THREE BIOPHYSICAL CONSTRAINTS DETERMINE THE VARIATION OF STRUCTURAL DIVERGENCE AMONG RESIDUES IN ENZYME EVOLUTION

Julian Echave
Instituto de Ciencias Físicas (ICIFI-CONICET)
Universidad Nacional de San Martín
San Martín, Buenos Aires, Argentina
jechave@unsam.edu.ar

Mathilde Carpentier
Institut de Systématique, Évolution, Biodiversité (ISYEB, UMR 7205)
CNRS-MNHN-SU-EPHE-UA, Sorbonne Université
Paris, France
mathilde.carpentier@mnhn.fr

October 7, 2025

ABSTRACT

Structural divergence varies across residue positions in homologous enzyme families, forming residue-dependent structural divergence profiles. The evolutionary constraints that shape these profiles and the underlying physical and biological principles remain unclear. To address this question, we develop a Mutation-Stability-Activity (MSA) model of enzyme structure evolution. In MSA, mutations arise randomly and fix with probabilities that depend on their effects on folding stability, $\Delta\Delta G$, and catalytic activation energy, $\Delta\Delta G^{\ddagger}$, scaled by two selection-strength parameters, a_S and a_A. To calculate $\Delta\Delta G$ and $\Delta\Delta G^{\ddagger}$, and predict mutant structures, we use the coarse-grained Linearly Forced Elastic Network Model (LFENM). Applied to 34 enzyme families, MSA recapitulates observed structural divergence profiles. To identify which constraints shape these profiles, we compare four progressively complex nested models: uniform divergence (M0), mutation model (MM), mutation-stability model (MS), and the full mutation-stability-activity model (MSA). Each successive model provides improved fit to observations, demonstrating that mutation, stability, and activity constraints influence structural divergence profiles. Decomposing MSA predictions shows that the relative contributions of the mutation, stability, and activity components vary widely among enzyme families. We find that contributions vary among families because underlying factors vary: mutation contributions depend on internal flexibility heterogeneity, while stability and activity contributions depend on their respective external selection strengths (a_S, a_A). MSA provides quantitative, family-level estimates of these selection strengths, enabling future studies to link them to context-level properties such as expression, metabolic role, and specificity.

1 Introduction

Protein structures evolve more slowly than sequences, enabling comparisons between distantly related proteins whose sequences have diverged beyond recognition [1–4]. This property has made structural data valuable for phylogenetics when sequence comparisons fail [5–17]. Recent advances in machine learning have dramatically expanded structural data availability through improved prediction methods [18–22], shifting the bottleneck from data scarcity to our incomplete mechanistic understanding of structural evolution. While some mechanistic models of protein structural evolution exist [23, 24, 8, 25, 10, 12, 26], this understanding remains limited.

While whole-protein structural divergence has been well-studied using global metrics [1–4], the variation of structural divergence among residues within proteins remains largely unexplored. This contrasts sharply with sequence evolution, where rate variation among sites is a central problem that led to our understanding of evolutionary constraints [27]. Here, we address this gap by developing a mechanistic biophysical framework to examine how evolutionary constraints shape residue-dependent structural divergence profiles in functionally conserved enzymes.

Our previous analysis of 34 functionally conserved monomeric enzyme families provided the empirical foundation for this investigation [28]. Using an empirical modelling approach, we demonstrated that structural divergence varies significantly among residues and found evidence suggesting roles for both functional and non-functional constraints. However, such interpretations remain somewhat speculative without establishing the underlying biophysical causes. In this work, we move from empirical patterns to theory, employing a model grounded in first principles to identify the specific constraints governing structural evolution.

To establish the mechanistic basis of these constraints, we extend the Mutation-Stability-Activity (MSA) model—originally developed for enzyme sequence evolution [29, 30]—by combining it with the Linearly Forced Elastic Network Model (LFENM) to predict structural divergence patterns. This integration enables us to develop an analytical framework that systematically quantifies the relative contributions of different evolutionary constraints. By comparing nested models of increasing complexity—from mutation effects alone to the full mutation-stability-activity model—and decomposing the predictions into constraint-specific components, we can identify which biophysical forces shape structural evolution patterns and determine their relative importance across enzyme families.

2 Methods

Our approach combines evolutionary modelling with biophysical calculations to predict residue-dependent structural divergence patterns. We first describe the Mutation-Stability-Activity (MSA) evolutionary framework and its integration with the Linearly Forced Elastic Network Model (LFENM) for calculating mutational effects on structure, stability, and activity. We then detail our implementation for generating structural divergence profiles, including mutational scanning, Bayesian parameter estimation, and profile prediction. Finally, we describe how we obtain observed structural divergence profiles from enzyme families and our analytical framework for nested model comparisons and constraint decomposition.

2.1 Theoretical framework: Mutation-Stability-Activity model

We apply the Mutation-Stability-Activity (MSA) evolutionary model [29] in combination with the Linearly Forced Elastic Network Model (LFENM) [23, 24] to predict structural divergence patterns in enzyme evolution. This MSA-LFENM combination has been used previously to study sequence evolution [29, 30]; here we extend its application to predict how structural divergence patterns emerge from the interplay of mutation, stability selection, and activity selection.

Mutation-selection process

The MSA model represents evolution as a mutation-selection process consisting of repeated mutation-selection steps. The population is assumed to be monoclonal (only one wild-type genotype exists at any given time), which is valid for low mutation rates (where the time between mutation events is much longer than the time required for fixation or loss) [31]. At each evolutionary time-step, we denote the wild-type protein structure as $\mathbf{r}_{\text{wt}}$. When a mutation is introduced, it creates a mutant structure $\mathbf{r}_{\text{mut}}$ that either disappears from the population or becomes fixed, replacing the wild type. As substitutions (fixed mutations) accumulate over time, the protein structure progressively diverges from its ancestral state.

The fixation process is governed by the fixation probability function. In the MSA model, this fixation probability depends on mutational effects on protein stability and catalytic activity [29]. Specifically, it is given by:

$$p_{\text{fix}}(\Delta\Delta\text{G}, \Delta\Delta\text{G}^{\ddagger}) = \frac{1}{N}\min(1, e^{-\,\text{a}_\text{S}\,\Delta\Delta\text{G}}) \times \min(1, e^{-\,\text{a}_\text{A}\,\Delta\Delta\text{G}^{\ddagger}}) \tag{1}$$

where N is the population size, $\Delta\Delta\text{G} = \Delta\text{G}_{\text{mut}} - \Delta\text{G}_{\text{wt}}$ is the mutational change in protein stability (negative values indicate stabilization), and $\Delta\Delta\text{G}^{\ddagger} = \Delta\text{G}^{\ddagger}{}_{\text{mut}} - \Delta\text{G}^{\ddagger}{}_{\text{wt}}$ is the mutational change in activation energy (negative values indicate lower activation energy and, therefore, increased catalytic activity).

The parameters a_S and a_A control selection strength against destabilizing and deactivating mutations, respectively. The maximum fixation probability is $1/N$, corresponding to neutral drift. (When mutant and wild type have equal fitness, each individual lineage has equal probability of eventual fixation, so a single mutant in a population of N has probability $1/N$ of fixing.) This maximum is achieved when mutations don't compromise stability or activity—that is, when $\Delta\Delta\text{G} \leq 0$ and $\Delta\Delta\text{G}^{\ddagger} \leq 0$—or when selection is absent ($\text{a}_\text{S} = \text{a}_\text{A} = 0$). For deleterious mutations ($\Delta\Delta\text{G} > 0$

or $\Delta\Delta G^{\ddagger} > 0$), the fixation probability decreases exponentially with both the magnitude of the deleterious effect and the corresponding selection parameter.

Linearly Forced Elastic Network Model

To complete the evolutionary model, we need a way to calculate mutant structures and the mutational effects on stability and activation free energy. For this purpose we use the Linearly Forced Elastic Network Model (LFENM) [23].

LFENM represents the wild-type protein as an elastic network and mutants as perturbed elastic networks [23]. The wild type protein is a network of nodes, that represent residues, connected by harmonic springs, that represent interactions. Mathematically, the wild-type's energy is

$$V_{\text{wt}}(\mathbf{r}) = \frac{1}{2}\sum_{i<j} k_{ij}(d_{ij} - l_{ij})^2 \tag{2}$$

where d_{ij} is the distance between residues i and j in conformation $\mathbf{r}$, l_{ij} is the equilibrium spring length, and k_{ij} is the spring constant.

