Electronic Supplementary Material (ESI) for Integrative Biology.

Supporting Information for the paper: The maximum penalty criterion for ridge regression: application to the calibration of the force constant in elastic network models

Yves Dehouck[(1)] and Ugo Bastolla[(2)]

[(2)] Machine Learning Group, Université Libre de Bruxelles (ULB),

1050 Brussels, Belgium. E-mail: ydehouck@ulb.ac.be

[(1)] Centro de Biologia Molecular "Severo Ochoa", CSIC-UAM Cantoblanco,

28049 Madrid, Spain. E-mail: ubastolla@cbm.csic.es

Supporting Text

Supporting Figures

Supporting Text

Ridge regression

In ridge regression, if $\mathbf{y}$ is the N-dimensional vector of dependent variables to be fitted, $\mathbf{X}$ the $N \times P$ matrix of independent variables, and $\mathbf{a}$ the P-dimensional vector of the fit parameters, the fit can be written in matrix notation as $\mathbf{y} \approx \mathbf{y}^{(\mathrm{fit})} = \mathbf{X}\mathbf{a}$. The objective function to be minimised is:

$$G^{(\mathrm{ns})}(\mathbf{a}, \mathbf{X}, \mathbf{y}, \Lambda) = E(\mathbf{a}, \mathbf{X}, \mathbf{y}) + \Lambda \left(\mathbf{a} \cdot \mathbf{a}\right) \,, \tag{S1}$$

where $\mathbf{x} \cdot \mathbf{y}$ denotes the scalar product, and Λ is the ridge parameter, i.e. the Lagrange multiplier that constraints the scale of the normalized fit parameters. The error of the fit is given by $E(\mathbf{a}, \mathbf{X}, \mathbf{y}) = (\mathbf{X}\mathbf{a} - \mathbf{y}) \cdot (\mathbf{X}\mathbf{a} - \mathbf{y}) = \mathbf{a} \cdot \mathbf{C}\mathbf{a} - 2\mathbf{a} \cdot \mathbf{X}^T\mathbf{y} + \mathbf{y} \cdot \mathbf{y}$. Note that both the dependent and independent variables are normalized in such a way that $\sum_i (X_{ik})^2 = \sum_i (y_i)^2 = 1$ $(k = 1 \dots P)$, so that Λ does not depend on the scale of the variables. The solution of this quadratic minimization problem can be written explicitly as:

$$\mathbf{a}^{(\mathrm{ns})} = \left(\mathbf{C} + \Lambda \mathbf{I}\right)^{-1} \left(\mathbf{X}^T\mathbf{y}\right) \,. \tag{S2}$$

where $\mathbf{C} = \mathbf{X}^T\mathbf{X}$ is the symmetric $P \times P$ covariance matrix of the normalized independent variables. In the following, we shall often use the P real and positive eigenvalues λ_α and eigenvectors $\mathbf{u}^\alpha$ of the covariance matrix. Since the dependent variables are normalized, the eigenvalues satisfy the condition $\sum_\alpha \lambda_\alpha = P$. Furthermore, we define the normalized projections of the dependent variable $\mathbf{y}$ over the eigenvectors $\mathbf{u}^\alpha$ of the covariance matrix as $y^\alpha = (\mathbf{X}^T\mathbf{y} \cdot \mathbf{u}^\alpha)$, i.e. $y^\alpha = \sum_k \left(\sum_i X_{ik} y_i\right) u_k^\alpha$.

When the Tykhonov parameter Λ increases above the scale set by the eigenvalues, the parameters of the fit tend to zero and so does the fitted dependent variable $\mathbf{y}^{(\mathrm{fit})} = \mathbf{X}\mathbf{a}$. Protocols for ridge regression typically address this problem by not penalizing the offset of the fit, and choosing its value in such a way that $\mathbf{y}^{(\mathrm{fit})}$ is properly scaled with respect to $\mathbf{y}$. However, this procedure modifies the relationship between the explanatory variables. For instance, in the B-factor fit, the offset has to be interpreted as the component of the fit due to translations, and not penalizing the offset would thus have the effect of artificially increasing the contribution of translations, which would then be treated differently than the other degrees of freedom.

