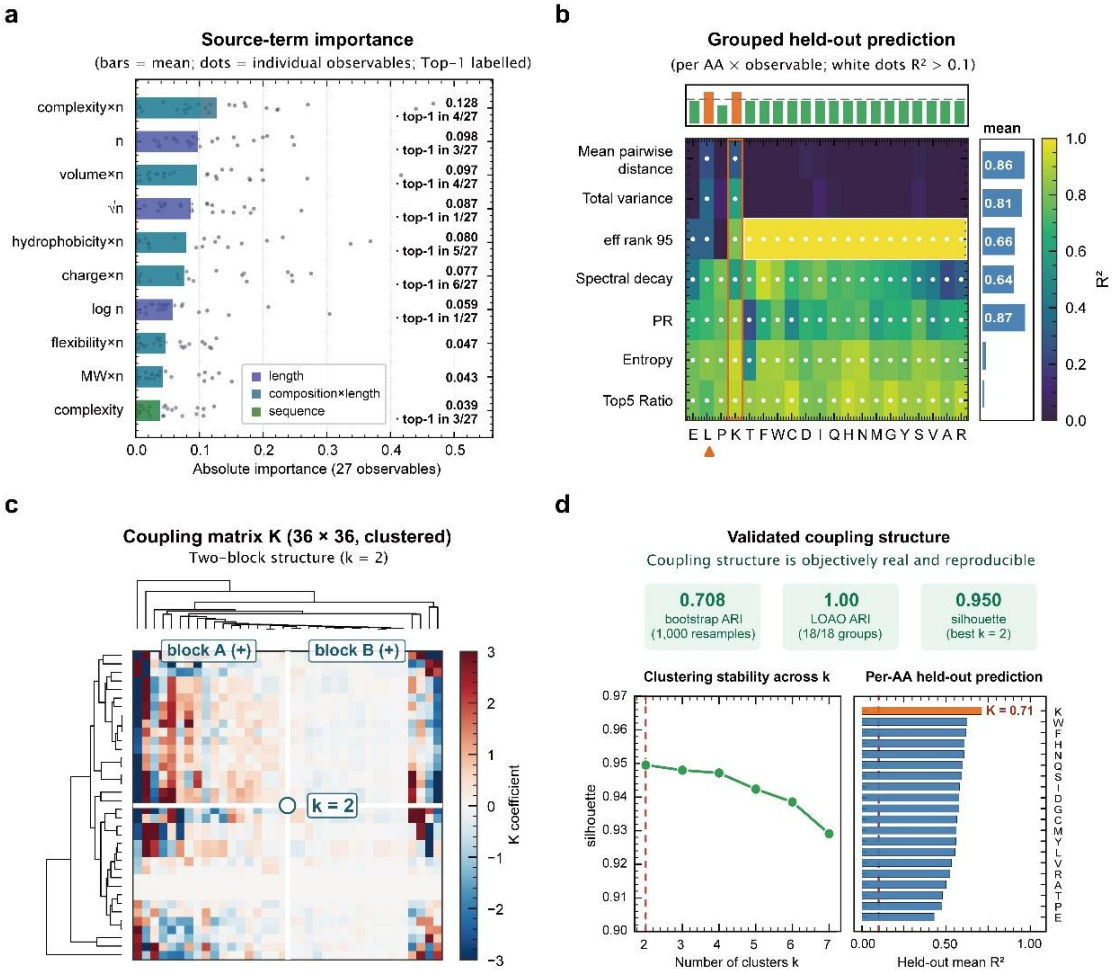


Figure 3 | Biological source variables predict an effective geometric response field. a, Importance of the leading source variables across 27 prespecified geometric observables. **b,** Grouped held-out prediction of the per-amino-acid-by-observable response; 73.6% of responses achieved $R^2 > 0.1$, the strongest per-amino-acid mean response reached $R^2 = 0.712$, and the strongest observable-level mean response reached $R^2 = 0.873$. **c,** The 36×36 response-coupling matrix shows a reproducible two-block organization; bootstrap clustering yielded mean ARI $= 0.708$ with median ARI $= 0.696$, and leave-one-amino-acid-out analyses yielded ARI $= 1.0$across 18 groups. **d,** Stability and descriptor-reduction analyses assess whether the coupling organization is recovered by the prespecified sequence-, physicochemical- and graph-based models.

4. Structured protein-state paths show reduced mean Gaussian transport

An ordered protein-state path was represented as

$$\gamma = (S_0, S_1, \ldots, S_N),\ P_t^\star = \left(\iota_{r,t}\right)_{\#} \widehat{P}_{S_t}^{(r)}$$

where all state measures were mapped into the same prespecified comparison space. The path may represent a controlled chain-length progression, a mutation sequence, a generated sequence of candidates or another ordered ensemble construction. Path order is part of the definition, but it is not interpreted as physical time unless an independent dynamical process is available.

The directly tested path statistic was the mean adjacent-state Gaussian transport,

$$\bar{W}_{2,\mathrm{G}}(\gamma) = \frac{1}{N}\sum_{t=0}^{N-1} W_{2,\mathrm{G}}(P_t^\star, P_{t+1}^\star) \qquad \text{(R11)}$$

This Gaussian-surrogate statistic captures changes in both ensemble location and covariance structure in the common representation space. We tested the conditional low-transport hypothesis

$$\mathbb{E}\left[\bar{W}_{2,\mathrm{G}}(\gamma_{\mathrm{structured}}) \mid \mathcal{C}\right] < \mathbb{E}\left[\bar{W}_{2,\mathrm{G}}(\gamma_{\mathrm{null}}) \mid \mathcal{C}\right] \qquad \text{(R12)}$$

where $\mathcal{C}$ denotes the matched endpoint, transition-number, operation, representation and admissibility constraints.

We first examined controlled sequence-structured paths in which chain length increased incrementally. In the joint-PCA representation, these paths showed substantially lower inter-state transport than matched null paths (***Fig. 4a***). In the primary comparison, the mean path transport was 11.75 for the structured paths and 22.90 for the null paths, corresponding to Cohen's $d = -1.54$ and $P = 8.4 \times 10^{-68}$ (***Fig. 4b***). An independent heteropolymer analysis reproduced the same direction of effect, although with a smaller magnitude: $W_{2,G} = 3.07$ for the structured paths versus 3.46 for the null paths ($\mathrm{d} = -0.20$, $\mathrm{P} = 2.3 \times 10^{-6}$, ***Fig. 4c***). The direct NPZ analysis showed the same direction, with mean adjacent-state transport of 7.62 for ordered paths versus 9.29 for randomized paths ($P = 4.2 \times 10^{-9}$; $n = 673$ ordered and $n = 263$ randomized paths, ***Fig. 4d***).

The comparison remained significant across the four prespecified null constructions after Holm correction, supporting the robustness of the transport difference to alternative path controls. The strongest interpretation is that sequence-structured paths occupy more coherent routes through the sampled protein representation space than the

corresponding null paths. This result does not imply that every transition is small or smooth: a globally low-transport path may contain localized sharp transitions, and local transition-sharpness measures should be interpreted as indicators of concentrated reorganization rather than as independent evidence for a global action minimum.

Structured state paths exhibited comparatively shorter distributional steps under the tested representation and matched-null design. A broader dimensionless action functional was defined as a theoretical extension:

$$\mathcal{A}_2[\gamma] = \sum_{t=0}^{N-1} \left[\frac{\widetilde{W}_{2,\mathrm{G},t}^{\,2}}{2\Delta s_t} + \frac{\Delta \widetilde{\boldsymbol{Y}}_t^{\top} \boldsymbol{\Lambda} \Delta \widetilde{\boldsymbol{Y}}_t}{2\Delta s_t} + \frac{\Delta s_t}{2} \left(V(P_t^{\star}) + V(P_{t+1}^{\star}) \right) \right] \qquad \text{(R13)}$$

This functional was not the primary empirical statistic, and its geometric-reorganization and feasibility terms were not independently validated under the same matched-null design. The present results therefore support a low-transport path relation: structured protein-state paths exhibited lower mean Gaussian transport than the tested matched null paths. Whether this relation extends to a complete low-action principle requires independent validation of a fully specified, dimensionless path functional under strictly matched transition, feasibility and representation constraints.

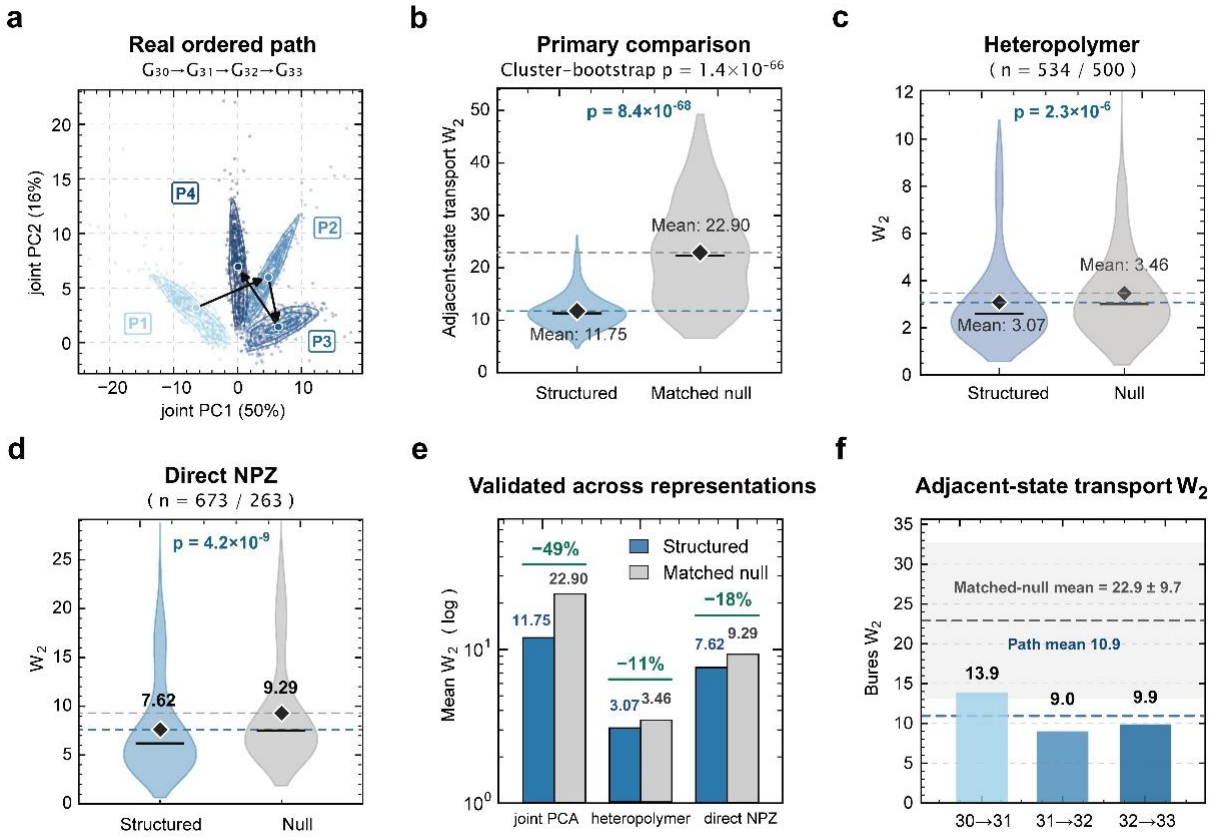


Figure 4 | Structured protein-state paths show reduced mean Gaussian transport. a, A real ordered path is represented as a sequence of empirical ensembles mapped into a joint PCA comparison space; arrows denote adjacent-state Gaussian Bures–Wasserstein transport. **b,** Mean adjacent-state transport was lower for structured paths than for matched null paths in the primary comparison (11.75 versus 22.90; Cohen's $d = -1.54$, $P = 8.4 \times 10^{-68}$). **c,** Independent heteropolymer analysis reproduced the same direction (3.07 versus 3.46; $d = -0.20$, $P = 2.3 \times 10^{-6}$). **d,** Direct NPZ analysis also reproduced the effect (7.62 versus 9.29; $P = 4.2 \times 10^{-9}$). **e,** The low-transport relation was evaluated across the prespecified null constructions and representation controls. **f,** Adjacent-state transport along the representative ordered path is compared with the matched-null distribution.

5. Unified effective-geometric organization

The analyses define distinct mathematical constructions and empirical associations:

$$\widehat{\boldsymbol{P}}_S^{(r)} \longrightarrow \boldsymbol{Y}_S^{(r)},$$

$$\widehat{\boldsymbol{P}}_S^{(r)} \longrightarrow \boldsymbol{G}_S^{(r)},$$

$$\left(\widehat{\boldsymbol{P}}_S^{(r)}, \mathcal{P}\right) \longrightarrow C_{\mathrm{geo}}^{(r)}(\mathcal{P} \mid S),$$

$$\boldsymbol{T}_S \dashrightarrow \boldsymbol{Y}_S^{(r)},$$

$$\{\mathsf{P}_t^{\star}\}_{t=0}^{N} \longrightarrow \overline{W}_{2,\mathrm{G}}^{(r)}(\gamma) \tag{R13}$$

Here $\mathsf{P}_t^{\star}$ denotes the pushforward of the native state measure into the common comparison space. Solid arrows denote mathematical construction or statistical estimation; the dashed arrow denotes grouped prediction. These mappings do not constitute a deterministic causal chain.

Results 1 identified reproducible covariance organization in aligned conformational ensembles. **Results 2** defined regularized perturbational atypicality and evaluated its association with real-mutant responses, cross-resolution scalar costs and mutational fitness. **Results 3** modelled biological source-to-geometry relations as representation-indexed predictive fields. **Results 4** showed that structured state paths exhibited lower mean Gaussian transport than matched null paths. Together, these analyses define a multiscale effective statistical framework connecting protein ensembles, perturbations and ordered state paths.

The extraction of $\boldsymbol{Y}_S^{(r)}$, estimation of the regularized local metric $\boldsymbol{G}_S^{(r)}$, calculation of the representative-displacement cost $C_{\mathrm{geo}}^{(r)}$ and evaluation of $\bar{W}_{2,\mathrm{G}}^{(r)}(\gamma)$ are mathematical constructions or statistical estimators within declared representations. Source-to-response prediction, mutationally associated perturbational atypicality and reduced mean path transport are empirical relations supported only within the tested systems, representations, perturbation classes and matched controls.

In particular, the path analysis directly establishes a **low-transport path relation**, not a complete minimum-action principle. A broader action functional incorporating geometric reorganization, intermediate-state feasibility and path parametrization remains a theoretical extension until all terms are independently specified and tested against strictly matched null paths.

The methodological contribution is threefold. First, protein states are represented as empirical distributions rather than isolated structures, allowing covariance spectra to serve as local statistical summaries. Second, perturbational atypicality is normalized by

the covariance of the unperturbed background, separating raw displacement from background-dependent geometric weighting. Third, ordered state measures are compared through common-space distributional transport, providing a path-level statistic that is distinct from a physical action. These contributions define a representation-aware statistical framework.

The unified result is therefore not a deterministic chain in which sequence causes geometry and geometry uniquely determines biological fate. Rather, sequence-level constraints, structures, ensembles and trajectories constitute distinct but mathematically connected layers of protein state space. Sequence variables provide predictive information about ensemble organization; individual structures locate sampled states; ensemble covariance supplies a local statistical geometry; and ordered ensembles reveal how state-space transport is distributed along a path. This formulation preserves biological context while avoiding the identification of representation geometry with physical spacetime, inverse covariance with a molecular force-constant matrix, geometric cost with free energy, or transport length with a physical action.

The framework provides a representation-aware language for relating sequence variation, ensemble heterogeneity, mutational response and ordered state paths. Its value lies in generating testable hypotheses rather than assigning literal physical meaning to every geometric construct. Independent ensembles, prospective perturbations and preregistered path comparisons will determine which relations transfer to biological measurements.

Discussion

This study should be understood as a bold but preliminary attempt in theoretical physic-biology or AI-augmented biology, rather than as a completed contribution to conventional biophysics. It seeks to formulate and begin testing a possible unified framework for relating sequence, structure, ensembles and ordered state paths within biology. The proposed GRP framework is exploratory, mathematically incomplete and necessarily dependent on declared representations, sampling procedures and empirical controls. The present manuscript should therefore be regarded as an initial **draft** of a broader theoretical programme, whose proposed relations and evaluation criteria require further development, independent datasets, prospective perturbations and experimental validation before they can be considered established biological laws.

At the ontological level, the deepest contribution of GRP (General Relativity of Protein) is a shift in the primary object of molecular biology: ***The object of study is no longer the molecule represented in isolation, but the state space that expresses the molecule's physically and biologically accessible possibilities***. A protein is not adequately identified with a sequence, a single structure or even an isolated conformational ensemble. Rather, A state space constrained jointly by thermos-dynamics, molecular dynamics, sequence, environment and biological context. Sequence indexes this space; a structure identifies one point or local configuration within it; an ensemble defines a probability distribution over accessible states; and a trajectory describes an ordered path through those states. This ontological shift does not deny the reality of sequences, structures or ensembles. Instead, it places them at different descriptive levels of the same object. The sequence specifies the molecular system and its constraints, the structure records a particular realization, the ensemble captures the distribution of possible realizations, and the trajectory describes how those possibilities are traversed or transformed. ***The fundamental scientific task of future molecular biology is therefore not merely to catalogue these objects, but to identify the causal relations and mathematical-physical laws connecting the corresponding state spaces***.