Introducing a random mutation at a given residue is modelled by adding random perturbations to the lengths of the springs that connect the mutated site with its neighbours [23]. As a result of this perturbation, there is a change of the energy landscape. The mutant's energy is

$$V_{\text{mut}}(\mathbf{r}) = \frac{1}{2}\sum_{i<j} k_{ij}(d_{ij} - l_{ij}^{\text{mut}})^2 \tag{3}$$

where l_{ij}^{mut} are the equilibrium spring lengths for the mutant, defined as $l_{ij}^{\text{mut}} = l_{ij} + \delta l_{ij}$ for springs involving the mutated residue, and $l_{ij}^{\text{mut}} = l_{ij}$ for all other springs. The perturbations δl_{ij} are independently picked from a normal distribution of mean zero and standard deviation σ_{mut}, which is chosen small enough to ensure the perturbative approximation remains valid. (Note that this approach does not model mutations between actual amino-acid residues, but does capture the correct distribution of effects caused by actual mutations without requiring detailed side-chain modelling [23, 24, 26].)

LFENM calculations of mutational effects

The LFENM model provides formulae to calculate the mutational changes of structure, stability, and activation energy.

These calculations require only two components: the network matrix $\mathbf{K}$, which is the Hessian of the wild-type energy function V_{wt}, and a force vector $\mathbf{f}$ that encodes the spring perturbations. The force vector has the following structure: for each perturbed spring connecting the mutated residue j to a neighbour i, residue i experiences a force of magnitude $f_{ij} = k_{ij}\delta l_{ij}$ directed along the contact, and residue j experiences an equal and opposite reaction force. The total force on the mutated residue j is the vector sum of reaction forces from all its perturbed neighbours. The vector $\mathbf{f}$ is assembled from these individual force components [24, 26]. Using these components, we now present the three key formulae

First, we consider the calculation of the structural change, $\Delta\mathbf{r}^0$. From Eq.(3), expanding the mutant's energy up to second order around its equilibrium conformation, the following equation can be derived [23, 24]:

$$\Delta\mathbf{r}^0 = \mathbf{C}\mathbf{f} \tag{4}$$

where $\Delta\mathbf{r}^0 = \mathbf{r}^0_{\text{mut}} - \mathbf{r}^0_{\text{wt}}$ is the structural change upon mutation, $\mathbf{f}$ is the force vector described above, and $\mathbf{C}$ is the wild-type protein's variance-covariance matrix, obtained as the pseudo-inverse of $\mathbf{K}$ (since $\mathbf{K}$ is singular due to translational and rotational invariance).

Second, we consider the calculation of the stability change, $\Delta\Delta G$. The LFENM $\Delta\Delta G$ is the difference between the minimum energies of the mutant (Eq.(3)) and wild-type (Eq.(2)). This can be written [29]:

$$\Delta\Delta G = \frac{1}{2}\sum_{i\sim j} k_{ij}\delta l_{ij}^2 - \frac{1}{2}(\Delta\mathbf{r}^0)^T\mathbf{K}(\Delta\mathbf{r}^0) \tag{5}$$

where the summation is over contacts involving the mutated site (the only non-zero δl_{ij} terms). The first term represents the local strain energy that results from perturbed contacts, while the second term accounts for the decrease in energy due to a global relaxation of the mutant's structure.

Finally, we consider the calculation of activation free energy change $\Delta\Delta G^{\ddagger}$. Briefly, assuming that the wild-type enzyme has a pre-organized active site with optimal catalytic geometry, mutations alter the active site conformation

away from this optimum. The activation free energy change represents the energetic cost for the mutant to restore the optimal geometry (for derivation details, see [29]):

$$\Delta\Delta G^{\ddagger} = \frac{1}{2}(\Delta \mathbf{r}_a^0)^T \mathbf{K}_{aa}^{\text{eff}} (\Delta \mathbf{r}_a^0) \tag{6}$$

where $\Delta \mathbf{r}_a^0$ is the active-site subset of the structural change $\Delta \mathbf{r}^0$, and $\mathbf{K}_{aa}^{\text{eff}}$ is the effective stiffness matrix for active site residues, which is the inverse of $\mathbf{C}_{aa}$, the block of the variance-covariance matrix that corresponds to active-site coordinates.

2.2 Computational implementation: MSA structural divergence profiles

In this section we describe our computational implementation of MSA for predicting residue-dependent structural divergence profiles. The calculation pipeline takes a protein structure as input and proceeds through four main stages:

1. Set up the elastic network model for the wild-type structure to enable LFENM calculations of mutational effects
2. Perform a mutational scan across all residue positions, calculating how each mutation affects protein structure ($\Delta \mathbf{r}^0$), stability ($\Delta\Delta G$), and activation energy ($\Delta\Delta G^{\ddagger}$) using LFENM
3. Estimate the posterior distribution of MSA model parameters (a_S and a_A) from observed structural divergence data using Bayesian inference
4. Calculate MSA predicted profiles by averaging predictions across the posterior sample of parameters

Wild-type elastic network model setup

Specifying the elastic network model for the wild-type, amounts to specifying the parameters of the energy function shown in Eq. (2), and calculating the network matrix $\mathbf{K}$ and it's the variance-covariance matrix $\mathbf{C}$, necessary for the calculations of mutational changes structure (Eq. (4)), stability (Eq. (5)), and activity (Eq. (6)).

There are several popular elastic network models of proteins [32–35]. However, this choice is not a major issue, because LFENM predictions are very robust with respect to the specific model used [26]. Here, we use the parameters derived by Ming and Wall [35], which are somewhat more realistic than other simpler models, and that work well for MSA predictions of sequence patterns [29, 30]. Nodes are placed at the residue alpha carbons, C_α; spring lengths are the distances in the equilibrium native structure: $l_{ij} = d_{ij}^0$; and spring force constants depend on whether contacts are along the sequence or not: $k_{ij} = 189$ kcal/mol for sequence neighbours, and $k_{ij} = 4.5$ kcal/mol for other contacts within cut-off distance $R_0 = 10.5$ Å.

Mutational scanning calculations

After setting up the wild-type elastic network model, we perform a mutational scan. For each residue j we generate ten independent mutants ($m = 1, \ldots, 10$). Each mutant is produced by perturbing the lengths of the springs connected to residue j, excluding its sequence neighbours ($j \pm 1$) that represent the covalent backbone. The perturbations δl_{ij} are drawn independently from a normal distribution with mean 0 and standard deviation $\sigma_{\text{mut}} = 0.3$ Å. This amplitude is small enough to avoid unphysical conformations, and its exact value does not affect results because our structural-divergence measures remove any uniform scaling.

For every mutation (j, m) we compute the corresponding force vector $\mathbf{f}$ and then evaluate the mutational effects: the structural change $\Delta \mathbf{r}^0(j, m)$ (Eq. (4)), the stability change $\Delta\Delta G(j, m)$ (Eq. (5)), and the activation-energy change $\Delta\Delta G^{\ddagger}(j, m)$ (Eq. (6)).

This procedure samples the typical range of structural, stability, and activation-energy changes caused by amino-acid substitutions without requiring explicit sequence information or detailed side-chain modelling.

Bayesian parameter estimation

We infer the stability and activity selection strengths, $\mathrm{a_S}$ and $\mathrm{a_A}$, using a Bayesian framework.

Forward model. As a first step we specify a forward model—that is, a computational procedure that predicts the residue-dependent structural-divergence profile for any given pair of parameter values.

To efficiently compute structural divergence profiles for different parameter values, we exploit a key property of evolutionary structural divergence: as substitutions accumulate over evolutionary time, the number of substitutions scales all residue RMSD values uniformly by the same factor F(n), where n is the total number of substitutions. This uniform scaling makes the ratios $\mathrm{RMSD}(i)/\,\mathrm{RMSD}(j)$ independent of the total number of substitutions, since the scaling factor F(n) cancels out in any ratio. Since we seek only the relative pattern of structural divergence

among residues—which doesn't depend on the number of substitutions—we may as well use n = 1. We exploit this independence to avoid repeated simulations: a naïve approach would require, for every trial pair of parameter values, simulating many multi-substitution evolutionary trajectories and recalculating structural displacements. Instead, we perform the mutation scan once—independently of $\mathrm{a_S}$ and $\mathrm{a_A}$, to obtain the mutation-induced displacement vectors $\Delta\mathbf{r}^0(j,m)$, and for each parameter pair we merely compute weighted averages of these pre-computed displacements as described below.