Rescaled ridge regression

We present a modified protocol for ridge regression, in which the offset of the fit is penalized as any other variable, but the fitted dependent variable $\mathbf{X}\mathbf{a}$ remains correctly scaled and does not tend to zero. For that purpose, all parameters $a_k^{(\mathrm{ns})}$ are multiplied by a constant scalar ν, to obtain the rescaled parameters $a_k^{(\mathrm{sc})}$. This transformation does not modify the physical interpretation of the fit. The optimal scale (in the minimum squares sense) is obtained with $\nu = (\mathbf{y} \cdot \mathbf{X}\mathbf{a}^{(\mathrm{ns})})/(\mathbf{X}\mathbf{a}^{(\mathrm{ns})} \cdot \mathbf{X}\mathbf{a}^{(\mathrm{ns})}) = (\mathbf{y} \cdot \mathbf{X}\mathbf{a}^{(\mathrm{ns})})/(\mathbf{a}^{(\mathrm{ns})} \cdot \mathbf{C}\mathbf{a}^{(\mathrm{ns})})$. Equivalently, the scaled parameters must satisfy the constraint:

$$F(\mathbf{a}^{(\mathrm{sc})}, \mathbf{X}, \mathbf{y}) \equiv (\mathbf{a}^{(\mathrm{sc})} \cdot \mathbf{X}^T\mathbf{y}) - (\mathbf{a}^{(\mathrm{sc})} \cdot \mathbf{C}\mathbf{a}^{(\mathrm{sc})}) = 0 \,. \tag{S3}$$

In order to keep the analytic treatibility, we impose the constraint Eq.(S3) through a new Lagrange multiplier μ, so that the rescaled ridge regression is still formulated as the minimization of a quadratic function of the parameters:

$$\begin{aligned} G'^{(\mathrm{sc})}(\mathbf{a},\mathbf{X},\mathbf{y},\Lambda,\mu) &= E(\mathbf{a},\mathbf{X},\mathbf{y}) + (1-\mu)\Lambda(\mathbf{a}\cdot\mathbf{a}) + \mu F(\mathbf{a},\mathbf{X},\mathbf{y}) \\ &= (1-\mu)\left[(\mathbf{a}\cdot\mathbf{C}\mathbf{a}) - 2\nu(\mathbf{a}\cdot\mathbf{X}^T\mathbf{y}) + \Lambda(\mathbf{a}\cdot\mathbf{a})\right] + (\mathbf{y}\cdot\mathbf{y})\,, \end{aligned} \tag{S4}$$

with

$$\nu = \frac{1-\mu/2}{1-\mu}\,. \tag{S5}$$

The non-scaled fit can be obtained as a particular case by setting $\nu(\Lambda) \equiv 1$, which implies $\mu = 0$. Note that, to obtain the above form that is computationally convenient, we have redefined the Tychonov parameter as $(1-\mu)\Lambda$, which coincides with the usual definition for the non-scaled fit when $\mu = 0$. Since the term $(\mathbf{y}\cdot\mathbf{y})$ is constant, minimizing the objective function $G'^{(\mathrm{sc})}$ is equivalent to minimizing $G'^{(\mathrm{sc})}/(1-\mu)$, and the solution of this problem is thus given by $\mathbf{a}^{(\mathrm{sc})} = \nu\left(\mathbf{C}+\Lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}\right)$, that is

$$a_k^{(\mathrm{sc})} \equiv \nu(\Lambda) a_k^{(\mathrm{ns})} = \nu(\Lambda)\sum_\alpha \frac{y^\alpha u_k^\alpha}{\lambda_\alpha + \Lambda}\,, \tag{S6}$$

From the condition Eq.(S3) we explicitly determine the proportionality factor ν as

$$\nu(\Lambda) = \frac{\mathbf{X}^T\mathbf{Y}\cdot\mathbf{a}^{(\mathrm{ns})}}{\mathbf{a}^{(\mathrm{ns})}\cdot\mathbf{C}\mathbf{a}^{(\mathrm{ns})}} = 1 + \Lambda\eta(\Lambda) \tag{S7}$$

$$\eta(\Lambda) = \frac{\sum_\alpha \frac{(y^\alpha)^2}{(\lambda_\alpha+\Lambda)^2}}{\sum_\alpha \lambda_\alpha \frac{(y^\alpha)^2}{(\lambda_\alpha+\Lambda)^2}}\,. \tag{S8}$$