In this view, a protein - and more broadly, any biological molecule - is not a point embedded in a pre-existing space, but a **structured space of possibilities**. In AI-mediated representation spaces, we hypothesize that ***these distributions may concentrate near comparatively low-dimensional manifolds*** *or structured covariance subspaces*. Their geometry can therefore become an object of quantitative investigation rather than a target of visualization alone. GRP consequently shifts the emphasis of molecular biology from static entities to distributions, from points to states and paths, and from molecular identity to organized possibility. In its broader form, this programme could extend beyond proteins to DNA, RNA, molecular complexes, virtual cells and other biological systems that can be represented and experimentally interrogated through computational models. The long-term implication is that ***biological laws may eventually be formulated as geometric relations on the state***

spaces and transformations underlying biological systems. Such a programme may require the language of differential geometry, tensors and probability-based metrics. This is a methodological and theoretical aspiration, and the immediate contribution of GRP is to formulate biological objects in a way that makes their distributions, local scales, perturbations and paths mathematically expressible and empirically testable.

This interpretation also provides a direct theoretical context for the three universal biological principles identified in ***ProtGenesis***[9]. Their common character is observational: short-order protein genesis shows local linear similarity and approximately additive organization in the relevant sequence–representation coordinates. Under GRP, such behaviour is expected when the accessible protein states occupy a locally low-dimensional effective manifold. Within a sufficiently small neighbourhood, the manifold is approximated by a locally flat statistical subspace, so nearby state changes can be represented to first order by linear coordinates and combined approximately additively. The ProtGenesis observations therefore provide a concrete biological manifestation of the local geometric regime proposed by GRP, while GRP supplies a general explanation for why short-order similarity and additive composition can emerge from constrained protein-state organization.

At the epistemological level, a central insight of GRP is that ***AI can function as a scientific instrument rather than merely as a black-box prediction tool***. AI transforms discrete, variable-length and strongly interacting biological states into representations that can be sampled, quantified and mathematically analyzed. The resulting scientific chain can be expressed as

$$\text{Reality} \longrightarrow \text{AI representation} \longrightarrow \text{geometric structure} \longrightarrow \text{invariant} \longrightarrow \text{candidate physical law} \longrightarrow \text{reality}$$

This epistemic programme is not limited to biology. In the present study, AI provides a new means of observing and measuring aspects of protein-state organization, enabling geometric descriptions of conformational distributions, perturbations and paths. AI did not create protein space. ***It made previously difficult-to-observe aspects of protein space computationally accessible - much as Galileo's telescope revealed, rather than created, the moons of Jupiter.***

The real biological world and the AI representation world should therefore be treated as ***effective mirror spaces***, not as ontologically identical domains. The biological space possesses physical causality, whereas the AI space possesses representational and predictive efficacy. Their scientific significance lies in the experimentally testable relations that connect them: a structure inferred in an AI representation becomes scientifically meaningful when it corresponds to measurable conformational properties, perturbation responses, functional outcomes or path probabilities across independent models and experiments.

We understand GRP as a form of **structural realism**, or more precisely, **AI-observation-mediated structural realism**. The biological world cannot be transferred into a computer without loss. Experiments, simulations, generative models and language models provide different, incomplete and task-dependent projections of that world, each with its own resolution, inductive biases and failure modes. GRP therefore does not seek one uniquely correct representation. It seeks **relationships and structures that remain stable across representations, models, resolutions and biological systems, invariants**. A relation confined to one network layer or coordinate basis remains a model-specific fact; it approaches the status of an empirical law only when it can be translated into measurable biological quantities and retains its functional form under independent representations and experiments.

Within this framework, four coupled but non-identical spaces should be distinguished. The biological space $\mathcal{B}$ contains sequences, atomic structures, molecular environments, energies, dynamics and functional outcomes. The AI representation space $\mathcal{R}_{\mathrm{AI}}$ is shaped by training data, model architecture and learning objectives. The geometric feature space $\mathcal{G}$ contains covariances, spectra, local metrics, response operators and topological summaries. The path space $\mathcal{P}$ contains ordered state sequences or distributional curves, together with statistical transport or action-like functionals. These spaces are related by effective mappings, for example,

$$\mathcal{B} \xrightarrow{\Phi_{\mathrm{AI}}} \mathcal{R}_{\mathrm{AI}} \xrightarrow{\Phi_{\mathrm{geo}}} \mathcal{G},\ \gamma\colon [0,1] \to \mathcal{P}$$

but they should not be treated as identical, isomorphic or interchangeable. ***Their relationship is better understood as an incomplete, task-dependent and scale-limited correspondence whose adequacy must be judged by cross-representation stability, independent simulation and experimental intervention.***

This framework also distinguishes three forms of causality. The biological space contains physical causality: sequence changes alter molecular interactions, environments reshape energy landscapes and state occupancies, and molecular dynamics generate conformational transitions. The AI space contains representational causality: architecture, training objectives and data distributions determine which relations are preserved, amplified or compressed in the representation. Finally, **AI-guided experimental design can establish a predictive-interventional causal chain** by using model-derived hypotheses to select and modify biological systems, followed by independent measurement of the predicted response. AI therefore mediates the transition from representation to prediction and from prediction to experimentally testable intervention. In the present framework, the mappings from empirical distributions to geometric summaries, from distributions to local metrics, and from metrics to normalized perturbation costs are mathematical constructions or statistical estimations. The source-to-geometry relation is predictive, and the observed low-

transport relation for structured paths is an empirical association relative to matched null models; none of these relations should be interpreted as a completed causal theory without prospective intervention and independent experimental confirmation.

The central claim of GRP is consequently not that AI representations are reality, nor that every geometric quantity extracted from an embedding has physical meaning. Rather, GRP proposes a scientific programme in which AI helps discover candidate collective structures, geometry expresses those structures in a coordinate-aware and quantitatively testable form, and experiment determines which of them belong to nature. ***The ultimate objective is not to make biology dependent on AI, but to use AI to identify relations that can eventually be reformulated as experimentally measurable and, where justified, mathematically and physically grounded laws of biological systems***.

Reproducibility and computation

All analyses used fixed configuration files, recorded random seeds and version-controlled scripts. The computational workflow stored state-level descriptors, ensemble identifiers, covariance estimates, perturbation-level costs, path-level transport values, null-path identifiers and held-out predictions. Numerical decompositions, matrix inversions, covariance regularization, transport calculations and failed-run handling followed prespecified stability rules. The repository release associated with this manuscript records the model versions, dataset versions, software environment, dependency versions, configuration files and source files used to generate each reported table and figure. Each reported result is linked to a machine-readable source file and a corresponding regeneration script.

Data availability

All redistributable code, processed data products and analysis workflows are available at https://github.com/Travis13197/GRP. The analyses reported in this manuscript were generated from release **GRP1.0**. External datasets and model outputs that cannot be redistributed are identified by their source metadata, version information and licensing restrictions. For these resources, the repository provides the derived tables, preprocessing scripts, configuration files and regeneration procedures permitted by the original data licenses.

Figure-specific source data, reproduction files and figure descriptions are provided in the figures directory at https://github.com/Travis13197/GRP/tree/bc40dfc/figures. For each figure, the directory provides the data files used to generate the plotted quantities together with a description of the panels, variables, preprocessing steps and reported statistical summaries. External datasets or model outputs that cannot be redistributed are identified by their original sources, version information and licensing restrictions; where redistribution is restricted, the repository provides the permitted metadata, derived tables and regeneration procedures.

Author contributions

Chuanyang Liu conceived the central scientific hypothesis and overall framework of GRP, designed the multilevel organization of the study, formulated the principal research questions, developed the empirical and theoretical architecture, constructed the key mathematical objects and evaluation criteria, and performed the computational analyses and experimental or empirical investigations. ***Chuanyang Liu*** also determined the data sources, ensemble-construction procedures, representation analyses, statistical designs, control strategies and biological interpretation, and prepared the manuscript.

GPT-5.6 (sol) assisted with selected mathematical derivations and formula

development and performed a comprehensive consistency review of the equations, notation, dimensional relationships, logical deductions and definitions of the proposed evaluation metrics. GPT-5.6 (sol) also assisted with language editing and structural clarification. The AI system did not determine data inclusion, generate or execute the computational analyses, perform experiments, select the final statistical results or make the primary scientific decisions and conclusions.

Funding

This work received no specific funding from public, commercial or not-for-profit funding agencies.

Reference

1 Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223-230 (1973). https://doi.org/10.1126/science.181.4096.223

2 Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. *Annual Review of Physical Chemistry* **48**, 545-600 (1997). https://doi.org/10.1146/annurev.physchem.48.1.545

3 Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838-844 (2007). https://doi.org/10.1038/nature06410

4 Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128-132 (2005). https://doi.org/10.1038/nature03199

5 Atilgan, A. R. *et al.* Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal* **80**, 505-515 (2001). https://doi.org/10.1016/S0006-3495(01)76033-X

6 Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021). https://doi.org/10.1073/pnas.2016239118

7 Candido, S. *et al.* Language Modeling Materializes a World Model of Protein Biology. *bioRxiv*, 2026.2006.2003.729735 (2026). https://doi.org/10.64898/2026.06.03.729735

8 Heinzinger, M. *et al.* Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* **6**, lqae150 (2024).

9 Liu, C. *et al.* Universal physical principles govern the deterministic genesis of protein structure. *bioRxiv*, 2026.2002.2020.706798 (2026). https://doi.org/10.64898/2026.02.20.706798

10 Lewis, S. *et al.* Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **389**, eadv9817 (2025). https://doi.org/10.1126/science.adv9817

Materials and Methods

Part I: Empirical State Measures and Intrinsic Covariance Geometry

Scope and object of analysis

GRP represents a protein state by a probability measure over declared observations, including conformations, mutations, amino-acid substitutions, generated candidates and learned representations. A sampling or enumeration rule defines the observations, and a representation map places them in an analysis space. The resulting geometry depends on the ensemble construction, representation, preprocessing and finite-sample estimator.

Let $\mathcal{X}_S$ denote the native observation space of state S, and let

$$\Phi_{r,S}: \mathcal{X}_S \to \mathcal{H}_r,\ x_{Sj}^{(r)} = \Phi_{r,S}(X_{Sj}) \qquad \text{(M1)}$$

denote the representation map at level r. The space $\mathcal{H}_r$ may contain Cartesian coordinates, learned embeddings, mutation features or generated-sequence features. Distinct representation levels need not be isometric or invertible.

BioEmu[10] ensembles are treated as model-defined conformational distributions. ProstT5[8], ProtT5 and ESM-C[7] provide representation maps for declared sequence or structural observations. Mutation and generation ensembles are measures over specified alternatives. These measures carry no automatic equilibrium interpretation. Covariance concentration describes organization in the declared effective space.

Representation of protein observations

Let S denote a protein state under a declared ensemble-construction rule and representation specification. Let

$$X_{S1}, \ldots, X_{SM_S} \in \mathcal{X}_S$$

denote M_S state-specific observations. An observation may be a conformation, sequence variant, amino-acid substitution, generated candidate, structural state or learned model output. At representation level r, define

$$\Phi_{r,S}: \mathcal{X}_S \longrightarrow \mathcal{H}_r, \qquad \mathbf{x}_{Sj}^{(r)} = \Phi_{r,S}(X_{Sj}) \qquad \text{(M2)}$$

When $\mathcal{H}_r$ is finite dimensional, $\mathbf{x}_{Sj}^{(r)} \in \mathbb{R}^{D_{r,S}}$. The ambient dimension may depend on the representation and state. Cartesian $C\alpha$ and full-atom coordinates define conformational representations; ProstT5, ProtT5 and ESMC define learned sequence or structure representations; mutation and generation analyses define additional ensemble spaces through their declared feature maps. Covariance, transport and displacement calculations are performed only within a common vector or metric space, or after an

explicit map into such a space.

For a protein with N_S represented atoms and Cartesian coordinates, an individual conformation is written as

$$\mathbf{X}_{Sj} = \begin{bmatrix} \mathbf{x}_{Sj,1}^{\top} \\ \vdots \\ \mathbf{x}_{Sj,N_S}^{\top} \end{bmatrix} \in \mathbb{R}^{N_S \times 3}, \qquad j = 1, \dots, M_S \qquad \text{(M3)}$$

where M_S is the number of sampled conformations and $\mathbf{x}_{Sj,a} \in \mathbb{R}^3$ is the coordinate of atom a in sample j. The vectorized Cartesian representation is

$$\mathbf{x}_{Sj} = \mathrm{vec}(\mathbf{X}_{Sj}) \in \mathbb{R}^{D_S}, \qquad D_S = 3N_S \qquad \text{(M4)}$$

For $C\alpha$ representations, N_S is the number of residues. For full-atom representations, N_S depends on the atom topology of the sequence. Consequently, states with different lengths, atom counts or representation resolutions do not automatically belong to a common fixed-dimensional Euclidean space. Direct coordinate differences, cross-state covariance matrices and ordinary Wasserstein distances require an explicitly declared correspondence or a map into a common comparison space.

The primary spectral analysis does not require that all proteins share one coordinate vector space. Instead, each ensemble is analyzed in its own aligned coordinate representation, and the resulting scalar spectral summaries are compared across states. This distinction is important: comparisons of participation entropy, spectral-decay parameters or other scalar summaries across different chain lengths are not equivalent to computing a coordinate-wise displacement between proteins of different lengths.

Rigid-body quotient and Kabsch alignment

Cartesian protein coordinates contain arbitrary global translation and rotation. These degrees of freedom are not internal conformational variables and must be removed before estimating covariance structure. Formally, Cartesian conformations are considered modulo the action of the rigid-body group $SE(3)$. In practice, a representative coordinate frame is obtained through a Procrustes-type alignment based on optimal translations and rotations.

Let $\mathbf{X}_{Sj} \in \mathbb{R}^{N_S \times 3}$ be a sampled conformation and let $\mathbf{M}_S \in \mathbb{R}^{N_S \times 3}$ be the current reference conformation or iterative ensemble mean. Define the centroids

$$\bar{\mathbf{x}}_{Sj} = \frac{1}{N_S} \sum_{a=1}^{N_S} \mathbf{x}_{Sj,a}, \qquad \bar{\mathbf{m}}_S = \frac{1}{N_S} \sum_{a=1}^{N_S} \mathbf{m}_{S,a} \qquad \text{(M5)}$$

and the centered coordinate matrices

$$\mathbf{X}_{Sj}^{c} = \mathbf{X}_{Sj} - \mathbf{1}_{N_S}\bar{\mathbf{x}}_{Sj}^{\top}, \qquad \mathbf{M}_{S}^{c} = \mathbf{M}_{S} - \mathbf{1}_{N_S}\bar{\mathbf{m}}_{S}^{\top}$$

The optimal Kabsch rotation is defined by

$$\widehat{\mathbf{R}}_{Sj} = \underset{\mathbf{R}\in SO(3)}{\operatorname{argmin}} \left\| \mathbf{X}_{Sj}^{c}\mathbf{R} - \mathbf{M}_{S}^{c} \right\|_{F}^{2} \qquad \text{(M6)}$$

This alignment is applied only to Cartesian conformational representations. Learned embeddings, mutation representations and sequence-generation ensembles are not subjected to rigid-body alignment unless their representation map explicitly produces Cartesian coordinates.

Let

$$\left(\boldsymbol{X}_{Sj}^{c}\right)^{\top}\boldsymbol{M}_{S}^{c} = \boldsymbol{U}_{Sj}\boldsymbol{\Sigma}_{Sj}\boldsymbol{V}_{Sj}^{\top}$$

be the singular-value decomposition of the cross-covariance matrix. Under the row-coordinate convention used here, the proper-rotation solution is

$$\widehat{\boldsymbol{R}}_{Sj} = \boldsymbol{U}_{Sj}\boldsymbol{D}_{Sj}\boldsymbol{V}_{Sj}^{\top}, \;\; \boldsymbol{D}_{Sj} = \operatorname{diag}\left(1, 1, \det\left(\boldsymbol{U}_{Sj}\boldsymbol{V}_{Sj}^{\top}\right)\right) \qquad \text{(M7)}$$

This construction gives $\widehat{\boldsymbol{R}}_{Sj} \in SO(3)$ and excludes improper reflections.

The aligned conformation is

$$\widetilde{\boldsymbol{X}}_{Sj} = \boldsymbol{X}_{Sj}^{c}\widehat{\boldsymbol{R}}_{Sj} + \mathbf{1}_{N_S}\bar{\boldsymbol{m}}_{S}^{\top} \qquad \text{(M8)}$$

Here, $\bar{\boldsymbol{m}}_{S}$is the centroid of the reference conformation $\boldsymbol{M}_{S}$. All subsequent Cartesian covariance calculations use $\widetilde{\boldsymbol{x}}_{Sj} = \operatorname{vec}\left(\widetilde{\boldsymbol{X}}_{Sj}\right)$.

When an iterative generalized Procrustes reference is used, it is defined by

$$\widehat{\mathbf{M}}_{S} \in \underset{\substack{\mathbf{M}\in\mathbb{R}^{N_S\times 3} \\ \mathbf{1}_{N_S}^{\top}\mathbf{M}=\mathbf{0} \\ \mathbf{R}_j\in SO(3),\; j=1,\ldots,M_S}}{\arg\min} \frac{1}{M_S}\sum_{j=1}^{M_S}\left\| \mathbf{X}_{Sj}^{c}\mathbf{R}_j - \mathbf{M} \right\|_{F}^{2} \qquad \text{(M9)}$$

The centring constraint removes translational non-identifiability. The remaining global

rotational freedom is fixed by the initialization or reference-frame convention. When a fixed reference is used, all conformations within the compared ensemble are aligned to that reference. For perturbation analyses, background and perturbed ensembles use the same reference frame. Near-degenerate coordinate configurations may yield unstable rotations; alignment sensitivity is therefore included in the resampling analysis where relevant.

After alignment, the vectorized conformation is

$$\tilde{\mathbf{x}}_{Sj} = \mathrm{vec}\left(\tilde{\mathbf{X}}_{Sj}\right) \in \mathbb{R}^{D_S}, \qquad D_S = 3N_S \qquad \text{(M10)}$$

All coordinate-based covariance calculations in **Part I** use $\tilde{\mathbf{x}}_{Sj}$, not laboratory-frame coordinates. This restriction applies only to Cartesian conformational representations.