For a given pair of parameter values $(\mathrm{a_S}, \mathrm{a_A})$ we calculate the fixation probability $p_{\text{fix}}(j,m;\mathrm{a_S},\mathrm{a_A})$ of every mutant (j,m) using Eq. (1), together with the mutational effects $\Delta\Delta\mathrm{G}(j,m)$ and $\Delta\Delta\mathrm{G}^{\ddagger}(j,m)$ obtained in the mutation-scan step. These probabilities are then normalised to give the composition of the evolved ensemble—the relative abundance of each mutant:

$$P(j,m;\mathrm{a_S},\mathrm{a_A}) = \frac{p_{\text{fix}}(j,m;\mathrm{a_S},\mathrm{a_A})}{Z}, \tag{7}$$

where the normalising constant Z is the sum of p_{fix} over all mutants.

From this ensemble composition we calculate the expected structural divergence. For each residue i, the root-mean-square deviation is the probability-weighted displacement:

$$\mathrm{RMSD}(i;\mathrm{a_S},\mathrm{a_A}) = \sqrt{\sum_{j,m} P(j,m;\mathrm{a_S},\mathrm{a_A}) \left\|\Delta\mathbf{r}_i^0(j,m)\right\|^2}, \tag{8}$$

where $\Delta\mathbf{r}_i^0(j,m)$ is the three-dimensional mutation-induced displacement vector of residue i obtained in the mutational-scan step.

To focus on residue-to-residue variation we use the centred log–root-mean-square displacement:

$$\mathrm{nlRMSD}_{\text{MSA}}(i) = \log \mathrm{RMSD}(i;\mathrm{a_S},\mathrm{a_A}) - \langle \log \mathrm{RMSD} \rangle_{\text{protein}}. \tag{9}$$

Equations (7)–(9) together constitute the forward model, mapping any pair of parameters $(\mathrm{a_S}, \mathrm{a_A})$ to the predicted profile of centred log–RMSD values.

Bayesian inference. To estimate the parameters we combine the forward model with the data using Bayes' rule, giving the posterior distribution of parameters:

$$p(\mathrm{a_S}, \mathrm{a_A} \mid \text{data}) \propto L(\mathrm{a_S}, \mathrm{a_A})\, p(\mathrm{a_S}, \mathrm{a_A}), \tag{10}$$

where $L(\mathrm{a_S}, \mathrm{a_A})$ is the likelihood of the observed centred log-RMSD profile. The log-likelihood used in computation assumes normally distributed residuals:

$$\log L(\mathrm{a_S}, \mathrm{a_A}) = \sum_i \log \mathcal{N}\big(\mathrm{nlRMSD}_{\text{obs}}(i) - \mathrm{nlRMSD}_{\text{MSA}}(i;\mathrm{a_S},\mathrm{a_A}) \mid 0, \sigma^2\big)\,, \tag{11}$$

where $\mathcal{N}(\cdot \mid 0, \sigma^2)$ is the density of a normal distribution with mean zero and variance σ^2.

We assigned weakly informative uniform priors:

$$\mathrm{a_S} \sim \mathrm{Uniform}(0, 10), \qquad \log_2\!\big(\mathrm{a_A} + 1\big) \sim \mathrm{Uniform}(0, 12). \tag{12}$$

Stability selection strengths across enzyme families vary within roughly one order of magnitude, whereas activity selection spans several orders [29, 30]. Accordingly we place a direct uniform prior on $\mathrm{a_S}$, while for $\mathrm{a_A}$ we work on a logarithmic scale. The transformation $\log_2(\mathrm{a_A} + 1)$ keeps the mapping defined even when $\mathrm{a_A}$ is near zero.

Posterior sampling uses a Metropolis–Hastings Markov chain Monte Carlo algorithm (we ran 2,500 iterations with a 500-iteration burn-in and checked convergence in runs extended to 10,000 iterations). At each step the algorithm proposes a new pair of parameters $(\mathrm{a_S}, \mathrm{a_A})$ and decides whether to accept it based on the ratio of posterior probabilities of the proposed and current states— that is, the product of likelihood and prior. Because the prior is uniform on $\mathrm{a_S} \in [0, 10]$ and $\log_2(\mathrm{a_A} + 1) \in [0, 12]$, any proposal outside these ranges is rejected automatically, and within these bounds the prior contributes equally, so acceptance is determined by the likelihood ratio. The resulting Markov chain provides the posterior sample of $(\mathrm{a_S}, \mathrm{a_A})$ used for subsequent analyses.

Throughout Results and figures, reported values of $\mathrm{a_S}$ and $\mathrm{a_A}$ correspond to posterior means from the MCMC sample, unless otherwise noted.

Posterior-averaged predictions

Having obtained an MCMC sample from the posterior of $(\mathrm{a_S}, \mathrm{a_A})$, we calculate the model's residue-dependent structural-divergence profile by averaging predictions over these sampled parameter values:

$$\mathrm{nlRMSD}_{\text{MSA}}(i) = \big\langle \mathrm{nlRMSD}_{\text{MSA}}(i;\mathrm{a_S},\mathrm{a_A}) \big\rangle_{\text{posterior sample}}. \tag{13}$$

This posterior-sample mean constitutes the MSA model's primary output for all subsequent analyses.

2.3 Observed structural divergence profiles

This section describes how we obtain observed residue-dependent structural divergence profiles from evolutionary data, to serve as targets for MSA model validation. We describe the enzyme families dataset, structural alignment and superposition procedures, and quantification of residue-level structural divergence.

Dataset of enzyme families

We used a dataset of 34 functionally conserved, monomeric, single-domain enzyme families previously curated from the Mechanism and Catalytic Site Atlas (M-CSA) database [28]. Each family consists of a reference enzyme (for which mechanism and active site information has been experimentally determined) and a set of homologues. The dataset construction process, detailed in our previous work, applied sequential filtering to ensure functional conservation and structural quality, focusing on monomeric, single-domain proteins to avoid complications from inter-domain or inter-chain interactions. The final dataset comprises 34 enzyme families with broad structural and functional diversity: 27 α/β, 5 all-β, and 2 all-α folds across major CATH architectures. Functionally, hydrolases dominate (20 families), with transferases, lyases, isomerases, and oxidoreductases also represented. Family sizes range from 4-23 proteins, with reference enzymes spanning 98-447 residues.

Structural alignment and superposition

Homologous sequences were aligned using MUSCLE v5 [36] via the bio3d R package [37]. We chose sequence over structural alignment because functionally conserved families retain sufficient sequence similarity for reliable alignment while providing an alignment basis independent of structural considerations for studying structural divergence. Following sequence alignment, protein structures were superimposed by minimizing the global (protein-level) RMSD of Cα coordinates, with each homologue aligned to its family's reference enzyme.

Calculation of observed structural divergence profiles

Structural divergence was quantified at each residue position using the residue-level root mean square deviation of Cα coordinates across homologous structures. For residue i in a family of N_{hom} homologous structures:

$$\text{RMSD}_{\text{obs}}(i) = \sqrt{\frac{1}{N_{\text{hom}}} \sum_{p=1}^{N_{\text{hom}}} ||\Delta \mathbf{r}_i^p||^2} \tag{14}$$

where $\Delta \mathbf{r}_i^p$ is the deviation of residue i in structure p from its average position.

We applied logarithmic transformation and centring to match the format used for model predictions:

$$\text{nlRMSD}_{\text{obs}}(i) = \log(\text{RMSD}_{\text{obs}}(i)) - \langle \log(\text{RMSD}_{\text{obs}}) \rangle_{\text{protein}} \tag{15}$$

where the second term is the mean log-RMSD across all residues in the protein. These centred log-RMSD profiles serve as the observational targets for model validation and parameter fitting.