The interplay of the constraints imposed by the two Lagrange multipliers implies that, contrary to $a_k^{(\mathrm{ns})}$, the rescaled fit parameters $a_k^{(\mathrm{sc})}$ do not tend to zero when Λ increases.

$$a_{k,\infty}^{(\mathrm{sc})} \equiv \lim_{\Lambda\to\infty} a_k^{(\mathrm{sc})} = \eta_\infty \sum_\alpha y^\alpha u_k^\alpha = \sum_i X_{ik} y_i\,, \text{ with} \tag{S9}$$

$$\eta_\infty \equiv \lim_{\Lambda\to\infty} \eta(\Lambda) = \frac{\sum_\alpha (y^\alpha)^2}{\sum_\alpha (y^\alpha)^2 \lambda_\alpha}\,. \tag{S10}$$

These limit values of the parameters $a_k^{(\mathrm{sc})}$ are independent of the correlation matrix $\mathbf{C}$, except for their scale η_∞ that depends on the eigenvalues λ_α. Therefore, when Λ increases, the information on the correlations between the predictor variables is progressively lost, and their relative weights in the fit become more and more strongly determined by their individual correlations with the dependent variable.

Penalty term

An alternative formulation of the rescaled ridge regression problem is obtained by imposing a constraint on the Euclidian distance D between the parameters $\mathbf{a}$ and adequately chosen reference values of these parameters, $\mathbf{a}^\circ$:

$$D(\mathbf{a},\mathbf{a}^\circ) = (\mathbf{a}-\mathbf{a}^\circ)\cdot(\mathbf{a}-\mathbf{a}^\circ)\,. \tag{S11}$$

The limit for infinite Λ of the reference parameters and the actual parameters must coincide, so that the term D vanishes in this limit and the reference parameters are enforced. The objective function to be minimized is then formulated as:

$$\begin{aligned} G^{(\mathrm{sc})}(\mathbf{a}, \mathbf{a}^\circ, \mathbf{X}, \mathbf{y}, \Lambda, \mu) &= E(\mathbf{a}, \mathbf{X}, \mathbf{y}) + (1-\mu)\Lambda D(\mathbf{a}, \mathbf{a}^\circ) + \mu F(\mathbf{a}, \mathbf{X}, \mathbf{y}) \\ &= (1-\mu)\left[\mathbf{a}\cdot\mathbf{C}\mathbf{a} - 2\nu'\mathbf{a}\cdot\mathbf{X}^T\mathbf{y} + \Lambda\left(\mathbf{a}\cdot\mathbf{a} - 2\mathbf{a}\cdot\mathbf{a}^\circ + \mathbf{a}^\circ\cdot\mathbf{a}^\circ\right)\right] + (\mathbf{y}\cdot\mathbf{y}), \end{aligned} \tag{S12}$$

where $\nu' = (1-\mu/2)/(1-\mu)$. The minimization of $G^{(\mathrm{sc})}$ yields exactly the same result as the minimization of $G'^{(\mathrm{sc})}$ (Eq. S4), i.e. $\mathbf{a}^{(\mathrm{sc})} = \nu\left(\mathbf{C} + \Lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}\right)$, provided that the reference parameters $\mathbf{a}^\circ$ are chosen as

$$\mathbf{a}^\circ = \xi(\Lambda)\mathbf{X}^T\mathbf{y}, \tag{S13}$$

where $\xi(\Lambda)$ satisfies the equation

$$\frac{1-\mu/2}{1-\mu} + \Lambda\xi(\Lambda) = \nu(\Lambda) = 1 + \Lambda\eta(\Lambda). \tag{S14}$$

Note that the value of the scale parameter is still determined by Eq.(S8), but we are free to choose the multiplier $\mu(\Lambda)$ and the scale of the reference parameters $\xi(\Lambda)$ as it is most convenient. We discuss below two natural ways to define the reference parameters $\xi(\Lambda)$.

