



HAL
open science

Explaining Conformational Diversity in Protein Families through Molecular Motions

Valentin Lombard, Sergei Grudinin, Elodie Laine

► **To cite this version:**

Valentin Lombard, Sergei Grudinin, Elodie Laine. Explaining Conformational Diversity in Protein Families through Molecular Motions. 2024. hal-04442287

HAL Id: hal-04442287

<https://hal.univ-grenoble-alpes.fr/hal-04442287v1>

Preprint submitted on 6 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Explaining Conformational Diversity in Protein Families through Molecular Motions

Valentin Lombard¹, Sergei Grudin^{2,*}, and Elodie Laine^{1,3,*}

¹Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France

²Université Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

³Institut Universitaire de France (IUF)

*corresponding author(s): Sergei Grudin (sergei.grudin@univ-grenoble-alpes.fr), Elodie Laine (elodie.laine@sorbonne-universite.fr)

ABSTRACT

Proteins play a central role in biological processes, and understanding their conformational variability is crucial for unraveling their functional mechanisms. Recent advancements in high-throughput technologies have enhanced our knowledge of protein structures, yet predicting their multiple conformational states and motions remains challenging. This study introduces Dimensionality Analysis for protein Conformational Exploration (DANCE) for a systematic and comprehensive description of protein families conformational variability. DANCE accommodates both experimental and predicted structures. It is suitable for analysing anything from single proteins to superfamilies. Employing it, we clustered all experimentally resolved protein structures available in the Protein Data Bank into conformational collections and characterized them as sets of linear motions. The resource facilitates access and exploitation of the multiple states adopted by a protein and its homologs. Beyond descriptive analysis, we assessed classical dimensionality reduction techniques for sampling unseen states on a representative benchmark. This work improves our understanding of how proteins deform to perform their functions and opens ways to a standardised evaluation of methods designed to sample and generate protein conformations.

Introduction

Proteins orchestrate all biological processes, and their malfunctions often result in disease. In recent years, high-throughput technologies have greatly improved our knowledge of their amino acid sequences and 3D shapes¹⁻⁴. While reaching the single-structure frontier⁵, these advances have also highlighted the complexities of how proteins move and deform to carry out their biological functions^{6,7}. They have stimulated a renewed interest in the modeling of protein and protein complex multiple conformational states⁸. In particular, the success of the protein structure prediction neural network AlphaFold2⁹ has inspired innovative strategies for modifying or repurposing it toward exploring protein conformational space. These approaches involve forced sampling¹⁰, modulation of input multiple sequence alignment content and depth^{11,12}, or guidance with state-annotated templates^{13,14}. Although they have achieved promising results for specific protein families, systematic assessments have revealed limitations^{15,16}. In addition, studies sampling from low-dimensional representations or manifolds learned from observed or simulated conformations¹⁷⁻¹⁹ have underscored the difficulty in predicting new, completely unseen states and the importance of high-quality data for training or benchmarking.

Experimental techniques like X-ray crystallography, cryogenic-electron microscopy (cryo-EM), and nuclear magnetic resonance spectroscopy (NMR) are essential for capturing protein functional states^{6,20}. The Protein Data Bank (PDB)⁴ offers access to multiple structural states for various proteins, solved independently in different conditions, oligomeric states, and with diverse cofactors and molecular partners. Researchers have actively engaged in efforts to collect, cluster, curate, represent, visualise, and functionally annotate these states²⁰⁻²³. These endeavours have provided valuable insights into the biologically meaningful conformational space for specific protein families such as protein kinases²⁴, RAS isoforms²⁵, ABC (ATP Binding Cassette) transporters²⁶, and G-protein coupled receptors (GPCRs)²⁷. However, producing or validating functional annotations for structural states involves a substantial amount of manual intervention. Despite the wealth of experimentally resolved protein conformational variability, its full exploitation remains an ongoing challenge.

Ideally, one would like to comprehensively describe protein conformational variability with low-dimensional representations or manifolds amenable to visualisation and interpretation. Principal Component Analysis (PCA) serves as a convenient and robust means to reduce the dimensionality of a dataset, capturing maximum variability^{28,29}. The principal components extracted from a conformational ensemble define 3D directions for every atom, and motions along them allow navigating the conformational space³⁰. PCA has proven useful for extracting structural transitions from sparse disconnected low-energy

38 structural states^{31–36}. Unlike more complex non-linear dimensionality reduction techniques, it offers the advantage of not
39 depending on numerous adjustable parameters and provides a straightforward geometrical interpretation.

40 Here, we describe a PDB-wide analysis of protein conformational variability across various levels of sequence homology.
41 Our fully-automated computational pipeline, named Dimensionality Analysis for protein Conformational Exploration (DANCE),
42 systematically compiles collections of aligned protein conformations and extracts their principal components. We interpret
43 the representation space defined by the main principal components as the *linear motion manifold* underlying the observed
44 conformations. We provide estimates of the intrinsic dimensionality of these motion manifolds. To assess generative methods,
45 we introduce a benchmark set comprising ten conformational collections representing therapeutic targets with substantial
46 functional transitions. Additionally, we provide baseline performances from classical linear and non-linear manifold learning
47 techniques.

48 DANCE is versatile, handling both experimental and predicted structures with varying amino acid sequences. It adopts
49 an unbiased approach, avoiding predetermined protein or domain definitions when building the conformational collections.
50 Considering the complete context of input protein chains enables a thorough examination of inter-domain motions. Furthermore,
51 DANCE accommodates uncertainty from unresolved protein regions without assuming potential conformations. It introduces a
52 weighting scheme to mitigate the imbalanced coverage of variables.

53 We provide several databases of conformational collections representing the whole PDB as well as detailed information
54 about the benchmark on Figshare. In addition, DANCE's source code is available at: [https://github.com/
55 PhyloSoFS-Team/DANCE](https://github.com/PhyloSoFS-Team/DANCE).

56 Methods

57 Overview of DANCE

58 DANCE takes as input a set of protein 3D structures (in Crystallographic Information File or CIF format) and outputs a
59 set of protein- or protein family-specific conformational collections or ensembles (in CIF or PDB format). It first clusters
60 and superimposes the input structures based on the similarities found in their corresponding amino acid sequences. It then
61 determines the set of principal components sufficient to explain the variability observed within each conformational ensemble.
62 The algorithm unfolds in six main steps depicted in **Fig. 1**.

- 63 • **a- Extraction of sequences.** The first step extracts the one-letter amino acid sequences of all polypeptidic chains
64 contained in the input CIF files. In case of multiple models, DANCE retains only the first one. The names of the residues
65 with resolved 3D coordinates are taken from the `_atom_site.label_comp_id` column. Residues missing from the protein
66 structure are included as lowercase letters in the sequence if they are defined in the `_entity_poly_seq` category. This
67 information will help in clustering and aligning the sequences (see below). Otherwise, they are replaced by the "X"
68 symbol. The "X" symbol is also used for unknown amino acid types and for modified amino acids without a close natural
69 neighbour. Sequences comprising less than 5 non-"X" residues are then filtered out.
- 70 • **b- Clustering of the sequences.** DANCE clusters sequences using MMseqs2³⁷. The users can choose the desired levels
71 of sequence similarity and coverage, both set to 80% by default. The coverage is bidirectional by default. This step
72 outputs a TSV file specifying the clusters.
- 73 • **c- Multiple sequence alignments.** DANCE then aligns the sequences within each cluster using MAFFT³⁸ with default
74 parameters. It further removes all the columns containing only Xs or gaps, and reorders the sequences according to their
75 PDB codes.
- 76 • **d- Extraction of structures.** DANCE extracts 3D coordinates of the backbone atoms N, C, C α , and the O atom, of
77 all polypeptidic chains contained in the input CIF files. It reconstructs missing O atoms based on the other atom's
78 coordinates. It disregards residues with missing backbone atoms and chains shorter than 5 residues.
- 79 • **e- Generation of the conformational collections.** DANCE then uses the sequence clusters defined in (b) to group
80 conformations and the residue matching provided by (c) to superimpose them. The superimposition puts their centers of
81 mass to zero and then aims at determining the optimal least-squares rotation matrix minimizing the Root Mean Square
82 Deviation (RMSD) between any conformation and a reference conformation (see below). This is achieved through the
83 ultrafast Quaternion Characteristic Polynomial method^{39,40}. The users can choose to account for all the atoms in the
84 superimposition, or only the C α atoms. Optionally, the users can filter out the conformations with too few (less than 5 by
85 default) residues aligning to the reference. As a post-processing step, DANCE reduces structural redundancy. Namely,
86 it removes any conformation *A* deviating by less than rms_{cut} Å from another one *B*, provided that the sequence of *A* is
87 identical to or included in that of *B*. The value of rms_{cut} is 0.1 Å by default and is customizable by the users. Finally,

DANCE saves the conformational ensemble as a multi-model file in PDB or CIF format. Notice that the models can display different amino acid sequences. DANCE also outputs the corresponding multiple sequence alignments (MSA) in FASTA format, and the matrix of all-to-all pairwise RMSDs.

- **f- Extraction of linear motions.** DANCE performs PCA on the 3D coordinates from each collection. This dimensionality reduction technique identifies orthogonal linear combinations of the variables, namely the Cartesian coordinates, maximally explaining their variance (see below). These linear combinations, which we refer to as principal components or PCA modes, represent directions in the 3D space for every atom. Deforming the protein structure using these components produce motions that connect the conformations observed in the collection. For the sake of simplicity, we directly refer to the principal components as to *linear motions*, although they may not represent actual physical motions undergone by the protein. Furthermore, we estimate the *intrinsic dimensionality* of the linear motion manifold underlying an ensemble’s conformational variability as the number of principal component explaining essentially all its positional variance. The higher the dimensionality – the more complex the linear motions.

