ANA, accurate analysis of changes of protein cavity volumes in predefined directions and flexibility

German P. Barletta§*, Matias Barletta†, Tadeo E. Saldaño§, and Sebastian Fernandez-Alberti§.
§ Universidad Nacional de Quilmes/CONICET, Roque Sáenz Peña 352, B1876BXD Bernal, Argentina.
† Lionix Evolve, Costa Rica, http://www.lionix.com/

* To whom correspondence should be addressed.

Abstract

Dynamics of protein cavities associated with protein fluctuations and conformational plasticity are essential for their biological function. NMR ensembles, Molecular Dynamics (MD) simulations combined with Principal Component Analysis (PCA), and Normal Mode Analysis (NMA) provide appropriate frameworks to explore functionally relevant protein dynamics and cavity changes relationships. Herein, we present ANA (Analysis of Null Areas), an accurate and efficient software to define and calculate changes of cavity volumes within ensembles of protein conformations. For this purpose, ANA uses an efficient combination of algorithms that results robust to numerical differentiations, allowing the quantification of changes in cavity volumes and flexibility due to protein structural distortions performed on predefined biologically relevant directions, e.g, directions of largest contribution to protein fluctuations (PCA modes) obtained by MD simulations or ensembles of NMR structures, collective NMA modes or any other direction of motion associated with specific conformational changes. We found that changes in the cavity volume and flexibility can contribute to differentiate functional conformers within the protein native state. Besides, we have explored the gradual changes of cavity volume and flexibility associated with protein ligand binding by analyzing sets of previously validated structures that connect the apo and holo conformations. Finally, a comparison study of the extent of variability between protein backbone structural distortions, and changes in cavity volumes and flexibilities within an ensemble of NMR structures is shown.

I. Introduction

Cavity rearrangements due to protein thermal fluctuations and conformational changes play key roles in protein functions. Native state of proteins, represented by an ensemble of conformers in dynamics equilibrium, frequently involves changes in cavity volumes with subsequent impacts on ligand affinity and specificity. Despite the main role of atomic packing in protein stability [1,2], cavities provide the additional conformational flexibility for their biological function. Perturbations of cavities with homologous ligands can induce discrete conformational changes that lead to conformations lower in energy relative to the ground state [3]. Besides, cavities have been proposed as sources of conformational transitions to high-energy states in equilibrium with the ground state ensemble [4]. That is, despite that protein cavities can be thought of as destabilizing packing defects, they can play important roles in modulating molecular flexibility and, therefore, in facilitating functionally important protein motions.[5] On the contrary, cavity-filling mutations lowers flexibility of the binding protein, possibly reducing conformational entropic penalty of binding and, therefore, generating high affinity variants of a binding protein. [6]

The connection between protein fluctuations and cavity changes has been the subject of a large variety of previous works [7,8]. Most of them are focused on cavity rearrangements during ligand binding and the effect of mutations on cavity shapes and sizes and their ultimate impact on ligand affinity. The flexibility of cavities are commonly investigated by studying changes among ensembles of structures from sources such as molecular dynamics snapshots, crystallography under different conditions, NMR ensembles, and homologous protein structures. On one hand, databases of conformational diversity in the native state of proteins (CoDNaS [9] ; PDBFlex [10]) provide non-redundant collections of three-dimensional structures for the same sequence where each structure could be taken as a snapshot of the conformational ensemble of the protein. In this way, variations of cavities among conformers have been revealed as part of conformational mechanisms related with biological functions and shared by subsets of proteins with common flexibility features, degree of conformational diversity and disordered regions. [11] On the other hand, numerous computational methods have been developed to provide extended explorations of the conformational space of proteins [12]. These methods are mainly based on Molecular Dynamics (MD) techniques including different accelerated conformational sampling techniques [13-15] and Normal modes analysis (NMA) using elastic network models (ENM) [16–20]. Cavity volumes of the collected set of structures can be calculated using a variety of available softwares [21-23] and information concerning cavity rearrangements due to protein structural distortions associated with specific protein conformational transitions can be analyzed.