Empirical protein-state measure

For state S at representation level r, the empirical measure is

$$\widehat{P}_S^{(r)} = \sum_{j=1}^{M_S} w_{Sj}\, \delta_{\boldsymbol{x}_{Sj}^{(r)}},\ w_{Sj} \geq 0,\ \sum_{j=1}^{M_S} w_{Sj} = 1 \qquad \text{(M11)}$$

Uniform sampling corresponds to $w_{Sj} = 1/M_S$. Non-uniform weights may represent declared sampling probabilities, mutation priors, generation probabilities or reweighting factors, provided that they are fixed independently of downstream outcomes. For any measurable function $f: \mathcal{H}_r \to \mathbb{R}$,

$$\int_{\mathcal{H}_r} f(\boldsymbol{x})\, d\widehat{P}_S^{(r)}(\boldsymbol{x}) = \sum_{j=1}^{M_S} w_{Sj}\, f\left(\boldsymbol{x}_{Sj}^{(r)}\right) \qquad \text{(M12)}$$

The empirical-measure construction applies to conformational, mutational, generative and learned-representation ensembles. For Cartesian conformational ensembles, $\boldsymbol{x}_{Sj}^{(r)}$ denotes the aligned coordinate representation. For ProstT5, ProtT5 or ESM-C ensembles, it denotes the declared sequence-, residue- or structure-level embedding after the specified pooling or masking operation. For a complete 20-amino-acid substitution ensemble at a site, the support consists of the admissible substitutions or their mapped representations, with weights defined by the declared enumeration or probability model.

The empirical mean is

$$\widehat{\boldsymbol{\mu}}_S^{(r)} = \sum_{j=1}^{M_S} w_{Sj}\, \boldsymbol{x}_{Sj}^{(r)} \qquad \text{(M13)}$$

Define centered observations by

$$\boldsymbol{y}_{Sj}^{(r)} = \boldsymbol{x}_{Sj}^{(r)} - \widehat{\boldsymbol{\mu}}_S^{(r)} \qquad \text{(M14)}$$

The covariance of the weighted empirical measure is

$$\widehat{\boldsymbol{C}}_{S,\mathrm{emp}}^{(r)} = \sum_{j=1}^{M_S} w_{Sj}\, \boldsymbol{y}_{Sj}^{(r)} \boldsymbol{y}_{Sj}^{(r)\top} \qquad \text{(M15)}$$

For uniformly sampled independent observations, the conventional sample covariance is

$$\widehat{\boldsymbol{C}}_{S,\mathrm{samp}}^{(r)} = \frac{1}{M_S - 1} \sum_{j=1}^{M_S} \boldsymbol{y}_{Sj}^{(r)}\, \boldsymbol{y}_{Sj}^{(r)\top} \qquad \text{(M16)}$$

The correction $1 - \sum_j w_{Sj}^2$ is used only when the weights represent a sampling scheme for which a weighted unbiased covariance estimator is justified. Mutation priors, generation probabilities and enumeration weights define an empirical measure but are not automatically sampling weights. Covariance geometry is therefore based on the declared covariance estimator, which must be fixed separately for each analysis.

The covariance matrix satisfies

$$\hat{\mathbf{C}}_S = \hat{\mathbf{C}}_S^{\top}, \qquad \hat{\mathbf{C}}_S \succeq 0$$

Its rank is bounded by

$$\mathrm{rank}\big(\hat{\mathbf{C}}_S\big) \leq \min(D_S, M_S - 1)$$

Therefore, when D_S is large relative to M_S, the covariance is necessarily singular or poorly estimated in the ambient coordinate space. **Part I** uses the covariance spectrum primarily as a descriptive representation of variance organization. Any inverse covariance or perturbation metric belongs to the subsequent local-geometry analysis and requires an explicitly declared regularization procedure.

Local statistical subspace and covariance spectrum

For the primary uniformly weighted Cartesian analysis, define the centered data matrix

$$\boldsymbol{Y}_S = \begin{bmatrix} \boldsymbol{y}_{S1}^\top \\ \vdots \\ \boldsymbol{y}_{SM_S}^\top \end{bmatrix} \in \mathbb{R}^{M_S \times D_S} \qquad \text{(M17)}$$

Its singular-value decomposition is

$$\boldsymbol{Y}_S = \boldsymbol{U}_S \boldsymbol{D}_S \boldsymbol{V}_S^\top \qquad \text{(M18)}$$

where $\boldsymbol{D}_S = \mathrm{diag}(d_{S1}, \ldots, d_{Sr_S})$ contains the positive singular values and $\boldsymbol{V}_{S,+}$ denotes the corresponding right-singular-vector matrix.

The covariance is

$$\widehat{\boldsymbol{C}}_S = \frac{1}{M_S - 1} \boldsymbol{Y}_S^\top \boldsymbol{Y}_S \qquad \text{(M19)}$$

Let $r_S = \mathrm{rank}(\widehat{\boldsymbol{C}}_S)$. Then

$$\widehat{\boldsymbol{C}}_S = \boldsymbol{V}_{S,+} \boldsymbol{\Lambda}_{S,+} \boldsymbol{V}_{S,+}^\top, \ \boldsymbol{\Lambda}_{S,+} = \mathrm{diag}(\lambda_{S1}, \ldots, \lambda_{Sr_S}) \qquad \text{(M20)}$$

with

$$\lambda_{Si} = \frac{d_{Si}^2}{M_S - 1}, \ \lambda_{S1} \geq \cdots \geq \lambda_{Sr_S} > 0 \qquad \text{(M21)}$$

If $d_S \leq r_S$ modes are retained, define

$$\boldsymbol{Q}_S = \boldsymbol{V}_{S,+}[:, 1{:}\, d_S] \in \mathbb{R}^{D_S \times d_S} \qquad \text{(M22)}$$

$$\boldsymbol{z}_{Sj} = \boldsymbol{Q}_S^\top \boldsymbol{y}_{Sj} \in \mathbb{R}^{d_S} \qquad \text{(M23)}$$

and

$$\widehat{\boldsymbol{C}}_{S,d} = \boldsymbol{Q}_S^\top \widehat{\boldsymbol{C}}_S \boldsymbol{Q}_S = \mathrm{diag}(\lambda_{S1}, \ldots, \lambda_{Sd_S}) \qquad \text{(M24)}$$

The primary spectral analysis uses the uniformly weighted sample covariance $\widehat{\boldsymbol{C}}_{S,\mathrm{samp}}^{(r)}$. In the primary Cartesian analysis, this matrix is denoted simply by $\widehat{\boldsymbol{C}}_S$. Weighted empirical covariances are used only when the corresponding analysis explicitly specifies sampling or enumeration weights. The covariance estimator is fixed before eigendecomposition and is not changed according to downstream biological outcomes.

Normalized spectral summaries use a leading-mode index set

$$I_S^{\mathrm{sum}} = \{1, \ldots, d_S^{\mathrm{sum}}\},\ d_S^{\mathrm{sum}} \le r_S$$

Participation ratio, entropy, anisotropy and relative dimensionality use this same index set. The spectral-decay interval I_S^{fit} may be narrower and is reported separately. No second spectral index convention is introduced later in **Part I**. The explained-variance rank is defined relative to the same leading-mode spectrum:

$$r_\eta(S) = \min\left\{ k \le d_S^{\mathrm{sum}} : \frac{\sum_{i=1}^{k} \lambda_{Si}}{\sum_{i=1}^{d_S^{\mathrm{sum}}} \lambda_{Si}} \ge \eta \right\} \qquad \text{(M25)}$$

Normalized covariance spectrum

Define

$$V_{\mathrm{cov}}(S) = \sum_{i \in I_S^{\mathrm{sum}}} \lambda_{Si} \qquad \text{(M26)}$$

$$p_{Si} = \frac{\lambda_{Si}}{V_{\mathrm{cov}}(S)},\ \sum_{i \in I_S^{\mathrm{sum}}} p_{Si} = 1 \qquad \text{(M27)}$$

The participation ratio is

$$\mathrm{PR}(S) = \left(\sum_{i \in I_S^{\mathrm{sum}}} p_{Si}^2 \right)^{-1} \qquad \text{(M28)}$$

The participation entropy and entropy effective rank are

$$H_{\mathrm{part}}(S) = -\sum_{i \in I_S^{\mathrm{sum}}} p_{Si} \log p_{Si} \qquad \text{(M29)}$$

$$\mathrm{erank}_{\mathrm{ent}}(S) = \exp\left[H_{\mathrm{part}}(S)\right] \qquad \text{(M30)}$$

The leading-mode fraction and normalized anisotropy are

$$a_S = p_{S1} \qquad \text{(M31)}$$

$$a_S^{\mathrm{norm}} = \frac{d_S^{\mathrm{sum}} a_S - 1}{d_S^{\mathrm{sum}} - 1},\ d_S^{\mathrm{sum}} = \mid I_S^{\mathrm{sum}} \mid \qquad \text{(M32)}$$

The normalized effective covariance dimensionality is

$$d_{\mathrm{rel}}(S) = \frac{\mathrm{PR}(S)}{d_S^{\mathrm{sum}}} \qquad \text{(M33)}$$

A regularized covariance log-volume may be defined by

$$L_{\mathrm{cov}}^{(\zeta)}(S) = \frac{1}{2}\log\det\left(\widehat{\boldsymbol{C}}_{S,d} + \zeta \boldsymbol{I}_{d_S}\right),\ \zeta > 0 \qquad \text{(M34)}$$

where ζ has the same squared coordinate units as the covariance. The Gaussian differential entropy is defined only for a positive-definite retained covariance. Both quantities depend on coordinate scale and retained dimension and are reported only as representation-specific statistical summaries.

Under a d_S-dimensional Gaussian approximation with positive-definite retained covariance, the differential entropy is

$$h_G(S) = \frac{1}{2}\log\left[(2\pi e)^{d_S}\det\left(\widehat{\boldsymbol{C}}_{S,d}\right)\right] \qquad \text{(M35)}$$

This quantity concerns the retained covariance subspace and is distinct from the entropy of the normalized spectrum when $I_S^{\mathrm{sum}} \neq \{1, \ldots, d_S\}$. Both quantities are representation-dependent statistical summaries rather than thermodynamic entropies.

Chain-length dependence of spectral organization

The finite-range spectral-decay model is

$$\lambda_{Si} = A_S i^{-\alpha_S}\exp(\epsilon_{Si}),\ A_S > 0,\ i \in I_S^{\mathrm{fit}} \qquad \text{(M36)}$$

Equivalently,

$$\log\lambda_{Si} = \log A_S - \alpha_S \log i + \epsilon_{Si} \qquad \text{(M37)}$$

The model is finite-range and does not imply exact scale invariance or a universal exponent. Zero eigenvalues are excluded rather than replaced by arbitrary positive values.

For a strictly positive response $Y_S^{(a)} > 0$,

$$\log Y_S^{(a)} = \beta_0^{(a)} + \beta_n^{(a)} \log n_S + \boldsymbol{b}^{(a)\top} \tilde{\boldsymbol{c}}_S^{\mathrm{AA}} + u_{\mathrm{system}(S)}^{(a)} + \varepsilon_S^{(a)} \qquad \text{(M38)}$$

For signed or non-positive responses, the model is

$$Y_S^{(a)} = \beta_0^{(a)} + \beta_n^{(a)} \log n_S + \boldsymbol{b}^{(a)\top} \tilde{\boldsymbol{c}}_S^{\mathrm{AA}} + u_{\mathrm{system}(S)}^{(a)} + \varepsilon_S^{(a)} \qquad \text{(M39)}$$

To compare two response slopes, the responses are placed on a declared common scale or modelled in separate units with an explicitly defined interaction contrast. With $I_k = \mathbf{1}_{\{k=2\}}$, the stacked model is

$$Y_{Sk} = \beta_0 + \gamma I_k + \beta_n \log n_S + \delta I_k \log n_S + \boldsymbol{b}_k^{\top} \tilde{\boldsymbol{c}}_S^{\mathrm{AA}} + u_{\mathrm{system}(S)} + \varepsilon_{Sk} \qquad \text{(M40)}$$

The direct slope contrast is $\Delta_{\mathrm{slope}} = \delta$. The interaction, rather than separate tests of the two slopes, evaluates whether the responses have different conditional chain-length associations. Response transformations are fixed before model fitting.

Part II: Covariance-Normalized Perturbation Geometry

Scope and statistical status

Part II defines a local geometric response for perturbations of protein states. The central object is a covariance-normalized displacement measured relative to the conformational variability of the unperturbed background state. The purpose of this construction is to distinguish a large displacement occurring along a direction that is commonly sampled by the background ensemble from a smaller displacement occurring along a direction that is locally constrained.

This part of the framework contains three logically distinct layers. The first layer is mathematical: aligned conformational ensembles define a covariance matrix, covariance regularization defines a positive-definite local reference form, and the inverse regularized covariance defines a quadratic displacement cost. The second layer is statistical: the resulting cost is estimated from finite samples and is therefore subject to alignment uncertainty, covariance-estimation error, sampling dependence and representation dependence. The third layer is biological: the cost is tested against independently measured mutational outcomes and real-mutant ensemble responses. These empirical tests determine whether the statistical geometry captures a reproducible component of perturbational constraint.

The local geometric cost is not identified with a molecular force, a Hessian-derived stiffness, a free-energy difference or a thermodynamic action. Its interpretation is restricted to the declared representation, alignment protocol, covariance estimator and perturbation class. In the current implementation, the primary geometry uses Kabsch-aligned conformational coordinates and Ledoit–Wolf covariance shrinkage. The covariance-normalized cost is therefore an effective statistical quantity constructed from the sampled ensemble rather than a coordinate-free physical energy.

The earlier comparison based on improvements in whitened directional consistency and coefficient of variation is not treated as primary evidence. Those apparent improvements can arise mechanically from the whitening transformation itself and were therefore withdrawn as direct validation of the local metric. The primary empirical support instead comes from real-mutant directional tests, cross-resolution correspondence and the association between geometric cost and independently measured deep-mutational-scanning fitness.

Aligned perturbation states

Let S denote an unperturbed protein state and let $\mathcal{P}$ denote a prespecified perturbation operator. The perturbed state is written as

$$S^{\mathcal{P}} = \mathcal{P}(S) \qquad \text{(M41)}$$

The operator $\mathcal{P}$ may represent a residue substitution, a controlled sequence perturbation or another explicitly defined state transformation. In the primary biological validation, $\mathcal{P}$ denotes a real protein mutation. Controlled synthetic perturbations may be used for diagnostic analyses, but they are not automatically equivalent to natural or experimentally measured mutations.

For a state S, let

$$\mathbf{X}_{Sj} \in \mathbb{R}^{N_S \times 3}, \qquad j = 1, \ldots, M_S$$

denote the sampled Cartesian conformations after the atom or residue correspondence has been fixed. The corresponding conformations of the perturbed state are

$$\mathbf{X}_{Sj}^{\mathcal{P}} \in \mathbb{R}^{N_{S^{\mathcal{P}}} \times 3}, \qquad j = 1, \ldots, M_{S^{\mathcal{P}}}$$

When the perturbation preserves the represented residue or atom correspondence, the two states can be mapped into a common coordinate space. Let

$$\iota_{S,\mathcal{P}} : \mathcal{X}_S \cup \mathcal{X}_{S^{\mathcal{P}}} \longrightarrow \mathcal{H}_{S,\mathcal{P}}$$

denote the prespecified common-space map. This map may use residue correspondence, atom correspondence, masking or another explicitly defined alignment rule. A direct coordinate displacement is defined only after both states have been represented in the same space. If the perturbation changes the number of represented residues or atoms, the unmatched degrees of freedom must be handled by a prespecified masking or correspondence rule; they cannot be silently omitted after inspecting the perturbation result.

After removal of global translation and rotation, denote the aligned vectorized conformations by

$$\tilde{\mathbf{x}}_{Sj} = \mathrm{vec}\left(\widetilde{\mathbf{X}}_{Sj}\right), \qquad \tilde{\mathbf{x}}_{Sj}^{\mathcal{P}} = \mathrm{vec}\left(\widetilde{\mathbf{X}}_{Sj}^{\mathcal{P}}\right) \qquad \text{(M42)}$$

For coordinate-based perturbation analyses, the background and perturbed ensembles are mapped into the same atom- or residue-correspondence space and aligned to one common prespecified reference coordinate system. Their representative means are subtracted only after this joint coordinate convention has been established. Independently chosen state-specific reference frames are not used for direct displacement calculations because their arbitrary relative rotation would contaminate $\Delta \boldsymbol{x}_{\mathcal{P}|S}$. For non-Cartesian mutation or learned-representation ensembles, the corresponding common representation map replaces rigid-body alignment.

Let $m(S)$ denote a prespecified representative of the state-specific empirical distribution. In the primary coordinate implementation, this representative is the aligned empirical mean,

$$\widehat{\boldsymbol{\mu}}_S = \frac{1}{M_S}\sum_{j=1}^{M_S}\tilde{\mathbf{x}}_{Sj} \qquad \text{(M43)}$$

and likewise

$$\widehat{\boldsymbol{\mu}}_{S^{\mathcal{P}}} = \frac{1}{M_{S^{\mathcal{P}}}}\sum_{j=1}^{M_{S^{\mathcal{P}}}}\tilde{\mathbf{x}}_{Sj}^{\mathcal{P}} \qquad \text{(M44)}$$

The representative-state displacement is

$$\Delta\mathbf{x}_{\mathcal{P}|S} = \widehat{\boldsymbol{\mu}}_{S^{\mathcal{P}}} - \widehat{\boldsymbol{\mu}}_S \qquad \text{(M45)}$$

Equation (M38) defines the displacement between empirical ensemble representatives. It does not describe a particular microscopic transition and should not be interpreted as the mean of a physical trajectory unless a transition process has been independently defined.