2.4 Analysis

Once we have obtained observed and model structural divergence profiles, we analysed them in the following four steps:

1. Model validation by comparison with observed structural divergence profiles
2. Nested model comparison to assess whether each constraint contributes to explaining the data
3. Decomposition analysis to quantify the relative importance of different constraints
4. Identification of enzyme properties that determine why constraint importance varies among families

Model validation

We validate MSA predictions by comparing them with observed structural divergence profiles and with predictions from our previous empirical model M12 [28]. M12 is a shape-constrained generalized additive model that smooths observed structural divergence based on local similarity in residue flexibility and distance to the active site. With its many fitted parameters, M12 provides a highly flexible fit that effectively removes observational noise, serving as a benchmark for the underlying signal. Model performance is quantified using Pearson correlation coefficients between predicted and observed profiles across all enzyme families. We also calculate correlations between MSA and M12 predictions to assess whether the parsimonious mechanistic approach captures similar patterns to the flexible empirical model.

Nested model comparison

To determine whether all three MSA constraints influence structural divergence patterns, we examine four nested model variants that systematically add components: M0, MM, MS, and MSA. The baseline null model M0 represents uniform structural divergence across all residues, which corresponds to $\mathrm{nlRMSD}_{\text{M0}}(i) = 0$ since we use centred profiles. The mutation model MM includes only mutational effects by setting both selection parameters to zero. The mutation-stability model MS adds stability selection by setting only the activity selection parameter to zero. The full MSA model includes all three constraints.

To obtain predictions from each nested model, we use the posterior sample of parameters from the full MSA fit. For each parameter pair in this sample, we calculate the corresponding nested model prediction by setting the appropriate parameters to zero (e.g., $\mathrm{a_S} = \mathrm{a_A} = 0$ for MM, or $\mathrm{a_A} = 0$ for MS). We then average these predictions over the entire posterior sample. This approach avoids potential compensation effects that could arise from refitting simplified models where remaining parameters might inflate to compensate for missing constraints.

We quantify the explanatory power of each model using explained deviance D^2:

$$D^2(M) = 1 - \frac{\mathrm{Var}(\mathrm{nlRMSD}_{\text{obs}} - \mathrm{nlRMSD}_{\text{M}})}{\mathrm{Var}(\mathrm{nlRMSD}_{\text{obs}})} \tag{16}$$

where M represents the model being evaluated, $D^2 = 0$ for the baseline model and $D^2 = 1$ for perfect fit. We test the contribution of each constraint by examining the increments ΔD^2 in the progression M0 $\rightarrow$ MM $\rightarrow$ MS $\rightarrow$ MSA. Positive increments at each step would indicate that the added constraint contributes explanatory power beyond what the simpler models already capture.

Decomposition analysis

To quantify the individual contributions of each constraint to structural divergence patterns, we decompose MSA predictions into additive components for each residue position. This decomposition isolates the effect of each constraint by calculating how much each adds to the overall structural divergence pattern:

$$\begin{aligned} &\mathrm{nlRMSD}_{\text{MSA}}(i) = \phi_{\mathrm{M}}(i) + \phi_{\mathrm{S}}(i) + \phi_{\mathrm{A}}(i) \\ \text{where} \quad &\phi_{\mathrm{M}}(i) = \mathrm{nlRMSD}_{\text{MM}}(i) \\ &\phi_{\mathrm{S}}(i) = \mathrm{nlRMSD}_{\text{MS}}(i) - \mathrm{nlRMSD}_{\text{MM}}(i) \\ &\phi_{\mathrm{A}}(i) = \mathrm{nlRMSD}_{\text{MSA}}(i) - \mathrm{nlRMSD}_{\text{MS}}(i) \end{aligned} \tag{17}$$

Here i denotes the residue position, and the components represent the mutation-only contribution, the additional effect of stability selection, and the additional effect of activity selection, respectively.

This decomposition approach is necessary because ΔD^2 values from the nested comparison, while establishing that constraints contribute, cannot isolate the individual contribution of each constraint due to interaction effects.

Two problems arise with variance-based measures like D^2. First, D^2 variation among enzymes depends on observational noise, not just model accuracy. Second, interaction effects contaminate the measures: since D^2 is based on squared deviations, interaction terms arise between components. (For example, from the definitions above $\mathrm{nlRMSD}_{\text{MS}}(i) = \phi_{\mathrm{M}}(i) + \phi_{\mathrm{S}}(i)$, so $D^2(\text{MS}) = D^2(\phi_{\mathrm{M}}) + D^2(\phi_{\mathrm{S}}) + 2\mathrm{Cor}(\phi_{\mathrm{M}}, \phi_{\mathrm{S}})$, where Cor denotes correlation. The increment $\Delta D^2 = D^2(\text{MS}) - D^2(\text{MM})$ therefore includes an interaction term, preventing ΔD^2 from measuring the stability constraint's contribution cleanly.)

To avoid these issues, we quantify component contributions using the standard deviations of the individual components across residue positions, $\mathrm{SD}(\phi_{\mathrm{M}})$, $\mathrm{SD}(\phi_{\mathrm{S}})$, and $\mathrm{SD}(\phi_{\mathrm{A}})$, which directly measure each constraint's contribution to profile variation without interaction effects. Relative contributions are calculated by dividing each component's standard deviation by the sum of all component standard deviations.

Determinants of constraint contributions

We examine which enzyme properties explain the variation in $\mathrm{SD}(\phi_{\mathrm{M}})$, $\mathrm{SD}(\phi_{\mathrm{S}})$, and $\mathrm{SD}(\phi_{\mathrm{A}})$ across enzyme families.

For mutation contributions, previous work shows that mutational structural effects correlate strongly with protein flexibility patterns [26]. We therefore examine the correlation between $\mathrm{SD}(\phi_{\mathrm{M}})$ and flexibility heterogeneity $\mathrm{SD}(\mathrm{lRMSF})$, where $\mathrm{lRMSF}(i) = \log \sqrt{\mathrm{Tr}(\mathbf{C}_{ii})}$ is the logarithm of the root mean square fluctuation of residue i.

For stability and activity contributions, previous findings indicate that selection contributions depend on their respective selection strengths [30]. We therefore examine correlations between $\mathrm{SD}(\phi_{\mathrm{S}})$ and the stability selection parameter $\mathrm{a_S}$, and between $\mathrm{SD}(\phi_{\mathrm{A}})$ and the activity selection parameter $\mathrm{a_A}$.

We use Spearman correlation coefficients to assess these relationships.

3 Results and Discussion

We use the Mutation-Stability-Activity MSA model to study the roles of mutation, stability, and activity constraints in shaping residue-dependent structural divergence profiles. First, we validate MSA by comparison with the observations and with predictions of the empirical model M12 of [28]. Second, we compare a progression of nested models to evaluate MSA predictions against observations, and assess whether all constraints contribute to the predictions. Third, we decompose MSA structural divergence profiles into a sum of mutation, stability, and activity components, to quantify their relative contributions. Finally, we study what enzyme properties account for the variation of constraint balance among enzyme families.

3.1 MSA captures empirical patterns mechanistically

We compared MSA's performance to our previous empirical model M12 to establish whether our mechanistic biophysical model captures the same patterns as a flexible empirical approach. M12 is a shape-constrained additive model that smooths observed structural divergence based on residue flexibility and distance to the active site, using many fitted parameters to flexibly capture underlying patterns.

MSA achieves comparable accuracy to the empirical model, despite having only two parameters. Correlations between predictions and observed profiles average $R(\text{MSA}, \text{obs}) = 0.66 \pm 0.02$ for MSA versus $R(\text{M12}, \text{obs}) = 0.69 \pm 0.02$ for M12 across 34 enzyme families (ranges $0.43 - 0.83$ and $0.45 - 0.86$, respectively; Fig. 1A). Furthermore, MSA and M12 profiles are highly similar to each other, with correlations averaging $R(\text{MSA}, \text{M12}) = 0.94 \pm 0.01$ (range $0.8 - 0.99$; Fig. 1B).

