ON THE VARIATION OF STRUCTURAL DIVERGENCE AMONG RESIDUES IN ENZYME EVOLUTION

Julian Echave
Instituto de Ciencias Físicas (ICIFI-CONICET),
Universidad Nacional de San Martín
Martín de Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina
jechave@unsam.edu.ar

Mathilde Carpentier
Institut de Systématique, Evolution, Biodiversité (ISYEB, UMR 7205)
CNRS-MNHN-SU-EPHE-UA, Sorbonne Université
57 rue Cuvier, 75005 Paris, France
mathilde.carpentier@mnhn.fr

ABSTRACT

Structural divergence varies among protein residues. Unlike the classic problem of substitution rate variation, this structural divergence variation has been largely ignored. Here we show that in enzymes structural divergence increases with both residue flexibility and distance from the active site. Although these factors are correlated, we demonstrate through modelling that the pattern arises from two independent types of constraints, non-functional and functional. Their relative importance varies across enzyme families: as functional constraints increase from 4% to 85%, non-functional constraints decrease from 96% to 15%, reshaping the divergence pattern. This analysis overturns two accepted views in protein evolution: First, evolutionary variation has been thought to mirror protein dynamics generally, but we show this similarity exists only when non-functional constraints dominate. Second, active site structural conservation has been attributed to functional constraints alone, but we show it stems largely from their location in rigid regions where non-functional constraints are high.

1 Introduction

Machine learning methods have improved the speed and accuracy of protein structure prediction, making structural data more accessible [1, 2, 3, 4, 5]. Protein structures evolve more slowly than sequences, a long-recognized property that enables comparisons between distantly related proteins whose sequences have diverged beyond recognition [6, 7, 8, 9]. This property makes structural data particularly valuable for structural phylogenetics, providing an alternative for studying evolutionary relationships when sequences can no longer provide meaningful comparisons [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. However, the use of structural data in structural phylogenetics and other evolutionary studies is limited by an incomplete understanding of protein structural evolution.

Most studies of structural evolution focus on whole-protein structural divergence, without examining how structural divergence varies within proteins[6, 7, 8, 9]. This is a major limitation of our current understanding. The variation of amino acid substitution rates among residues is one of the classic problems of protein sequence evolution [23], which led to the concept of evolutionary constraints as the causes of the observed variation in substitution rates. In contrast, the analogous problem in structure evolution, the variation of structural divergence among residues, has received little attention. Here, we examine how evolutionary constraints determine the variation of structural divergence among residues within enzymes.

To anticipate key issues that we will develop in this work, we introduce class A beta lactamases as an illustrative case of how structural divergence varies among residues and how this variation reflects the influence of evolutionary constraints. Their residue-level structural divergence (Fig. 1a) shows a conserved core surrounded by variable regions. This pattern

correlates with both conformational flexibility (Fig. 1b), a potential indicator of non-functional constraints, and distance from the active site (Fig. 1c), which may reflect functional constraints. The correlation between flexibility and distance complicates disentangling their respective contributions. Below, we examine these issues in detail to understand their roles in shaping structural divergence.

Flexibility serves as a proxy for the influence of non-functional constraints. Evolutionary variation among homologous structures resembles the conformational variation of individual proteins during dynamics [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. For example, residue structural divergence (evolution) correlates with residue flexibility (dynamics), with flexible residues diverging more rapidly (Fig. 1) [33]. Biophysical models attribute this relationship to the effects of non-functional constraints[26, 29, 33]. Specifically, flexible residues are more structurally divergent because their conformations are more sensitive to mutations and because these residues more readily accept substitutions under selection for protein stability. The correlation between evolution and dynamics, explained by non-functional constraints, validates the use of flexibility as their indicator.

In enzymes, distance from the active site serves as a proxy for the influence of functional constraints. Enzyme active sites are structurally conserved among homologous enzymes, suggesting the presence of functional constraints [35, 36]. While studies on the structural divergence of non-active site residues remain limited, sequence-based analyses indicate that functional constraints extend beyond the active site, with their impact diminishing gradually with distance [37, 38]. The maximal conservation at active sites and the gradual decay of functional constraints with distance validate the use of distance as their indicator.

The correlation between residue flexibility and distance from the active site prevents a straightforward attribution of structural divergence to functional and non-functional constraints. Active sites are typically located in low-flexibility regions, leading to an increase in flexibility with increasing distance. Consequently, when we observe how structural divergence depends on flexibility or distance, we cannot determine whether the effects arise from functional or non-functional constraints, since our measurements are correlated. To disentangle the independent contributions of these constraints, both flexibility and distance must be analysed simultaneously.

Here we analyse how structural divergence varies among residues within enzymes. To determine how functional and non-functional constraints influence structural divergence, we examine how it relates to residue flexibility and distance from the active site, accounting for their interdependence.

2 Results

To investigate the constraints influencing structural divergence, we first curated a dataset of functionally conserved monomeric single-domain enzyme families from the Mechanism and Catalytic Site Atlas (M-CSA) database [39, 40], where each family includes a reference enzyme and its homologues. This approach yielded 34 enzyme families, representing a diverse range of structures and functions. (Detailed dataset properties are provided in Section 4.1).

Next, we calculated three measures at each residue position, as illustrated in Figure 1: (1) Root Mean Square Deviation (`RMSD`), using the C_α coordinates of aligned and superimposed homologous enzymes, quantifying residue-level structural divergence; (2) Root Mean Square Fluctuation (`RMSF`), using a simple C_α-based coarse- grained Elastic Network Model of the reference enzyme–`RMSF` quantifies the flexibility of individual residues–; (3) Distance from the active site (d), defined as the distance between a residue's C_α atom and the C_α atom of the nearest active-site residue.

Given the wide variation in structural divergence between conserved and variable regions, we primarily use logarithmic transformations (`lRMSD`, `lRMSF`) for their better statistical properties for modelling. We also employ normalized versions (`nlRMSD`, `nlRMSF`) that express residue-level divergence and flexibility relative to the protein average (see Section 4 for detailed definitions). We specify whether we use the raw, logarithmic, or normalized-logarithmic versions of `RMSD` and `RMSF` only when necessary. In contexts where the precise metric does not affect the interpretation, such as general discussions of trends (e.g., 'divergence increases with flexibility'), we may refer to these variables without explicitly specifying the measure used.

Finally, we analysed how structural divergence (`RMSD`) relates to flexibility (`RMSF`) and distance to active sites (d).

2.1 Variation of structural divergence among residues

Structural divergence varies widely among residues. Specifically, `RMSD` differs markedly between conserved and variable residues. In class A beta-lactamases, the Q90/Q10 `RMSD` ratio, calculated using the 90th and 10th percentiles, is 6.6 (Figure 2a); the evolutionary deformation of variable regions is almost 7 times faster than that of conserved regions. This wide variation is consistent across all enzyme families: Q90/Q10 `RMSD` ratios range from 3.6 to 14.0 (mean = 7.7 ± 2.5; Figure 2d; and Supplementary Figures S3a to S36a).