MD simulations combined with Principal Component Analysis (PCA), and NMA not only provide ensembles of protein structures that can be subjected to subsequent structural analysis, but also they enlighten us with the main directions of protein motions that are closely related to the protein's biological function. Within their framework, the complexity of protein dynamics can be decomposed into decoupled individual contributions that can be ultimately associated to specific cavity changes [24–26]. The analysis of variations of cavity volumes in these preselected directions of motions can reveal the connections between protein structural dynamics, critically tied to their biological function, and concomitant cavity changes with direct impact on ligand affinities or enzymatic catalytic capacities. The evaluation of changes of cavity volumes in the direction of specific PCA or NMA collective coordinates requires an efficient software to calculate partial derivatives of the cavity volumes with respect to these predefined coordinates. Nevertheless, currently available softwares for the evaluation of cavity volumes do not result robusts for numerical differentiations [27]. In order to carry out this task, we present ANA (Analysis of Null Areas), an accurate and efficient software for cavity volume calculations that not only represents an alternative for fast tracking of cavity volumes along large conformational samplings provided by NMA and/or snapshots obtained during MD trajectories, but also results

robust for numerical differentiations. That is, ANA has been developed to be sensitive enough to detect changes in the cavity volumes due to relatively small structural distortions and localized/delocalized rearrangements of residues lining their surfaces. Furthermore, ANA allows the evaluation of changes of the cavity volume with respect to the change of the predefined coordinates like PCA or NMA modes or any other direction of motion associated to a specific conformational change. The robustness of ANA for numerical differentiations allows us to characterize the dynamics of a protein cavity in terms of its Volume Gradient Vector, that is, its direction of maximum change, and to quantify its flexibility in terms of the variation of the potential energy in this direction.

II. Methods

A. ANA Overview

ANA (Analysis of Null Areas) has been developed to deal with large ensembles of protein conformations, provided either experimentally (NMR, X-ray crystallography) or theoretically (MD simulations, and NMA among others). It is particularly suitable to calculate partial derivatives of the cavity volumes with respect to predefined directions of motion (e.g. PCA and NMA modes or any other direction of motion associated to a specific conformational change). For this purpose, ANA makes use of a combination of algorithms that can be summarized as follows:

***Cavity definition, Included Area and Convex Hull*.** ANA defines the Included Area (IA), that is, the area that lines a specific cavity volume, by using points on atoms corresponding to a list of residues provided by the user. The Convex Hull (CH) of these points, $S \in R^d$, is defined as the smallest convex set in R^d containing S.[28] The algorithm yields the IA as a convex polyhedron fixed on the molecule subject to updates at each new structure of a given ensemble (see Figure 2). ANA will search for cavities inside this convex polyhedron.

Delaunay Triangulation and Cavity Volume. After IA calculation, ANA performs a Delaunay Triangulation using the CGAL [29,30] libraries. The Delaunay triangulation of a set of points $S \in R^3$ is defined as a triangulation such that no point in S is inside the circumsphere of any tetrahedron of the triangulation. This results in a set of empty tetrahedrons which are used to obtain the volume of the selected cavity of the protein. After the tessellation, ANA discards those tetrahedrons that do not intersect the IA. At this point, ANA supports two levels of precision upon user request. At the Low Precision (LP) level, the cavity volume is calculated considering the volumes of the whole intersecting tetrahedrons. Only tetrahedrons with volume larger than the volume of a spherical probe with radius equal to certain threshold (1.4 Å as default) and with at least one vertex inside the Convex Hull are retained. Besides, the volumes of spheres centered on each atom and radius equal to their corresponding van der Waals radius are subtracted to obtain the final volume of the cavity. Nevertheless, if a High Precision (HP) level is required, further refinements of the cavity volume are performed by evaluating the intersections between these tetrahedrons and the IA. Only the section of the tetrahedrons that lines inside the IA is kept at the HP level. A comparison between the application of LP and HP levels is displayed in **Figure S1**.