The covariance-normalized analysis uses the local statistical subspace estimated from the unperturbed background state S. Let

$$\mathbf{Q}_S \in \mathbb{R}^{D\times d_S}$$

contain the retained orthonormal covariance modes of the background ensemble, where D is the dimension of the common coordinate space and $d_S \le D$ is the retained statistical dimension. The tangent-coordinate perturbation is

$$\delta\mathbf{z}_{\mathcal{P}|S} = \mathbf{Q}_S^{\top}\Delta\mathbf{x}_{\mathcal{P}|S} \in \mathbb{R}^{d_S} \qquad \text{(M46)}$$

Here $\boldsymbol{Q}_S$ contains the retained covariance modes of the unperturbed background. This subspace is an empirical linear representation and is not assumed to be a physical tangent space.

The corresponding projection into the retained ambient subspace is

$$\Delta\mathbf{x}_{\mathcal{P}|S}^{\parallel} = \mathbf{Q}_S\delta\mathbf{z}_{\mathcal{P}|S} = \mathbf{Q}_S\mathbf{Q}_S^{\top}\Delta\mathbf{x}_{\mathcal{P}|S} \qquad \text{(M47)}$$

The component orthogonal to the retained statistical subspace is

$$\mathbf{r}_{\mathcal{P}|S} = (\mathbf{I}_D - \mathbf{Q}_S\mathbf{Q}_S^{\top})\Delta\mathbf{x}_{\mathcal{P}|S} \qquad \text{(M48)}$$

Because $\mathbf{Q}_S\mathbf{Q}_S^{\top}$ is an orthogonal projector,

$$\Delta\mathbf{x}_{\mathcal{P}|S} = \Delta\mathbf{x}_{\mathcal{P}|S}^{\parallel} + \mathbf{r}_{\mathcal{P}|S}$$

and

$$\left(\Delta\mathbf{x}_{\mathcal{P}|S}^{\parallel}\right)^{\top}\mathbf{r}_{\mathcal{P}|S} = 0 \qquad \text{(M49)}$$

The retained-subspace fraction is defined, when the denominator is non-zero, by

$$R_{\parallel}(\mathcal{P}|S) = \frac{\left\|\Delta\mathbf{x}^{\parallel}_{\mathcal{P}|S}\right\|_2^2}{\left\|\Delta\mathbf{x}_{\mathcal{P}|S}\right\|_2^2} = \frac{\left\|\delta\mathbf{z}_{\mathcal{P}|S}\right\|_2^2}{\left\|\Delta\mathbf{x}_{\mathcal{P}|S}\right\|_2^2}, \qquad 0 \le R_{\parallel} \le 1 \qquad \text{(M50)}$$

If $\Delta\mathbf{x}_{\mathcal{P}|S} = 0$, the ratio is undefined and is excluded under a prespecified rule. It is not assigned an arbitrary value.

Local covariance and Ledoit–Wolf regularization

The centered background observations in the retained local statistical coordinates are

$$\mathbf{z}_{Sj} = \mathbf{Q}_S^{\top}\left(\tilde{\mathbf{x}}_{Sj} - \hat{\boldsymbol{\mu}}_S\right) \qquad \text{(M51)}$$

The unregularized sample covariance is

$$\hat{\mathbf{C}}_S = \frac{1}{M_S - 1}\sum_{j=1}^{M_S} \mathbf{z}_{Sj}\,\mathbf{z}_{Sj}^{\top} \qquad \text{(M52)}$$

The matrix $\hat{\mathbf{C}}_S$ is symmetric positive semidefinite, but it may be singular or ill-conditioned when the number of sampled conformations is not large relative to d_S, when some modes have negligible variance or when sampling noise produces unstable small eigenvalues.

The primary analysis uses Ledoit–Wolf shrinkage toward an isotropic target. Let

$$\mathbf{T}_S = \tau_S \mathbf{I}_{d_S}, \qquad \tau_S = \frac{1}{d_S}\operatorname{tr}\left(\hat{\mathbf{C}}_S\right) \qquad \text{(M53)}$$

The regularized covariance is

$$\hat{\mathbf{C}}_{S,\mathrm{LW}} = (1 - \alpha_S)\hat{\mathbf{C}}_S + \alpha_S \mathbf{T}_S, \qquad 0 \le \alpha_S \le 1 \qquad \text{(M54)}$$

where α_S is the shrinkage intensity estimated from the covariance data according to the Ledoit–Wolf procedure. The scalar α_S is estimated without using the final perturbation response, mutational fitness or path statistic.

If $\tau_S > 0$and $\alpha_S > 0$, then

$$\hat{\mathbf{C}}_{S,\mathrm{LW}} \succ 0$$

This follows because $\hat{\boldsymbol{C}}_S \succeq 0$ and $\boldsymbol{T}_S \succ 0$, so for every non-zero $\mathbf{u}$,

$$\mathbf{u}^{\top}\hat{\mathbf{C}}_{S,\mathrm{LW}}\mathbf{u} = (1 - \alpha_S)\mathbf{u}^{\top}\hat{\mathbf{C}}_S\mathbf{u} + \alpha_S\tau_S\|\mathbf{u}\|_2^2 > 0 \qquad \text{(M55)}$$

The shrinkage estimator stabilizes the inverse covariance by reducing the influence of poorly estimated small eigenvalues. It also changes the estimated covariance spectrum

and therefore does not perform exact whitening of the empirical sample covariance. Regularization is consequently a bias–variance trade-off: it sacrifices some fidelity to the sample covariance in exchange for a more stable local quadratic form.

A general additive regularization may be written as

$$\hat{\mathbf{C}}_{S,\varepsilon} = \hat{\mathbf{C}}_S + \varepsilon \mathbf{H}_S,\ \varepsilon > 0,\ \mathbf{H}_S \succ 0 \qquad \text{(M56)}$$

This is a general theoretical form and is not numerically identical to Ledoit–Wolf shrinkage under the naive identification $\varepsilon = \alpha_S \tau_S$. For $0 \le \alpha_S < 1$,

$$\hat{\mathbf{C}}_{S,\mathrm{LW}} = (1-\alpha_S)\left[\hat{\mathbf{C}}_S + \frac{\alpha_S \tau_S}{1-\alpha_S}\mathbf{I}_{d_S}\right] \qquad \text{(M57)}$$

Thus, the corresponding additive parameter inside the brackets is

$$\varepsilon_{\mathrm{eff},S} = \frac{\alpha_S \tau_S}{1-\alpha_S}$$

together with the multiplicative factor $1-\alpha_S$. The operational estimator used for the local metric is the Ledoit–Wolf covariance itself, not an unscaled additive surrogate. The boundary case $\alpha_S = 1$ gives $\hat{\mathbf{C}}_{S,\mathrm{LW}} = \tau_S \mathbf{I}_{d_S}$.

Let

$$\hat{\mathbf{C}}_{S,\mathrm{LW}} = \mathbf{U}_S \,\mathrm{diag}\left(\tilde{\lambda}_{S1}, \ldots, \tilde{\lambda}_{Sd_S}\right) \mathbf{U}_S^{\top}$$

be its eigendecomposition. Under isotropic shrinkage,

$$\tilde{\lambda}_{Si} = (1-\alpha_S)\lambda_{Si} + \alpha_S \tau_S. \qquad \text{(M58)}$$

Thus, all regularized eigenvalues are shifted away from zero toward the common target scale. This transformation changes both the total variance and the relative spectral allocation. The regularized covariance should therefore be used for the local metric, whereas unregularized or separately defined spectra should be used for **Results 1** descriptors unless the analysis explicitly states otherwise.

Local effective metric

The inverse regularized covariance defines the local metric whenever $\hat{\boldsymbol{C}}_{S,\mathrm{LW}} \succ 0$. Under isotropic shrinkage, positive definiteness follows from $\tau_S > 0$ and $\alpha_S > 0$. Zero-variance states are excluded from inverse-covariance analysis. The covariance and displacement carry squared and first-order coordinate units, respectively, so the resulting quadratic form is dimensionless when the regularization target uses the same units.

$$\mathbf{G}_S = \hat{\mathbf{C}}_{S,\mathrm{LW}}^{-1} \qquad \text{(M59)}$$

Because $\hat{\mathbf{C}}_{S,\mathrm{LW}} \succ 0$, $\mathbf{G}_S$ is symmetric positive definite. It therefore defines a valid inner product on the retained local statistical coordinate space,

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{G}_S} = \mathbf{u}^\top \mathbf{G}_S \mathbf{v} \qquad \text{(M60)}$$

and the corresponding squared norm,

$$\|\mathbf{u}\|_{\mathbf{G}_S}^2 = \mathbf{u}^\top \mathbf{G}_S \mathbf{u} \qquad \text{(M61)}$$

The covariance-normalized geometric perturbation cost is

$$C_{\mathrm{geo}}^{\parallel}(\mathcal{P}|S) = \delta \mathbf{z}_{\mathcal{P}|S}^\top \mathbf{G}_S \delta \mathbf{z}_{\mathcal{P}|S} \qquad \text{(M62)}$$

Equivalently,

$$C_{\mathrm{geo}}^{\parallel}(\mathcal{P}|S) = \left\| \mathbf{W}_S \delta \mathbf{z}_{\mathcal{P}|S} \right\|_2^2 \qquad \text{(M63)}$$

where the symmetric whitening operator is

$$\mathbf{W}_S = \mathbf{G}_S^{1/2} = \hat{\mathbf{C}}_{S,\mathrm{LW}}^{-1/2} \qquad \text{(M64)}$$

The equality in Eq. (M63) follows from

$$\mathbf{W}_S^\top \mathbf{W}_S = \mathbf{G}_S$$

If the local covariance eigenvectors are $\mathbf{U}_S$, then

$$C_{\mathrm{geo}}^{\parallel}(\mathcal{P}|S) = \sum_{i=1}^{d_S} \frac{\left(\mathbf{u}_{Si}^\top \delta \mathbf{z}_{\mathcal{P}|S}\right)^2}{\tilde{\lambda}_{Si}} \qquad \text{(M65)}$$

This expression makes the interpretation explicit. A displacement is penalized according to its squared amplitude divided by the regularized background variance in the corresponding covariance direction. Directions with large background variability contribute less to the cost, whereas directions with small regularized variance contribute more.

The cost is a squared Mahalanobis-type distance in the local statistical coordinates. It is not a distance between two full probability measures, because it is computed from a chosen representative displacement, typically the difference between empirical means. A distributional distance such as Wasserstein distance belongs to a different level of analysis and should not be conflated with C_{geo}.

A full local cost incorporating the discarded normal component may be defined as

$$C_{\mathrm{geo}}^{\mathrm{full}}(\mathcal{P}|S) = C_{\mathrm{geo}}^{\parallel}(\mathcal{P}|S) + \omega_{\perp,S} \left\| \mathbf{r}_{\mathcal{P}|S} \right\|_2^2 \qquad \text{(M66)}$$

where $\omega_{\perp,S} \geq 0$ is a prespecified normal-component weight. This quantity is not the primary local metric used for the current **Results 2** validation unless the normal component and its scale have been independently defined. Introducing $\omega_{\perp,S}$ after observing a desired association would create a post hoc composite statistic and is therefore prohibited.

Regularized whitening and its limitations

For an unregularized positive-definite covariance matrix,

$$\mathbf{W}_S = \hat{\mathbf{C}}_S^{-1/2}$$

the transformed background coordinates satisfy

$$\mathrm{Cov}(\mathbf{W}_S \mathbf{z}_S) = \mathbf{I}_{d_S} \qquad \text{(M67)}$$

For Ledoit–Wolf regularization, however,

$$\mathrm{Cov}(\mathbf{W}_S \mathbf{z}_S) = \hat{\mathbf{C}}_{S,\mathrm{LW}}^{-1/2} \hat{\mathbf{C}}_S \hat{\mathbf{C}}_{S,\mathrm{LW}}^{-1/2} \qquad \text{(M68)}$$

which is not generally the identity matrix. Under the common eigenbasis of isotropic shrinkage,

$$\mathrm{Cov}(\mathbf{W}_S \mathbf{z}_S) = \mathbf{U}_S \,\mathrm{diag}\left(\frac{\lambda_{S1}}{\tilde{\lambda}_{S1}}, \ldots, \frac{\lambda_{Sd_S}}{\tilde{\lambda}_{Sd_S}}\right) \mathbf{U}_S^{\top} \qquad \text{(M69)}$$

The regularized transform is therefore better described as covariance normalization or shrinkage-stabilized whitening than as exact whitening. The distinction matters because inverse-covariance weighting may amplify noise in directions with very small empirical variance. The shrinkage estimator reduces, but does not eliminate, this risk.

The regularization intensity is determined independently of the final biological outcome. Sensitivity analyses may vary the shrinkage rule or compare the operational estimator with admissible alternative covariance estimators, but such comparisons must use identical perturbation identities, backgrounds, coordinate maps and test partitions. A method cannot be declared superior solely because it produces a larger biological association.

Coordinate covariance of the scalar cost

Let

$$\boldsymbol{z}' = \boldsymbol{A}\boldsymbol{z},\ \boldsymbol{A} \in GL(d_S) \qquad \text{(M70)}$$

be an invertible linear reparameterization of the local coordinates. If the covariance and the reference quadratic form transform covariantly,

$$\boldsymbol{C}' = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^{\top},\ \boldsymbol{H}' = \boldsymbol{A}\boldsymbol{H}\boldsymbol{A}^{\top} \qquad \text{(M71)}$$

then

$$\boldsymbol{C}' + \varepsilon\boldsymbol{H}' = \boldsymbol{A}(\boldsymbol{C} + \varepsilon\boldsymbol{H})\boldsymbol{A}^{\top} \qquad \text{(M72)}$$

Using

$$(\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^{\top})^{-1} = \boldsymbol{A}^{-\top}\boldsymbol{M}^{-1}\boldsymbol{A}^{-1} \qquad \text{(M73)}$$

we obtain

$$\begin{aligned}\boldsymbol{z}'^{\top}(\boldsymbol{C}' + \varepsilon\boldsymbol{H}')^{-1}\boldsymbol{z}' &= \boldsymbol{z}^{\top}\boldsymbol{A}^{\top}\boldsymbol{A}^{-\top}(\boldsymbol{C} + \varepsilon\boldsymbol{H})^{-1}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{z} \\ &= \boldsymbol{z}^{\top}(\boldsymbol{C} + \varepsilon\boldsymbol{H})^{-1}\boldsymbol{z}\end{aligned} \qquad \text{(M74)}$$

The quadratic cost is invariant under an invertible linear reparameterization when the covariance and regularization form are transformed by the same congruence map. This result applies to the general form $(\boldsymbol{C} + \varepsilon\boldsymbol{H})^{-1}$. Ledoit–Wolf shrinkage toward an identity target is invariant under orthogonal basis changes and consistent uniform changes of coordinate units, but not under an arbitrary linear transformation followed by redefinition of the identity target. Scalar-cost comparisons therefore use a fixed representation and regularization convention. Directional comparisons additionally require an explicit common basis or transport map.

Directional diagnostics and the withdrawn control analysis

For completeness, one may define a covariance-normalized response vector

$$\mathbf{v}_{\mathcal{P}|S} = \mathbf{W}_S \delta\mathbf{z}_{\mathcal{P}|S} \qquad \text{(M75)}$$

If several backgrounds share the same perturbation identity, pairwise directional consistency can be written as

$$\mathrm{Cons}_{\mathrm{geo}}(\mathcal{P}) = \frac{2}{m(m-1)} \sum_{a<b} \frac{\mathbf{v}_{\mathcal{P}|S_a}^{\top}\mathbf{v}_{\mathcal{P}|S_b}}{\left\|\mathbf{v}_{\mathcal{P}|S_a}\right\|_2 \left\|\mathbf{v}_{\mathcal{P}|S_b}\right\|_2} \qquad \text{(M76)}$$

provided that all vectors have been expressed in a common coordinate system and have non-zero norm. The analogous raw statistic is obtained by replacing $v_{\mathcal{P}|S}$ with a transported raw displacement.

The raw reference cost is

$$C_{\mathrm{raw}}(\mathcal{P} \mid S) = \| \delta \mathbf{z}_{\mathcal{P}|S} \|_2^2$$

whereas the normalized cost is

$$C_{\mathrm{geo}}(\mathcal{P} \mid S) = \delta \mathbf{z}_{\mathcal{P}|S}^{\top} \boldsymbol{G}_S \delta \mathbf{z}_{\mathcal{P}|S} \qquad \text{(M77)}$$

Both costs are computed in the same retained subspace before their background-level coefficients of variation are compared.

Likewise, for scalar costs $C_u(\mathcal{P} \mid S)$, $u \in \{raw, geo\}$ a background coefficient of variation can be defined as

$$\mathrm{CV}_u(\mathcal{P}) = \frac{\mathrm{sd}_S[C_u(\mathcal{P}|S)]}{\mathrm{mean}_S[C_u(\mathcal{P}|S)]} \qquad \text{(M78)}$$

when the denominator is strictly positive.

These quantities are retained only as diagnostic definitions. The differences

$$\Delta_{\mathrm{cons}} = \mathrm{Cons}_{\mathrm{geo}} - \mathrm{Cons}_{\mathrm{raw}}, \qquad \text{(M79)}$$

and

$$\Delta_{\mathrm{CV}} = \mathrm{CV}_{\mathrm{raw}} - \mathrm{CV}_{\mathrm{geo}} \qquad \text{(M80)}$$

are not used as primary evidence for the biological validity of the local metric. The whitening operation itself can mechanically increase directional similarity or reduce relative variation under certain constructions. Consequently, the previous Cons/CV headline claims were withdrawn and should not be presented as an independent validation of covariance-normalized geometry.