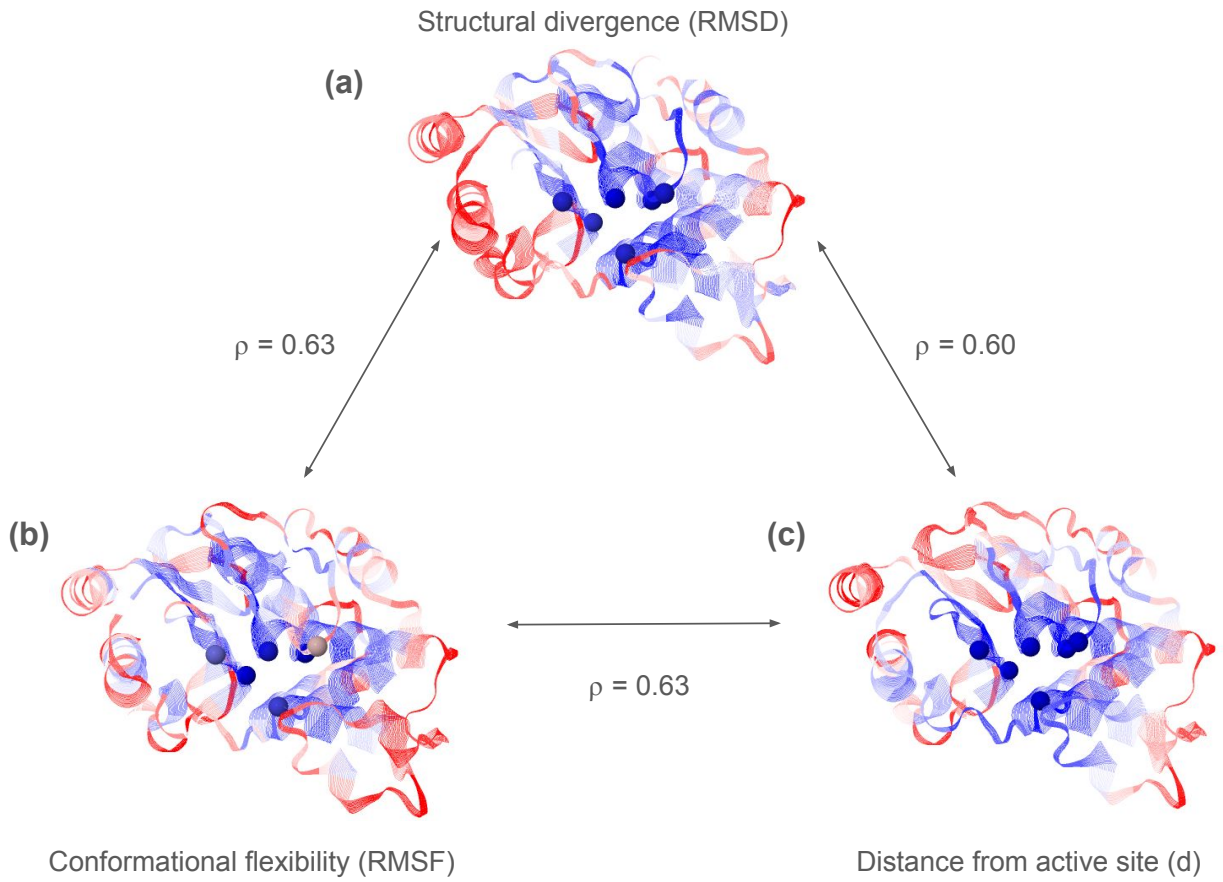

Figure 1: **Patterns of residue-level structural divergence, flexibility, and distance to the active site in class A beta-lactamases (M-CSA ID = 2).** All measurements use C_α coordinates. Values are shown as ranks mapped onto the 3D structure of the reference enzyme (PDB code: 1btl), coloured from blue (low) to red (high). **a**: Structural divergence at each residue position, quantified by the root mean square deviation (`RMSD`) of residue coordinates across homologous structures. **b**: Local flexibility of each residue, measured by root mean square fluctuation (`RMSF`) in the reference enzyme during protein dynamics. **c**: Each residue's distance in the reference enzyme to the closest active-site residue (d). Active-site C_α atoms are shown as spheres. Spearman correlation coefficients between these patterns are displayed on the edges of the triangle.

To investigate the determinants of this variation, we analysed the dependence of `RMSD` on two properties: residue flexibility (`RMSF`), linked to non-functional constraints, and distance to the active site (d), associated with functional constraints.

Structural divergence increases with flexibility. As shown in Figure 2b, for class A beta-lactamases, residue `RMSD` increases with residue `RMSF`, with a Spearman's correlation coefficient $\rho = 0.63$ ($p < 0.001$). This trend is observed in all enzyme families, with $\rho(\texttt{RMSD}, \texttt{RMSF})$ ranging from 0.36 to 0.85 and averaging 0.63 ± 0.11 (Figure 2e; Figures S3b to S36b).

Structural divergence increases with distance to the active site in most enzyme families. In class A beta-lactamases, residue `RMSD` increases with d ($\rho = 0.60$, $p < 0.001$; Figure 2c). This trend is observed in most enzyme families, with $\rho(\texttt{RMSD}, d)$ averaging 0.47 ± 0.20. Of the 34 families studied, 32 show positive correlations ranging from 0.16 to 0.78 (Figure 2f; Figures S3c to S36c). Only two families deviate from this pattern: M-CSA ID 749 shows a slight negative correlation ($\rho = -0.12$, $p = 0.03$; Fig. S30c) and M-CSA ID 258 shows no correlation ($\rho = 0$; Fig. S16c). (These cases will be discussed in Section 3.5.)

If flexibility (`RMSF`) and distance (d) were independent variables, the observed divergence-flexibility dependence could be attributed to non-functional constraints, and the observed divergence-distance dependence to functional constraints.

We could then conclude that both constraints are influencing the observed patterns of variation of structural divergence within enzymes.

However, flexibility and distance are not independent. There is a positive correlation between RMSF and d across enzyme families, with an average $\rho = 0.5 \pm 0.2$. Across all families except one, correlations are positive, ranging from 0.16 to 0.85 (Figure S2a). These correlations arise as a consequence of active site location: when active sites are located in regions of low flexibility, as is typical, a correlation between flexibility and distance emerges. This mechanism is reflected in the correlation ($R = 0.84$) between active site average flexibility and $\rho(\texttt{RMSF}, d)$ (Figure S2b). In M-CSA ID 362, where the active site is located in a flexible region, there is no correlation between flexibility and distance ($\rho = -0.04$, p = 0.7).

Because of this interdependence, the observed divergence-flexibility and divergence-distance trends cannot be directly interpreted as evidence of purely non-functional or functional constraints, as each trend may reflect contributions from both. The observed divergence-flexibility dependence could reflect a direct effect of non-functional constraints, or an indirect effect, via the positive flexibility-distance correlation, of functional constraints, or a combination of both. Likewise, the observed divergence-distance trend could reflect a direct influence of functional constraints, an indirect effect of non-functional constraints via their correlation with flexibility, or a combination. Further analysis is needed to disentangle the independent contributions of functional and non-functional constraints to structural divergence.

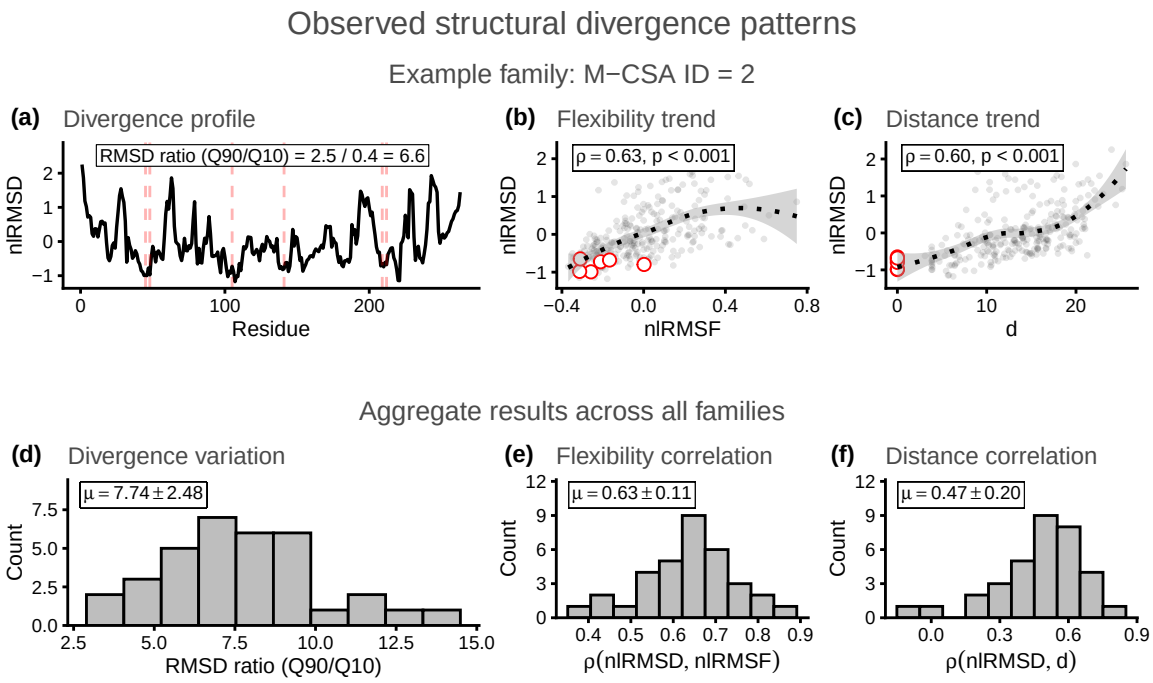

Figure 2: **Observed structural divergence patterns. a**: Residue-specific structural divergence (nlRMSD) vs. residue number for class A beta-lactamases (M-CSA ID = 2), with ratio of 90th to 10th percentile RMSD values shown. Active-site residues are indicated by vertical dashed red lines. **b**: nlRMSD vs. residue flexibility (nlRMSF) for the same family. **c**: nlRMSD vs. distance to the closest active-site residue (d) for the same family. In **b** and **c**, grey points represent individual residues, active-site residues are marked in red, and the black dotted line represents a LOESS fit to the data points; ρ is Spearman's correlation coefficient, with corresponding p-values provided. **d**: Distribution of Q90/Q10 RMSD ratios across all enzyme families. **e,f**: Distributions of Spearman's correlation coefficients across all families between nlRMSD and: **e**) nlRMSF; **f**) d. μ indicates the mean $\pm$ standard deviation for each distribution.