Alternative definitions of the Maximum Penalty fit

We proposed a criterion for choosing the ridge parameter Λ that is based on the "entropic" contribution to the free energy of the fit, or penalty term,

$$\mathrm{Penalty}(\Lambda) = (1-\mu)\Lambda((\mathbf{a}-\mathbf{a}^\circ)\cdot(\mathbf{a}-\mathbf{a}^\circ)). \tag{S15}$$

The penalty term is equal to zero both at $\Lambda = 0$, where all the information of the correlation between explanatory variables is retained, and for $\Lambda \to \infty$, where $\mathbf{a}$ and $\mathbf{a}^\circ$ reach the same limit and the information of the correlation matrix is lost (except for the scaling factor). In between, the penalty term reaches a maximum, and we hypothesize that this maximum corresponds to a possibly optimal choice of the ridge parameter Λ. We call "maximum penalty fit" ridge regression with this choice of Λ.

In non-scaled ridge regression (Eq. S1), μ and all reference parameters $\mathbf{a}^\circ$ are equal to 0. The definition of the MP fit is thus unequivocal, as the value of Λ identified by the MP criterion is the one that maximizes the penalty term, i.e. $\Lambda_{\mathrm{MP}} = \max_\Lambda(\Lambda(\mathbf{a}^{(\mathrm{ns})}\cdot\mathbf{a}^{(\mathrm{ns})}))$. With the more general definition of rescaled ridge regression (Eq. S13), there is some flexibility in the definition of the penalty term, and thus of the MP criterion, depending on the choice of the reference parameters $\mathbf{a}^\circ$. In addition to the definition retained in the main text, we also tested three slightly different definitions of the MP criterion, which are presented below.

MP criterion. The criterion presented in the main text consists in choosing reference parameters independent of Λ, which requires that ξ does not depend on Λ. To recover the correct infinite Λ limit, it must hold $\xi(\Lambda) \equiv \xi_\infty = \eta_\infty$, so that $\mathbf{a}^\circ = \mathbf{a}^{(\mathrm{sc})}_\infty$:

$$\text{MP}: \qquad \xi \equiv \lim_{\Lambda\to\infty} \eta(\Lambda) = \eta_\infty = \frac{\sum_\alpha (y^\alpha)^2}{\sum_\alpha (y^\alpha)^2 \lambda_\alpha}\,, \tag{S16}$$

$$\mu(\Lambda) = \frac{2\Lambda\,(\eta(\Lambda) - \eta_\infty)}{1 + 2\Lambda\,(\eta(\Lambda) - \eta_\infty)}\,. \tag{S17}$$

In this way, the reference parameters are equal to the infinite Λ limit of the fit parameters, i.e. $\mathbf{a}^\circ = \mathbf{a}_\infty \equiv \lim_{\Lambda\to\infty} \mathbf{a}^{(\text{sc})}$. Thus, $\Lambda_{\text{MP}} = \max_\Lambda \left((1-\mu)\Lambda((\mathbf{a}^{(\text{sc})} - \mathbf{a}_\infty)\cdot(\mathbf{a}^{(\text{sc})} - \mathbf{a}_\infty))\right)$. Note that the reference parameters are exactly equal to those that would be obtained if the explanatory variables were uncorrelated ($\Lambda_\alpha \equiv 1$). This gives a more direct interpretation to the penalisation term in rescaled ridge regression, i.e. the error cannot be minimized at the cost of having parameters values too different from those that would be obtained if the predictor variables were not correlated to each other.

MP* criterion. A first variation is obtained as an approximation of the MP criterion, in which the reference parameters are identical, $\mathbf{a}^\circ = \mathbf{a}_\infty$, but the dependence of μ on Λ is considered negligible in the range of interest of Λ values. We have then, $\Lambda_{\text{MP}*} = \max_\Lambda \left(\Lambda((\mathbf{a}^{(\text{sc})} - \mathbf{a}_\infty)\cdot(\mathbf{a}^{(\text{sc})} - \mathbf{a}_\infty))\right)$, which can be easier to implement than the MP criterion but gives highly similar results, as shown in Fig. S1.