Real-mutant directional validation

The direct biological validation uses real mutant-associated conformational responses. Let $\mathcal{M}$ denote a set of experimentally defined mutations and let $\Delta\mathbf{x}_{\mu|S}$ denote the representative displacement associated with mutation μin protein background S. The covariance-normalized response is

$$\mathbf{v}_{\mu|S} = \mathbf{W}_S \mathbf{Q}_S^{\top} \Delta\mathbf{x}_{\mu|S} \qquad \text{(M81)}$$

The null reference is generated from the background ensemble under a prespecified randomization or bootstrap procedure. The randomization must preserve the relevant covariance estimation, perturbation magnitude and sampling structure while removing the specific mutational direction being tested.

Let $\mathcal{A}_1$ and $\mathcal{A}_2$ denote two prespecified directional criteria. In general form, each criterion is a ratio

$$T_k = \frac{\widehat{\Psi}_k^{\mathrm{mut}}}{\widehat{\Psi}_k^{\mathrm{null}}}, \qquad k \in \{1,2\} \qquad \text{(M82)}$$

where $\widehat{\Psi}_k^{\mathrm{mut}}$ is the observed statistic for real mutant responses and $\widehat{\Psi}_k^{\mathrm{null}}$ is the corresponding statistic under the matched null procedure.

Each T_k is interpreted according to its prespecified directional alternative. A ratio above or below one is favourable only when that direction follows from the definition of Ψ_k. The null statistic, soft-direction subset, aggregation level and handling of zero denominators are specified with the corresponding analysis. The criteria remain separate unless a multivariate combination rule is fixed before evaluation.

Inference is performed at the independent-protein or independent-system level. Mutations from the same protein are not automatically independent biological replicates, and conformational frames from the same ensemble are not independent mutational observations. Protein-level clustering, hierarchical bootstrap or an equivalent grouped procedure is required to avoid pseudoreplication.

The real-mutant analysis is the primary direct test of whether the covariance-normalized local representation captures a reproducible component of perturbational geometry. It does not establish that every mutation follows the same local law, nor does it prove that the normalized response vector is a physical force or a background-independent molecular coordinate.

Cross-resolution geometric correspondence

To assess whether the geometric cost reflects structure beyond a single coordinate resolution, define matched costs from Cα and full-atom representations,

$$C_{\mathrm{geo}}^{\mathrm{C}\alpha}(\mu|S), \qquad C_{\mathrm{geo}}^{\mathrm{FA}}(\mu|S) \qquad \text{(M83)}$$

using the same mutation identity, protein background and prespecified perturbation representative. Because the two representations generally have different dimensions and covariance structures, the comparison is made at the level of matched scalar costs rather than by requiring vector-level identity.

A cross-resolution correspondence may be quantified by Pearson or Spearman correlation,

$$\rho_{\mathrm{res}} = \mathrm{corr}\left(C_{\mathrm{geo}}^{\mathrm{C}\alpha}, C_{\mathrm{geo}}^{\mathrm{FA}}\right) \qquad \text{(M84)}$$

with confidence intervals estimated by resampling independent perturbation groups or proteins. The representation-specific costs must be computed without selecting

preprocessing or regularization parameters to maximize ρ_{res}.

Association with deep-mutational-scanning fitness

Let $F_{\mu|S}$ denote the DMS score for mutation μin protein S. The pooled descriptive association is

$$\rho_{\mathrm{DMS,pool}} = \mathrm{Spearman}\left(\{C_{\mathrm{geo}}(\mu \mid S)\}_{S,\mu}, \{F_{\mu|S}\}_{S,\mu}\right)$$

Protein-specific associations are

$$\rho_{\mathrm{DMS},S} = \mathrm{Spearman}\left(\{C_{\mathrm{geo}}(\mu \mid S)\}_{\mu\in\mathcal{M}_S}, \{F_{\mu|S}\}_{\mu\in\mathcal{M}_S}\right)$$

The pooled coefficient is reported descriptively. Confidence intervals and population-level inference use protein-level resampling or a hierarchical model.

For a baseline model B and a geometry-augmented model $B + \mathrm{geo}$, the incremental predictive contribution can be written as

$$\Delta R^2_{\mathrm{geo}} = R^2(B + \mathrm{geo}) - R^2(B), \qquad \text{(M85)}$$

provided that both models are evaluated on identical held-out mutations and that model selection is performed only within the training partition. A positive $\Delta R^2_{\mathrm{geo}}$ indicates additional predictive information under that evaluation design.

When mutation-level observations are clustered within proteins, confidence intervals and hypothesis tests must be obtained through protein-level resampling, clustered permutation or hierarchical modelling.

Per-residue resolution

Site-level analysis uses a separately defined site-restricted statistic rather than an additive decomposition of the global Mahalanobis cost. Let $\boldsymbol{S}_i$ select the coordinates assigned to site i. Define

$$\delta\boldsymbol{z}_i = \boldsymbol{S}_i\delta\boldsymbol{z},\ \boldsymbol{C}_{S,i} = \boldsymbol{S}_i\widehat{\boldsymbol{C}}_{S,\mathrm{LW}}\boldsymbol{S}_i^{\top} \qquad \text{(M86)}$$

If $\boldsymbol{C}_{S,i}$ is not positive definite, the same prespecified shrinkage rule used for the global covariance is applied:

$$\boldsymbol{C}_{S,i}^{(\mathrm{reg})} = \boldsymbol{C}_{S,i} + \varepsilon_i \boldsymbol{H}_i,\ \boldsymbol{H}_i \succ 0$$

The site-restricted cost is

$$C_{\mathrm{geo},i} = \delta \boldsymbol{z}_i^{\top} \left(\boldsymbol{C}_{S,i}^{(\mathrm{reg})} \right)^{-1} \delta \boldsymbol{z}_i \qquad \text{(M87)}$$

This statistic omits cross-site covariance terms and is analysed separately from the global cost. Site-level associations are computed within protein and summarized at the protein or region level.

Statistical interpretation and limits of Part II

Part II establishes a covariance-normalized local perturbation construction. A perturbation is first represented as a displacement between matched ensemble representatives, projected into the covariance subspace of the unperturbed background and evaluated using the inverse of a shrinkage-stabilized covariance matrix. This construction is mathematically equivalent to a regularized Mahalanobis quadratic form and to an ordinary Euclidean squared norm after the corresponding linear transformation.

The construction does not establish a physical metric in molecular configuration space. Its numerical value depends on the representation, alignment, covariance estimator, shrinkage intensity, retained dimension and perturbation representative. Coordinate covariance of the scalar quadratic form holds only when the covariance and reference form transform consistently. The normalized vector itself is gauge- and representation-dependent and requires explicit alignment before directional comparison.

The direct empirical evidence is restricted to three claims. First, real-mutant responses satisfy the prespecified directional criteria relative to their matched null distributions. Second, Cα- and full-atom-derived costs exhibit a weak but positive cross-resolution correspondence. Third, geometric cost is weakly and negatively associated with DMS fitness across the evaluated proteins, while a stiffness-related component contributes limited orthogonal information beyond sequence baselines.

These results support the interpretation that local covariance geometry captures a reproducible but incomplete component of mutational constraint. They do not support the stronger claims that covariance-normalized geometry universally outperforms sequence models, that the geometric cost is a molecular energy or free energy, that the normalized response is a physical force, or that the withdrawn Cons/CV improvements independently validate the metric. The appropriate conclusion is therefore a bounded one: the inverse local covariance provides a statistically well-defined representation of perturbational atypicality, and selected components of this representation are biologically informative under the tested mutation and assay conditions.

Part III: Biological Constraint Fields and Source-to-Geometry Relations

Scope and methodological status

Part III defines an empirical source-to-geometry framework for testing whether biological descriptors predict the organization of protein conformational ensembles. The framework maps a declared biological source representation to a response representation composed of effective geometric observables computed from the corresponding ensemble. The resulting object is a conditional statistical field over descriptor space. It is not a fundamental physical field equation, a causal law, or a coordinate-independent tensor field over all protein conformations.

The source-to-geometry relation is defined at the level of comparable descriptors rather than as a single matrix-valued field acting on all proteins in one common Cartesian space. Proteins of different chain lengths possess coordinate spaces of different dimensions, and their covariance matrices therefore do not automatically act on a common vector space. In the present construction, cross-system inference is performed primarily on scalar or low-dimensional geometric summaries that are defined consistently within each declared representation. A matrix-valued local response operator is introduced only within a fixed source and response coordinate system or within an explicitly defined common subspace.

The biological interpretation of the field is deliberately restricted. A source variable may be statistically associated with an ensemble response because of direct sequence effects, shared biological constraints, system identity, latent variables, sampling structure or unmeasured confounding. Accordingly, the field defines a predictive relation between biological descriptors and effective geometry. It does not, by itself, establish that a source variable causally induces a particular geometric response.

Biological source space

Let $\mathcal{T}$ denote the biological descriptor space and let S index a sequence-defined protein system. The source representation is written as

$$\mathbf{T}_S = \begin{bmatrix} \mathbf{T}_S^{\mathrm{seq}} \\ \mathbf{T}_S^{\mathrm{phys}} \\ \mathbf{T}_S^{\mathrm{ctx}} \\ \mathbf{T}_S^{\mathrm{func}} \end{bmatrix} \in \mathcal{T} \subseteq \mathbb{R}^p \qquad \text{(M88)}$$

The sequence-derived component T_S^{seq} may contain chain length, amino-acid composition and sequence-level pattern descriptors. The physicochemical component T_S^{phys} may contain fixed sequence-derived hydrophobicity, charge and charge-

patterning descriptors. The contextual component T_S^{ctx} may contain independently defined disorder, secondary-structure, domain or compactness descriptors when these are available without using the target response. The functional component T_S^{func} may contain independently defined functional, interface or environmental annotations.

The source vector is a declared predictive representation rather than a complete description of a protein. Its components, coding procedures, transformations and missing-value rules are fixed before final model fitting. Continuous variables are standardized using parameters estimated from the training systems. If $\boldsymbol{\mu}_T^{\mathrm{tr}}$ and $\mathbf{D}_T^{\mathrm{tr}}$ denote the training mean and diagonal scale matrix, the standardized source vector is

$$\widetilde{\mathbf{T}}_{\mathbf{S}} = (\mathbf{D}_T^{\mathrm{tr}})^{-1}(\mathbf{T}_S - \boldsymbol{\mu}_T^{\mathrm{tr}}) \qquad \text{(M89)}$$

The transformation is applied unchanged to validation and test systems. No test-system mean, variance or distributional property is used to redefine the source coordinate system.

The source variables are separated according to provenance. Sequence-intrinsic variables are calculated directly from the sequence or from fixed residue-level physicochemical tables. External variables are obtained from independent annotations. Ensemble-derived structural descriptors, such as contact density, radius of gyration, solvent exposure or secondary-structure occupancy, are not treated as independent biological sources when the response is calculated from the same ensemble. Including such variables in a model predicting the corresponding response would create information leakage and would convert the analysis into a joint descriptive model rather than a source-to-response model.

This distinction is particularly important for variables that may be defined in more than one way. A side-chain descriptor based on fixed residue identity, standard atom counts or a prespecified residue-level table can be used as a sequence source variable. A descriptor calculated from the sampled three-dimensional side-chain conformations cannot be used as an independent source variable for predicting full-atom ensemble geometry. The provenance of every source column is therefore recorded before model fitting.

Amino-acid composition and identifiable coordinates

Let

$$\mathbf{c}_S = (c_{S1}, \ldots, c_{SK})^{\top}$$

denote the amino-acid composition vector, where

$$c_{Sa} \geq 0, \ \sum_{a=1}^{K} c_{Sa} = 1 \qquad \text{(M90)}$$

Because the components lie on a simplex, the K raw composition variables are linearly

dependent. A model containing all composition fractions together with an intercept is therefore not identifiable.

Let

$$\boldsymbol{W} = \left[\mathbf{1}_K, \boldsymbol{w}_h, \boldsymbol{w}_q\right]$$

after removal of any linearly dependent columns. Let $r_W = \mathrm{rank}(\boldsymbol{W})$. The composition basis $\boldsymbol{B}_{\mathrm{AA}} \in \mathbb{R}^{K\times(K-r_W)}$ has orthonormal columns spanning $\ker(\boldsymbol{W}^\top)$:

$$\boldsymbol{W}^\top \boldsymbol{B}_{\mathrm{AA}} = \mathbf{0},\ \boldsymbol{B}_{\mathrm{AA}}^\top \boldsymbol{B}_{\mathrm{AA}} = \boldsymbol{I}_{K-r_W} \qquad \text{(M91)}$$

The retained composition coordinates are

$$\tilde{\boldsymbol{c}}_S^{\mathrm{AA}} = \boldsymbol{B}_{\mathrm{AA}}^\top \boldsymbol{c}_S \qquad \text{(M92)}$$

Hydrophobicity and charge are

$$h_S = \boldsymbol{w}_h^\top \boldsymbol{c}_S,\ q_S^{\mathrm{chg}} = \boldsymbol{w}_q^\top \boldsymbol{c}_S \qquad \text{(M93)}$$

This construction prevents the interpreted physicochemical directions from being duplicated by the retained composition coordinates.

Let $\mathbf{Z}$ denote the full design matrix containing an intercept and all source covariates. If the design matrix has $p_Z$columns, coefficient-level identifiability requires

$$\mathrm{rank}(\boldsymbol{Z}) = p_Z$$

Regularization may yield unique predictions under rank deficiency, while individual coefficients remain non-identifiable. Consequently, predictions, response subspaces and invariant contrasts are interpreted preferentially over individual coefficients.

Geometric response space

Let $\mathcal{Y}_r$ denote the response space associated with representation r. For state S,

$$\boldsymbol{Y}_S^{(r)} = \left[Y_{S1}^{(r)}, \ldots, Y_{Sq_r}^{(r)}\right]^\top \in \mathcal{Y}_r \qquad \text{(M94)}$$

When multivariate modelling is required, response coordinates are standardized using training-only parameters:

$$\widetilde{\boldsymbol{Y}}_S^{(r)} = \left(\boldsymbol{D}_{Y,r}^{\mathrm{tr}}\right)^{-1}\left(\boldsymbol{Y}_S^{(r)} - \boldsymbol{\mu}_{Y,r}^{\mathrm{tr}}\right) \qquad \text{(M95)}$$

The same transformation is applied to held-out systems. Responses from incompatible representation spaces are not concatenated without an explicit common response map. Response components derived from the same covariance spectrum are assigned to one testing family. Participation ratio, spectral entropy and entropy effective rank are mathematically related summaries and are not counted as independent biological responses.

Conditional source-to-geometry field

Because different representation levels need not be isomorphic, the source-to-geometry relation is a representation-indexed family:

$$\boldsymbol{F}_r(\boldsymbol{t}) = \mathbb{E}\left[\widetilde{\boldsymbol{Y}}_S^{(r)} \mid \widetilde{\boldsymbol{T}}_S = \boldsymbol{t}\right] \qquad \text{(M96)}$$

A representation-independent field is asserted only if an explicit response-space map $\psi_r : \mathcal{Y}_r \to \mathcal{Y}_\star$ has been defined. Otherwise, each $\boldsymbol{F}_r$ is an effective field in its declared representation.

Define the conditional residual

$$\boldsymbol{\eta}_{r,S} = \widetilde{\boldsymbol{Y}}_S^{(r)} - \boldsymbol{F}_r\left(\widetilde{\boldsymbol{T}}_S\right) \qquad \text{(M97)}$$

which satisfies

$$\mathbb{E}\left[\boldsymbol{\eta}_{r,S} \mid \widetilde{\boldsymbol{T}}_S\right] = \boldsymbol{0} \qquad \text{(M98)}$$

If $\boldsymbol{F}_r$ is differentiable near $\boldsymbol{t}_0$,

$$\boldsymbol{F}_r(\boldsymbol{t}) = \boldsymbol{F}_r(\boldsymbol{t}_0) + \boldsymbol{K}_r(\boldsymbol{t}_0)(\boldsymbol{t} - \boldsymbol{t}_0) + \boldsymbol{R}_{r,2}(\boldsymbol{t}; \boldsymbol{t}_0) \qquad \text{(M99)}$$

where

$$\boldsymbol{K}_r(\boldsymbol{t}_0) = D\boldsymbol{F}_r(\boldsymbol{t}_0) \qquad \text{(M100)}$$

The local affine approximation is

$$\widetilde{\boldsymbol{Y}}_S^{(r)} = \boldsymbol{b}_{r,0} + \boldsymbol{K}_r(\boldsymbol{t}_0)\widetilde{\boldsymbol{T}}_S + \boldsymbol{\varepsilon}_{r,S}^{\mathrm{lin}} \qquad \text{(M101)}$$

with

$$\boldsymbol{b}_{r,0} = \boldsymbol{F}_r(\boldsymbol{t}_0) - \boldsymbol{K}_r(\boldsymbol{t}_0)\boldsymbol{t}_0 \qquad \text{(M102)}$$

and

$$\boldsymbol{\varepsilon}_{r,S}^{\mathrm{lin}} = \boldsymbol{\eta}_{r,S} + \boldsymbol{R}_{r,2}\left(\widetilde{\boldsymbol{T}}_S; \boldsymbol{t}_0\right) \qquad \text{(M103)}$$

A globally fitted linear model estimates one affine approximation over the sampled descriptor domain. It is not identified with a constant Jacobian unless global linearity is assumed.

Heterogeneous systems and hierarchical coupling

Let c index a declared system, context or biological group. A shared-plus-context-specific response model is

$$\widetilde{\boldsymbol{Y}}_{S,c}^{(r)} = \boldsymbol{b}_{r,0,c} + \left(\boldsymbol{K}_{r,0} + \Delta\boldsymbol{K}_{r,c}\right)\widetilde{\boldsymbol{T}}_{S,c} + \boldsymbol{\varepsilon}_{r,S,c} \qquad \text{(M104)}$$

The context-specific operators satisfy

$$\boldsymbol{K}_{r,c} = \boldsymbol{K}_{r,0} + \Delta\boldsymbol{K}_{r,c} \qquad \text{(M105)}$$

with the centring constraint

$$\sum_c w_c\,\Delta\boldsymbol{K}_{r,c} = \boldsymbol{0},\ w_c \geq 0,\ \sum_c w_c = 1$$

The shared operator represents the potentially transferable component within representation r, whereas $\Delta\boldsymbol{K}_{r,c}$ represents context-specific deviation. Singular-vector analysis is performed after one fixed source and response standardization.