Choosing a reference

We choose the reference conformation for the superimposition as the one with the amino acid sequence most representative of the MSA. For this, we first determine the consensus sequence s^* by identifying the most frequent symbol at each position. We consider "X" symbols as equivalent to gaps. Hence, each position is described by a 21-dimensional vector giving the frequencies of occurrence of the 20 amino acid types and of the gaps. In case of ambiguity, we prefer an amino acid over a gap and a more frequent amino acid over a less frequent one. Then, we compute a score for each sequence s in the MSA reflecting its similarity to s^* and expressed as,

$$\text{score}(s) = \sum_{i=1}^P \sigma(s_i, s_i^*), \quad (1)$$

where P is the number of positions in the MSA and $\sigma(s_i, s_i^*)$ is the substitution score between the amino acid s_i at position i in sequence s and the consensus symbol s_i^* at position i . We use the substitution matrix BLOSUM62 and we set the gap score to $\min_{a,b}(\sigma(a,b)) - 1 = -5$. MAFFT also uses BLOSUM62 for generating the MSAs.

Judging the quality of the MSA

We compute the identity level of an MSA as the average percentage of sequence pairs sharing the same amino acid in a column, and the coverage as the percentage of positions having less than 20% of gaps. In addition, we evaluate the global quality of the MSA with a sum-of-pairs score, with $\sigma_{\text{match}} = 1$ and $\sigma_{\text{mismatch}} = \sigma_{\text{gap}} = -0.5$. We normalise the raw sum-of-pairs scores by dividing them by the maximum expected values. The final score for an MSA is thus expressed as,

$$\text{score}_{\text{rel}}(\text{MSA}) = \frac{\text{score}(\text{MSA})}{\binom{n}{2} L_{\text{eff}}}, \quad (2)$$

where is the raw MSA score, n is the number of chains or sequences, and L_{eff} is the effective length of the MSA, computed as,

$$L_{\text{eff}} = \max_{s \in \mathcal{S}} \sum_{i=1}^{L(s)} \mathbb{I}\{s_i \in \mathcal{A}\}, \quad (3)$$

where \mathcal{S} is the set of sequences comprised in the MSA, $L(s)$ is the length of the aligned sequence s , and \mathcal{A} is the 20-letter amino acid alphabet (e.g., excluding gap characters).

Extracting linear motions

The Cartesian coordinates of each conformational ensemble can be stored in a matrix R of dimension $3m \times n$, where m is the number of positions in the associated MSA and n is the number of conformations. Each position is represented by a C- α atom. We compute the covariance matrix as,

$$C = \frac{1}{n-1} R^c (R^c)^T = \frac{1}{n-1} (R - \bar{R})(R - \bar{R})^T, \quad (4)$$

where \bar{R} is obtained by averaging the coordinates over the conformations. Alternatively, the users can choose to center the data on the reference conformation. The covariance matrix is a $3m \times 3m$ square matrix, symmetric and real.

The PCA consists in decomposing C as $C = VDV^T$ where V is a $3m \times 3m$ matrix where each column defines an eigenvector or a PCA mode that we interpret as a linear motion. D is a diagonal matrix containing the eigenvalues. The sum of the

118 eigenvalues $\sum_{k=1}^{3m} \lambda_k$ amounts to the total positional variance of the ensemble. The portion of the total variance explained by the
 119 k th eigenvector or linear motion is estimated as $\frac{\lambda_k}{\sum_{k=1}^{3m} \lambda_k}$.

In addition, we estimate the collectivity^{41,42} of the k th eigenvector as,

$$\text{coll}(\mathbf{v}_k) = \frac{1}{m} \exp \left(- \sum_{i=1}^{3m} v_{ki}^2 \log v_{ki}^2 \right). \quad (5)$$

120 If $\text{coll}(\mathbf{v}_k) = 1$, then the corresponding motion is maximally collective and has all the atomic displacements identical. In case
 121 of an extremely localised motion, where only one single atom is affected, the collectivity is minimal and equals to $1/m$.

We also apply PCA to the correlation matrix computed by normalising the covariance matrix as,

$$\text{Cor}_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i}} \sqrt{C_{j,j}}}. \quad (6)$$

122 In that case, the sum of the eigenvalues $\sum_{k=1}^{3m} \lambda_k$ amounts to 1.

123 **Handling missing data**

As stated above, the conformations in a collection may have different lengths reflected by the introduction of gaps in the associated MSA. We fill these gaps with the coordinates of the conformation used to center the data (average conformation, by default). In doing so, we avoid introducing biases through reconstruction of the missing coordinates. Moreover, this operation results in low variance for highly gapped positions, thus limiting their contribution to the extracted motions. To go further and explicitly account for data uncertainty, we implemented a weighting scheme. Specifically, DANCE assigns confidence scores to the residues and include them in the structural alignment step and the PCA. The confidence score of a position i reflects its coverage in the MSA, $w_i = \frac{1}{n} \sum_S \mathbb{1}_{a_i^S \neq "X"}$, where "X" is the symbol used for gaps. The structural alignment of the j th conformation onto the reference conformation amounts to determining the optimal rotation that minimises the following function⁴³,

$$E = \frac{1}{\sum_i w_i} \sum_i w_i (r_{ij}^c - r_{i0}^c)^2, \quad (7)$$

124 where r_{ij}^c is the i th centred coordinate of the j th conformation and r_{i0}^c is the i th centred coordinate of the reference conformation.
 125 The resulting aligned coordinates are then multiplied by the confidence scores prior to the PCA.

126 **Implementation details**

127 We implemented DANCE in C/C++ and Python. It relies on the C++ GEMMI library⁴⁴ to parse the CIF files and manipulate
 128 the structures. It runs MMseqs2 through the following command: `cluster DB clusterDB tmp -cov-mode 0 -c $cov -min-seq-id`
 129 `$id`. It launches MAFFT with the options `auto`, `amino` and `preservecase`. The multiple sequence alignment and structure
 130 superimposition steps are parallelized. For the PCA, we use the singular value decomposition (SVD) implemented in NumPy⁴⁵
 131 on the R matrix directly. SVD is computationally more advantageous when $3m \gg n$, which is typically the case of our data,
 132 since we only compute the required number of n components.

133 **Application and extension of DANCE**

134 DANCE is applicable to experimental 3D structures as well as predicted 3D models, as long as they comply with the CIF
 135 standards.

136 **Describing conformational variability over the whole PDB**

137 We applied DANCE to all 748 297 protein chains with experimentally resolved 3D structures available in the PDB, as of June
 138 2023. We downloaded all the PDB entries in CIF format from the RCSB⁴⁶. We replaced the raw CIF files with their updated
 139 and optimised versions from PDB-REDO whenever possible⁴⁷. It took about 2.25 hours to run DANCE on the whole PDB
 140 on a desktop computer with Intel Xeon W-2245 @ 3.90GHz and 32Go of RAM (**Table S1**). The most time consuming steps
 141 are the extraction and superimposition of the 3D structures to create the conformational ensembles. We ran DANCE at eight
 142 different levels of sequence similarity, designated as l_{cov}^{id} , where id and cov are the sequence identity and coverage thresholds,
 143 correspondingly, and range from 50 to 80%. For investigating how the ensembles transformed across levels, we focused on the
 144 18 616 conformational ensembles detected in the most relaxed set up, namely at 30% identity and 50% coverage (l_{50}^{30}). For each
 145 ensemble, we extracted its reference protein chain and we traced back the conformational ensembles to which it belonged upon
 146 progressively applying stricter thresholds.

147 **Focusing on the ABC superfamily**

148 We extended DANCE usage beyond the single-chain and sequence-similarity paradigms to describe the conformational
149 variability of ABC (ATP Binding Cassette) transporters. We retrieved a set of 354 ABC protein experimental 3D structures
150 from <https://abc3d.hegelab.org>²⁶. They correspond to functionally relevant states annotated as biological units in
151 the PDB. In most of these structures, several polypeptidic chains, typically 2 or 4, encode the two nucleotide-binding domains
152 (NBDs) and two transmembrane domains (TMDs) of the ABC architecture. In addition, some structures contain several ABC
153 protein copies or some ABC protein cellular partners (small molecules, substrate peptides, interacting proteins). We chose
154 the murine ABC transporter P-glycoprotein (5KOYA) as reference for the subsequent analysis. Its 1182-residue long single
155 polypeptidic chain the full-length transporter architecture.

156 To cope with the high sequence divergence of the ABC superfamily, we relied on structural similarity for grouping and
157 matching the ABC conformations. Specifically, we used the method Foldseek⁴⁸ to identify structures sharing significant
158 similarity with the reference and align them. We performed a first screen by querying the reference against all individual chains
159 (1 244 in total) and defined significant hits as those with an e-value lower than 10.0. Then, for each structure, we estimated
160 an upper bound on its coverage of the reference by summing up the reference residue ranges appearing in the alignments
161 associated with its significant hits. We filtered out the structures with coverage upper bounds lower than 90%. We performed
162 a second screen by querying the reference against the 209 remaining structures defined as monomers by concatenating their
163 chains. We identified two structures (5NIK, 5NIL) spanning less than 90% of the reference. Permuting their chains did not
164 increase their coverage and thus we removed them. To further detect potentially suboptimal chain orderings, we computed
165 reference to target residue span ratios. We identified one structure, namely 7AHD, with a highly imbalanced ratio of 1.6. Such
166 a high value is indicative of large parts of the reference that could not be aligned to the target structure. Permuting the four
167 chains (A,B,C,D) of 7AHD into (A,D,B,C) led to a more balanced ratio of 0.86. We did not observe discrepancies for other
168 structures and thus we retained their original chain ordering. Finally, we removed the structures with low-quality alignments,
169 *i.e.*, with more than 200 gaps or with a continuous gapped region of more than 60 positions.

170 Among the 195 structures finally selected, 4F4C, 7SHN and 7AHD contained unknown or unrecognized amino acids which
171 we removed. We ran Foldseek one more time to generate a structure similarity-based multiple sequence alignment centred on
172 the reference 5KOYA. We trimmed the alignment and the 3D structures by removing the residues inserted with respect to the
173 reference. We gave the trimmed alignment and 3D coordinate files as input to DANCE, starting directly from step *d* (see the
174 overview of DANCE algorithm above). For consistency and comparison purposes, we asked DANCE to center the data on
175 the reference. To mitigate the impact of potential alignment errors, we applied weights reflecting position-specific confidence
176 scores (see above, *Handling missing data*). DANCE structural redundancy reduction step removed 7 conformations, resulting
177 in an ensemble of 188 conformations.