MP_Λ criterion. Another possibility is to set $\mu = 0\;\forall\Lambda$ or, more generally, to set μ independent of Λ, which is only compatible with $\mu = 0$, so that we do not have to explicitly impose the constraint on the scale of the parameters. With (Eq. S14), this implies that ξ depends on Λ as $\xi(\Lambda) = \eta(\Lambda)$. The resulting criterion is defined by the equations

$$\text{MP}_\Lambda: \qquad \xi(\Lambda) = \eta(\Lambda) \equiv \frac{\sum_\alpha \frac{(y^\alpha)^2}{(\lambda_\alpha+\Lambda)^2}}{\sum_\alpha \lambda_\alpha \frac{(y^\alpha)^2}{(\lambda_\alpha+\Lambda)^2}}\,, \tag{S18}$$

$$\mu = 0\,. \tag{S19}$$

This particular choice of the reference parameters $\mathbf{a}^\circ = \mathbf{a}_\Lambda = \eta(\Lambda)\mathbf{X}^T\mathbf{y}$ automatically ensures the optimal scaling of the $\mathbf{a}^{(\text{sc})}$. The objective function can thus be expressed as $G^{(\text{sc})} = E + \Lambda((\mathbf{a} - \mathbf{a}_\Lambda)\cdot(\mathbf{a} - \mathbf{a}_\Lambda))$, without the need for an additional constraint on the scale (i.e. $\mu = 0$). The drawback is that ξ, and hence the reference parameters, depend on Λ. We found that the MP_Λ criterion produces results that are generally similar, although slightly poorer on average, than those based on the above definition of MP with $\xi = \xi_\infty$.

MP_{ns} criterion. The last alternative is obtained by setting $\xi = 0$, so that the reference parameters $\mathbf{a}^\circ$ are all equal to zero. In that case, the penalty term does not reach a maximum value in rescaled ridge regression, because $(\mathbf{a}^{(\text{sc})}\cdot\mathbf{a}^{(\text{sc})})$ does not tend to zero when $\Lambda\to\infty$. It does however reach a maximum value in non-scaled ridge regression. A possibility is thus to choose the parameter Λ that maximizes the penalty in the non-scaled fit, $\Lambda_{\text{MP}_{\text{ns}}} = \max_\Lambda \left(\Lambda(\mathbf{a}^{(\text{ns})}\cdot\mathbf{a}^{(\text{ns})})\right)$, and apply it to the rescaled fit. In

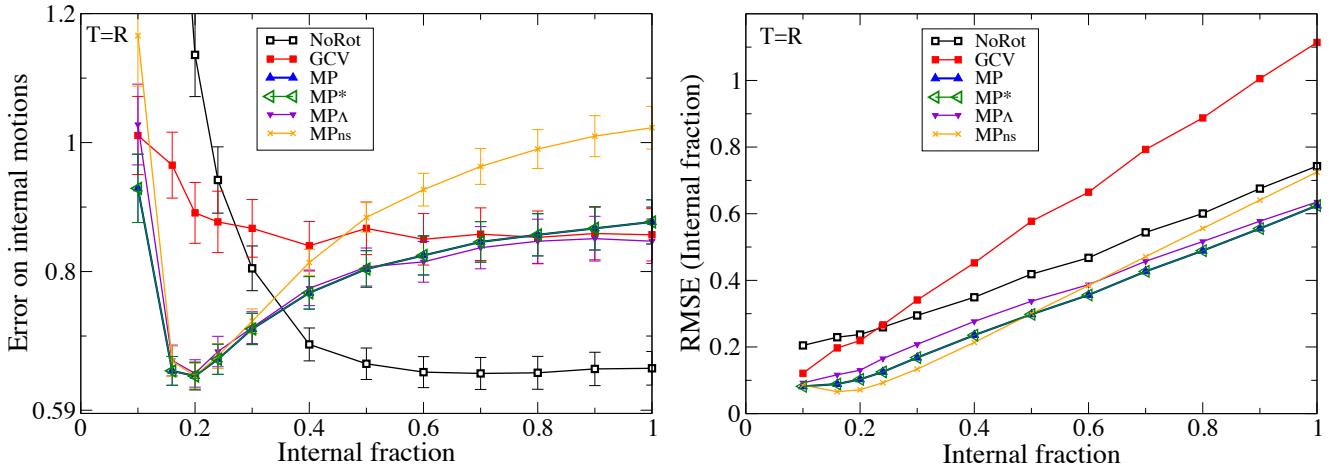
non-scaled ridge regression, $\mu = 0$, since we do not impose the optimal scale. We have thus