2.2 Independent contributions of non-functional and functional constraints

To evaluate whether non-functional and functional constraints independently contribute to residue-specific structural divergence, we compared three non-linear models: M_1, M_2, and M_{12}. These models, formulated as shape-constrained

additive models (SCAMs), represent lRMSD as a function of different predictors: M_1 depends on lRMSF, M_2 on d, and M_{12} on both. The models use monotone-increasing functions to reflect expected direct effects: structural divergence is expected to increase with flexibility due to weakening non-functional constraints and with distance due to weakening functional constraints. While most subsequent analyses focus on normalized structural divergence ($\texttt{nlRMSD} = \texttt{lRMSD} - \langle \texttt{lRMSD} \rangle$), we fit the models to unnormalized lRMSD values to preserve information about mean structural divergence across families. Details of the SCAM formulations and modelling approach are provided in the Methods section.

We fit these models to residue lRMSD values for each enzyme family separately. Model selection relied on the Akaike Information Criterion (AIC), which balances model complexity and goodness of fit. To evaluate model performance, we used explained deviance, a measure analogous to explained variance in linear models but well-suited for non-linear relationships.

The findings demonstrate that functional and non-functional constraints make independent contributions to the variation in structural divergence among residues. In class A beta-lactamases (Figure 3a), M_{12} is the best model according to AIC, with values of 447.7, 454.9, and 410.7 for M_1, M_2, and M_{12}, respectively. Explained deviance values—0.36 for M_1, 0.35 for M_2, and 0.46 for M_{12}—further confirm that M_{12} outperforms the single-variable models, indicating that both nlRMSF and d contribute independently to nlRMSD. This pattern holds across most enzyme families, with AIC identifying M_{12} as the best model in 27 of 34 cases (Figure 3d; Figures S3d to S36d). On average, M_{12} explains more deviance (0.49 ± 0.13) than M_1 (0.42 ± 0.14) or M_2 (0.28 ± 0.13) (Figure 3e), demonstrating that M_{12} consistently fits better across enzyme families. These results confirm that both nlRMSF and d independently contribute to structural divergence.

After evaluating the models' fit to the observed nlRMSD values, we next assessed their ability to reproduce the $\texttt{nlRMSD} \sim \texttt{nlRMSF}$ and $\texttt{nlRMSD} \sim d$ trends. To quantify accuracy, we used two root mean square error (RMSE) metrics: RMSE(flex), which evaluates how well a model captures the divergence-flexibility trend, and RMSE(dist), which assesses its fit to the divergence-distance trend.

In class A beta-lactamases, M_1 cannot explain the observed divergence-distance trend, and M_2 cannot explain the observed divergence-flexibility trend. Figures 3b and 3c show the performance of each model for this family. M_1, which models nlRMSD as a function of nlRMSF, accurately reproduces the divergence-flexibility trend (Figure 3b, RMSE(flex) = 0.04) but predicts an overly flat divergence-distance trend (Figure 3c, RMSE(dist) = 0.20). This misfit leads to overestimation of structural divergence near the active site and underestimation at greater distances. Conversely, M_2, which depends only on d, accurately reproduces the divergence-distance trend (Figure 3c, RMSE(dist) = 0.04) but predicts a flat divergence-flexibility trend (Figure 3b, RMSE(flex) = 0.18), overestimating divergence in rigid regions and underestimating it in flexible regions.

In contrast, M_{12}, which incorporates both nlRMSF and d, accurately reproduces both trends, with RMSE(flex) = 0.04 and RMSE(dist) = 0.04 (Figures 3b and 3c). This performance demonstrates that M_{12} effectively captures both divergence-flexibility and divergence-distance trends, outperforming the single-variable models.

The superior performance of M_{12} extends beyond beta-lactamases and is consistent across the entire dataset of enzyme families. As shown in Figures 3f and 3g, and in panels e and f of Figures S3 to S36, M_{12} reliably captures both divergence-flexibility and divergence-distance trends across families, unlike the single-variable models M_1 and M_2.

The single-variable models, M_1 and M_2, each perform well for one trend but fail for the other. M_1 accurately reproduces the divergence-flexibility trend, with a mean RMSE(flex) of 0.05 ± 0.02, but fails to capture the divergence-distance trend, yielding a higher mean RMSE(dist) of 0.18 ± 0.09. Conversely, M_2 accurately reproduces the divergence-distance trend, with a mean RMSE(dist) of 0.07 ± 0.05, but fails to capture the divergence-flexibility trend, resulting in a mean RMSE(flex) of 0.29 ± 0.12.

The double-variable model M_{12} accurately describes the dependence of structural divergence on both flexibility and distance, offering a more comprehensive explanation than either single-variable model. This is supported by its mean RMSE(flex) of 0.05 ± 0.02 and mean RMSE(dist) of 0.08 ± 0.06, demonstrating that M_{12} captures both trends effectively.

Together, these results show that the two-variable model M_{12} effectively captures residue-specific structural divergence. Its higher explained deviance and superior fit to both the divergence-flexibility and divergence-distance trends, compared to single-variable models, indicate that lRMSF and d contribute independently to structural divergence. Under the assumption that lRMSF and d serve as proxies for non-functional and functional constraints, respectively, the findings imply that non-functional and functional constraints provide independent contributions to the variation of structural divergence among residues.