We compared this ensemble with those generated by DANCE default sequence similarity-based end-to-end procedure
applied to the whole PDB. More specifically, we took the ensembles generated at l_{80}^{80} and l_{50}^{30} and containing 5KOYA and we
rebuilt them with DANCE, applying the 5KOYA centering and the uncertainty weighting scheme. We estimated the similarity
between the ensembles' motion subspaces as the Root Mean Square Inner Product (RMSIP)^{49,50}. The latter measures the
overlap between all pairs of the *l* first PCA modes and is defined as,

$$\text{RMSIP} = \sqrt{\frac{1}{l} \sum_{i=1}^l \sum_{j=1}^l (\mathbf{v}_i^{\mathcal{S}_A} \cdot \mathbf{v}_j^{\mathcal{S}_B})^2}, \quad (8)$$

178 where $\mathbf{v}_i^{\mathcal{S}_A}$ and $\mathbf{v}_j^{\mathcal{S}_B}$ are the *i*th and *j*th PCA modes extracted from the conformational ensembles \mathcal{S}_A and \mathcal{S}_B , and *l* is the
179 number of modes considered for the comparison. Moreover, we monitored the distance between the geometric centres of the
180 two NBDs defined by the C- α atoms of residues numbered 346-596 and 929-1182, respectively, in the reference 5KOYA.

181 **Benchmarking for the generation of unseen conformations**

182 **Linear PCA**

We further investigated whether the extracted principal components could be useful to predict unseen conformations. Given a
set of *l* PCA modes computed from the coordinates *R*, we generate a new conformation $\mathbf{r}_{\text{pred}}^*$ as,

$$\mathbf{r}_{\text{pred}}^* = \mathbf{p}^* V_l^T + \bar{\mathbf{r}}, \quad (9)$$

183 where the matrix $V_k \in \mathbb{R}^{3m \times l}$ contains the modes, $\bar{\mathbf{r}} \in \mathbb{R}^{3m}$ is the average conformation, and $\mathbf{p}^* \in \mathbb{R}^l$ is a point in the *l*-dimensional
184 representation space defined by the modes. The coordinates of \mathbf{p}^* specify the amplitudes of the modes.

185 **Nonlinear kernel PCA**

186 The manifold underlying our data is *a priori* non-linear. This motivated us to investigate whether non-linear methods could
 187 achieve better reconstructions than linear PCA. We focused on the widely used kernel Principal Component Analysis (kPCA)⁵¹.
 188 The intuition behind kPCA is to map the input data points to a higher dimensional space where they will be linearly separable by
 189 a classical PCA. The mapping function $\phi : \mathbb{R}^{3m} \rightarrow \mathbb{R}^M$ is not known. Instead of explicitly calculating it, we use a kernel function
 190 $k(\mathbf{r}_i, \mathbf{r}_j) = \phi(\mathbf{r}_i)^T \phi(\mathbf{r}_j)$, where \mathbf{r}_i and \mathbf{r}_j are two conformations. We chose the radial basis function (RBF) kernel, defined as
 191 $k(\mathbf{r}_i, \mathbf{r}_j) = e^{-\frac{d(\mathbf{r}_i, \mathbf{r}_j)^2}{2\sigma^2}}$, where $d(\mathbf{r}_i, \mathbf{r}_j)$ is the Euclidean distance between the two conformations \mathbf{r}_i and \mathbf{r}_j . We explored different
 192 values of the hyperparameter σ . For sufficiently small values, *i.e.*, $\frac{1}{2\sigma^2}d(\mathbf{r}_i, \mathbf{r}_j)^2 \ll 1$, the RBF kernel becomes effectively linear,
 193 since, in this case, $k(\mathbf{r}_i, \mathbf{r}_j) \approx 1 - \frac{1}{2\sigma^2}d(\mathbf{r}_i, \mathbf{r}_j)^2$.

Thus, given the input coordinates R representing n conformations, we computed the corresponding RBF kernel matrix K of dimension $n \times n$ and decomposed it using the classical PCA. The resulting principal components $\{v_1, v_2, \dots, v_n\}$ can then be expressed as

$$v_j = \sum_{i=1}^n a_{ji} \phi(\mathbf{r}_i). \quad (10)$$

194 They allow extracting nonlinear features but they cannot be combined straightforwardly to generate new conformations. Instead,
 195 for generative purposes, we need to learn an inverse transform function that maps points in the l -dimensional representation
 196 space defined by the components back to the input space. This problem is known as the *pre-image problem*. To solve it, we used
 197 kernel ridge regression of the input coordinates R on their low-dimensional projections in the representation space as described
 198 in^{52,53} and implemented in the scikit-learn Python library⁵⁴. The contribution of the L2-norm regularisation is controlled
 199 through the hyperparameter α . More technically, α connects the squared L2-norm between a point in the representation space
 200 and its reconstruction with the squared L2-norm of the kernel weights used for the reconstruction.

201 **Leave-one-cluster-out cross-validation procedure**

202 We assessed the predictive performance of PCA and kPCA with a *leave-one-out* cross-validation procedure. Since the
 203 conformations are not evenly distributed within an ensemble, we grouped them into clusters prior to the evaluation. We
 204 performed the clustering in the l -dimensional PCA representation space, where l is the minimal number of linear components
 205 sufficient to explain 90% of the ensemble's total positional variance. We used the k -means clustering⁵⁵ with $k = l + 2$.

Given a clustered ensemble, we systematically tested the ability of the principal modes inferred from $l + 1$ clusters to predict the conformations belonging to the held-out cluster. We reconstructed each test conformation \mathbf{r}^* from its projection \mathbf{p}^* in the l -dimensional representation space. For the classical PCA, we computed the projection as,

$$\mathbf{p}^* = (\mathbf{r}^* - \bar{\mathbf{r}})V_l. \quad (11)$$

For the kPCA, the projection onto the principal component v_j is expressed as,

$$\phi(\mathbf{r}^*)v_j = \sum_{i=1}^n a_{ji} \phi(R)^T \phi(\mathbf{r}^*) = \sum_{i=1}^n a_{ji} K(R, \mathbf{r}^*). \quad (12)$$

206 We evaluated the reconstruction error as the RMSD between the predicted conformation $\mathbf{r}_{\text{pred}}^*$ and the original conformation \mathbf{r}^* .

207 **Distance to the training set**

208 We estimated the difficulty of reconstructing a given conformation by computing its distance to the convex hull defined by the
 209 conformations used for training in the l -dimensional representation space. Setting the number of clusters in the training set to
 210 $l + 1$ ensures that the convex hull will be a polytope of dimension at least l . For instance, in 1 dimension, we need at least 2
 211 affine-independent points to define a 1-polytope. The explicit computation of the convex hull of n points in l dimensions is
 212 an operation whose complexity is of the order of $O(n^{l/2})$ ⁵⁶ and rapidly becomes computationally infeasible as the value of
 213 l increases. Nevertheless, the calculation of the distance of a given point to the hull does not require computing the convex
 214 hull explicitly and is a much simpler computational problem. It can be solved in quasilinear time with quadratic programming
 215 (QP). Here, we used the efficient and exact QP simplex solver proposed in⁵⁷ and implemented in the Computational Geometry
 216 Algorithms Library (CGAL)⁵⁸. It takes advantage of the low dimensionality of the representation space by observing that the
 217 closest features of two l -polytopes are always determined by at most $l + 2$ points.

In order to compare distances across systems of different sizes, we scale them by the number of positions m ,

$$d^{\text{norm}} = \frac{d}{\sqrt{m}}. \quad (13)$$

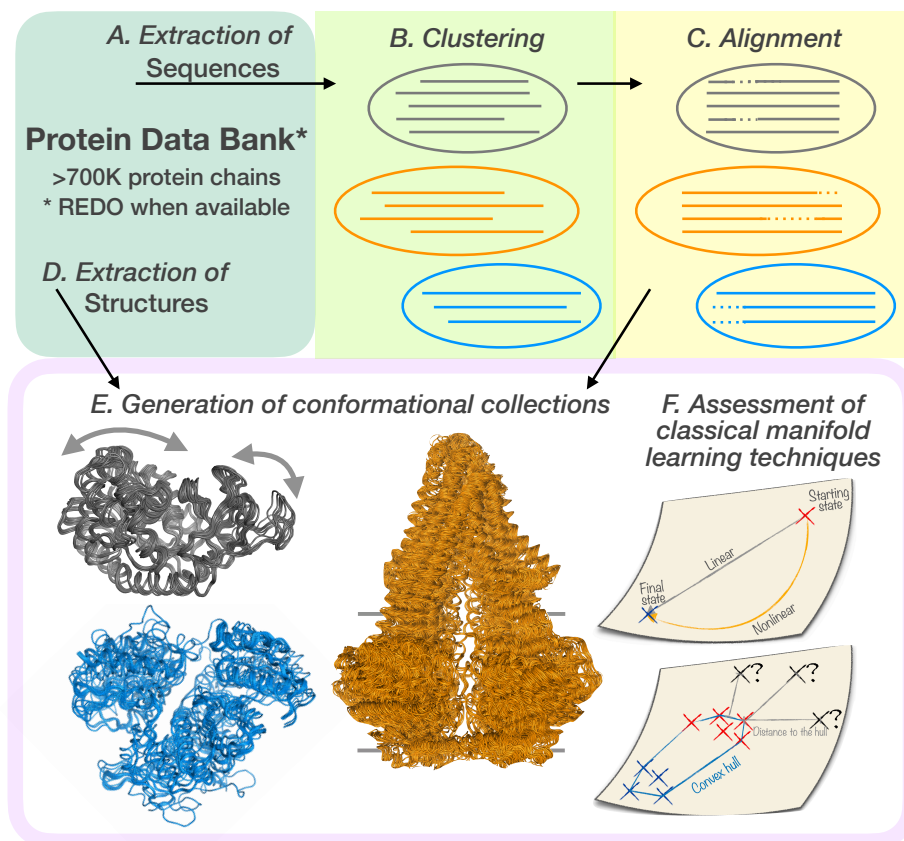


Figure 1. Outline of the study. Our approach, DANCE, exploits both amino acid sequences and 3D coordinates. We applied it to all experimentally determined protein-containing 3D structures from the PDB. Alternatively, users can provide a custom set of experimental structures or predicted models. DANCE first concentrates on sequences. It extracts them from the input structures (A) and clusters them with MMseqs2 based on user-defined similarity and coverage thresholds (B). For each cluster, It generates a multiple sequence alignment using MAFFT (C). It then extracts all 3D coordinates (D), groups the conformations according to the clusters identified in B and superimposes them to generate conformational ensembles (E). The superimposition aims at minimizing the Root Mean Square Deviation to a chosen reference, using the alignments produced by C for mapping the residues. The examples of the bacterial enzymes adenylate kinase (in grey, reference PDB code: 1AKEA) and MurD (in blue, 1E0DA), and the murine ABC transporter P-glycoprotein (5KOYB) are depicted. The arrows indicate adenylate kinase’s main motion. The horizontal lines behind the P-glycoprotein indicate the boundaries for the membrane bilayer. Finally, DANCE summarises conformational diversity through Principal Component Analysis (F). We further assessed the ability of classical manifold learning techniques to reconstruct and extrapolate conformations.