$$\mathrm{MP}_{\mathrm{ns}}: \qquad \xi = 0 \tag{S20}$$

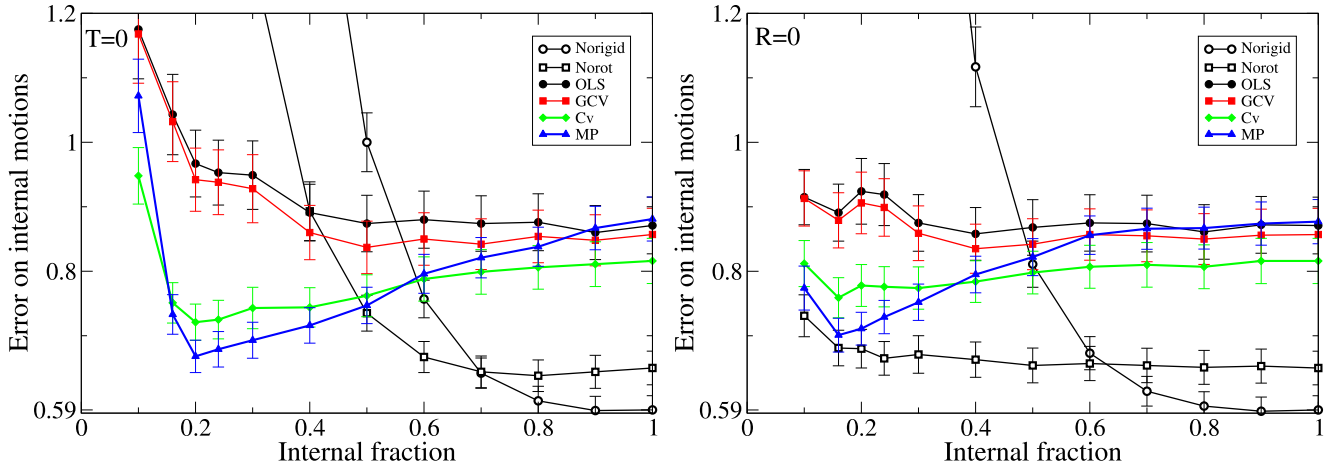
$$\mu = 0\,. \tag{S21}$$

We found that this definition tends to select larger values of Λ, and thus to impose a stronger regularisation of the fit. It did yield particularly good results in some cases, but generally performs worse than the MP criterion with $\xi = \xi_\infty$.