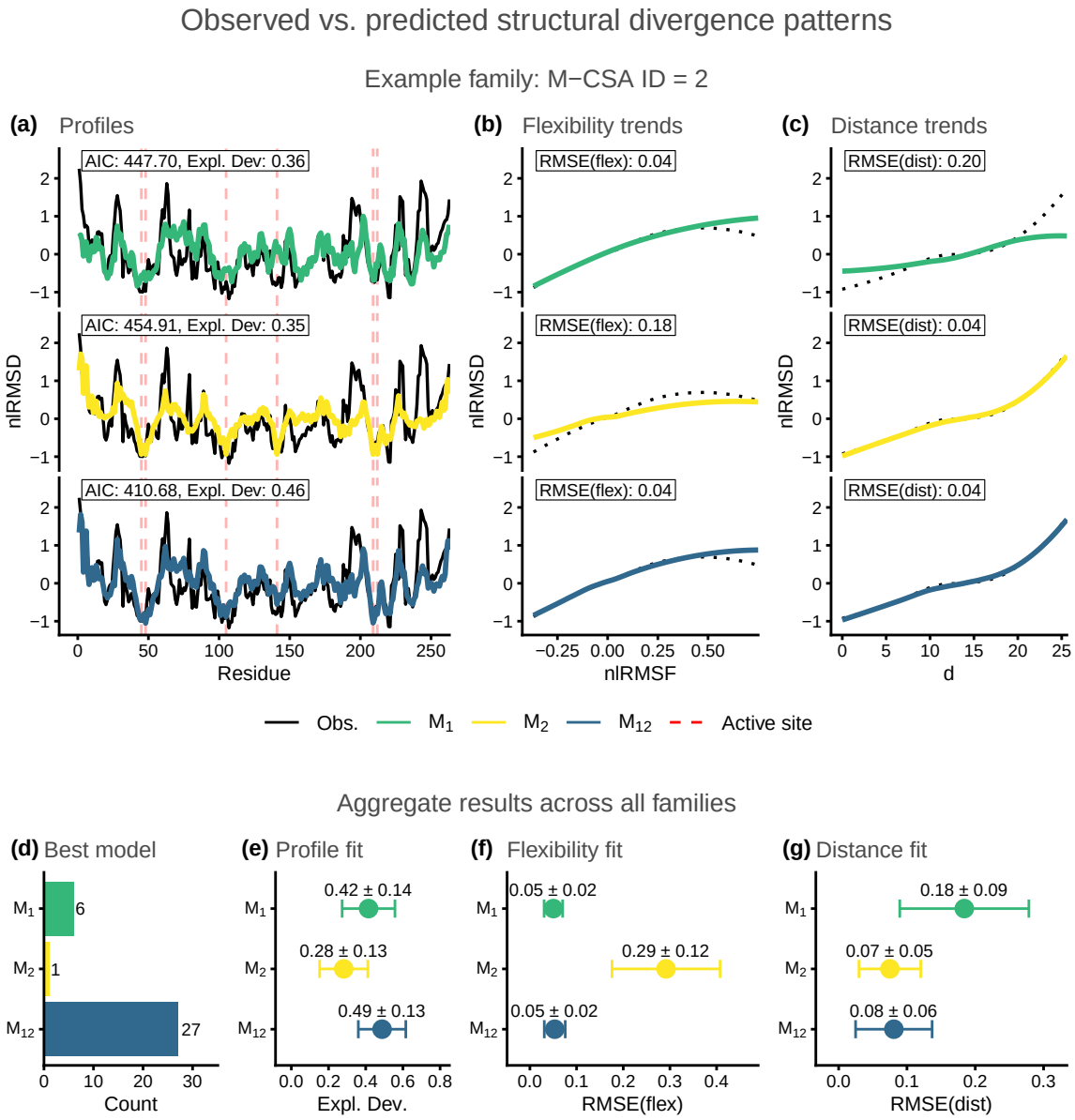

Figure 3: **Observed vs. predicted structural divergence patterns. a**: Residue-specific `nlRMSD` profiles for class A beta-lactamases (M-CSA ID = 2), showing observed values (black) and model predictions: M_1 (green), M_2 (yellow), and M_{12} (blue). Active-site residues are indicated by vertical dashed red lines. AIC and explained deviance values for each model are shown. **b**: Divergence-flexibility trends for class A beta-lactamases. **c**: Divergence-distance trends for class A beta-lactamases. In **b** and **c**, black dotted lines show LOESS fits to observed data, coloured lines show LOESS fits to model predictions, and RMSE values quantify the differences between these trends. **d**: Number of enzyme families (bar chart) for which each model is optimal (lowest AIC). **e**: Explained deviance for each model across enzyme families. **f**: RMSE(flex) across enzyme families, measuring fit to divergence-flexibility trends. **g**: RMSE(dist) across enzyme families, measuring fit to divergence-distance trends. In **e**–**g**, points and error bars show mean ± standard deviation.

2.3 Relative contributions of non-functional and functional constraints

Having established that both functional and non-functional constraints contribute independently to structural divergence patterns, we next analyzed their individual contributions through two complementary approaches. Model M12 first provides a functional decomposition, expressing structural divergence as the sum of components s_1 (flexibility) and s_2 (distance). Figure 4a illustrates this decomposition for class A beta-lactamases: the residue-dependent profile (blue line) splits into flexibility (green) and distance (yellow) components. This decomposition extends to the observed trends (Figures 4b-c): while flexibility most directly influences the divergence-flexibility relationship and distance the divergence-distance relationship, each trend reflects both components due to the correlation between these variables.

To complement this functional decomposition, we quantified the overall contribution of each variable using the Shapley method. This approach distributes the explained deviance among predictors by averaging their marginal contributions across all possible predictor subsets, ensuring fair attribution even when predictors are correlated (see Methods). For class A beta-lactamases, with a total explained deviance of 0.46, flexibility and distance contribute Shapley values of 0.23 and 0.22, respectively, representing nearly equal relative contributions of 51% and 49%.

The relative importance of flexibility and distance varies widely among enzyme families. While class A beta-lactamases show nearly equal Shapley contributions, this balance is not typical across the dataset. The explained deviance of model M12 averages 0.49 ± 0.12 across families, with flexibility contributing 0.31 ± 0.12 and distance 0.18 ± 0.10. Thus, on average, flexibility accounts for 64% of the explained variance and distance for 36%. Excluding MCSA ID 749, which shows no functional contribution and is discussed as a special case in Section 3.5, the distance contribution increases from 4% to 85% across families (with flexibility contribution accordingly decreasing from 96% to 15%).

To understand why the contributions of flexibility and distance to structural divergence vary among enzymes, we examined several properties of each family's reference enzyme: its size, the distribution of residue flexibilities (mean, standard deviation, range), the distribution of distances to the active site (mean, standard deviation, range), and the correlation between flexibility and distance. We also considered two family-level characteristics, the number of homologous proteins and the overall structural divergence. Using stepwise regression to identify the most relevant among these predictors, we found that only two properties account for most of the variation: $\texttt{sd}(\texttt{lRMSF})$, which measures how much flexibility varies among residues within a protein, and $\langle d \rangle$, the mean distance to the active site.

The flexibility contribution increases with $\texttt{sd}(\texttt{lRMSF})$ ($R^2 = 0.44$, p < 0.001, Figure 4e). When flexibility varies more widely among residues within a protein, it becomes a stronger determinant of structural divergence patterns. The distance contribution decreases with $\langle d \rangle$ ($R^2 = 0.36$, p < 0.001, Figure 4f). In larger or more elongated proteins, where many residues lie far from the active site, distance-based constraints play a diminished role in shaping structural divergence patterns. Both predictors reflect aspects of protein architecture: $\texttt{sd}(\texttt{lRMSF})$ depends on the pattern of residue-residue contacts, while $\langle d \rangle$ reflects protein size, shape, and active site location. Together, they explain 38% of the variation in relative contributions between families, as determined by multiple regression. The remaining unexplained variation suggests that additional factors beyond these architectural properties, possibly including varying degrees of selection pressure, also influence the balance between functional and non-functional constraints in each enzyme family.

2.4 Model performance across some example cases

To complement our general analysis of structural divergence patterns, we examine five specific cases that span the range of observed behaviors. Using the distribution of enzyme families based on their flexibility and distance contributions (Figure 5), we selected cases representing the full range of relative distance contributions to structural divergence: the aldo/keto reductase family (MCSA ID 858; Figure S32) shows minimal contribution (RSC(d) = 0.13); a metallo-beta-lactamase (MCSA ID 15, RSC(d) = 0.37; Figure S4), a class-A beta- lactamase (MCSA ID 2, RSC(d) = 0.49; Figure S3), and a TrpC family member (MCSA ID 252, RSC(d) = 0.63; Figure S14) show intermediate values; and the ribonuclease U2 family (MCSA ID 908; Figure S34) shows the strongest distance contribution (RSC(d) = 0.85).

Table 1 shows several trends across the five cases. As the relative contribution of distance increases, protein size decreases (from 322 to 107 residues) and mean distance to active site shortens (from 15.3Å to 9.8Å). These changes align with the shift from flexibility to distance contributions. The spread of local flexibility, $sd(\texttt{lRMSF})$, which we showed drives flexibility contributions, shows no consistent trend in these cases. This pattern illustrates our earlier finding that the spread of flexibility and mean distance from the active site explain only 38% of variation in relative contributions between constraints, with the remaining variation potentially arising from other factors such as the variation among families of selection pressure.

The explained deviance of each model reflects which type of constraint dominates (Table 2, columns Expl. Dev.). When the distance contribution is low (MCSA ID 858), M_1 performs well and M_2 poorly, with M_{12} matching M_1's