218 This normalisation also allows relating distances in the representation space with RMS deviations in the 3D Cartesian space.
 219 Indeed, let us consider an ensemble of conformations exhibiting a purely one-dimensional motion. Any two conformations
 220 distant by an RMSD of 1 Å in the original 3D space will be separated by a normalised distance of 1 Å in the one-dimensional
 221 representation space.

222 Results

223 We used DANCE to chart the experimentally resolved conformational diversity of protein families (Fig. 1). We explored eight
 224 levels of sequence similarity (*sim*) and coverage (*cov*), denoted as l_{cov}^{sim} , to group the ~750K chains included in the PDB as of
 225 June 2023 (Fig. S1A and Table S2). In the most conservative set up, namely l_{80}^{80} , less than 3% of the conformations remain

226 isolated (**Fig. S1A**, *singletons*). Most of the conformational collections (or ensembles) are associated with multiple sequence
227 alignments of high quality across all levels (**Fig. S1B**). Sequence identity and coverage are more widely distributed in more
228 relaxed conditions, but the median values always remain very high, above 0.95 (**Fig. S1C-D**).

229 **Experimentally resolved conformations lie on low-dimensional manifolds**

230 Only one or two linear principal components suffice to explain almost half of the ensembles' conformational diversity (**Fig.**
231 **2A**). We interpret these components as directions of motion, and by simplification, we will denote them as linear motions in the
232 following (see *Methods*). In the overwhelming majority of cases, less than eight linear motions explain more than 90% of the
233 total positional variance. These observations hold true across all sequence identity and coverage levels. They indicate that the
234 conformational states captured by experimental techniques for a protein or a protein family lie on a low-dimensional manifold.
235 This low dimensionality is only partially determined by the cardinality of the ensembles (**Fig. S2A-B**). Almost 30% of the most
236 highly populated ensembles (>50 conformations) detected at l_{80}^{80} can be comprehensively described with less than three linear
237 motions (**Fig. S2C**). This proportion increases up to 46% in the most relaxed conditions, namely at l_{50}^{30} (**Fig. S2D**).

238 The bacterial adenylate kinase gives an example of a one-dimensional motion underlying its 42 conformations (**Fig. 1E**,
239 in grey). One can easily classify the conformations by visual inspection into two main states, open and closed, deviating by
240 about 7 Å. The bacterial enzyme MurD (**Fig. 1E**, in blue) and the murine ABC transporter P-glycoprotein (**Fig. 1E**, in orange)
241 also exhibit low-dimensional opening-closing motions. In particular, the P-glycoprotein's collection reveals a rich spectrum of
242 intermediate conformations between the open and closed forms (**Fig. 1E**, in orange). The main motion involves about 70%
243 of the protein and modulates the volume of the transporter's internal cavity within the lipid bilayer up to over 6,000 Å³⁵⁹. It
244 explains about 80% of the total positional variance on its own. The remaining variability is mostly due to rotations of the
245 nucleotide binding domains with respect to the transmembrane helical bundles and to loop deformations.

246 **A few protein families display huge conformational expansion upon relaxing the sequence selection criteria**

248 To investigate how the conformational ensembles transformed with sequence similarity, we systematically backtracked the
249 18 616 representative protein chains identified at l_{50}^{30} across more stringent levels (see *Methods*). The fragment antigen-binding
250 regions display the largest growth between the most stringent and most relaxed sequence selection criteria (**Fig. 2**). For
251 instance, while the Fab6785 light chain's ensemble at l_{80}^{80} comprises a bit less than 300 conformations, it expands up to over
252 12 500 conformations at l_{50}^{30} (**Fig. 2B**, PDB id: 4QHUH). With the largest number of conformations at l_{80}^{80} , the HIV-1 capsid
253 protein's ensemble however displays a relatively limited expansion across the different levels, from 3 334 to 3 391 (**Fig. 2B**,
254 3J345). Bovine trypsin and its close homologs give an example of an extensively characterized subfamily, with 470 different
255 conformations detected at l_{80}^{80} . This ensemble expands by more than 5 folds, aggregating different serine proteases, upon
256 relaxing the criteria to l_{50}^{30} (**Fig. 2B**, PDB id: 1TAWA). Likewise, the Beta-2-microglobulin and its close homologs have a large
257 body of 1 465 conformations at l_{80}^{80} , growing further up to 2 025 conformations at l_{50}^{30} by including other immunoglobulins
258 (**Fig. 2B**, 7MX4B). By contrast, the reconstructed ancestral tyrosine kinase AS, a common ancestor of Src and Abl, has only 2
259 conformations available in the PDB and no close homologs. At l_{50}^{30} , it serves as representative for a huge ensemble of over
260 4 000 protein kinase conformations (**Fig. 2B**, 4UEUA). Apart from these over-represented protein families or superfamilies, the
261 ensembles generally gain only a few conformations, with a median value of 4.

262 **Family expansion may lead to an apparent motion simplification**

263 As an ensemble grows, the gained conformations may lie on the same motion manifold, defined by the subset of principal
264 components explaining the variance, or give rise to new motions represented by new components (**Fig. 2C**). The bacterial
265 long-chain flavodoxin exemplifies the second scenario (**Fig. 2D-F**, in black). At l_{80}^{80} , it undergoes a one-dimensional motion
266 describing the transition between a compact state and a partially unfolded conformation (**Fig. S3**). Upon relaxing sequence
267 similarity to l_{50}^{30} , the ensemble roughly doubles in size (**Fig. 2F**) and the newly added conformations exhibit complex
268 deformations of the FMN binding pocket. As a result, five more linear motions are required to explain the positional variance
269 (**Fig. 2D**). Hence, in this case, the motions get more complex when considering more distant homologs.

270 The emergence of new motions does not however systematically lead to an increased motion complexity. The murine MCL1
271 gives an illustrative example of apparent motion simplification upon expansion (**Fig. 2D-F**, in red, and **Fig. 2G**). At l_{80}^{80} , almost
272 30 components are needed to explain the variability observed over the couple of hundreds conformations in the ensemble. They
273 represent local deformations of the inter-helical loops and the extremities (**Fig. 2G** and **Fig. S3**). Extending the ensemble to
274 distant members of the Bcl-2 family brings in about 50 new conformations (**Fig. 2F**). They reveal a new extended state the
275 protein BAX adopts upon assembling into domain-swapped dimers⁶⁰. The large amplitude transition between the compact
276 conformation and the extended one takes a big part in the variance, resulting in a drastically reduced motion complexity (**Fig.**
277 **2D**). The benzaldehyde lyase BAL gives another example (**Fig. 2D-F**, in blue) where the transition to a new state, adopted