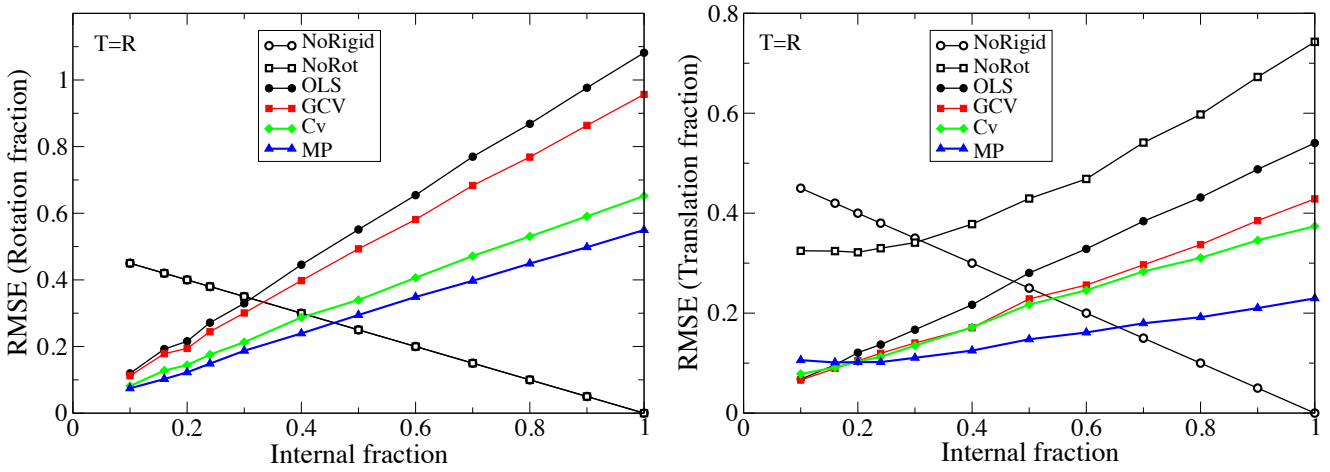
Supporting figures



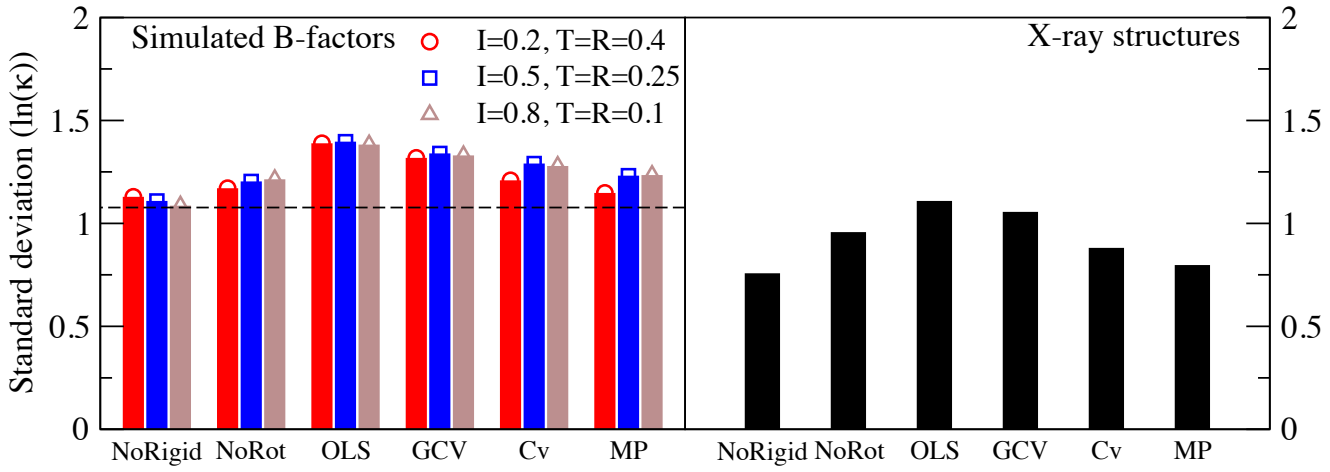
Supporting Figure S1: Performances of the alternative definitions of the MP fit. The error on the estimated profile of internal fluctuations (left), and the error on the estimated fraction of internal fluctuations (right) are given against the simulated fraction of internal fluctuations, I.



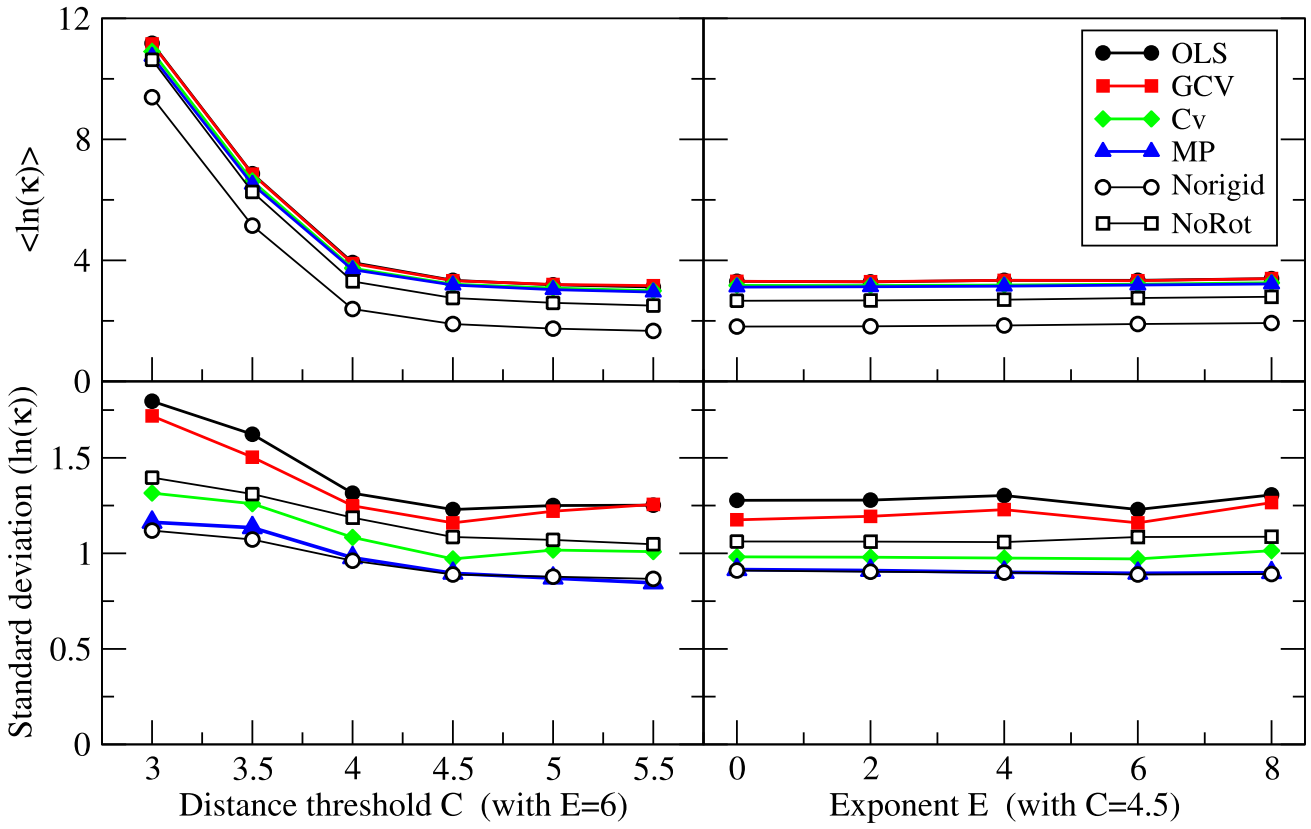
Supporting Figure S2: Error of the fitted internal motions when $T = 0$ or $R = 0$. E^{int} (see Eq. 22 in the main text) is given as a function of the fraction of internal motions I, in the simulated sets with either $T = 0$ (left) or $R = 0$ (right), for different types of fit. The lowest possible value of the error, $E^{\text{int}} = 0.59$, is obtained with the NoRigid fit on the NMR dataset ($I = 1.0$). One can see that for the $T = 0$ set with no translations the NoRot and NoRigid fits behave rather similarly.