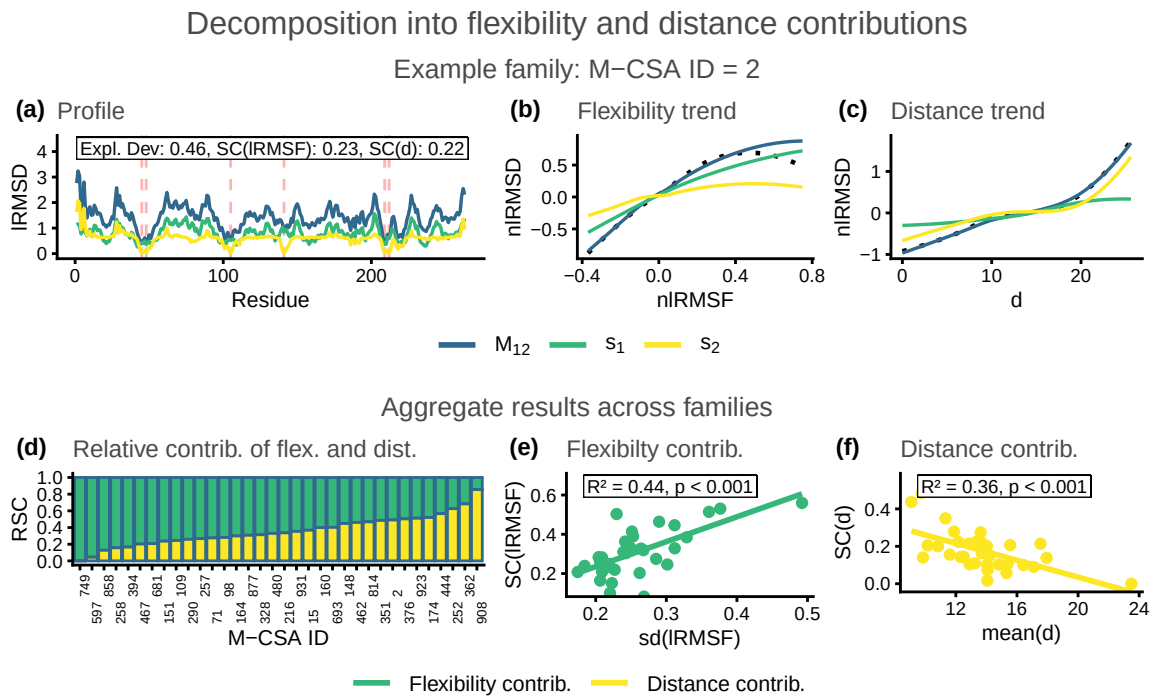

Figure 4: **Decomposition of structural divergence into flexibility and distance contributions.** **a**: Residue-dependent `lRMSD` profile for class A beta-lactamases (M-CSA ID = 2), showing M_{12} predictions (blue, constant term removed) decomposed into flexibility component s_1 (green) and distance component s_2 (yellow). Active-site residues are indicated by vertical dashed red lines. The model's explained deviance and Shapley contributions (quantifying each variable's contribution to structural divergence) are shown. **b**: Flexibility trend (`nlRMSD` vs. `nlRMSF`) for the same family. Black dotted line shows LOESS fit to observations, coloured lines show LOESS fits to M_{12} predictions (blue) and its components s_1 (green) and s_2 (yellow). **c**: Distance trend (`nlRMSD` vs. d) for the same family. Lines as in **b**. **d**: Stacked bars showing the relative Shapley contributions of flexibility ($RSC(\texttt{lRMSF})$) and distance ($RSC(d)$) across enzyme families, sorted by increasing s_2 (these relative contributions add up to 1). **e**: Relationship between the flexibility contribution ($SC(\texttt{lRMSF})$) and the standard deviation of residue flexibilities within each enzyme. **f**: Relationship between the distance contribution ($SC(d)$) and the mean residue distance to active site within each enzyme. Linear fits with corresponding R^2 and p-values are shown in **e** and **f**.

performance. Through the intermediate cases (MCSA 15, 2, and 252), neither single-variable model suffices, and M_{12} outperforms both. When the distance contribution is high (MCSA ID 908), M_2 performs well and M_1 poorly, with M_{12} matching M_2's performance. (Although M_{12} explains slightly more deviance than M_2 in this case, M_2 is selected as the best model by AIC due to its lower complexity.)

Each model's ability to reproduce the divergence-flexibility and divergence-distance trends reveals their limitations (Table 1, columns RMSE(flex) and RMSE(dist)). M_1 accurately captures the flexibility trend but, except when the distance contribution is minimal, fails to account for the distance trend. Similarly, M_2 captures the distance trend but, except when the distance contribution is maximal, fails to account for the flexibility trend. Only M_{12} reproduces both trends across all cases, demonstrating that both constraints are needed to explain the observed patterns of variation.

The cases considered here illustrate how the balance of constraints determines model performance. The accuracy of single-variable models depends on which type of constraint dominates, while M_{12} succeeds across all cases by accounting for both flexibility and distance contributions. This confirms our general finding that both non-functional and functional constraints influence structural divergence patterns.

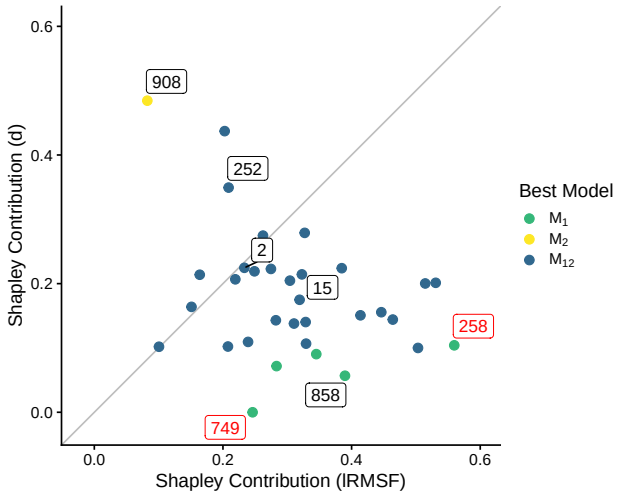

Figure 5: **Map of enzyme families based on their Shapley contributions.** Each point represents an enzyme family positioned according to its Shapley contributions (explained deviance) from flexibility ($SC(\texttt{lRMSF})$, x-axis) and distance ($SC(d)$, y-axis). The diagonal line indicates equal contributions from both variables. Colours indicate the best model (M_1, M_2, or M_{12}) according to AIC, with M_1 dominating at high flexibility contributions, M_2 at high distance contributions, and M_{12} at intermediate values. Five representative cases (MCSA IDs 858, 15, 2, 252, and 908, in black) and two outlier cases (258 and 749, in red) are labeled.

Table 1: Flexibility and distance contributions of example families

MCS-A	PDB	Length	sd(lRMSF)	$< d >$	RSC(lRMSF)	RSC(d)	SC(lRMSF)	SC(d)
858	1mrq	322	0.25	15.3	0.87	0.13	0.39	0.06
15	1znb	225	0.33	12.2	0.63	0.37	0.38	0.22
2	1btl	263	0.22	13.3	0.51	0.49	0.23	0.22
252	1igs	245	0.21	11.3	0.37	0.63	0.21	0.35
908	1rtu	107	0.27	9.8	0.15	0.85	0.08	0.48

MCS-A: case ID; PDB: reference PDB ID; Length: number of residues; sd(lRMSF): spread of lRMSF for reference; $< d >$: average distance to active site; RSC(lRMSF) and RSC(d): Relative Shapley Contributions of lRMSF and d, respectively; SC: corresponding Shapley Contributions.

Table 2: Variation of model performance across example families

				Expl. Dev.			RMSE(flex)			RMSE(dist)		
M-CSA	PDB	RSC(d)	Best	M_1	M_2	M_{12}	M_1	M_2	M_{12}	M_1	M_2	M_{12}
858	1mrq	0.13	M_1	0.45	0.11	0.45	0.05	0.41	0.05	0.19	0.08	0.19
15	1znb	0.37	M_{12}	0.51	0.35	0.61	0.08	0.32	0.06	0.21	0.06	0.04
2	1btl	0.49	M_{12}	0.36	0.35	0.46	0.04	0.18	0.04	0.2	0.04	0.04
252	1igs	0.63	M_{12}	0.37	0.51	0.56	0.04	0.14	0.04	0.26	0.04	0.04
908	1rtu	0.85	M_2	0.16	0.56	0.57	0.09	0.08	0.05	0.48	0.07	0.07

RSC(d): Relative Shapley Contribution of distance to structural divergence; Best: model with lowest AIC; Expl. Dev.: deviance explained by each model; RMSE(flex): Root Mean Square Error between observed and predicted divergence-flexibility trends; RMSE(dist): RMSE between observed and predicted divergence-distance trends.

2.5 Active site structural divergence

Having shown that both types of constraints influence structural divergence, we examined whether they similarly affect the structural conservation of active sites. The common assumption is that active sites are conserved solely due to functional constraints. However, their location in rigid regions suggests an alternative explanation: limited flexibility, i.e., non-functional constraints, could also account for their conservation.

To test this, we compared the structural divergence of active site residues against three baselines: the protein's average divergence (M_0) and the predictions of models M_1 and M_{12}.

Active sites were more conserved than the protein average: using M_0 as a baseline, active-site residues showed significantly lower divergence (-0.62 ± 0.04, Fig.6).

Non-functional constraints account for much of this conservation. Using M_1 predictions, which account for flexibility, the residual shifted closer to zero (-0.21 ± 0.04), indicating that rigidity explains a significant portion of the observed conservation. This result challenges the assumption that functional constraints alone are responsible.