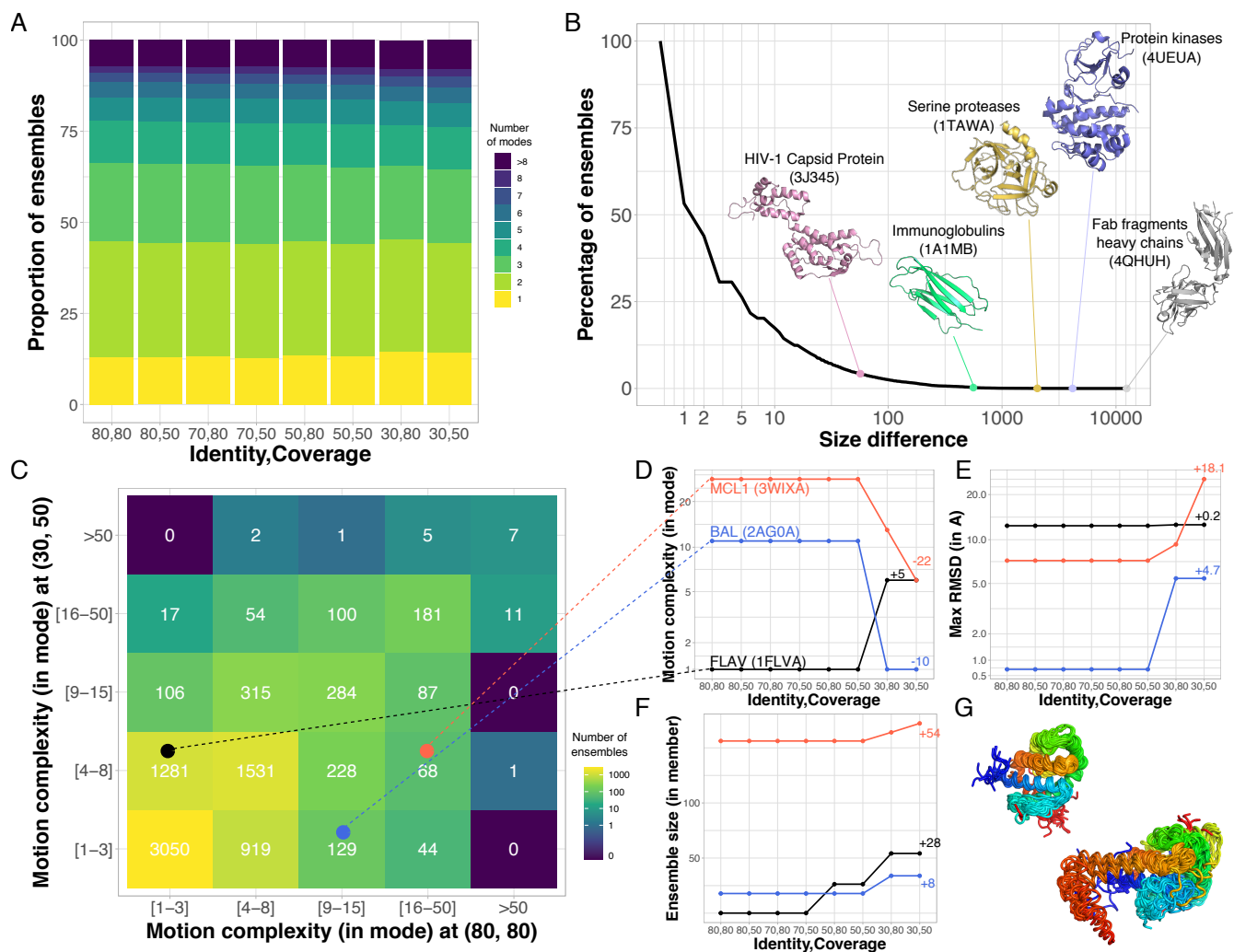


Figure 2. Evolution of protein conformational diversity across sequence similarity levels. **A.** Proportion of conformational ensembles requiring n linear PCA modes to explain 90% of their total positional variance, with n varying from 1 to 8. The number of modes n is an indicator of motion complexity. Singletons and pairs are excluded. **B.** Cumulative distribution of the number of conformations gained from the most stringent level, namely 1_{80}^{80} , with 80% sequence similarity and coverage, to the most permissive one, 1_{30}^{30} , with 30% similarity and 50% coverage. The 3D structures of the reference protein chains are depicted for a few ensembles. **C.** Comparison of motion complexity between the most stringent and most relaxed set ups. We considered only the cases where the ensemble at 1_{30}^{30} is bigger than the corresponding one at 1_{80}^{80} . Singletons and pairs are excluded. **D-G.** Detailed evolution of three ensembles marked by colored dots in panel C. **D.** Motion complexity expressed as a number of modes. The names and PDB codes of the reference chains are indicated. **E.** Motion amplitude, measured as the maximum RMSD between any two conformations (in Å). **F.** Conformational collection size. **G.** Conformational diversity observed for the Bcl-2 family. On the top left, the 54 conformations comprised in the MCL1 ensemble at 1_{80}^{80} . At the bottom right, the 218 additional conformations at 1_{30}^{30} . The color code indicates the position in the sequence, from the N-terminus in blue to the C-terminus in red.

278 by the distant homolog actinobacterial 2-hydroxyacyl-CoA lyase⁶¹, dominates the variance (**Fig. S3**). The conformational
279 variability transforms from small ($<1\text{\AA}$) seemingly random fluctuations to a one-dimensional motion.

280 Overall, about a third of the ensembles undergo an apparent motion simplification upon expansion (**Fig. 2C** and **Fig. S4A**).
281 They likely represent protein families where distant homologs exhibit novel distinct states. The larger the deviations of these
282 novel states with respect to the other ones, the higher the contribution of the corresponding motions to the variance. To mitigate
283 this variance-dependent effect, we repeated the analysis on the correlation matrix. The latter estimates the extent to which the
284 residues move in the same direction, regardless of the magnitude of their displacements. We found that the motion complexity
285 still decreases in over 20% of the ensembles (**Fig. S4B**). This result indicates that motion simplification does not merely reflect
286 larger transitions "hiding" smaller rearrangements. A substantial fraction of protein families show evidence of more concerted
287 residue movements between more distant homologs.

288 **Beyond single chains and sequence similarity, the ABC superfamily as a case study**

289 We explored the possibility of using DANCE to chart the conformational variability of remote homologs with low sequence
290 similarity and variable chain composition. We focused on the ABC (ATP Binding Cassette) transporter superfamily. The ABC
291 architecture comprises two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs) encoded by one or
292 several polypeptidic chains (**Fig. 3A**). The NBDs are highly conserved across species and families, whereas the TMDs exhibit
293 various scaffolds associated with heterogeneous transport functions²⁶. We considered a collection of a few hundreds ABC
294 protein experimental 3D structures²⁶, taking the single-chain murine P-glycoprotein as reference (**Fig. 3A**, 5KOYA).

295 We bypassed DANCE sequence extraction, clustering and alignment steps and directly gave it a pre-computed alignment
296 built from structural similarities as input (see *Methods*). Relying on structure rather than sequence similarity and considering
297 various oligomeric states provided a more comprehensive description of ABC transporters' functional motions and states (**Fig.**
298 **3** and **Movies S1-2**). The resulting ensemble comprises 188 conformations encompassing 295 protein chains, some of which
299 have sequence identity below 30% or coverage lower than 50% (**Fig. 3A**). A set of 25 linear motions are required to explain
300 the positional variance. By comparison, the sequence similarity-based 5KOYA-containing collection generated by DANCE at
301 l_{50}^{30} contains only 71 conformations explained by only four linear motions. These motions are essentially identical to those
302 extracted from the 61 conformations at l_{80}^{80} (**Fig. 3B**, RMSIP = 0.99).

303 Despite having different motion complexities, the sequence- and structure-based conformational collections have largely
304 overlapping motion subspaces (**Fig. 3B**, RMSIP ~ 0.7). In particular, they all share the same most contributing motion
305 describing the transition between the transporter inward-closed and inward-open forms (**Fig. S5**). This functional transition
306 controls the substrate access to the transporter's central binding pocket. It explains 45 to 70% of the variance on its own and
307 involves over two-thirds of the residues. The structure similarity-based collection represents a quasi-continuum of increasingly
308 open states (**Fig. 3C**, in blue, and **Movie S1**) between two extreme dimeric forms, one from the human lysosomal cobalamin
309 exporter ABCD4 where the two NBDs are in contact and the other from *Salmonella typhimurium*'s lipid A transporter MsbA
310 with a widely open cavity. The overwhelming majority of conformations are regularly spaced by inter-NBD distance increments
311 smaller than 1 \AA . By contrast, the sequence similarity-based collections populate sparse regions of this continuous transition,
312 with a high concentration of semi-open and open states (**Fig. 3C**, in pink and red, and **Movie S2**).

313 **Classical manifold learning techniques can generate highly accurate conformations**

314 Beyond describing the observed conformational variability, we evaluated the ability of two classical manifold learning
315 techniques, namely the linear PCA and the non-linear kernel PCA (kPCA), to generate unseen conformations. To do so, we
316 identified a set of ten conformational ensembles with very different degrees of motion complexity (**Fig. 4A** and **Table S3**). They
317 comprise between 20 and over 3 300 conformations and their reference chains contain 80 to 1 200 residues. They represent
318 proteins or protein families displaying substantial ($\geq 5\text{\AA}$) and functionally relevant conformational changes, namely adenylate
319 kinase (ADK)^{62,63}, MurD^{19,64}, the calcium pump ATPase^{65,66}, the ABC transporters^{26,67}, the small heat shock protein α B
320 crystallin (Crys)^{68,69}, the heat shock protein HSP90^{70,71}, calmodulin (CALM)^{72,73}, kinases (KIN)^{74,75}, RAS^{25,76}, and the
321 HIV capsid protein (CAP)^{77,78}. Most of them have been extensively characterized by experimental structure determination
322 techniques or computational methods for simulating protein dynamics. Targeting their motions or their specific conformations
323 bears a therapeutic interest.

324 Within each ensemble, we first learned low-dimensional representations of a subset of conformations used as training
325 samples. We then projected the test conformations, not seen during training, to the learned representation space, and mapped
326 the projections back to the original 3D Cartesian space. The mapping is determined analytically in the case of linear PCA
327 and learned in the case of kPCA (see *Methods*). We evaluated the quality of the 3D reconstructions by computing their RMS
328 deviations from the original conformations. We found that both PCA and kPCA yield some high-accuracy reconstructions, with
329 an RMSD error below 1.5 \AA , for all proteins (**Fig. 4B**). The error distribution width varies from one protein to another and does
330 not depend on motion complexity. For instance, all reconstructed conformations of HSP90 deviate by less than 2 \AA from the
331 original ones, while the reconstruction error can be as high as 8 \AA for MurD. The nonlinear kPCA performs significantly better