The covariance of the residual response may be written as

$$\mathrm{Cov}\left(\widetilde{\mathbf{Y}}_{S,c}^{(r)} \mid \widetilde{\mathbf{T}}_{S,c}\right) = \boldsymbol{\Sigma}_{\mathrm{res},r,c} \qquad \text{(M106)}$$

If frame-level sampling uncertainty and between-system biological variation are both present, $\boldsymbol{\Sigma}_{\mathrm{res},r,c}$ is not interpreted as a single homogeneous noise source. Its estimation

must respect the nested structure in which conformational observations are conditional within protein systems.

The response operator may be decomposed by singular-value analysis,

$$\mathbf{K}_{r,0} = \mathbf{U}_{r,K}\mathbf{\Sigma}_{r,K}\mathbf{V}_{r,K}^{\top} \qquad \text{(M107)}$$

The right singular vectors define combinations of source descriptors associated with large predicted responses, whereas the left singular vectors define corresponding combinations of geometric responses. This decomposition is coordinate dependent under arbitrary rescaling, but it is useful for comparing response subspaces after a fixed standardization convention.

Estimation of the conditional field

Let $\boldsymbol{Z}_{\mathrm{tr}} \in \mathbb{R}^{m\times(p+1)}$ contain an intercept and standardized source variables, and let $\widetilde{\boldsymbol{Y}}_{\mathrm{tr}}^{(r)} \in \mathbb{R}^{m\times q_r}$ contain standardized responses. Under full column rank,

$$\widehat{\boldsymbol{B}}_{r,\mathrm{OLS}} = (\boldsymbol{Z}_{\mathrm{tr}}^{\top}\boldsymbol{Z}_{\mathrm{tr}})^{-1}\boldsymbol{Z}_{\mathrm{tr}}^{\top}\widetilde{\boldsymbol{Y}}_{\mathrm{tr}}^{(r)} \qquad \text{(M108)}$$

Ridge estimation uses

$$\widehat{\boldsymbol{B}}_{r,\lambda} = (\boldsymbol{Z}_{\mathrm{tr}}^{\top}\boldsymbol{Z}_{\mathrm{tr}} + \lambda\boldsymbol{P})^{-1}\boldsymbol{Z}_{\mathrm{tr}}^{\top}\widetilde{\boldsymbol{Y}}_{\mathrm{tr}}^{(r)} \qquad \text{(M109)}$$

where $\boldsymbol{P} \succeq 0$, the intercept is unpenalized by setting $P_{00} = 0$, and the remaining diagonal or block structure of $\boldsymbol{P}$ is fixed before model fitting. The penalty parameter $\lambda$is selected within the training data only. The response operator is the transpose of the non-intercept coefficient block:

$$\widehat{\boldsymbol{K}}_r = \widehat{\boldsymbol{B}}_{r,\mathrm{source}}^{\top} \in \mathbb{R}^{q_r\times p}$$

For differentiable nonlinear estimators,

$$\widehat{\boldsymbol{K}}_r(\boldsymbol{t}_0) = D\widehat{\boldsymbol{F}}_r(\boldsymbol{t}_0) \qquad \text{(M110)}$$

For nonlinear estimators, $\widehat{\boldsymbol{F}}_r$ belongs to a prespecified model class. Its derivative is interpreted as a local statistical derivative in the declared descriptor space, not as an intrinsic derivative on the discrete sequence space.

Grouped validation and predictive risk

Let $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{val}}$ and $\mathcal{D}_{\mathrm{test}}$ denote partitions formed at the independent protein or system level. All conformational frames from a protein system remain within the same partition. Source standardization, response standardization, feature selection, dimension reduction, model selection and hyperparameter tuning are performed within the training partition.

For a held-out system S, the prediction is

$$\widehat{\widetilde{\boldsymbol{Y}}}_S^{(r)} = \widehat{\boldsymbol{F}}_{r,\mathrm{tr}}\left(\widetilde{\boldsymbol{T}}_S\right)$$

For response component a, the conventional held-out coefficient of determination is

$$R^2_{a,\mathrm{test}} = 1 - \frac{\sum_{S\in\mathcal{D}_{\mathrm{test}}}\left(\tilde{Y}^{(r)}_{S,a} - \hat{\tilde{Y}}^{(r)}_{S,a}\right)^2}{\sum_{S\in\mathcal{D}_{\mathrm{test}}}\left(\tilde{Y}^{(r)}_{S,a} - \bar{\tilde{Y}}^{(r)}_{\mathrm{test},a}\right)^2} \qquad \text{(M111)}$$

where

$$\bar{\tilde{Y}}^{(r)}_{\mathrm{test},a} = \frac{1}{|\,\mathcal{D}_{\mathrm{test}}\,|}\sum_{S\in\mathcal{D}_{\mathrm{test}}}\tilde{Y}^{(r)}_{S,a} \qquad \text{(M112)}$$

A training-mean denominator defines a different skill score and is reported separately.

For response component $a$and representation r, the root mean squared error and mean absolute error are

$$\mathrm{RMSE}^{(r)}_a = \left[\frac{1}{|\,\mathcal{D}_{\mathrm{test}}\,|}\sum_{S\in\mathcal{D}_{\mathrm{test}}}\left(\tilde{Y}^{(r)}_{S,a} - \hat{\tilde{Y}}^{(r)}_{S,a}\right)^2\right]^{1/2} \qquad \text{(M113)}$$

$$\mathrm{MAE}^{(r)}_a = \frac{1}{|\,\mathcal{D}_{\mathrm{test}}\,|}\sum_{S\in\mathcal{D}_{\mathrm{test}}}\left|\tilde{Y}^{(r)}_{S,a} - \hat{\tilde{Y}}^{(r)}_{S,a}\right| \qquad \text{(M114)}$$

For a predictive loss L, the group-weighted risk is computed by averaging the loss over independent groups rather than over conformational frames.

$$\hat{\mathcal{R}}_{\mathrm{test}} = \frac{1}{G}\sum_{g=1}^{G} L_g \qquad \text{(M115)}$$

where g indexes independent groups. This prevents systems with many

conformational observations from dominating inference about biological transfer.

The full source model is compared with prespecified baselines. These include an intercept-only model, a chain-length-only model, an identifiable-composition model, a restricted physicochemical model and a permutation model. When two models are evaluated on identical outer folds, the nonlinear improvement is

$$\Delta R_a^2 = R_{a,\text{nonlinear}}^2 - R_{a,\text{linear}}^2 \qquad \text{(M116)}$$

This quantity measures additional held-out predictive performance under a fixed evaluation design. It does not by itself establish mechanistic nonlinearity.

For a loss R in which lower values indicate better prediction, the transfer gap is

$$\Delta_{\text{transfer}} = R_{\text{between}} - R_{\text{within}} \qquad \text{(M117)}$$

The between-system and within-system losses must use the same response transformation, loss function and independent grouping unit. A positive gap indicates reduced transfer to systems that were not represented in the corresponding training group.

Coupling-structure stability

Because the coupling matrix depends on sampled systems, source coordinates, response coordinates and model fitting, it is treated as an estimated random object. Let

$$\widehat{\mathbf{K}}^{(b)}$$

denote the estimate from bootstrap replicate b. Matrix-level stability can be summarized by

$$d_K^{(b)} = \left\| \widehat{\mathbf{K}}^{(b)} - \widehat{\mathbf{K}}^{(0)} \right\|_F \qquad \text{(M118)}$$

where $\widehat{\mathbf{K}}^{(0)}$ is the estimate from the full training set. This distance is meaningful only after the source and response coordinates have been fixed and standardized consistently.

Because individual entries are coordinate dependent, subspace stability is also assessed. If $\mathbf{U}^{(b)}$ and $\mathbf{U}^{(0)}$ span the leading response subspaces of two coupling estimates, their principal angles satisfy

$$\cos\theta_j = \sigma_j \left[\left(\mathbf{U}^{(b)}\right)^{\top} \mathbf{U}^{(0)} \right] \qquad \text{(M119)}$$

where σ_j denotes the j-th singular value of the cross-subspace matrix. Small principal angles indicate that the response subspace is stable even when individual coefficients vary.

If the coupling structure is represented by discrete clusters, reproducibility between two

partitions is quantified by the adjusted Rand index,

$$\mathrm{ARI} = \frac{\mathrm{Index} - \mathrm{ExpectedIndex}}{\mathrm{MaxIndex} - \mathrm{ExpectedIndex}} \qquad \text{(M120)}$$

The ARI measures agreement between partitions after correction for chance agreement under the selected cluster-size structure. It is a reproducibility statistic, not a measure of predictive accuracy, causal validity or physical necessity. An internal separation statistic, such as a silhouette coefficient, may be reported as a descriptive complement but does not establish that the inferred groups are ontologically discrete biological classes.

For leave-one-amino-acid-out analysis, let $\hat{\mathcal{C}}_{(-a)}$ denote the coupling organization estimated after excluding amino-acid category a, and let $\hat{\mathcal{C}}_{\mathrm{full}}$ denote the organization estimated from the complete source set. The corresponding stability statistic is

$$\mathrm{ARI}_{(-a)} = \mathrm{ARI}\left(\hat{\mathcal{C}}_{(-a)}, \hat{\mathcal{C}}_{\mathrm{full}}\right) \qquad \text{(M121)}$$

This procedure tests whether the inferred organization depends disproportionately on any single amino-acid category. It does not establish that an excluded category has no biological influence.

Operational test of descriptor reduction

The reduction analysis evaluates whether the observed coupling structure can be reconstructed from progressively richer biological descriptor families. Let

$$\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots \subseteq \mathcal{X}_L \qquad \text{(M122)}$$

denote a prespecified sequence of predictor classes, beginning with simple baselines and progressing through composition, physicochemical, contextual and graph-based representations. Each predictor class is evaluated using identical independent-system partitions.

Let $R^2_{\ell,\mathrm{OOS}}$ denote the held-out predictive score for predictor class $\mathcal{X}_\ell$. The incremental predictive contribution of the additional descriptor class is

$$\Delta R^2_\ell = R^2_{\ell,\mathrm{OOS}} - R^2_{\ell-1,\mathrm{OOS}} \qquad \text{(M123)}$$

The term *operationally irreducible* is used when the coupling organization remains stable under resampling and exclusion tests, while progressively richer members of the prespecified descriptor family fail to reproduce it under grouped held-out evaluation. This statement is explicitly relative to the tested descriptor family and validation design. It does not imply absolute mathematical irreducibility or exclude future representations containing currently unmeasured biological variables.

The biological interpretation of this test is that a stable ensemble-level organization may

contain collective information not encoded in the selected low-dimensional source descriptors. Such information may arise from distributed sequence patterns, higher-order composition structure, domain organization or other variables absent from the declared source vector. The method therefore distinguishes failure of a descriptor family from failure of the field formalism itself.

Graph-level representation and held-out reduction

For graph-based reduction, let

$$\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S, \mathbf{X}_S, \mathbf{A}_S) \qquad \text{(M124)}$$

denote the graph representation of protein system S. The node set $\mathcal{V}_S$ represents residues or other declared structural units, $\mathcal{E}_S$ represents sequence or spatial relationships, X_S contains node features and A_S contains adjacency or edge features derived from the prespecified representation.

A graph predictor is a function

$$\widehat{\widetilde{\mathbf{Y}}}_S = \Psi_\vartheta(\mathcal{G}_S) \qquad \text{(M125)}$$

where the parameter vector ϑ is selected exclusively within the training systems. The multivariate held-out score is computed using a prespecified positive semidefinite response-weight matrix $\boldsymbol{W}_Y$:

$$R^2_{\text{graph,OOS}} = 1 - \frac{\sum_{S\in\mathcal{D}_{\text{test}}} \left\| \boldsymbol{W}_Y^{\frac{1}{2}} \left(\widetilde{\boldsymbol{Y}}_S - \widehat{\widetilde{\boldsymbol{Y}}}_S \right) \right\|_2^2}{\sum_{S\in\mathcal{D}_{\text{test}}} \left\| \boldsymbol{W}_Y^{\frac{1}{2}} \left(\widetilde{\boldsymbol{Y}}_S - \bar{\widetilde{\boldsymbol{Y}}}_{\text{test}} \right) \right\|_2^2} \qquad (M126)$$

The response transformation and $\boldsymbol{W}_Y$ are fixed within the training partition and applied identically to graph and non-graph models. Without this convention, high-variance response components can dominate the score.

The graph construction, node and edge features, output dimensionality, response weighting and independent-system grouping are fixed before evaluation. A negative score indicates that the tested graph representation performs worse than the declared mean-response baseline. The score is interpreted as evidence about the adequacy of the tested graph descriptor family, not as a proof that no graph representation could predict the response.

Affine reparameterization of the field

For invertible affine transformations

$$\boldsymbol{t}' = \boldsymbol{A}_T\boldsymbol{t} + \boldsymbol{a}_T,\ \boldsymbol{y}' = \boldsymbol{A}_Y\boldsymbol{y} + \boldsymbol{a}_Y \qquad \text{(M127)}$$

the response field transforms as

$$\boldsymbol{F}'_r(\boldsymbol{t}') = \boldsymbol{A}_Y\boldsymbol{F}_r[\boldsymbol{A}_T^{-1}(\boldsymbol{t}' - \boldsymbol{a}_T)] + \boldsymbol{a}_Y \qquad \text{(M128)}$$

and its Jacobian transforms as

$$\boldsymbol{K}'_r = \boldsymbol{A}_Y\boldsymbol{K}_r\boldsymbol{A}_T^{-1} \qquad \text{(M129)}$$

Matrix rank is preserved. Singular values require fixed source and response metrics and are preserved under compatible orthogonal transformations.

Secondary response directions

For a differentiable conditional field, the response direction associated with a continuous source variable t_a is

$$\boldsymbol{k}_a(\boldsymbol{t}) = \frac{\partial \boldsymbol{F}}{\partial t_a}(\boldsymbol{t}). \qquad \text{(M130)}$$

For a linear field, $\boldsymbol{k}_a$ is the corresponding column of $\boldsymbol{K}$. For a nonlinear estimator, it is a local derivative evaluated within the support of the training distribution. A reference-averaged direction may be defined as

$$\bar{\boldsymbol{k}}_a = \int \boldsymbol{k}_a\,(\boldsymbol{t})\, d\nu_{\mathrm{ref}}(\boldsymbol{t}) \qquad \text{(M131)}$$

where ν_{ref} is a prespecified distribution over source states in the training domain.

For two source variables $a$and b, let

$$\boldsymbol{\Sigma}_{\boldsymbol{Y},\delta} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}} + \delta\boldsymbol{H}_{\boldsymbol{Y}},\ \delta > 0,\ \boldsymbol{H}_{\boldsymbol{Y}} \succ 0 \qquad \text{(M132)}$$

be a training-derived regularized response covariance. Their response-space angular separation is

$$\cos\theta_{ab} = \frac{\bar{\boldsymbol{k}}_a^{\top}\,\boldsymbol{\Sigma}_{\boldsymbol{Y},\delta}^{-1}\,\bar{\boldsymbol{k}}_b}{\sqrt{\bar{\boldsymbol{k}}_a^{\top}\,\boldsymbol{\Sigma}_{\boldsymbol{Y},\delta}^{-1}\,\bar{\boldsymbol{k}}_a}\sqrt{\bar{\boldsymbol{k}}_b^{\top}\,\boldsymbol{\Sigma}_{\boldsymbol{Y},\delta}^{-1}\,\bar{\boldsymbol{k}}_b}} \qquad \text{(M133)}$$

This angle measures separation in the declared standardized response geometry. It is not interpreted as evidence of causal or mechanistic independence between the corresponding biological variables. It is a secondary geometric descriptor of the fitted response map.

Uncertainty estimation

Let $b = 1, \ldots, B$ index bootstrap replicates sampled at the independent protein or system level. Each replicate repeats the complete model-fitting procedure and produces

$$\widehat{\mathbf{F}}^{(b)}, \quad \widehat{\mathbf{K}}^{(b)}, \quad \widehat{\mathcal{R}}^{(b)} \qquad \text{(M134)}$$

When conformational sampling uncertainty is included, conformational observations are resampled conditionally within each selected protein system using an appropriate dependence-aware procedure. Frames from a single ensemble are never resampled as independent biological systems.

For a scalar statistic θ, a percentile interval is obtained from

$$\left[\hat{q}_{\alpha/2}\left(\{\hat{\theta}^{(b)}\}_{b=1}^{B}\right), \hat{q}_{1-\alpha/2}\left(\{\hat{\theta}^{(b)}\}_{b=1}^{B}\right)\right] \qquad \text{(M135)}$$

For coupling matrices, entry-wise intervals are treated as descriptive because they depend on coordinate parameterization. The primary uncertainty summaries are based on response-subspace angles, matrix singular structure, cluster agreement and grouped held-out prediction.

When multiple response components are tested, the tested family is defined before analysis and multiplicity is controlled at the family level. Algebraically related response components are not counted as independent hypotheses merely because they are stored in separate columns of $\mathbf{Y}_S$.

Biological meaning of the construction

The biological source vector represents constraints that can be specified before analysing the target ensemble, whereas the response vector represents how the sampled ensemble is organized under those constraints. Chain length describes the size of the sequence system, composition describes the distribution of residue classes, physicochemical descriptors encode fixed sequence-level tendencies and independently defined contextual variables describe additional biological conditions. Their relationship to the response is evaluated through a conditional mean rather than through an assumed mechanistic equation.