While flexibility explains much of active site conservation, adding distance to the active site (d) as a predictor in M_{12} brought the residuals even closer to zero (0.07 ± 0.03, $p = 0.03$, Fig.6), consistent with an additional role for functional constraints.

Thus, active sites are more conserved than average, but their conservation cannot be attributed solely to functional constraints. These results show that much of their conservation is due to their location in rigid regions, indicating the need to reconsider the assumption that active site conservation arises exclusively from functional constraints.

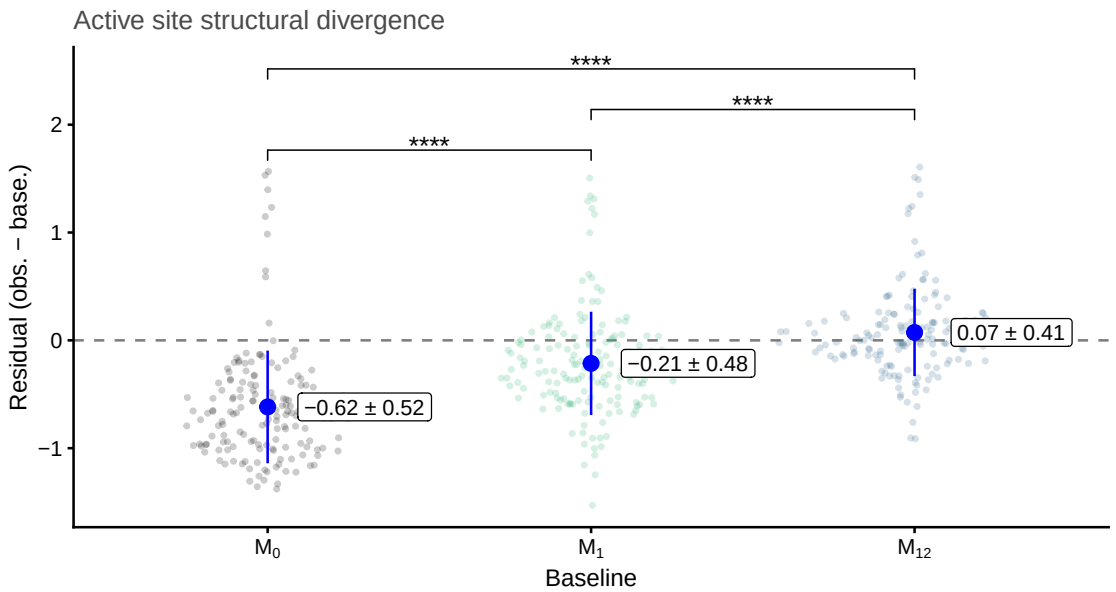

Figure 6: **Structural divergence of active site residues**. Distribution of residuals showing how much more conserved active sites are compared to three baselines: the protein average (M_0), and predictions by models M_1 and M_{12}. Individual points show residuals for each residue, with mean ± SD shown in blue. Paired t-tests comparing means between models (matching same residue across models) showed significant differences (**** $p < 0.0001$).

3 Discussion

3.1 Patterns of structural divergence and their constraints

Our analysis revealed two trends in how structural divergence varies among residues: divergence increases with both flexibility and distance from active sites (Section 2.1). While these trends suggest the action of non-functional and functional constraints respectively, interpreting them requires careful analysis because they are interrelated through the correlation between flexibility and distance.

Previous models have shown how mutational effects and stability requirements can create the divergence-flexibility trend without invoking functional selection [33]. The divergence-distance trend, in turn, suggests functional constraints that extend beyond active sites with diminishing strength [37, 38]. However, because active sites are typically located in rigid regions, flexibility and distance are correlated, making it impossible to determine from these trends alone the independent contributions of each type of constraint.

Model M12 (Section 2.2) addresses this problem by separating structural divergence into independent components: s1(flexibility) and s2(distance). These terms isolate the independent contributions of non-functional and functional constraints. The model shows that each observed trend actually combines direct and indirect effects (Section **??**): the divergence-flexibility trend includes both the direct effect of non-functional constraints (s1) and, through the flexibility-distance correlation, an indirect contribution from functional constraints (s2). Similarly, the divergence-distance trend combines the direct effect of functional constraints (s2) with indirect effects of non-functional ones (s1). The model's success in reproducing both observed trends while outperforming single-variable models (Section 2.2) demonstrates that both types of constraints shape structural evolution independently. The relative strength of these constraints determines how structural divergence depends on flexibility and distance.

3.2 Variation in constraint balance across enzymes

The balance between non-functional and functional constraints varies substantially across enzyme families. The contribution of functional constraints ranges from 4% to 85% of explained structural divergence, with non-functional constraints correspondingly decreasing from 96% to 15% (Figure 4d). These shifts in constraint balance reshape how structural divergence depends on flexibility and distance.

What determines which constraints dominate in a particular enzyme family? Our analysis reveals that protein architecture influences constraint balance through two properties: the spread of flexibility among residues and the mean distance to active sites (Section 2.3). When flexibility varies widely within a protein structure, it becomes a stronger predictor of structural divergence patterns (Figure 4e). Conversely, in larger or more elongated proteins where residues tend to lie far from active sites, distance-based constraints become less influential (Figure 4f). Together, these architectural properties explain 38% of the differences in constraint balance among enzyme families.

The modest explanatory power of architecture indicates that other factors significantly influence how structural divergence is shaped. This is further supported by our observation that enzymes with similar architectural properties can show different patterns of structural divergence (Section 2.4). Previous sequence-based studies suggest that differences in functional selection pressure among enzymes could explain part of this remaining variation [38]. Testing this hypothesis will require mechanistic biophysical models that can explicitly account for both protein architecture and varying selection regimes.

3.3 Relationship between evolutionary and dynamical ensembles

Previous work has stressed the similarity between sets of structures related by homology with sets of conformations reached by the same protein during its dynamics [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34] (see Introduction). If non-functional constraints were the only ones operating, then we would expect evolutionary and dynamical ensembles to be similar [26, 29, 33]. However, the present results show that structural divergence patterns depend not only on flexibility but also on distance from the active site. As functional constraints increase, their contribution to structural divergence becomes stronger, making evolutionary patterns increasingly different from those predicted by flexibility alone. The similarity between evolutionary and dynamical ensembles thus exists only to the extent that non-functional constraints dominate over functional ones. This finding revises our understanding of protein evolution by showing that the correspondence between evolutionary variation and protein dynamics is not universal but depends on the balance between these constraints.

3.4 Structural conservation of active sites

Active sites are more structurally conserved between homologous enzymes than between random proteins [35], and more conserved than the protein average [36]. This structural conservation is routinely interpreted as evidence of functional constraints, without questioning this assumption. However, our analysis in Section 2.5 challenges this interpretation. Before concluding that functional constraints drive structural conservation of active sites, we must first account for non-functional constraints. Given that active sites are typically located in rigid regions, could their structural conservation simply reflect high non-functional constraints? Our results show that indeed, much of their enhanced conservation is explained by their lack of flexibility. This finding revises the common assumption about active site structural evolution being subject mainly to functional constraints. Nevertheless, after accounting for non-functional

constraints, we find that active sites are still more structurally conserved than expected. This additional conservation, beyond what rigidity predicts, provides the first clear evidence that functional constraints do contribute to active site structural conservation. The active site analysis thus reinforces our general finding that both non-functional and functional constraints shape structural divergence at every residue position in proteins, including active-site residues, while also demonstrating how their relative contributions can be rigorously separated.

3.5 Exceptions and special cases

Two enzyme families deviate from the general relationship between structural divergence and distance to active sites described in Section 2.1. MCSA ID 258, an L1 metallo-beta-lactamase from Stenotrophomonas maltophilia (according to M-CSA and PDB annotations), shows no correlation between divergence and distance to the active site. As a hospital-derived pathogen conferring broad-spectrum antibiotic resistance, it may have undergone recent adaptive evolution under selective pressures that differ from those leading to active site conservation in other enzymes.

MCSA ID 749, a cytochrome c peroxidase, shows a negative correlation between divergence and distance, violating our assumption that functional constraints decrease with distance from active sites. Examination of the structures in the PDB reveals this apparent divergence actually reflects conformational changes between calcium-bound and calcium-free states rather than evolutionary differences. This explains why this family shows no contribution from distance-based constraints (Section 2.3): what we measure is not evolutionary divergence but conformational variation.

While we could have detected and filtered out these cases by examining active site structural differences, doing so would have introduced circularity in our analysis of active site conservation. Instead, their detection as statistical outliers serves to validate our analytical approach by showing it can identify cases that violate its assumptions.