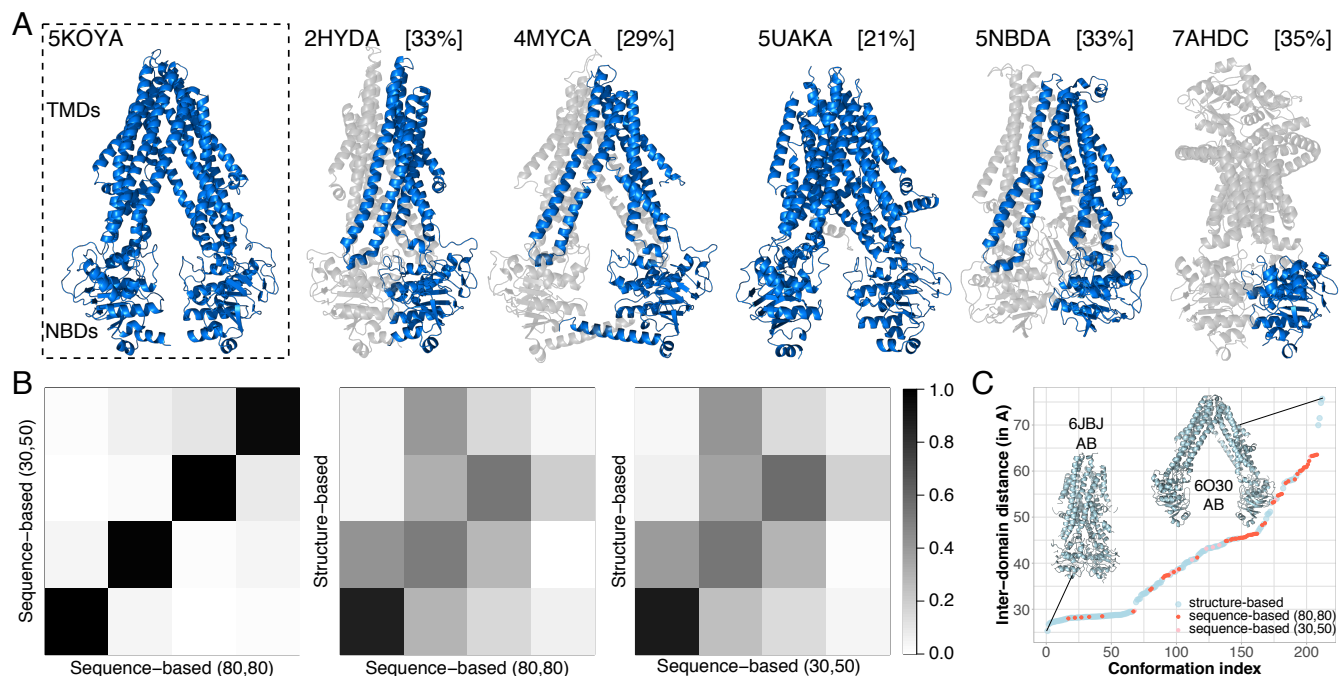


Figure 3. ABC transporters' conformational variability. **A.** Examples of protein structures from the ABC structure similarity-based conformational collection. The reference chain (5KOYA) is on the left, where we indicate the location of the two NBDs (~ 500 residues) and two TMDs (~ 700 residues). Within each of the other structures, we highlight one chain in marine, give its percentage of identity with the reference in squared brackets, and display the remaining chains in transparent grey. The six marine chains were assigned to six different collections by DANCE's default sequence similarity-based end-to-end protocol at l_{50}^{30} . **B.** Comparison of motion subspaces extracted from the sequence-based ensembles at l_{80}^{80} (61 conformations) and l_{50}^{30} (71 conformations) and the structure-based one (188 conformations). Each matrix shows the absolute pairwise scalar products computed for the first four PCA modes. The corresponding RMSIP are 0.99, 0.71 and 0.73. **C.** Distance between the geometric centres of the two NBDs (in Å). The conformations are ordered along the x -axis from the most closed one to the most open one.

332 than the linear PCA for all proteins from the benchmark. It allows increasing the percentage of high-quality reconstructions
333 (RMSD error < 1.5 Å) from 68 to 82% for MurD and from 1 to 33% for CALM (Table S4). Nevertheless, the reconstruction
334 accuracy of kPCA varies greatly depending on the values of the two hyperparameters controlling the kernel width and the
335 amount of regularisation (Fig. S6). The optimal values vary from one system to another and determining them *a priori* is not
336 trivial (Table S5). The applicability of non-linear techniques is thus limited by the choice of the adjustable parameters.

337 Reconstruction accuracy strongly depends on the distance to the training set

338 The quality of the predictions strongly correlates with the distance between the test conformation and the training set's convex
339 hull in the low-dimensional representation space (Fig. 4C). The linear PCA produces highly accurate reconstructions, with
340 an RMSD error smaller than 1.5 Å, only for conformations distant by less than 3 Å from the training set. We observed a
341 similar tendency for kPCA (Fig. S6). This dependence can be appreciated by visualising how the conformations cluster in the
342 representation space (Fig. 4D). For instance, the most poorly reconstructed MurD conformation forms a singleton located far
343 away from all other conformations, particularly along the first most important principal component (Fig. 4D, dark dot). For this
344 protein, the kPCA performs substantially better than the PCA thanks to a better reconstruction of the most populated cluster Fig.
345 4D, light squares). In addition, the overwhelming majority of conformations lie outside of the training set's convex hull. This
346 observation agrees with a recent study showing that interpolation almost surely never happens with high dimensional datasets⁷⁹.
347 The 14 conformations located inside come from ADK, CALM, KIN, RAS and CAP and are all very well reconstructed, with
348 RMSD errors ranging from 0.1 to 1.7 Å.

349 Influence of data uncertainty handling and conformation-specific centring

350 We assessed the influence of accounting for uncertainty in the data with position-wise weights and centring the data to a
351 reference conformation (Fig. S8-11). In principle, both operations may impact the conformations' superimposition and, as a
352 consequence, their final coordinates, as well as the extracted motions (see *Methods*). In practice, 95% of the ~35 000 ensembles
353 at 1₈₀⁸⁰ – excluding singletons and pairs, are not substantially altered by introducing position-wise uncertainty weights (Fig. S8).
354 They display the same displacement amplitude (± 1 Å) and motion complexity (± 1 mode). When the weights are impactful,
355 they effectively lower the importance of large deviations in uncertain regions, *i.e.*, poorly covered by the conformations, and
356 prevent the associated motions, typically highly localised, from dominating the variance (Fig. S8, red dots). Hence, the
357 uncertainty weights tend to induce smaller deviations (Fig. S8A), increased motion complexities (Fig. S8B), and less dominant
358 and more collective main motions (Fig. S8C-D).

359 The choice of the reference conformation used for superimposing and centring the 3D coordinates has a much stronger
360 influence (Fig. S9). Only 43% of the 1₈₀⁸⁰ ensembles remain unaffected upon changing the reference. In this experiment, the
361 first reference is the multiple sequence alignment consensus (see *Methods*), while the second reference maximises the RMS
362 deviation from the first one. We expect this setup to yield the most contrasted resFig. S7ults. It almost never happened that
363 an ensemble consistently displayed a high motion complexity or a weakly contributing main motion for both references (Fig.
364 S9B-C). This result suggests that the ensembles exhibiting complex conformational rearrangements (e.g., loop deformations)
365 among a bulk of conformations also include a few conformations comparatively far from all the others. The motions simplify
366 when performing the PCA from the perspective of this minority. Normalising out the variance to focus on inter-residue
367 correlations attenuates this effect (Fig. S10).

368 Discussion

369 This work proposes a new perspective on the variability of protein 3D conformations. It provides the community with
370 conformational collections representing the multiple protein states available in the PDB and a fully automated versatile
371 computational pipeline to build custom collections. In doing so, it contributes to the representation and managing of multiple
372 conformational models of proteins. It enhances access and understanding of protein functional states and motions and facilitates
373 predictive methods benchmarking. Both DANCE pipeline and the produced PDB-wide data are readily usable in other studies.

374 We chose to rely on classical principal component analysis because of its intuitive geometrical interpretation. It allows
375 describing protein conformational variability with a limited set of orthogonal vectors interpretable as linear motions. We
376 provided estimates of motion complexity as the number of PCA components necessary to explain most of the observed
377 conformational variability. We found that a few linear motions suffice to explain most conformational collections. The high
378 complexity exhibited by a few protein families may reflect nonlinear structural deformations or seemingly random fluctuations.
379 For instance, protein kinases exhibit highly complex loop conformational rearrangements despite a well-conserved overall fold
380 and only two metastable functional states. Our analysis helps to identify such cases to prioritise their in-depth characterisation
381 with more sophisticated nonlinear dimensionality reduction techniques.

382 We designed DANCE for dealing primarily with single polypeptidic chains grouped based on sequence similarity. To
383 go further, we have provided a proof-of-concept application study of DANCE's usefulness for comprehensively describing