The high correlation between these fundamentally different approaches indicates that MSA provides a mechanistic foundation for patterns that were previously interpreted more speculatively. While our earlier M12 analysis suggested that structural divergence reflects both non-functional constraints (related to flexibility) and functional constraints (related to active site proximity), MSA now demonstrates that these patterns arise specifically from the interplay of mutation, stability, and activity selection pressures operating through well-defined biophysical mechanisms. The strong agreement between MSA and M12 profiles confirms that MSA successfully accounts for the observed gradients of increasing structural divergence with both flexibility and distance from the active site.

With MSA validated against empirical patterns, we can proceed to analyze the mechanistic basis of structural divergence. MSA's mechanistic foundation allows explicit consideration of three constraint components—mutation, stability, and activity—enabling identification of which constraints operate in each enzyme family and quantification of their relative contributions, which we investigate in the following sections.

[Figure 1 about here.]

3.2 Three constraints shape structural divergence profiles

To determine whether all three MSA constraints influence structural divergence, we compared four nested models of increasing complexity (see Methods). Starting from uniform divergence (M0), we progressively added mutational effects (MM), stability selection (MS), and finally activity selection (MSA). This systematic progression allows us to test whether each constraint contributes explanatory power beyond the previous constraints.

Three example families that illustrate different constraint patterns show explained deviance (ΔD^2) increases across nested models (Fig. 2A). The aldo/keto reductase family (M-CSA ID 858) shows explained deviance increasing from MM (0.09) to MS (0.56) with no further improvement to MSA (0.56), indicating stability selection dominance. The metallo-beta-lactamase family (M-CSA ID 15) shows moderate improvements at each step (MM: 0.23, MS: 0.40, MSA: 0.59), indicating balanced contributions. The ribonuclease U2 family (M-CSA ID 908) shows minimal improvement from MM (0.11) to MS (0.17) but substantial improvement to MSA (0.65), indicating activity selection dominance. These examples suggest that all three constraints may contribute to structural divergence, with substantial variation in their relative importance across enzyme families.

Across the full dataset, adding each constraint improves model performance on average (Fig. 2B). Mean D^2 values across all 34 families increase from 0.13 ± 0.01 with mutation effects alone, to 0.33 ± 0.02 when stability selection is added, to 0.44 ± 0.02 when activity selection is included. The incremental explained deviance (ΔD^2) values confirm that each step provides significant improvement in explained deviance on average (Fig. 2C): mutation adds 0.13 ± 0.01, stability selection adds 0.20 ± 0.02, and activity selection adds 0.11 ± 0.02, with each increment significant at p ¡ 0.001.

However, the wide ranges in ΔD^2 values reveal that individual families may deviate substantially from these averages (Fig. 2C): mutation contributions range from 0.07 to 0.38, stability contributions range from 0.02 to 0.47, and activity contributions range from 0.00 to 0.48. These ranges indicate that while all three constraints contribute in general, individual enzyme families may show minimal contribution from specific constraints.

These results establish that the three MSA constraints influence structural divergence profiles. The improvement from M0 to MM demonstrates that mutational effects create residue-dependent variation beyond uniform divergence, so structural divergence would be residue-dependent even in the absence of selection. The improvements from MM to MS and from MS to MSA demonstrate that both stability and activity constraints also contribute to the observed profiles, establishing that all three constraints need to be considered when analysing structural evolution in any given enzyme family.

The variation in ΔD^2 values among families suggests that the balance of constraints varies among families. However, while these ΔD^2 values might appear to measure individual contributions, they cannot isolate them due to interaction effects, requiring a different decomposition method.

[Figure 2 about here.]

3.3 Constraint contributions vary among enzymes

As established in the previous section, all three MSA constraints (mutation, stability, and activity constraints) affect residue-dependent structural divergence profiles. However, the ΔD^2 increments from the nested model comparison cannot serve as measures of each constraint's individual contribution due to interaction effects between components (see Methods for detailed explanation). To quantify these individual contributions across enzyme families, we decompose MSA predictions into constraint-specific components.

Briefly, we write $\mathrm{lRMSD}_{\mathrm{MSA}} = \phi_{\mathrm{M}} + \phi_{\mathrm{S}} + \phi_{\mathrm{A}}$, where the components represent the contributions of mutation, stability, and activity constraints, respectively. We quantify each component's contribution using its standard deviation across residue positions: $\mathrm{SD}(\phi_{\mathrm{M}})$, $\mathrm{SD}(\phi_{\mathrm{S}})$, and $\mathrm{SD}(\phi_{\mathrm{A}})$, where higher standard deviation indicates greater variation across residues and thus larger contribution to the overall divergence pattern.

The same three enzyme families examined in the nested model analysis illustrate different constraint contribution patterns (Fig. 3A). The aldo/keto reductase family (M-CSA ID 858) shows stability component dominance with relative contribution (relative standard deviation) of 0.60, mutation component contribution of 0.39, and minimal activity component contribution of 0.01. The metallo-beta-lactamase family (M-CSA ID 15) shows balanced contributions with relative contributions of 0.48 for mutation, 0.19 for stability, and 0.33 for activity. The ribonuclease U2 family (M-CSA ID 908) shows activity component dominance with relative contribution of 0.57, minimal stability component contribution of 0.09, and mutation component contribution of 0.34.

Across the full dataset, the absolute contributions show mutation averaging highest (0.27±0.01), followed by stability (0.19±0.02) and activity (0.12±0.02) (Fig. 3B). All components show substantial variation among families, but the mutation component maintains meaningful contributions across all families (range 0.19-0.50), whereas stability and activity components can have minimal impact in some families (ranges 0.03-0.41 and 0.01-0.46, respectively). To understand constraint balance patterns, we examine the relative contributions.

The relative contributions show mutation as the primary constraint, followed by stability, then activity (Fig. 3C). On average, the mutation component accounts for 0.47±0.01 of the total variation, the stability component contributes 0.33±0.03, and the activity component contributes 0.20±0.03.

However, all components show large variation among families, revealing diverse constraint patterns (Fig. 3C). At the low end, the mutation component is always present with substantial contributions (minimum 0.34), whereas stability and activity components can have minimal impact (minimums 0.07 and 0.01). At the high end, all three constraints can reach dominance levels (maximums 0.66, 0.60, and 0.57), indicating that any constraint can become the single dominant force shaping structural divergence patterns in individual families.

The decomposition analysis demonstrates that all three constraints contribute to structural divergence, but their relative importance varies considerably among enzyme families. The mutation constraint ranges from substantial to dominant contributions and is always present, whereas stability and activity constraints can have minimal impact in some families or become the dominant force in others. This variation in constraint balance among enzyme families raises the question of what determines these different patterns, which we examine in the following section.

[Figure 3 about here.]

Methodological note: Decomposition vs. nested model analysis

This decomposition approach reveals constraint patterns that differ markedly from those suggested by the ΔD^2 analysis. While the ΔD^2 increments suggested the stability constraint as having the largest effect (0.20±0.02), followed by mutation and activity (0.13±0.01 and 0.11±0.02, respectively) (Fig. 2C), the decomposition analysis reveals the mutation constraint as strongest (0.47±0.01), followed by stability (0.33±0.03) and activity (0.20±0.03) (Fig. 3C).

The difference arises because decomposition measures contributions as the SD of each individual component, which depends only on that component, whereas ΔD^2 increments capture improvements in fit that include interaction effects between components. Strong correlation between mutation and stability components inflates the apparent stability effect in the ΔD^2 analysis, while moderate correlations with activity similarly inflate the apparent activity effect.

3.4 Constraint balance is determined by protein architecture and selection strengths

The decomposition analysis revealed substantial variation in the contributions of mutation, stability, and activity components among enzyme families. Previous studies showed that mutational effects on structural patterns correlate with protein flexibility [26], while stability and activity selection contributions to sequence evolution depend on their respective selection parameters [30]. We examined whether similar relationships determine structural divergence patterns.

Each contribution is primarily determined by a single factor (Fig. 4). The contribution of the mutation component correlates strongly with flexibility heterogeneity, as shown by its strong correlation with $\mathrm{SD(lRMSF)}$ (Spearman correlation $\rho = 0.99$) (Fig. 4A). The contribution of the stability component is primarily determined by the stability selection parameter, supported by the high correlation between $\mathrm{SD}(\phi_\mathrm{S})$ and the posterior mean estimate of $\mathrm{a_S}$ ($\rho = 0.93$) (Fig. 4B). Similarly, the contribution of the activity component is primarily determined by the activity selection parameter, as indicated by the correlation between $\mathrm{SD}(\phi_\mathrm{A})$ and the posterior mean estimate of $\mathrm{a_A}$ ($\rho = 0.94$) (Fig. 4C).