The Jacobian $\mathbf{K}$ quantifies local sensitivity of effective ensemble geometry to changes in the continuous source descriptors. A large response along a source direction means

that small changes in that descriptor are associated with substantial changes in the selected response coordinates within the fitted model. It does not mean that the source variable is a molecular force or that the response is an energy derivative.

The shared-plus-context-specific model separates potentially transferable structure from system-dependent calibration. The source-to-geometry relation may therefore contain both common statistical organization and biological context dependence. This formulation is appropriate for protein systems because sequence constraints are distributed across residues and may depend on higher-order patterns rather than on isolated residue properties. The descriptor-reduction analysis tests this possibility without presupposing that an unsuccessful low-dimensional model falsifies the field construction.

Methodological boundaries

The field is defined on a continuous descriptor space constructed from declared biological variables, not directly on the discrete set of all amino-acid sequences. Its Jacobian is therefore the derivative of a fitted conditional response function in descriptor space. It is not an intrinsic derivative on sequence space and should not be interpreted as a universal response to arbitrary mutations.

The response vector is similarly representation dependent. A response based on Cα covariance, a full-atom covariance response and a learned sequence embedding response belong to different effective coordinate systems. Their predictive relations may be compared through held-out performance and appropriately transformed subspaces, but their coefficient matrices cannot be equated without an explicit common representation.

Finally, the conditional field is empirical. Its mathematical definition does not imply equilibrium thermodynamics, causal biological control, a globally defined physical metric or a universal constitutive law. The field formalism provides a controlled way to test whether prespecified biological descriptors predict effective ensemble geometry and whether that predictive organization remains stable across systems, representations and resampling procedures.

Part IV: Ordered Protein-State Paths and Probability-Measure Transport

Scope and methodological status

Part IV represents an ordered state path as a sequence of state-indexed empirical measures. Adjacent measures are compared only after mapping them into a common metric space. Path order may encode mutation order, chain-length progression, generated candidates or another declared construction; it carries no automatic physical-

time interpretation.

The primary empirical hypothesis concerns mean adjacent-state transport relative to matched null paths. The broader action functional combines transport, geometric-response changes and feasibility penalties as a theoretical extension.

Ordered protein-state trajectories

Let $\mathcal{S}$ denote the set of admissible protein states. An ordered state path is

$$\gamma = (S_0, S_1, \ldots, S_N) \qquad \text{(M136)}$$

where N is the number of transitions. The state-level measure is

$$\widehat{P}_t^{(r)} = \widehat{P}_{S_t}^{(r)} \qquad \text{(M137)}$$

The path order is part of the object. A path may represent a mutation sequence, a controlled chain-length progression, a generated sequence of candidates or another ordered ensemble construction. It does not represent physical time unless an independently calibrated dynamical process is available.

For transition t, define the constraint record

$$\mathcal{C}_t = \mathcal{C}_t = (S_t, S_{t+1}, \Delta s_t, \mathcal{O}_t, \mathcal{B}_t), \qquad \Delta s_t > 0$$

where $\mathcal{O}_t$ denotes the permitted operation class and $\mathcal{B}_t$ contains the remaining biological, structural or combinatorial constraints.

For each state, let

$$\iota_{r,t}: \mathcal{X}_{S_t} \longrightarrow \mathcal{X}_\star$$

map the native representation space into one common metric space. The pushforward measure is

$$P_t^\star = \left(\iota_{r,t}\right)_\# \widehat{P}_{S_t}^{(r)} \qquad \text{(M139)}$$

For measures with finite second moments in the same metric space,

$$W_2^2(P_t^\star, P_{t+1}^\star) = \inf_{\pi \in \Pi\left(P_t^\star, P_{t+1}^\star\right)} \int_{\mathcal{X}_\star \times \mathcal{X}_\star} d_\star(x, y)^2 \, d\pi(x, y) \qquad \text{(M140)}$$

If the common-space measures are approximated by

$$P_t^\star \approx \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

then

$$W_{2,\mathrm{G}}^2(P_t^\star, P_{t+1}^\star) = \| \boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1} \|_2^2 + \mathrm{tr}\left[\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_{t+1} - 2\left(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_{t+1}\boldsymbol{\Sigma}_t^{1/2}\right)^{1/2}\right] \qquad \text{(M141)}$$

This expression is exact for Gaussian measures on the same Euclidean space and is a Gaussian-surrogate statistic for non-Gaussian ensembles. Direct empirical transport is a separate estimator and is not interchangeable with $W_{2,\mathrm{G}}$.

Transition-level transport uncertainty

Let T_t denote the selected transport statistic for transition t, either $W_{2,\mathrm{G}}$ or a separately declared empirical transport estimator. For bootstrap replicate b,

$$\hat{T}_t^{(b)} = T\left(\hat{P}_t^{\star(b)}, \hat{P}_{t+1}^{\star(b)}\right) \qquad \text{(M142)}$$

where ensemble observations are resampled within both adjacent states using the declared dependence-aware procedure. The conditional variance is

$$\widehat{\mathrm{Var}}(\hat{T}_t) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{T}_t^{(b)} - \bar{T}_t\right)^2, \ \bar{T}_t = \frac{1}{B}\sum_{b=1}^{B}\hat{T}_t^{(b)} \qquad \text{(M143)}$$

These are conditional ensemble replicates rather than independent biological samples. Path-level uncertainty is estimated by resampling independent proteins or systems and repeating the conditional ensemble resampling within each selected system.

Standardization of transport and geometric response

Transport and geometric-response changes generally have different numerical scales. Let $s_W > 0$ be a transport scale estimated from training transitions. The standardized transport is

$$\widetilde{W}_{2,\mathrm{G},t} = \frac{W_{2,\mathrm{G}}(P_t^\star, P_{t+1}^\star)}{s_W} \qquad \text{(M144)}$$

Let $\boldsymbol{Y}_t \in \mathbb{R}^q$ be the response vector associated with state S_t. If $\boldsymbol{\mu}_Y^{\mathrm{tr}}$ and $\boldsymbol{D}_Y^{\mathrm{tr}}$ are the training mean and positive diagonal scale matrix, define

$$\widetilde{\mathbf{Y}}_t = (\mathbf{D}_Y^{\mathrm{tr}})^{-1}(\mathbf{Y}_t - \boldsymbol{\mu}_Y^{\mathrm{tr}}) \qquad \text{(M145)}$$

$$\Delta\widetilde{\boldsymbol{Y}}_t = \widetilde{\boldsymbol{Y}}_{t+1} - \widetilde{\boldsymbol{Y}}_t \qquad \text{(M146)}$$

The same training-derived quantities are used for structured paths, null paths and held-out paths. Standardization makes the declared action dimensionless; it does not give the action a physical interpretation.

Effective geometric action

A general dimensionless action extension is

$$\mathcal{A}_2[\gamma] = \sum_{t=0}^{N-1}\left[\frac{\widetilde{W}_{2,\mathrm{G},t}^{\,2}}{2\Delta s_t} + \frac{\Delta\widetilde{\boldsymbol{Y}}_t^{\top}\boldsymbol{\Lambda}\Delta\widetilde{\boldsymbol{Y}}_t}{2\Delta s_t} + \frac{\Delta s_t}{2}\left(V(P_t^{\star}) + V(P_{t+1}^{\star})\right)\right] \qquad \text{(M147)}$$

where $\boldsymbol{\Lambda} \succeq 0$ is a prespecified response-weight matrix and $V(P_t^{\star}) \geq 0$ is a prespecified dimensionless feasibility penalty. This functional is a theoretical extension and is not the primary empirical statistic unless all of its components are evaluated under the same matched-null design.

The feasibility penalty may be decomposed as

$$V(P_t^{\star}) = V_{\mathrm{dens}}(P_t^{\star}) + V_{\mathrm{struct}}(P_t^{\star}) + V_{\mathrm{topo}}(P_t^{\star}) + V_{\mathrm{constraint}}(P_t^{\star}) \qquad \text{(M148)}$$

If a density estimator is used, its effective negative log-density is

$$U_{\mathrm{eff},S}(x) = -\log \hat{p}_S\,(x) \qquad \text{(M149)}$$

$U_{\mathrm{eff},S}$ is not itself the feasibility penalty: its numerical value depends on the reference measure, coordinate system and density normalization, and it need not be non-negative. If it contributes to V_{dens}, a prespecified non-negative and training-derived transformation, including its location and scale parameters, must be declared. The present action extension does not interpret $U_{\mathrm{eff},S}$ as a physical potential or thermodynamic free energy.

Total-variation action

For irregular paths, a linear-penalty alternative is

$$\mathcal{A}_{\mathrm{TV}}[\gamma] = \sum_{t=0}^{N-1} \left[\widetilde{W}_{2,\mathrm{G},t} + \left\| \boldsymbol{\Lambda}^{1/2} \Delta \widetilde{\boldsymbol{Y}}_t \right\|_1 + \Delta s_t \; \bar{V}_t \right] \qquad \text{(M150)}$$

where

$$\bar{V}_t = \frac{V(P_t^\star) + V(P_{t+1}^\star)}{2} \qquad \text{(M151)}$$

This functional is not interchangeable with the quadratic action. Its ℓ_1 term depends on the selected response basis, so one fixed standardized response coordinate system is used for structured and null paths. The primary functional and any sensitivity functional are selected before final path ranking.

Action decomposition and ablation

The action decomposes as

$$\mathcal{A}_2[\gamma] = \mathcal{A}_{\mathrm{tr}}[\gamma] + \mathcal{A}_{\mathrm{geo}}[\gamma] + \mathcal{A}_{\mathrm{feas}}[\gamma] \qquad \text{(M152)}$$

with

$$\mathcal{A}_{\mathrm{tr}}[\gamma] = \sum_{t=0}^{N-1} \frac{\widetilde{W}_{2,\mathrm{G},t}^{\,2}}{2\Delta s_t} \qquad \text{(M153)}$$

$$\mathcal{A}_{\mathrm{geo}}[\gamma] = \sum_{t=0}^{N-1} \frac{\Delta \widetilde{\boldsymbol{Y}}_t^{\top} \boldsymbol{\Lambda} \Delta \widetilde{\boldsymbol{Y}}_t}{2\Delta s_t} \qquad \text{(M154)}$$

$$\mathcal{A}_{\mathrm{feas}}[\gamma] = \sum_{t=0}^{N-1} \frac{\Delta s_t}{2} \left[V(P_t^\star) + V(P_{t+1}^\star) \right] \qquad \text{(M155)}$$

Transport-only, geometry-only, feasibility-only and full-action variants are defined by retaining the corresponding terms. All weights, standardization parameters and feasibility rules remain fixed across structured and null paths.

Matched null-path construction

Let $\mathcal{N}(S_0, S_N, N, \mathcal{C})$ denote the conditional null distribution over paths with initial state S_0, terminal state S_N, N transitions and constraint set $\mathcal{C}$. A null path is sampled as

$$\gamma^0 \sim \mathcal{N}(S_0, S_N, N, \mathcal{C})$$

The constraint set may include system identity, sequence length, mutation count, mutation spectrum, composition budget, permitted operations, atom or residue correspondence, admissible state-space support and path-step convention. The exact null-path generator, including its randomization rule and admissibility checks, is provided in the repository. Structured and null paths use the same common-space map, transport estimator, ensemble-size rule, standardization, action weights, feasibility penalty and numerical tolerances.

Low-transport and low-action hypotheses

The primary empirical statistic is the mean adjacent-state Gaussian transport

$$\bar{W}_{2,\mathrm{G}}(\gamma) = \frac{1}{N}\sum_{t=0}^{N-1} W_{2,\mathrm{G}}(P_t^\star, P_{t+1}^\star) \qquad \text{(M157)}$$

The low-transport hypothesis is

$$\mathbb{E}\left[\bar{W}_{2,\mathrm{G}}(\gamma_{\mathrm{structured}}) \mid \mathcal{C}\right] < \mathbb{E}_{\gamma^0\sim\mathcal{N}}\left[\bar{W}_{2,\mathrm{G}}(\gamma^0) \mid \mathcal{C}\right] \qquad \text{(M158)}$$

The complete-action hypothesis is

$$\mathbb{E}[\mathcal{A}_2(\gamma_{\mathrm{structured}}) \mid \mathcal{C}] < \mathbb{E}_{\gamma^0\sim\mathcal{N}}[\mathcal{A}_2(\gamma^0) \mid \mathcal{C}] \qquad \text{(M159)}$$

The first hypothesis concerns mean adjacent-state transport. The second additionally contains geometric-response and feasibility terms. The two hypotheses are evaluated separately.

For an observed path, the standardized action score is

$$Z_{\mathcal{A}} = \frac{\mathcal{A}_2(\gamma_{\mathrm{bio}}) - \widehat{\mathbb{E}}_{\mathcal{N}}[\mathcal{A}_2]}{\sqrt{\widehat{\mathrm{Var}}_{\mathcal{N}}[\mathcal{A}_2]}} \qquad \text{(M160)}$$

provided that the null variance is positive. The action advantage is

$$\Delta_{\mathcal{A}} = \widehat{\mathbb{E}}_{\mathcal{N}}[\mathcal{A}_2] - \mathcal{A}_2(\gamma_{\mathrm{bio}}) \qquad \text{(M161)}$$

A one-sided randomization Pvalue for the low-action alternative is

$$p_{\mathcal{A}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left[\mathcal{A}_2(\gamma_b^0) \le \mathcal{A}_2(\gamma_{\mathrm{bio}})\right]}{B+1} \qquad \text{(M162)}$$

The validity of this test requires a conditionally valid null-path construction. For the primary low-transport test, the statistic is

$$T_{\mathrm{tr}}(\gamma) = \bar{W}_{2,\mathrm{G}}(\gamma)$$

For B exchangeable null paths, the one-sided randomization Pvalue is

$$p_{\mathrm{tr}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left[T_{\mathrm{tr}}(\gamma_b^0) \le T_{\mathrm{tr}}(\gamma_{\mathrm{structured}})\right]}{B+1}$$

Null paths generated from one endpoint pair are conditional replicates for estimating the null distribution and are not independent biological samples. Population-level uncertainty is estimated by resampling independent proteins, systems or endpoint groups.

Transition-level sharpness

Let a_t denote the contribution of transition $t$to the selected action functional. For the corresponding matched-null transition class, let

$$\mu_{0,t} = \mathbb{E}_{\mathcal{N}}[a_t],\ \sigma_{0,t} > 0 \qquad \text{(M163)}$$

denote the null mean and standard deviation. The transition sharpness index is

$$\mathrm{TSI}_t = \frac{a_t - \mu_{0,t}}{\sigma_{0,t}} \qquad \text{(M164)}$$

A positive TSI_tindicates that the transition contributes more action than expected for its matched null class, whereas a negative value indicates a comparatively economical transition. The TSI is a localization statistic. It is not independent evidence for or against the global path hypothesis.

A path may have

$$Z_{\mathcal{A}} < 0 \qquad \text{(M165)}$$

while containing one or more transitions with positive TSI_t. This is mathematically consistent because a globally economical path may contain localized geometric bottlenecks. Transition-level annotations may be used to classify such peaks as folding, expansion, side-chain repacking, interface switching or other prespecified transition

types. Inference remains at the path or protein level.

Constrained minimizers and normalized excess action

If a numerical optimization problem is explicitly defined, let Γ_{adm}denote the set of admissible paths satisfying the same endpoints, transition count, operation budget, correspondence rules and feasibility constraints as the biological and null paths. A constrained minimizer is

$$\gamma^{\star} \in \arg\min_{\gamma \in \Gamma_{\mathrm{adm}}} \mathcal{A}_2[\gamma] \qquad \text{(M166)}$$

The existence of a numerical minimizer depends on the compactness or discretization of Γ_{adm}, the continuity of the action and the numerical optimization procedure. If the path space is not compact or the optimization is approximate, $\gamma^{\star}$should be described as a numerically obtained admissible minimizer rather than as a mathematically guaranteed global minimum.

When the biological, optimized and null paths obey identical constraints, a normalized excess action may be defined as

$$E_{\mathrm{path}} = \frac{\mathcal{A}_2(\gamma_{\mathrm{bio}}) - \mathcal{A}_2(\gamma^{\star})}{\widehat{\mathbb{E}}_{\mathcal{N}}[\mathcal{A}_2] - \mathcal{A}_2(\gamma^{\star})} \qquad \text{(M167)}$$

provided that the denominator is strictly positive.

A value of zero indicates equality with the computed admissible reference, whereas a value of one indicates equality with the estimated null mean. Values above one are permitted. Under identical constraints and exact global minimization, a negative value is impossible. A negative value indicates approximate optimization, unequal preprocessing, inconsistent constraints or numerical error. Such cases are audited and are not truncated.

The existence of a numerical minimizer does not imply that the biological system causally solves an optimization problem. It only provides a reference point for comparing observed and null paths under an identical computational functional.

Coarse-graining between full-atom and Cα representations

Let $\mathcal{X}_{\mathrm{FA}}$ and $\mathcal{X}_{\mathrm{C}\alpha}$ denote full-atom and Cα representation spaces. A declared coarse-graining map is

$$\pi: \mathcal{X}_{\mathrm{FA}} \longrightarrow \mathcal{X}_{\mathrm{C}\alpha} \qquad \text{(M168)}$$

If the Cα ensemble is obtained by applying the same fixed coarse-graining map $\pi$to the same full-atom conformational samples, then

$$\widehat{P}_S^{C\alpha} = \pi_{\#}\widehat{P}_S^{\mathrm{FA}} \qquad (M169)$$

If the Cα and full-atom ensembles are sampled independently, this equality is not assumed; the two empirical measures are compared only after applying their declared common-space or matched-sample construction. In either case, coarse-graining does not imply that the Cα representation retains all full-atom information.

.