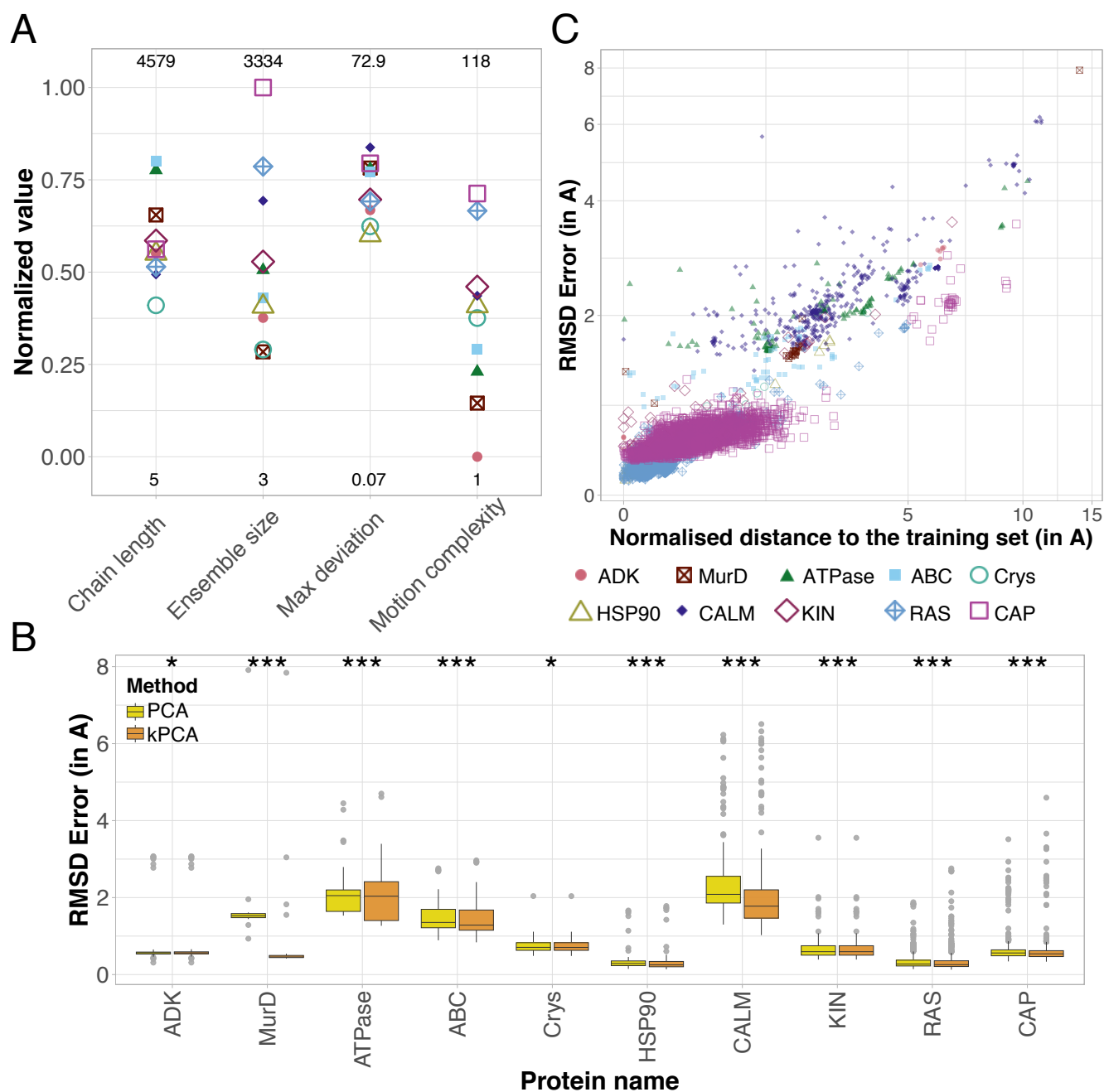


Figure 4. Assessment of classical manifold learning techniques. **A.** Properties of the benchmark set. For each property y , we computed its normalised value as $\frac{\log(y) - \min(\log(y))}{\max(\log(y)) - \min(\log(y))}$. The minimum and maximum are determined over the whole 180^2 database. They are given at the bottom and on the top, respectively. **B.** Distributions of the RMSD reconstruction errors (in Å) for each ensemble in the benchmark set. We systematically reconstructed each conformation through a leave-one-cluster-out cross-validation procedure (see *Methods*). We set the two hyper-parameters of the kPCA to the values yielding the best reconstruction, for each ensemble. The protein names in the x -axis are ordered according to motion complexity. The stars indicate the statistical significance of the better performance of kPCA compared to linear PCA (one-sided paired t-test; *: p -val $< 1e^{-2}$; ***: p -val $< 1e^{-5}$). **C.** RMSD reconstruction error in function of the distance to the training set's convex hull in the PCA representation space.

384 continuous motions shared across very distant homologs comprising different numbers of chains. We showed that ABC proteins
385 with a wide diversity of substrates and transport mechanisms share a highly collective high amplitude opening/closing motion
386 underlying their functioning. In addition, our work goes beyond a descriptive analysis by showing that classical manifold
387 learning techniques can generate plausible conformations in the vicinity of the training set. Our results can serve as baselines
388 for evaluating more sophisticated approaches.

389 DANCE superimposes the conformations onto representative references and describes conformational variability as a set
390 of linear motions of these references. This approach offers a multi-view perspective on a given collection of conformations,
391 easing interpretability and allowing for augmenting data in a learning context. Nevertheless, radical differences between
392 conformations, such as fold changes, might confound the superimposition. Another limitation comes from the dependency of
393 the superimposition on the multiple sequence alignment heuristic. Ambiguities arising from sequence similarities might result
394 in suboptimal 3D coordinates matching and, thus, in large deviations. Future improvements will explore multi-reference or
395 reference-free probabilistic frameworks and more refined accounts of data uncertainty^{80–84}.

396 Data availability

397 We provide public access to the conformational collections compiled by DANCE from the PDB at two levels of sequence
398 similarity, namely I_{80}^{80} and I_{50}^{30} on Figshare. This repository also contains the structural similarity-based ABC transporter
399 conformational collection along with the supplementary **Movies S1** and **S2**. In addition, we provide detailed information about
400 the benchmark set and the assessment of PCA and kPCA.

401 Code availability

402 DANCE source codes are written in C/C++ and Python and are publicly available on GitHub at <https://github.com/PhyloSofS-Team/DANCE>. This repository also contains a Python wrapper allowing users to seamlessly run DANCE full
403 pipeline. In addition, we provide example input 3D structures.
404

405 References

- 406 1. Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531,
407 [10.1093/nar/gkac1052](https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf) (2022). <https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf>.
- 408 2. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence
409 space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444, [10.1093/nar/gkab1061](https://academic.oup.com/nar/article-pdf/50/D1/D439/43502749/gkab1061.pdf) (2021). <https://academic.oup.com/nar/article-pdf/50/D1/D439/43502749/gkab1061.pdf>.
- 411 3. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids*
412 *Res.* **34**, D187–D191, [10.1093/nar/gkj161](https://academic.oup.com/nar/article-pdf/34/suppl_1/D187/3926611/gkj161.pdf) (2006). https://academic.oup.com/nar/article-pdf/34/suppl_1/D187/3926611/gkj161.pdf.
- 414 4. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
- 415 5. Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **20**, 170–173 (2023).
- 416 6. Miller, M. D. & Phillips, G. N. Moving beyond static snapshots: Protein dynamics and the protein data bank. *J. Biol.*
417 *Chem.* **296** (2021).
- 418 7. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
- 419 8. Kryshchak, A. *et al.* Breaking the conformational ensemble barrier: Ensemble structure modeling challenges in casp15.
420 *Proteins: Struct. Funct. Bioinforma.* **91**, 1903–1911 (2023).
- 421 9. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589, [10.1038/](https://doi.org/10.1038/s41586-021-03819-2)
422 [s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2) (2021).
- 423 10. Johansson-Åkhe, I. & Wallner, B. Improving peptide-protein docking with alphafold-multimer using forced sampling.
424 *Front. Bioinforma.* **2**, 85 (2022).
- 425 11. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and alphafold2. *Nature* 1–3
426 (2023).
- 427 12. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and
428 receptors with alphafold2. *Elife* **11**, e75751 (2022).
- 429 13. Faezov, B. & Dunbrack Jr, R. L. Alphafold2 models of the active form of all 437 catalytically-competent typical human
430 kinase domains. *bioRxiv* 2023–07 (2023).

- 431 **14.** Heo, L. & Feig, M. Multi-state modeling of g-protein coupled receptors at experimental accuracy. *Proteins: Struct. Funct. Bioinforma.* **90**, 1873–1885 (2022).
- 432
- 433 **15.** Chakravarty, D., Schafer, J. W., Chen, E. A., Thole, J. & Porter, L. Alphafold2 has more to learn about protein energy landscapes. *bioRxiv* 2023–12 (2023).
- 434
- 435 **16.** Chakravarty, D. & Porter, L. L. Alphafold2 fails to predict protein fold switching. *Protein Sci.* **31**, e4353 (2022).
- 436 **17.** Jing, B. *et al.* Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198* (2023).
- 437
- 438 **18.** Zheng, S. *et al.* Towards predicting equilibrium distributions for molecular systems with deep learning, [10.48550/ARXIV.2306.05445](https://doi.org/10.48550/ARXIV.2306.05445) (2023).
- 439
- 440 **19.** Ramaswamy, V. K., Musson, S. C., Willcocks, C. G. & Degiacomi, M. T. Deep learning protein conformational space with convolutions and latent interpolations. *Phys. Rev. X* **11**, 011052 (2021).
- 441
- 442 **20.** Ramelot, T. A., Tejero, R. & Montelione, G. T. Representing structures of the multiple conformational states of proteins. *Curr. Opin. Struct. Biol.* **83**, 102703 (2023).
- 443
- 444 **21.** Wankowicz, S. & Fraser, J. Comprehensive encoding of conformational and compositional protein structural ensembles through mmcif data structure. *ChemRxiv* [10.26434/chemrxiv-2023-ggd1w-v2](https://doi.org/10.26434/chemrxiv-2023-ggd1w-v2) (2023).
- 445
- 446 **22.** Ellaway, J. I. *et al.* Identifying protein conformational states in the pdb and comparison to alphafold2 predictions. *bioRxiv* 2023–07 (2023).
- 447
- 448 **23.** Varadi, M. *et al.* PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. *Protein Sci.* **31**, [10.1002/pro.4439](https://doi.org/10.1002/pro.4439) (2022).
- 449
- 450 **24.** Modi, V. & Dunbrack Jr, R. L. Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Res.* **50**, D654–D664 (2022).
- 451
- 452 **25.** Parker, M. I., Meyer, J. E., Golemis, E. A. & Dunbrack Jr, R. L. Delineating the RAS conformational landscape. *Cancer research* **82**, 2485–2498 (2022).
- 453
- 454 **26.** Tordai, H. *et al.* Comprehensive collection and prediction of abc transmembrane protein structures in the ai era of structural biology. *Int. J. Mol. Sci.* **23**, 8877 (2022).
- 455
- 456 **27.** Pándy-Szekeres, G. *et al.* Gpcrdb in 2023: state-specific structure models using alphafold2 and new ligand resources. *Nucleic Acids Res.* **51**, D395–D402 (2023).
- 457
- 458 **28.** Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. transactions royal society A: Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
- 459
- 460 **29.** Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin philosophical magazine journal science* **2**, 559–572 (1901).
- 461
- 462 **30.** Amadei, A., Linssen, A. B. & Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct. Funct. Bioinforma.* **17**, 412–425 (1993).
- 463
- 464 **31.** Maity, A., Majumdar, S. & Dastidar, S. G. Flexibility enables to discriminate between ligands: Lessons from structural ensembles of bcl-xl and mcl-1. *Comput. Biol. Chem.* **77**, 17–27 (2018).
- 465
- 466 **32.** Yao, X.-Q. *et al.* Navigating the conformational landscape of g protein-coupled receptor kinases during allosteric activation. *J. Biol. Chem.* **292**, 16032–16043 (2017).
- 467
- 468 **33.** Bakan, A. & Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci.* **106**, 14349–14354 (2009).
- 469
- 470 **34.** Yang, L., Song, G., Carriquiry, A. & Jernigan, R. L. Close correspondence between the motions from principal component analysis of multiple hiv-1 protease structures and elastic network modes. *Structure* **16**, 321–330 (2008).
- 471
- 472 **35.** Mestres, J. Structure conservation in cytochromes p450. *Proteins: Struct. Funct. Bioinforma.* **58**, 596–609 (2005).
- 473 **36.** Van Aalten, D. *et al.* Protein dynamics derived from clusters of crystal structures. *Biophys. J.* **73**, 2891–2896 (1997).
- 474 **37.** Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028, [10.1038/nbt.3988](https://doi.org/10.1038/nbt.3988) (2017).
- 475
- 476 **38.** Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. biology evolution* **30**, 772–780 (2013).
- 477