The strong correlation between the mutation contribution and flexibility heterogeneity arises because mutational structural effects at individual residues correlate nearly perfectly with their flexibility [26]. Consequently, the overall mutation contribution (measured by $\mathrm{SD}(\phi_\mathrm{M})$) must correlate with the heterogeneity of flexibility among residues (measured by $\mathrm{SD(lRMSF)}$), an intrinsic property of protein architecture. In contrast, the stability and activity contributions are governed by the selection parameters $\mathrm{a_S}$ and $\mathrm{a_A}$: stronger selection against destabilizing mutations (higher $\mathrm{a_S}$) amplifies the stability contribution to divergence profiles, while stronger selection against deactivating mutations (higher $\mathrm{a_A}$) amplifies the activity contribution, consistent with previous observations for sequence evolution [30].

[Figure 4 about here.]

4 Conclusion

We developed the Mutation–Stability–Activity (MSA) model to investigate the biophysical constraints that shape residue-dependent structural divergence profiles in enzyme evolution. MSA shows that these profiles result from three constraints operating in different proportions across enzyme families: mutation, stability, and activity constraints.

Decomposition reveals how these constraints operate and what determines their variation among families. The mutational component is set entirely by protein physics: a protein's structural response to external perturbations (mutations) mirrors its intrinsic thermal fluctuations (flexibility). Thus mutational contributions are largely determined by protein architecture. The stability and activity components likewise arise from physics, with $\Delta\Delta\mathrm{G}$ and $\Delta\Delta\mathrm{G}^\ddagger$ determining the residue-level effects, but their magnitude is scaled by the selection parameters $\mathrm{a_S}$ and $\mathrm{a_A}$. Variation among enzyme families in the contributions of these components could in principle reflect differences in the distributions of $\Delta\Delta\mathrm{G}$ or $\Delta\Delta\mathrm{G}^\ddagger$ across proteins; however, our analysis shows that the dominant driver is variation in $\mathrm{a_S}$ and $\mathrm{a_A}$. In this way, physics is necessary to generate the components, but family-to-family differences in their contributions are primarily explained by differences in selection parameters, which reflect each enzyme's biological context.

Having established this, MSA now makes it possible to investigate why $\mathrm{a_S}$ and $\mathrm{a_A}$ vary among enzymes. Their values should reflect context-level properties such as expression levels, position in metabolic networks, and specificity requirements, providing a basis for future work linking structural evolutionary patterns to enzyme biology.

5 Data availability

All processed data and code needed to reproduce the figures and tables in this paper are available at `https://github.com/jechave/msa_profiles_shared`.

Acknowledgments

This work was supported by CONICET (grant PIP-11220210100462) and regular ISYEB funding.

References

[1] Chothia, C., and A. Lesk, 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* 5:823–826.

[2] Hubbard, T., and T. Blundell, 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Engineering, Design and Selection* 1:159–171.

[3] Wood, T. C., and W. R. Pearson, 1999. Evolution of protein sequences and structures. *Journal of molecular biology* 291:977–995.

[4] Illergård, K., D. H. Ardell, and A. Elofsson, 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* 77:499–508.

[5] Eventoff, W., M. G. Rossmann, and C.-I. Brändén, 1975. The Evolution of Dehydrogenases and Kinase. *CRC Critical Reviews in Biochemistry* 3:111–140.

[6] Johnson, M. S., M. J. Sutcliffe, and T. L. Blundell, 1990. Molecular anatomy: Phyletic relationships derived from three-dimensional structures of proteins. *Journal of Molecular Evolution* 30:43–59.

[7] Bujnicki, J. M., 2000. Phylogeny of the Restriction Endonuclease-Like Superfamily Inferred from Comparison of Protein Structures. *Journal of Molecular Evolution* 50:39–44.

[8] Challis, C. J., and S. C. Schmidler, 2012. A stochastic evolutionary model for protein structure alignment and phylogeny. *Molecular Biology and Evolution* 29:3575–3587.

[9] Lundin, D., A. M. Poole, B.-M. Sjöberg, and M. Högbom, 2012. Use of Structural Phylogenetic Networks for Classification of the Ferritin-like Superfamily. *Journal of Biological Chemistry* 287:20565–20575.

[10] Golden, M., E. Garcia-Portugués, M. Sørensen, K. V. Mardia, T. Hamelryck, and J. Hein, 2017. A generative angular model of protein structure evolution. *Molecular Biology and Evolution* 34:2085–2100.

[11] Perron, U., A. M. Kozlov, A. Stamatakis, N. Goldman, and I. H. Moal, 2019. Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States. *Molecular Biology and Evolution* 36:2086–2103.

[12] Larson, G., J. L. Thorne, and S. Schmidler, 2020. Incorporating Nearest-Neighbor Site Dependence into Protein Evolution Models. *Journal of Computational Biology* 27:361–375.

[13] Malik, A. J., A. M. Poole, and J. R. Allison, 2020. Structural Phylogenetics with Confidence. *Molecular Biology and Evolution* 37:2711–2726.

[14] Moi, D., S. Nishio, X. Li, C. Valansi, M. Langleib, N. G. Brukman, K. Flyak, C. Dessimoz, D. De Sanctis, K. Tunyasuvunakool, J. Jumper, M. Graña, H. Romero, P. S. Aguilar, L. Jovine, and B. Podbilewicz, 2022. Discovery of archaeal fusexins homologous to eukaryotic HAP2/GCS1 gamete fusion proteins. *Nature Communications* 13:3880.

[15] Moi, D., C. Bernard, M. Steinegger, Y. Nevers, M. Langleib, and C. Dessimoz, 2023. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. *bioRxiv* .

[16] Puente-Lelievre, C., A. J. Malik, J. Douglas, D. Ascher, M. Baker, J. Allison, A. Poole, D. Lundin, M. Fullmer, R. Bouckert, H. Kim, M. Steinegger, and N. Matzke, 2023. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. *bioRxiv* .

[17] Garg, S. G., and G. K. A. Hochberg, 2025. A General Substitution Matrix for Structural Phylogenetics. *Molecular Biology and Evolution* 42:msaf124.

[18] Baek, M., F. DiMaio, et al., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373:871–876.

[19] Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

[20] Mirdita, M., K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, 2022. ColabFold: making protein folding accessible to all. *Nature Methods* 19:679–682.

[21] Varadi, M., S. Anyango, et al., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50:D439–D444.

[22] Abramson, J., J. Adler, et al., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630:493–500.

[23] Echave, J., 2008. Evolutionary divergence of protein structure: The linearly forced elastic network model. *Chemical Physics Letters* 457:413–416.

[24] Echave, J., and F. M. Fernández, 2010. A perturbative view of protein structural variation. *Proteins: Structure, Function and Bioinformatics* 78:173–180.

[25] Nelson, E. D., and N. V. Grishin, 2016. Evolution of off-lattice model proteins under ligand binding constraints. *Physical Review E* 94:022410.

[26] Marcos, M. L., and J. Echave, 2020. The variation among sites of protein structure divergence is shaped by mutation and scaled by selection. *Current Research in Structural Biology* 2:156–163.

[27] Echave, J., S. J. Spielman, and C. O. Wilke, 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics* 17:109–121.

[28] Echave, J., and M. Carpentier, 2025. On the variation of structural divergence among residues in enzyme evolution. *bioRxiv* .

[29] Echave, J., 2019. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Molecular Biology and Evolution* 36:613–620.

[30] Echave, J., 2021. Evolutionary coupling range varies widely among enzymes depending on selection pressure. *Biophysical Journal* 120:4320–4324.

[31] McCandlish, D. M., and A. Stoltzfus, 2014. Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology* 89:225–252.

[32] Hinsen, K., 1998. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics* 33:417–429.

[33] Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal* 80:505–515.

[34] Yang, L., G. Song, and R. L. Jernigan, 2009. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences* 106:12347–12352.