3.6 Applications to structural phylogenetics

Our findings have immediate applications for structural phylogenetics, a field of growing importance given the increased availability of protein structures through methods like AlphaFold. Understanding how structural divergence varies among residues suggests several possible improvements to existing approaches. For example, distance-based methods [18, 19, 20] could incorporate residue-specific weights based on flexibility and distance to active sites, while probabilistic models [13, 15, 16, 17, 21, 22] could account for varying structural divergence rates that depend on these factors.

3.7 Physical basis of protein structural evolution

Our empirical analysis revealed two types of constraints that independently shape structural divergence. These patterns suggest fundamental mechanisms in protein evolution. Non-functional constraints point to effects arising from basic protein physics: how different residues respond to mutations and how stability requirements influence where mutations are tolerated. These mechanisms would operate similarly across all proteins. Functional constraints, in contrast, point to effects arising from specific molecular interactions required for protein function, which would vary among proteins according to their particular roles.

Some biophysical models address aspects of these potential mechanisms: how mutational biases and stability selection shape structural divergence [26, 29, 33], how functional constraints operate through ligand binding [41], while extensions of stability-based models that incorporate selection on catalytic activity have been applied to sequence evolution patterns [42, 38]. However, a complete mechanistic understanding of our findings will require models that integrate all these elements: how mutations affect protein structure and how selection on both stability and function governs their fixation. Such models would provide a mechanistic understanding of the constraints that we identify empirically.

4 Materials and Methods

4.1 Dataset

Initial filtering

The dataset was derived from the M-CSA (Mechanism and Catalytic Site Atlas) database, accessed on August 21, 2023. M-CSA is a manually curated database of catalytic residues and enzyme mechanisms, accessible at `www.ebi.ac.uk/thornton-srv/m-csa/`[39, 40]. It contains 1002 entries and extends to over 131,807 structures through homologue searches in the Protein Data Bank. Homologues are identified using PHMMER with an E-value threshold of 10^{-6} [36]. Each M-CSA entry includes a reference enzyme structure with manually annotated catalytic residues.

We focused on the simplest case of monomeric single-domain enzymes, as multi-domain and multi-chain proteins introduce additional complexity through domain and chain interactions that require separate analysis.

Starting from 1002 entries, we applied sequential filtering steps (remaining entries after each step shown in parentheses):

- Removal of inconsistencies and duplicates (921)
- Filtering for entries with intra-chain active sites (795)
- Filtering for X-ray determined structures (786)
- Filtering for entries with intra-domain active sites (568) (Note that this step became redundant after filtering for single-domain proteins, as they cannot have inter-domain active sites.)
- Filtering for single CATH ID reference proteins (313)
- Filtering for monomeric biological assemblies (120)
- Filtering for entries with homologues with the same CATH ID as the reference (106)
- Within each remaining entry: for homologues only, when multiple PDB structures exist for the same UniProt ID, selecting the optimal structure based on sequence coverage and X-ray resolution

This filtering process resulted in a dataset of 106 enzyme families. Each entry comprises a manually annotated reference PDB chain and its homologues, totalling 1171 homologue structures. All reference enzymes are monomeric and single-domain.

Alignment, Superposition, and Further Filtering

For each entry, the reference and homologous chains were aligned using Muscle V (via 'pdb.aln' in the 'bio3d' R package) [43, 44]. Sequence alignment was chosen over structural alignment because (1) within families, sequences are similar enough for reliable alignments, and (2) it provides an independent basis for studying structural divergence.

Following alignment, structures were superimposed to minimize the Root Mean Square Deviation (`RMSD`). The dataset then underwent additional filtering:

- Removal of homologues with gaps or mutations at the active site
- Clustering-based removal of structural outliers and redundant structures. We removed clusters with RMSD > 6Å from the reference structure (to exclude likely alternative conformational states) and merged clusters with RMSD < 1Å (to avoid redundancy from near-identical structures)
- Clustering-based filtering of sequence identity, removing clusters with identity > 95% or < 25% relative to the reference structure
- Removal of families with fewer than 4 homologues

This process ensures the retained enzymes for a given M-CSA entry are non-redundant and likely belong to the same functional family, characterized by sequence and structural similarity and active sites with identical amino acids.

Dataset Properties

The final dataset contains 34 enzyme families: 27 alpha/beta, 5 all-beta, and 2 all-alpha structures. CATH architectures include Alpha-Beta Barrel, 3-Layer(aba) Sandwich, Alpha-Beta Complex, 4-Layer Sandwich, Distorted Sandwich, 2-Layer Sandwich, and Orthogonal Bundle. The EC distribution shows hydrolases (20 families) as most frequent, followed by transferases, lyases, isomerases, and oxidoreductases. Each family contains 4-23 proteins (mean 6.82, SD 3.98). Reference proteins range from 98 to 447 residues (mean 248.50, SD 91.31), and Id% ranges from 0.31 to 0.56 (mean 0.41, SD 0.06). `RMSD` ranges from 0.72Å to 2.88Å (mean 1.38, SD 0.45). Active sites comprise 1-8 residues (mean 4.74, SD 1.76), with active site `RMSD` from 1.22Å to 6.68Å (mean 2.26, SD 0.96). The highest active site `RMSD` occurs in MCSA ID 749, where the reference structure lacks a calcium ion present in homologues. See Figure S1 for distributions.

4.2 Residue-specific structural divergence: `RMSD`

Let $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N$ represent the structures of N homologous enzymes. Each $\mathbf{R}_p$ is a vector of Cartesian coordinates of the alpha carbons (C_α) in the native structure of enzyme p. These structures form a sample from the set of structures

that evolved from the proteins' common ancestor. We quantify the structural divergence of a residue i using its Root Mean Square Deviation, RMSD:

$$\texttt{RMSD}^i = \sqrt{\frac{1}{N}\sum_p ||\Delta \mathbf{R}_p^i||^2} \tag{1}$$

Here, $\Delta \mathbf{R}_p^i$ is the deviation of residue i's position vector in protein p from its average position across all proteins.

The direct use of RMSD presents two challenges. First, its wide variation range can obscure smaller variations in conserved regions. Second, the error in RMSD estimates is proportional to RMSD itself. To address these issues, we use the natural logarithm of RMSD, $\texttt{lRMSD} = \ln(\texttt{RMSD})$. This transformation converts multiplicative error into additive error (independent of measured value) and gives appropriate weight to small RMSD values in conserved regions.

We then define a normalized measure:

$$\texttt{nlRMSD}^i = \ln(\texttt{RMSD}^i) - \langle \ln(\texttt{RMSD}) \rangle \tag{2}$$

where $\langle \ln(\texttt{RMSD}) \rangle$ is the average logarithmic RMSD across all residues in the protein. This normalized logarithmic Root Mean Square Deviation, nlRMSD, provides a relative measure of structural divergence. Negative values indicate residues that are structurally less diverged than average, while positive values indicate more diverged residues.

Implementation uses MUSCLE v5 [43] for sequence alignment, followed by structural superposition using bio3d (version 2.4.4) [44]. Each structure is aligned to a reference by minimizing overall RMSD. We then calculate residue-specific values: RMSD (Eq.1, excluding gaps), its natural logarithm lRMSD, and the normalized version nlRMSD (Eq.2). We use lRMSD when analyzing absolute divergence and nlRMSD when comparing divergence patterns across proteins of different sizes or overall divergence levels.

4.3 Residue-specific Flexibility: RMSF

We correlate evolutionary divergence with dynamical fluctuations, building on prior research [33]. We designate one homologous enzyme as the reference. This enzyme fluctuates, producing an equilibrium ensemble of conformations described by the distribution function $\rho_{\texttt{dyn}}(\mathbf{r})$, where $\mathbf{r}$ represents any given conformation. The native structure is the average conformation: $\mathbf{R} = \langle \mathbf{r} \rangle$.

We quantify the flexibility of residue i using the Root Mean Square Fluctuation (RMSF):

$$\texttt{RMSF}^i = \sqrt{\langle ||\mathbf{r}^i - \mathbf{R}^i||^2 \rangle} = \sqrt{\int ||\mathbf{r}^i - \mathbf{R}^i||^2 \rho_{\texttt{dyn}}(\mathbf{r}) d\mathbf{r}} \tag{3}$$

Here, $\mathbf{r}^i$ is the position vector of residue i in conformation $\mathbf{r}$, and $\mathbf{R}^i$ is its position in the native conformation.