- 478 **39.** Theobald, D. L. Rapid calculation of rmsds using a quaternion-based characteristic polynomial. *Acta Crystallogr. Sect. A: Foundations Crystallogr.* **61**, 478–480 (2005).
479
- 480 **40.** Liu, P., Agrafiotis, D. K. & Theobald, D. L. Fast determination of the optimal rotational matrix for macromolecular
481 superpositions. *J. computational chemistry* **31**, 1561–1563 (2010).
- 482 **41.** Brüschweiler, R. Collective protein dynamics and nuclear spin relaxation. *The J. Chem. Phys.* **102**, 3396–3403 (1995).
- 483 **42.** Tama, F. & Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**,
484 1–6 (2001).
- 485 **43.** Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **32**, 922–923,
486 [10.1107/S0567739476001873](https://doi.org/10.1107/S0567739476001873) (1976).
- 487 **44.** Wojdyr, M. Gemmi: A library for structural biology. *J. Open Source Softw.* **7**, 4200, [10.21105/joss.04200](https://doi.org/10.21105/joss.04200) (2022).
- 488 **45.** Harris, C. R. *et al.* Array programming with numpy. *Nature* **585**, 357–362, [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2) (2020).
- 489 **46.** Burley, S. K. *et al.* RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules
490 for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and
491 energy sciences. *Nucleic Acids Res.* **49**, D437–D451, [10.1093/nar/gkaa1038](https://doi.org/10.1093/nar/gkaa1038) (2020). [https://academic.oup.com/nar/
492 article-pdf/49/D1/D437/35364241/gkaa1038.pdf](https://academic.oup.com/nar/article-pdf/49/D1/D437/35364241/gkaa1038.pdf).
- 493 **47.** Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The pdb_redo server for macromolecular structure model
494 optimization. *IUCrJ* **1**, 213–220 (2014).
- 495 **48.** van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *Biorxiv* 2022–02 (2022).
- 496 **49.** Skjærven, L., Yao, X.-Q., Scarabelli, G. & Grant, B. J. Integrating protein structural dynamics and evolutionary analysis
497 with bio3d. *BMC bioinformatics* **15**, 1–11 (2014).
- 498 **50.** Amadei, A., Ceruso, M. A. & Di Nola, A. On the convergence of the conformational coordinates basis set obtained by
499 the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct. Funct. Bioinforma.* **36**,
500 419–424 (1999).
- 501 **51.** Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*
502 **10**, 1299–1319, [10.1162/089976698300017467](https://doi.org/10.1162/089976698300017467) (1998).
- 503 **52.** Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A. & Schölkopf, B. Kernel dependency estimation. In Becker, S., Thrun, S.
504 & Obermayer, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 15 (MIT Press, 2002).
- 505 **53.** Weston, J., Schölkopf, B. & Bakir, G. Learning to find pre-images. In Thrun, S., Saul, L. & Schölkopf, B. (eds.) *Advances
506 in Neural Information Processing Systems*, vol. 16 (MIT Press, 2003).
- 507 **54.** Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- 508 **55.** Hartigan, J. A. & Wong, M. A. Algorithm as 136: A k-means clustering algorithm. *J. Royal Stat. Soc. Ser. C (Applied Stat.*
509 **28**, 100–108 (1979).
- 510 **56.** Chazelle, B. An optimal convex hull algorithm in any fixed dimension. *Discret. & Comput. Geom.* **10**, 377–409,
511 [10.1007/BF02573985](https://doi.org/10.1007/BF02573985) (1993).
- 512 **57.** Gärtner, B. & Schönherr, S. An efficient, exact, and generic quadratic programming solver for geometric optimization. In
513 *Proceedings of the Sixteenth Annual Symposium on Computational Geometry*, SCG '00, 110–118, [10.1145/336154.336191](https://doi.org/10.1145/336154.336191)
514 (Association for Computing Machinery, New York, NY, USA, 2000).
- 515 **58.** The CGAL Project. *CGAL User and Reference Manual* (CGAL Editorial Board, 2023), 5.6 edn.
- 516 **59.** Aller, S. G. *et al.* Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* **323**,
517 1718–1722 (2009).
- 518 **60.** Czabotar, P. E. *et al.* Bax crystal structures reveal how BH3 domains activate Bax and nucleate its oligomerization to
519 induce apoptosis. *Cell* **152**, 519–531, [10.1016/j.cell.2012.12.031](https://doi.org/10.1016/j.cell.2012.12.031) (2013).
- 520 **61.** Zahn, M. *et al.* Mechanistic details of the actinobacterial lyase-catalyzed degradation reaction of 2-hydroxyisobutyryl-coa.
521 *J. Biol. Chem.* **298** (2022).
- 522 **62.** Müller, C., Schlauderer, G., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: an energetic
523 counterweight balancing substrate binding. *Structure* **4**, 147–156 (1996).
- 524 **63.** Whitford, P. C., Miyashita, O., Levy, Y. & Onuchic, J. N. Conformational transitions of adenylate kinase: switching by
525 cracking. *J. molecular biology* **366**, 1661–1671 (2007).

- 526 **64.** Perdih, A., Kotnik, M., Hodoscek, M. & Solmajer, T. Targeted molecular dynamics simulation studies of binding and
527 conformational changes in e. coli murd. *PROTEINS: Struct. Funct. Bioinforma.* **68**, 243–254 (2007).
- 528 **65.** Stokes, D. L. & Green, N. M. Structure and function of the calcium pump. *Annu. Rev. Biophys. Biomol. Struct.* **32**,
529 445–468 (2003).
- 530 **66.** Kabashima, Y., Ogawa, H., Nakajima, R. & Toyoshima, C. What atp binding does to the ca²⁺ pump and how nonproductive
531 phosphoryl transfer is prevented in the absence of ca²⁺. *Proc. Natl. Acad. Sci.* **117**, 18448–18458 (2020).
- 532 **67.** Hopfner, K.-P. Invited review: Architectures and mechanisms of atp binding cassette proteins. *Biopolymers* **105**, 492–504
533 (2016).
- 534 **68.** De Jong, W. W., Leunissen, J. A. & Voorter, C. Evolution of the alpha-crystallin/small heat-shock protein family. *Mol.*
535 *biology evolution* **10**, 103–126 (1993).
- 536 **69.** Basha, E., O’Neill, H. & Vierling, E. Small heat shock proteins and α -crystallins: dynamic proteins with flexible functions.
537 *Trends biochemical sciences* **37**, 106–117 (2012).
- 538 **70.** Krukenberg, K. A., Street, T. O., Lavery, L. A. & Agard, D. A. Conformational dynamics of the molecular chaperone
539 hsp90. *Q. reviews biophysics* **44**, 229–255 (2011).
- 540 **71.** Li, J., Soroka, J. & Buchner, J. The hsp90 chaperone machinery: conformational dynamics and regulation by co-chaperones.
541 *Biochimica et Biophys. Acta (BBA)-Molecular Cell Res.* **1823**, 624–635 (2012).
- 542 **72.** Chin, D. & Means, A. R. Calmodulin: a prototypical calcium sensor. *Trends cell biology* **10**, 322–328 (2000).
- 543 **73.** Zhang, M., Tanaka, T. & Ikura, M. Calcium-induced conformational transition revealed by the solution structure of apo
544 calmodulin. *Nat. structural biology* **2**, 758–767 (1995).
- 545 **74.** Kornev, A. P. & Taylor, S. S. Dynamics-driven allostery in protein kinases. *Trends biochemical sciences* **40**, 628–647
546 (2015).
- 547 **75.** Modi, V. & Dunbrack Jr, R. L. Defining a new nomenclature for the structures of active and inactive kinases. *Proc. Natl.*
548 *Acad. Sci.* **116**, 6818–6827 (2019).
- 549 **76.** Simanshu, D. K., Nissley, D. V. & McCormick, F. Ras proteins and their regulators in human disease. *Cell* **170**, 17–33
550 (2017).
- 551 **77.** Sundquist, W. I. & Kräusslich, H.-G. Hiv-1 assembly, budding, and maturation. *Cold Spring Harb. perspectives medicine*
552 a006924 (2012).
- 553 **78.** Zhao, G. *et al.* Mature hiv-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497**,
554 643–646, [10.1038/nature12162](https://doi.org/10.1038/nature12162) (2013).
- 555 **79.** Balestrieri, R., Pesenti, J. & LeCun, Y. Learning in high dimension always amounts to extrapolation. *arXiv preprint*
556 *arXiv:2110.09485* (2021).
- 557 **80.** Ghosh, S. & Rigollet, P. Sparse multi-reference alignment: Phase retrieval, uniform uncertainty principles and the beltway
558 problem. *Foundations Comput. Math.* 1–48 (2022).
- 559 **81.** Bandeira, A. S. *et al.* Estimation under group actions: recovering orbits from invariants. *Appl. Comput. Harmon. Analysis*
560 (2023).
- 561 **82.** Abas, A., Bendory, T. & Sharon, N. The generalized method of moments for multi-reference alignment. *IEEE Transactions*
562 *on Signal Process.* **70**, 1377–1388 (2022).
- 563 **83.** Theobald, D. L. & Steindel, P. A. Optimal simultaneous superpositioning of multiple structures with missing data.
564 *Bioinformatics* **28**, 1972–1979 (2012).
- 565 **84.** Bandeira, A. S., Niles-Weed, J. & Rigollet, P. Optimal rates of estimation for multi-reference alignment. *Math. Stat. Learn.*
566 **2**, 25–75 (2020).

567 Acknowledgements

568 We are grateful to Juliana Bernardes, Pablo Chacon, Tamas Hegedus, Anatoli Juditsky, and the Elixir 3D-Bioinfo Community
569 members for insightful discussions and feedback. The Sorbonne Center for Artificial Intelligence (SCAI) provided a salary to
570 VL and computational resources. This work has also been partially supported by the European Research Council under the
571 European Union’s H2020 Framework Programme (2023–2028)/ ERC Grant agreement ID 101087830 awarded to EL.

572 **Author contributions statement**

573 S.G. and E.L. designed research. V.L. and S.G. carried out the implementation. V.L., E.L. and S.G. produced and analysed the
574 results. E.L. wrote the manuscript with support and feedback from all authors. S.G. and E.L. supervised the project.

575 **Competing interests**

576 The author(s) declare no competing interests.