[35] Ming, D., and M. E. Wall, 2005. Allostery in a coarse-grained model of protein dynamics. *Physical Review Letters* 95:198103.

[36] Edgar, R. C., 2022. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* 13:6968.

[37] Grant, B. J., L. Skjærven, and X. Yao, 2021. The Bio3D packages for structural bioinformatics. *Protein Science* 30:20–30.

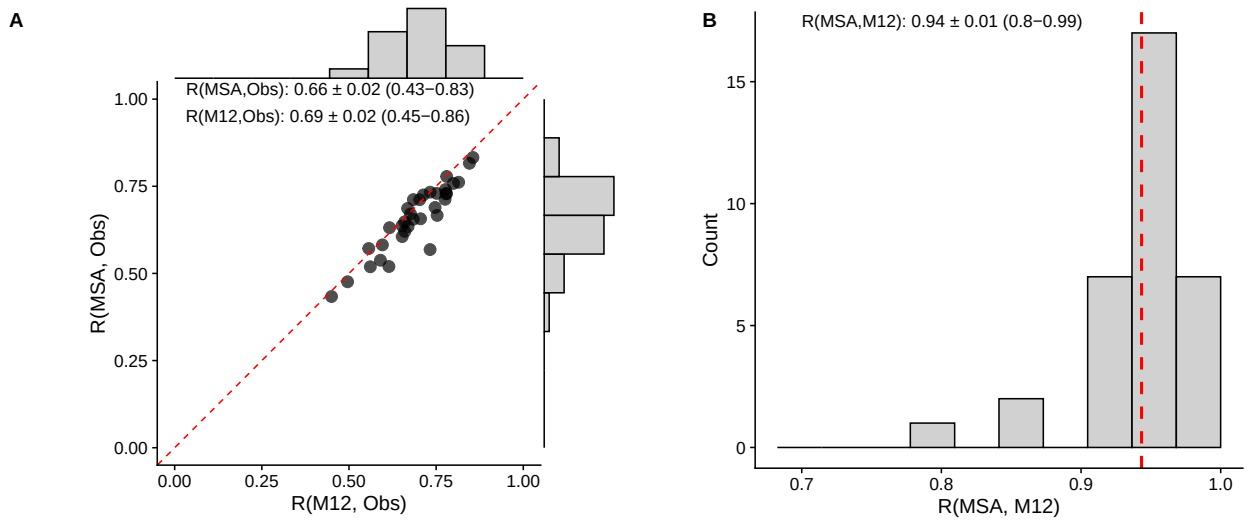

Figure 1: **Comparison of MSA and empirical model performance.** (A) Scatter plot comparing correlations between model predictions and observed structural divergence profiles, measured as centred log–RMSD (nlRMSD), across 34 enzyme families. The x-axis shows correlations between the empirical model (M12) and observed profiles, the y-axis shows correlations between MSA predictions and observed profiles. The dashed red line indicates perfect agreement. Marginal histograms show the distribution of prediction–observation correlations for each model. Annotations show mean $\pm$ SEM and range (min–max) for each model's correlations with observations. (B) Histogram showing the distribution of correlations between MSA and M12 predictions across all enzyme families. The red dashed line indicates the mean correlation. Annotation shows mean $\pm$ SEM and range (min–max).

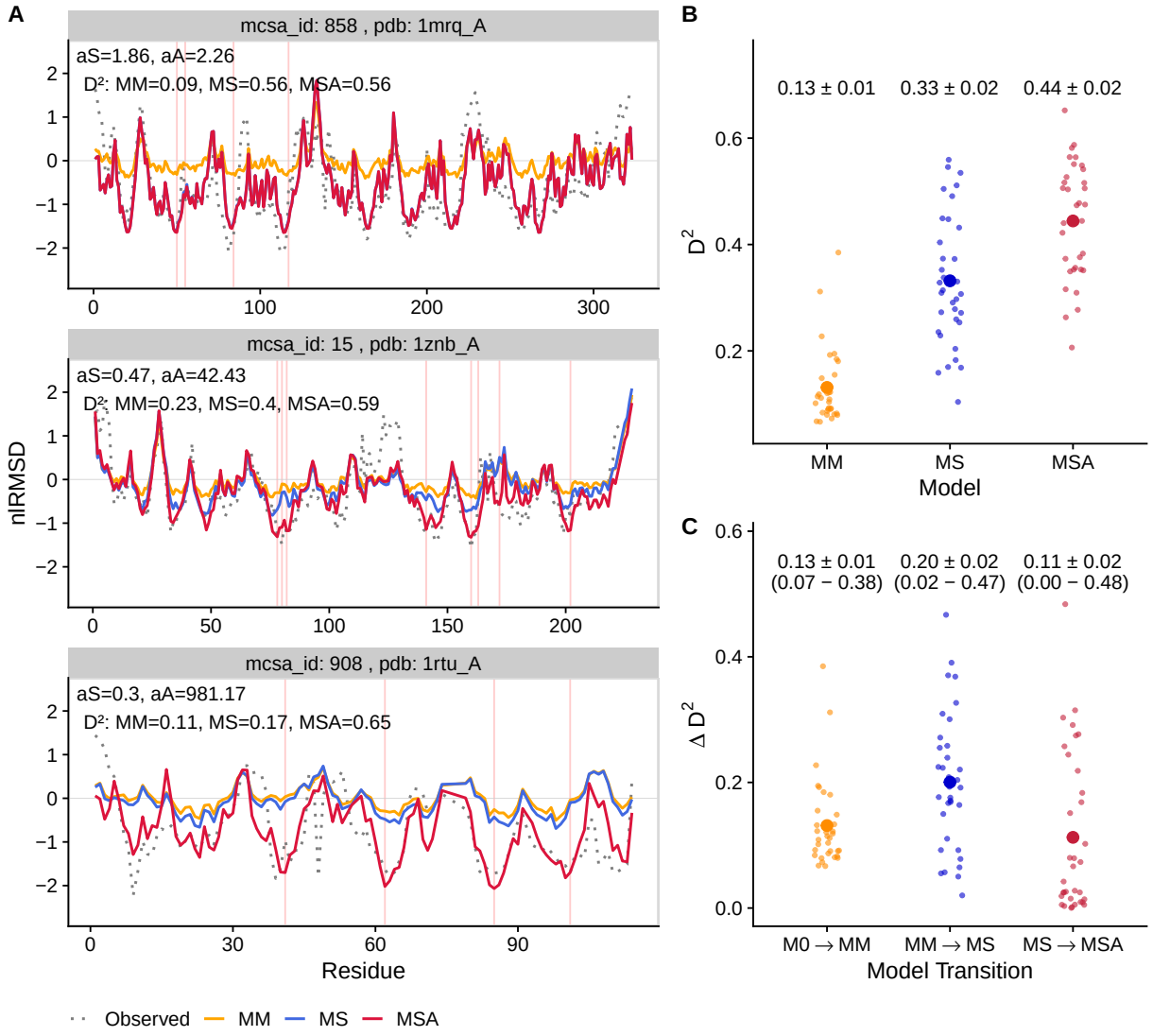

Figure 2: **Nested model comparison of structural divergence profiles across enzyme families.** (A) Residue-dependent structural divergence profiles for three enzyme families. Each panel shows observed centred log–RMSD (nlRMSD, gray dashed line) and predictions from M0 (non-residue-dependent uniform divergence, gray line), MM (mutation-only model, orange), MS (mutation–stability model, blue), and MSA (mutation–stability–activity model, red). Vertical red lines mark active site residues. Annotations show posterior mean estimates of $\mathrm{a_S}$ and $\mathrm{a_A}$ from MSA, together with the observed deviance explained by each model D^2. (B) Distribution of deviance explained by models MM, MS, and MSA across all enzyme families. (The deviance explained by M0 is, by definition, zero, and is not shown.) Small points represent individual enzyme families; large points indicate means over families. Means $\pm$ standard errors are shown above each distribution. (C) Distribution of incremental explained deviance from sequentially adding mutation (M0 $\rightarrow$ MM), stability (MM $\rightarrow$ MS), and activity (MS $\rightarrow$ MSA) constraints. Small points represent individual enzyme families; large points indicate means over families. Means $\pm$ standard errors and ranges (min–max) are shown above each distribution.