The equilibrium distribution function, $\rho_{\texttt{dyn}}(\mathbf{r})$, follows from basic statistical mechanics:

$$\rho_{\texttt{dyn}}(\mathbf{r}) = e^{-V(\mathbf{r})/k_B T} \tag{4}$$

where $V(\mathbf{r})$ is the potential energy, k_B the Boltzmann constant, and T the absolute temperature.

We obtain the potential energy using the Elastic Network Model (ENM) developed by Ming and Wall [45]. This coarse-grained model represents the protein as a network of nodes (residues) connected by springs (interactions). The potential energy landscape of the elastic network is:

$$V(\mathbf{r}) = \frac{1}{2}\sum_{i<j} k_{ij}(r_{ij} - R_{ij})^2 \tag{5}$$

where r_{ij} is the distance between residues i and j in conformation $\mathbf{r}$, and R_{ij} is the distance in the native conformation. The spring constant k_{ij} is defined as:

$$k_{ij} = \begin{cases} 0 & \text{if } R_{ij} > R_{ij}^{\texttt{cut}} \\ k & \text{if } R_{ij} \leq R_{ij}^{\texttt{cut}} \text{ and } |i-j| > 1 \\ a \cdot k & \text{if } |i-j| = 1 \end{cases} \tag{6}$$

where R_{ij} is the distance between residues i and j in the native conformation, $R_{ij}^{\texttt{cut}}$ is the cut-off distance, and a is a scaling factor for consecutive residues.

We compute RMSF by replacing Eq.5 into Eq.4, and applying the result in Eq.3. We used parameter values $R_{ij}^{\texttt{cut}} = 10.5$Å, $k = 4.5$, and $a = 42$, that were shown to yield accurate RMSF predictions [45].

As with structural divergence, we use both the natural logarithm of RMSF (lRMSF) and a normalized version:

$$\texttt{nlRMSF}^i = \ln(\texttt{RMSF}^i) - \langle \ln(\texttt{RMSF}) \rangle \tag{7}$$

Like nlRMSD, nlRMSF provides a relative measure, with negative and positive values indicating residues less or more flexible than average, respectively. We use lRMSF when analysing absolute flexibility and nlRMSF when comparing flexibility patterns across proteins.

4.4 Distance from the active site: d

We define d for each residue as the minimum distance between its C_α and those of active-site residues.

4.5 Divergence-flexibility and divergence-distance trends

For each enzyme family, we analyse how lRMSD varies with lRMSF and d. We quantify these relationships through smooth fits (which we refer to as "trends") using local polynomial regression (the loess function from R's stats package):

```
trend_lrmsf = loess(lrmsd ~ lrmsf)
trend_dactive = loess(lrmsd ~ dactive)
```
(8)

Here, lrmsd, lrmsf, and dactive correspond to their mathematical counterparts lRMSD, lRMSF, and d, respectively.

4.6 SCAM models of structural divergence

For each enzyme family, we fit three shape-constrained additive models using the scam function from R package scam:

M_{12}: uses both predictors (lRMSF and d):

```
m12 <- scam(lrmsd ~ s1(lrmsf, bs = "mpi") + s2(dactive, bs = "mpi"))
```
(9)

M_1: uses only flexibility (lRMSF):

```
m1 <- scam(lrmsd ~ s1(lrmsf, bs = "mpi"))
```
(10)

M_2: uses only distance (d):

```
m2 <- scam(lrmsd ~ s2(dactive, bs = "mpi"))
```
(11)

The models use monotone-increasing P-splines (bs = "mpi"), reflecting two assumptions: structural divergence increases with flexibility in the absence of functional constraints, and functional constraints decrease with distance from the active site.

We assess model performance using AIC (penalized likelihood criterion), explained deviance (a generalization of R^2 for non-linear models), and RMSE between predicted and observed trends. For a given model, RMSE(flex) is the Root Mean Square Error between observed and predicted divergence-flexibility trends, defined through LOESS fits (Eq.8). Similarly, RMSE(dist) is the RMSE between observed and predicted divergence-distance trends.

4.7 Shapley decomposition

To quantify the contributions of protein flexibility and distance to the active site to structural divergence, we applied Shapley's method to the three empirical models (M_1, M_2, and M_{12}). This method allocates the explained deviance between lRMSF and d, accounting for both their individual and combined effects.

For each enzyme family, we calculated the Shapley Contributions (SC) of lRMSF and d using the following formulas:

$$\begin{aligned} SC(\texttt{lRMSF}) &= 0.5D_1 + 0.5(D_{12} - D_2) \\ SC(d) &= 0.5D_2 + 0.5(D_{12} - D_1) \end{aligned} \tag{12}$$

where D_1, D_2, and D_{12} represent the deviance explained by M_1 (lRMSF only), M_2 (d only), and M_{12} (both variables).

Finally, we calculated the Relative Shapley Contributions (RSC):

$$\begin{aligned} RSC(\texttt{lRMSF}) &= \frac{SC(\texttt{lRMSF})}{D_{12}} \\ RSC(d) &= \frac{SC(d)}{D_{12}} \end{aligned} \tag{13}$$

These contributions sum to 1 and represent the relative importance of lRMSF and d in explaining structural divergence.

To identify the determinants of Shapley contributions SC and RSC, we used bidirectional stepwise linear regression, using the step function from base R.

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/enzyme_evo_str_analysis`. Additional figures and analyses are provided in Supplementary Material.

Acknowledgments

We thank Ioannis Riziotis for sharing the data of Riziotis et al. [36]. This work was supported by CONICET (grant PIP-11220210100462) and regular ISYEB funding.

References

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

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

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

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] 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* .

[21] 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* .

[22] Garg, S. G., and G. Hochberg, 2024. A general substitution matrix for structural phylogenetics. *bioRxiv* .

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

[24] Elber, R., and M. Karplus, 1987. Multiple Conformational States of Proteins: A Molecular Dynamics Analysis of Myoglobin. *Science* 235:318–321.

[25] Leo-Macias, A., P. Lopez-Romero, D. Lupyan, D. Zerbino, and A. R. Ortiz, 2005. An Analysis of Core Deformations in Protein Superfamilies. *Biophysical Journal* 88:1291–1299.

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

[27] Friedland, G. D., N.-A. Lakomek, C. Griesinger, J. Meiler, and T. Kortemme, 2009. A Correspondence Between Solution-State Dynamics of an Individual Protein and the Sequence and Conformational Diversity of its Family. *PLOS Computational Biology* 5:e1000393.

[28] Velázquez-Muriel, J. A., M. Rueda, I. Cuesta, A. Pascual-Montano, M. Orozco, and J. M. Carazo, 2009. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Structural Biology* 9:1–14.

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

[30] Mahajan, S., A. G. De Brevern, B. Offmann, and N. Srinivasan, 2014. Correlation between local structural dynamics of proteins inferred from NMR ensembles and evolutionary dynamics of homologues of known structure. *Journal of Biomolecular Structure and Dynamics* 32:751–758.

[31] Derényi, I., and G. J. Szöllősi, 2015. Effective temperature of mutations. *Physical Review Letters* 114:058101.

[32] Vetrivel, I., A. G. De Brevern, F. Cadet, N. Srinivasan, and B. Offmann, 2019. Structural variations within proteins can be as large as variations observed across their homologues. *Biochimie* 167:162–170.

[33] 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.

[34] Tang, Q. Y., and K. Kaneko, 2021. Dynamics-Evolution Correspondence in Protein Structures. *Physical Review Letters* 127:098103.

[35] Torrance, J. W., G. J. Bartlett, C. T. Porter, and J. M. Thornton, 2005. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *Journal of Molecular Biology* 347:565–581.

[36] Riziotis, I. G., A. J. Ribeiro, N. Borkakoti, and J. M. Thornton, 2022. Conformational Variation in Enzyme Catalysis: A Structural Study on Catalytic Residues. *Journal of Molecular Biology* 434:167517.

[37] Jack, B. R., A. G. Meyer, J. Echave, and C. O. Wilke, 2016. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLOS Biology* 14:e1002452.

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

[39] Ribeiro, A. J. M., G. L. Holliday, N. Furnham, J. D. Tyzack, K. Ferris, and J. M. Thornton, 2018. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research* 46:D618–D623.

[40] Ribeiro, A. J., J. D. Tyzack, N. Borkakoti, G. L. Holliday, and J. M. Thornton, 2020. A global analysis of function and conservation of catalytic residues in enzymes. *Journal of Biological Chemistry* 295:314–324.

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

[42] 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.

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

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

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