If π is differentiable near the state and

$$\boldsymbol{A}_S = D\pi \qquad (\mathrm{M170})$$

denotes its local Jacobian, then

$$\boldsymbol{\Sigma}_S^{C\alpha} \approx \boldsymbol{A}_S \boldsymbol{\Sigma}_S^{\mathrm{FA}} \boldsymbol{A}_S^{\top} \qquad (\mathrm{M171})$$

The relation is exact for a fixed linear projection in a common aligned frame and local for nonlinear, state-dependent, masked or alignment-coupled maps.

Degrees of freedom in the kernel of A_S, including many side-chain coordinates, are removed by coarse-graining. Consequently, effective rank, spectral entropy, curvature and chain-length exponents are not expected to be numerically conserved between full-atom and Cα representations.

After alignment, the full-atom coordinate vector is partitioned into backbone and side-chain coordinates, giving

$$\boldsymbol{\Sigma}_S^{\mathrm{FA}} = \begin{bmatrix} \boldsymbol{\Sigma}_{S,\mathrm{bb}} & \boldsymbol{\Sigma}_{S,\mathrm{bs}} \\ \boldsymbol{\Sigma}_{S,\mathrm{sb}} & \boldsymbol{\Sigma}_{S,\mathrm{ss}} \end{bmatrix} \qquad (\mathrm{M172})$$

The side-chain contribution to coordinate variance is

$$f_{\mathrm{sc}}(S) = \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_{S,\mathrm{ss}}\right)}{\mathrm{tr}\left(\boldsymbol{\Sigma}_S^{\mathrm{FA}}\right)} \qquad (\mathrm{M173})$$

provided the denominator is positive. Backbone–side-chain covariance coupling is summarized by

$$\chi_{\mathrm{bs}}(S) = \frac{\left\|\boldsymbol{\Sigma}_{S,\mathrm{bs}}\right\|_F}{\sqrt{\mathrm{tr}\left(\boldsymbol{\Sigma}_{S,\mathrm{bb}}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_{S,\mathrm{ss}}\right)}} \qquad (\mathrm{M174})$$

provided both trace terms are positive. This quantity is invariant to separate common scalar rescaling of the backbone and side-chain coordinate blocks. It is a statistical coupling measure, not a molecular interaction energy.

Canonical correlations between the backbone and side-chain blocks may be used as an

alternative block-coupling measure. Because both χ_{bs} and canonical correlations are derived from the same covariance partition, they are treated as complementary rather than independent evidence.

Representation-specific scaling

Let $r \in \{\mathrm{C}\alpha, \mathrm{FA}\}$, and define $I_{\mathrm{FA}} = \mathbf{1}_{\{r=\mathrm{FA}\}}$. For paired responses,

$$Y_{S,r} = b_0 + b_r I_{\mathrm{FA}} + b_n \log n_S + b_{rn} I_{\mathrm{FA}} \log n_S + \mathbf{b}_{\mathrm{AA}}^{\top} \tilde{\mathbf{c}}_S^{\mathrm{AA}} + u_S + \varepsilon_{S,r} \qquad \text{(M175)}$$

The coefficient b_{rn} directly tests the difference between the representation-specific chain-length slopes. Separate significance tests do not test this contrast.

A representation-dependent scaling reversal requires that the estimated slopes have opposite signs and that the confidence interval for their direct contrast excludes the null value. Such a reversal is described as an effective-geometric crossover between representations. It is not interpreted as a thermodynamic phase transition.

Persistent topology

Let $\mathcal{X}_S$ be a point cloud or weighted graph associated with state S. Persistent homology is computed using a prespecified metric, filtration function, simplicial-complex construction and maximum homological dimension. Let

$$D_{k,S}$$

denote the persistence diagram in homological dimension k. For $p > 0$, total persistence is

$$\mathrm{TP}_{k,p}(S) = \sum_{(b,d) \in D_{k,S}} (d - b)^p \qquad \text{(M176)}$$

The persistence lifetime $d - b$ is measured in filtration units and does not automatically represent a physical hole or cavity. Essential classes with infinite death time are excluded from total persistence or assigned a prespecified finite truncation. The same rule is applied to all states and null models. Persistence descriptors are interpreted only after subsampling, filtration and null-model stability checks.

Finite-scale multifractal analysis

For a scale ℓ, let $p_i(\ell)$ denote the empirical probability mass assigned to cell or neighbourhood i. For $q \neq 1$, the generalized dimension is formally defined by

$$D_q = \frac{1}{q-1} \lim_{\ell \to 0} \frac{\log[\sum_i p_i (\ell)^q]}{\log \ell} \qquad \text{(M177)}$$

provided that the limit exists. Finite protein ensembles do not provide access to this asymptotic limit. The empirical estimate is therefore obtained from the slope over a validated finite-scale interval:

$$\widehat{D}_q = \frac{1}{q-1} \widehat{\text{slope}} \left(\log \ell \,,\, \log \left[\sum_i p_i (\ell)^q \right] \right) \qquad \text{(M178)}$$

The scaling interval, occupancy threshold and neighbourhood construction are fixed before final comparison. A multifractal interpretation requires reproducible dependence of $\widehat{D}_q$ on q beyond finite-sample null models.

Graph construction and multiscale curvature

Let a state or conformational point cloud be represented by a graph

$$\mathcal{G}_{S,\ell} = (\mathcal{V}_S, \mathcal{E}_{S,\ell})$$

constructed using a prespecified distance, neighbourhood rule, edge-weight function and graph scale ℓ. For each node i, let m_i denote a probability measure over its local neighbourhood, constructed using a fixed laziness, diffusion or weighting parameter.

For an edge (i, j) with distance $d(i, j) > 0$, Ollivier–Ricci curvature is defined as

$$\kappa(i,j) = 1 - \frac{W_1(m_i, m_j)}{d(i,j)} \qquad \text{(M179)}$$

where W_1 is the first Wasserstein distance between the local neighbourhood measures. Node-level curvature is obtained by a prespecified weighted or unweighted aggregation over incident edges.

The sign of $\kappa(i,j)$ describes the relationship between local neighbourhood transport and edge distance within the chosen graph construction. Negative curvature indicates that neighbouring local measures are more separated in the Wasserstein sense than expected from the direct edge distance. This is a graph-geometric property and is not automatically equivalent to curvature of the molecular configuration manifold.

Curvature depends on graph density, neighbourhood scale, edge weights and node measures. Curvatures obtained from different graph definitions cannot be pooled directly without calibration. A multiscale model may be written as

$$\kappa_{S,\ell} = b_0 + b_\ell + b_c + b_{\ell c} + u_S + \varepsilon_{S,\ell} \qquad \text{(M180)}$$

where $b_{\ell c}$ is the scale-by-context interaction. A multiscale curvature claim requires

stability under admissible graph constructions and comparison with degree-, density- and sample-size-matched null graphs.

Opposite curvature signs at different graph scales are not mathematically contradictory. They indicate that local neighbourhood organization changes with scale and should be described as multiscale effective-geometric crossover rather than as a thermodynamic phase transition.

Effective density landscapes

The effective density landscape is

$$U_{\mathrm{eff},S}(x) = -\log \hat{p}_S\,(x) \qquad \text{(M181)}$$

Its numerical value depends on the density estimator, bandwidth or model architecture, regularization and reference measure. Landscape descriptors are evaluated only within regions supported by observed or generated samples. The quantity is not a physical potential energy or thermodynamic free energy without independent equilibrium calibration. Unsupported extrapolation regions receive no physical or biological interpretation. If $U_{\mathrm{eff},S}$ contributes to the feasibility penalty, its scale and weight are fixed from training data and evaluated in a prespecified ablation.

Learned representations and coordinate gauge

Learned representations can change under orthogonal rotations, neuron permutations, layerwise rescaling, model architecture and random initialization without changing their predictive information. Equality of individual embedding coordinates is therefore not a criterion for representational equivalence.

For the same matched observations represented by centered matrices

$$\boldsymbol{Z}^{(r)}, \boldsymbol{Z}^{(s)} \in \mathbb{R}^{m\times d}$$

with equal latent dimension d, orthogonal Procrustes alignment is

$$\boldsymbol{R}^{\star} = \underset{\boldsymbol{R}^{\top}\boldsymbol{R}=\boldsymbol{I}_d}{\arg\min} \left\|\boldsymbol{Z}^{(r)}\boldsymbol{R} - \boldsymbol{Z}^{(s)}\right\|_F^2 \qquad \text{(M182)}$$

If the native dimensions differ, an explicit dimension-reduction or common-space map is applied before alignment. The map is estimated without final biological labels or test outcomes. Representational similarity is assessed through invariant subspaces, canonical correlations, projector distance or held-out transfer rather than equality of individual coordinates.

For orthonormal subspace bases $\mathbf{Q}_r$ and $\mathbf{Q}_s$ of equal dimension, the projector distance is

$$d_{\mathrm{proj}}(\mathbf{Q}_r, \mathbf{Q}_s) = \frac{1}{\sqrt{2}} \|\mathbf{Q}_r \mathbf{Q}_r^\top - \mathbf{Q}_s \mathbf{Q}_s^\top\|_F \qquad \text{(M183)}$$

The two bases must be expressed in the same ambient inner-product space. If their native spaces differ, an explicit common representation map is applied before calculating the projector distance.

Cross-representation claims require stability of invariant scalar costs, aligned response subspaces, representational similarity or held-out predictive transfer. Equality of individual latent coordinates or neurons is neither necessary nor expected.

Cross-representation validation

Let $C_{S,\mathcal{P}}^{(r)}$ and $C_{S,\mathcal{P}}^{(s)}$ denote a matched scalar geometric quantity estimated in representations r and s. Agreement is evaluated using rank correlation, reliability-corrected correlation when reliability estimates are available and conditional association after adjustment for prespecified confounders.

For paired geometric costs, a hierarchical model is

$$C_{S,\mathcal{P}}^{(s)} = b_0 + b_1 C_{S,\mathcal{P}}^{(r)} + \mathbf{b}_Z^\top \mathbf{Z}_{S,\mathcal{P}} + u_S + u_{\mathcal{P}} + \varepsilon_{S,\mathcal{P}} \qquad \text{(M184)}$$

where $\mathbf{Z}_{S,\mathcal{P}}$ contains prespecified confounders, u_S is a state-level random effect and $u_{\mathcal{P}}$ is a perturbation-level random effect.

Cross-representation correspondence is supported only when the conditional slope, uncertainty interval and held-out predictive value remain stable under representation-specific uncertainty and admissible alignment choices. A statistically significant unadjusted correlation alone is insufficient evidence for representation equivalence.

Hierarchical inference

When transitions, paths, perturbations or geometric targets are nested within proteins or systems, inference is performed using hierarchical models or cluster-aware resampling. A generic nested effect is modelled as

$$\delta_{ij} = \mu_\delta + u_j + \varepsilon_{ij},\ u_j \sim \mathcal{N}\left(0, \tau_\delta^2\right),\ \varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\delta^2\right) \qquad \text{(M185)}$$

where j indexes independent proteins or systems and i indexes observations nested within them. Robust or non-Gaussian alternatives are used when residual diagnostics require them. Transitions and conformational frames are not treated as independent biological replicates.

No analysis treats transitions or conformational frames within one protein as independent biological replicates. Bootstrap resampling follows the hierarchy of the data: independent proteins or systems are sampled first, and conformational observations are resampled conditionally within the selected states. Covariance estimation, standardization, transport calculation, action evaluation and all model-fitting steps are repeated within each bootstrap replicate when computationally feasible.

Leakage prevention and held-out evaluation

Outcome-dependent transformations, global feature maps, response standardization, model selection, hyperparameter tuning, transport scales, joint-PCA maps and learned response subspaces are fitted within the training partition and applied unchanged to held-out systems. State-intrinsic unsupervised quantities required to characterize a held-out ensemble, such as its local alignment reference and covariance estimator, may be estimated from that ensemble provided that no labels, fitness values or path outcomes are used.

Grouped nested cross-validation follows the intended transfer claim. Sequence-related replicates, homologous systems, near-duplicate sequences and paths sharing endpoint pairs remain in the same outer partition whenever separation could leak system-specific information.

Missing data and quality control

Missing residues, unresolved atoms, incomplete ensembles and failed geometric estimates are handled using prespecified quality-control criteria. Analyses requiring atomwise correspondence use the common valid atom set or a fixed masking procedure. States are excluded only according to criteria defined independently of the final path or geometric outcome.

If source imputation is required, the imputation model is fitted within each training partition and applied unchanged to the corresponding test partition. Missingness indicators are included only when scientifically justified. Complete-case and prespecified imputation analyses are treated as sensitivity analyses rather than as alternative opportunities for outcome-dependent selection.

Minimum ensemble size, covariance condition number, transport convergence, graph connectivity, persistence stability and density-support coverage are monitored as quality indicators. The same quality criteria are applied to biological and null paths. Failed transport or optimization runs are not silently replaced by favourable values.

Multiple testing and decision criteria

Primary path families, transport metrics, action functionals, representation comparisons, topological descriptors, graph scales and multiscale tests are grouped into prespecified testing families. Within each family, false-discovery-rate control is applied using a declared procedure and target rate.

Statistical significance alone is insufficient to support a path law. A primary path comparison requires the predicted effect direction, a non-negligible effect size, uncertainty compatible with the prediction and stability under prespecified transport, standardization and null-path sensitivity analyses. Equivalence or invariance claims require a prespecified equivalence margin and an interval-based test rather than inference from a non-significant difference.

The low-transport criterion is a lower mean matched transport cost for the biological path. The low-action criterion is a lower complete action under the same endpoints, operation budget, step convention and feasibility constraints. These criteria are distinct and are not substituted for one another.

Falsification criteria and interpretation boundaries

The low-transport relation is unsupported if biological paths do not exhibit lower matched transport than valid null paths, if the effect disappears under admissible transport estimators or common-space constructions, or if it depends on unequal endpoint, step-count or state-space constraints.

The low-action relation is unsupported if the full action does not differ in the predicted direction under strictly matched null paths, if action weights or feasibility terms are selected using final path ranks, or if the result disappears under prespecified action ablations and alternative admissible transport estimators.

A transition-level sharpness pattern is not considered independent confirmation of a global path law. Likewise, a numerically optimized path is not evidence that biological systems causally solve an optimization problem. It is only a constrained computational reference.

Persistent topology, multifractal dimensions, graph curvature and effective density landscapes are representation- and construction-dependent descriptors. Their use requires stability under subsampling, graph or filtration perturbation and matched null models. Non-zero persistence, non-zero curvature or a rugged negative log-density landscape alone does not establish biological significance.

Part IV therefore therefore defines a bounded path framework:

$$\{\mathrm{P}_t^{\star}, \mathbf{Y}_t\}_{t=0}^{N} \longrightarrow \mathcal{A}_2[\gamma] \qquad \text{(M186)}$$

The primary empirical object is the mean adjacent-state transport $\bar{W}_{2,\mathrm{G}}(\gamma)$. The action $\mathcal{A}_2[\gamma]$ is a theoretical extension unless all terms are evaluated under the same matched-null design. Without calibrated time-resolved data, it is a dimensionless statistical functional rather than a mechanical or thermodynamic action.

Computational framework

Protein ensembles were constructed from conformational observations, mutation sets, generated candidates and learned representations. BioEmu supplied conformational observations. ProstT5, ProtT5 and ESM-C supplied representation maps applied to declared sequence or structural observations. Cartesian conformations were aligned before covariance analysis; learned and discrete ensembles were analysed through their native maps or an explicitly defined common representation. The analyses comprised ensemble covariance geometry, Ledoit–Wolf-regularized perturbation costs, grouped source-to-geometry prediction and probability-measure transport along ordered state paths. The model versions, sampling parameters, representation layers, preprocessing rules and final configuration are recorded in the repository release associated with the manuscript.

Data

The final dataset contained the sequence-defined systems, generated conformations and quality-controlled DMS mutations reported in the repository manifest associated with this manuscript. DMS observations were grouped by protein for inference. Conformational observations were used to estimate state-level measures and covariances and were not treated as independent biological replicates. Independent proteins or systems defined the primary inference and held-out evaluation units.

Ensemble generation and processing

Conformational ensembles were generated with the BioEmu model version and sampling parameters recorded in the repository configuration. Mutation ensembles were constructed from the declared admissible substitutions; the treatment of the wild-type state and the inclusion or exclusion of the complete 20-amino-acid alternative set were fixed before analysis. Generated candidate ensembles used fixed generation settings. ProstT5, ProtT5 and ESM-C representations used fixed model versions, layers, tokenization, masking and pooling rules. Cartesian conformations were aligned by the declared Kabsch or generalized Procrustes procedure. Learned and discrete ensembles were not subjected to rigid-body alignment. Every state was represented by an empirical measure in its declared analysis space, and cross-state transport was calculated only after construction of a common metric space.

Statistical analysis

Covariance-normalized perturbation costs were calculated from common-coordinate displacements between matched background and perturbed ensemble representatives using the Ledoit–Wolf covariance estimator. Real-mutant responses, cross-resolution scalar costs and DMS associations were evaluated with protein-level grouping or

hierarchical resampling. Directional-consistency and coefficient-of-variation statistics were retained as transformation diagnostics and were not treated as independent biological validation.

Source-to-geometry relations were estimated from identifiable biological descriptors using grouped linear or nonlinear models. Predictors derived from the target ensemble were excluded from the corresponding predictive model. Model selection and hyperparameter tuning were performed within training partitions, and outer grouped partitions were reserved for held-out evaluation.

Ordered protein-state paths were mapped into a common representation space. The primary path statistic was

$$\bar{W}_{2,\mathrm{G}}(\gamma) = \frac{1}{N}\sum_{t=0}^{N-1} W_{2,\mathrm{G}}(P_t^{\star}, P_{t+1}^{\star}) \qquad \text{(M187)}$$

Structured paths were compared with null paths matched for endpoints, transition number, operation constraints, representation and admissible state space. The reported path analysis tested reduced mean distributional transport. The complete action functional was retained as a theoretical extension.