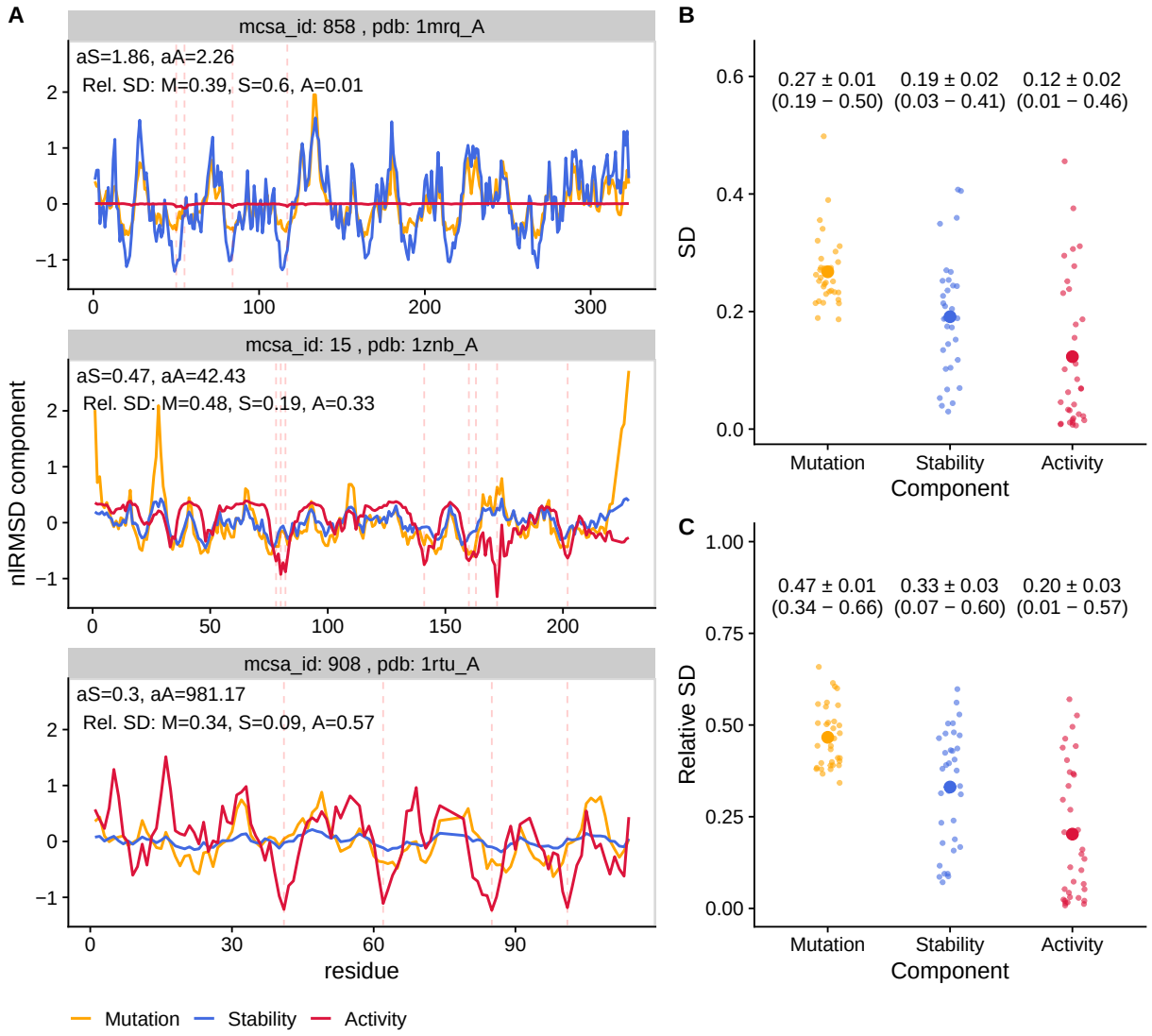

Figure 3: **Decomposition of divergence profiles into mutation, stability, and activity components.** MSA profiles are decomposed into a sum of three components ($\mathrm{nlRMSD} = \phi_{\mathrm{M}} + \phi_{\mathrm{S}} + \phi_{\mathrm{A}}$). The contribution of a given component to an enzyme's profile is quantified by the standard deviation of its contribution profile: $\mathrm{SD}(\phi_{\mathrm{M}})$ for mutation, $\mathrm{SD}(\phi_{\mathrm{S}})$ for stability, and $\mathrm{SD}(\phi_{\mathrm{A}})$ for activity. The relative contribution is defined as the component's SD divided by the sum of the SDs of all components. (A) Component profiles for three enzyme families showing mutation (orange), stability (blue), and activity (red) contributions. Vertical dashed red lines mark active site residues. Annotations show posterior mean estimates of $\mathrm{a_S}$ and $\mathrm{a_A}$, together with the relative SD of each component. (B) Distributions of component contributions (SD) across the dataset. (C) Distributions of relative component contributions (relative SD) across the dataset. In Panels B and C, small points represent individual families; large points indicate means over families. Means $\pm$ SEM and ranges (min–max) are shown above the distributions.

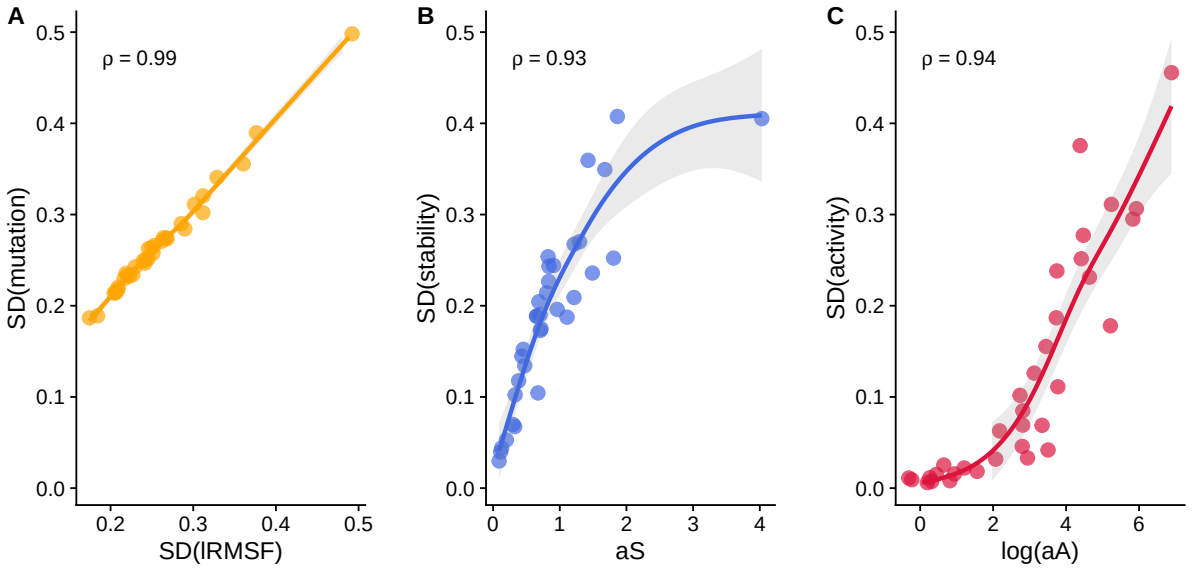

Figure 4: **Relationship between mutation, stability, and activity contributions and underlying enzyme properties.** (A) Contribution of the mutation component, $\mathrm{SD}(\phi_{\mathrm{M}})$, versus heterogeneity of residue flexibility, measured by the standard deviation of the logarithm of the root mean square fluctuation, $\mathrm{SD}(\mathrm{lRMSF})$. (B) Contribution of the stability component, $\mathrm{SD}(\phi_{\mathrm{S}})$, versus Bayesian estimate (posterior mean) of the stability selection parameter $\mathrm{a_S}$. (C) Contribution of the activity component, $\mathrm{SD}(\phi_{\mathrm{A}})$, versus Bayesian estimate (posterior mean) of the activity selection parameter $\log(\mathrm{a_A})$. Each point represents one enzyme family. Loess fits (lines) with confidence intervals (gray shading) and Spearman correlation coefficients are shown.