Supporting Figure S3: Error on the estimated fractions of rigid-body fluctuations. The RMSE of the fitted rotational (left) and translational fractions (right) is given as a function of the internal fraction I. The error on the rotational fraction is the same for the NoRot and NoRigid fits, since the estimated fraction of motion due to rotations is always null in both fits. The MP fit is the best of all variants of ridge regression except for the translation fraction at $I < 0.2$.



Supporting Figure S4: Variation of the fitted force constant for different types of fit. The standard deviation (across proteins) of the logarithm of the force constant, $\sigma_{\ln(\kappa)}$, is given for selected sets of simulated data (left) and for the X-ray dataset (right). We consider the logarithm because its fluctuations are better behaved, and it allows to eliminate the influence of multiplicative scale factors. In each set, the proteins for which at least one of the fits yielded a negative value of κ were omitted, for all fits. The dashed line corresponds to the value of $\sigma_{\ln(\kappa)}$ obtained with the NoRigid fit on the NMR dataset ($I = 1$). A value of $\sigma_{\ln(\kappa)} = 1$ corresponds to a multiplicative spread of $e^1 \approx 2.7$, i.e. for most proteins, the estimated force constant is smaller than 2.7 times and larger than (1/2.7) times the geometric mean.



Supporting Figure S5: Dependence on the ENM parameters. Average value (top) and standard deviation (bottom) of the logarithm of the force constant κ estimated for the proteins of the X-ray dataset, versus the distance cut-off C for exponent $E = 6$ (left) and versus the exponent E for distance cut-off $C = 4.5$Å(right). Proteins for which some fits generated a negative value of κ were not considered. When C increases, more pairs are considered to be interacting, and a smaller κ is thus necessary to maintain a similar amplitude of the atomic fluctuations. In contrast, κ shows little dependence on E, thanks to the choice of the reference distance $r_0 = 3.5$Å. Indeed, when E increases, the stiffness increases for pairs of residues with $r_{ij} < 3.5$Å, but decreases for pairs or residues with $r_{ij} > 3.5$Å, and these two effects compensate each other fairly well. Importantly, all types of fits show the same qualitative behavior, and the differences between the fits are mostly independent of the choice of the E and C parameters. This demonstrates the robustness of the results presented in the main text, which were obtained with $C = 4.5$Å and $E = 6$.