Non-Delaunay Displacements and changes of cavity volumes in predefined directions. ANA has been particularly developed to be robust for numerical differentiations. That is, it is sensitive enough to evaluate changes in the cavity volumes due to differential structural displacements in predefined directions like PCA or NMA modes. A structural displacement along a predefined direction is evaluated by displacing the set of tetrahedrons previously selected during the Delaunay triangulation. We call this procedure *Non-Delaunay Displacements (NDD)* since the new set of displaced tetrahedrons does not satisfy anymore the Delaunay conditions. NDD avoids discontinuities in the calculation of the cavity volume changes in predefined directions of motion. Because of that, ANA provides numerically stable gradients, suitable to explore protein fluctuations−cavity changes relationships[27], calculated using finite differences.

Volume Gradient Vector and cavity flexibility. The robustness of ANA for the calculation of partial derivatives allows the calculation of the volume gradient vector (VGV) of a protein cavity, that is, its direction of maximum change. ANA allows to obtain the VGV of a protein cavity in terms of Cartesian coordinates$\{x_i, y_i, z_i\}$or PCA or NMA modes $\{Q_i\}$ as

$$VGV = \frac{\partial V}{\partial \boldsymbol{Q}_i}\boldsymbol{Q}_i = \sum c_i \boldsymbol{Q}_i \qquad (1)$$

VGVcan be used to gain insight on the protein structural dynamics−cavity changes relationships since it provides information concerning which are the collective motions that modify the most the volume of the cavity [27]. Besides, ANA quantifies relative flexibilities of cavities by using the variation of the potential energy (ΔE_{VGV}) of a protein in the direction of VGVas

$$\Delta E_{VGV} = \sum \Delta E_{\boldsymbol{Q}_i} = \frac{1}{2} k_{VGV} \Delta X^2 \qquad (2)$$

with

$$\Delta E_{\boldsymbol{Q}_i} = \frac{1}{2} k_i c_i^2 \Delta X^2 \qquad (3)$$

and k_i being the force constant associated with the *i*th PCA or NMA mode. *ΔX* represents the displacement in the direction of VGVrelative the corresponding reference structure for PCA or NMA modes. $k_{VGV} = \sum k_i c_i^2$ represents the effective force constant in the direction of VGV. ANA uses the value of k_{VGV} an index of relative rigidity among different protein cavities. That is, the lower the value of k_{VGV}, the more flexible the cavity. More numerical details related to the calculation of k_i can be found at ANA's webpage https://ana.run/.

B. ANA Workflow

ANA is a C++ program that can be installed and run on Linux, Mac and Windows machines. In order to select the cavity of interest, ANA can initially explore all cavities present on a given protein structure and provide them on separated output files. It can also supply the list of residues lining these cavities in order to help define an appropriate IA. All output files are provided in an adequate format that allows their further visualization with biomolecular simulation programs like VMD or Pymol(see Quickstart section in https://ana.run/docs). Otherwise, a selected and depured set of residues lining a cavity of interest can be provided by the user in a configuration file.

ANA allows three different options of use: (a) *static*: calculation of a cavity volume in a single protein structure; (b) *dynamic*: calculation of cavity volumes for an ensemble of protein structures; (c) VGV& *flexibility*: calculation of the direction of maximum change of the cavity volume (VGV) in terms of the collective coordinates (PCA or NMA vectors) and the index of relative flexibility of the cavity. The workflows corresponding to these three options are displayed in **Figure 1**. Each option requires a different set of input files: (a) a pdb file and a configuration file with options such as: the list of residues (or atoms) lining the cavity, level of precision, minimum radius of the spherical probe used to accept void tetrahedrons as part of the protein cavity during the Delaunay triangulation, among others (see the complete list of options in https://ana.run/docs/config.html); (b) an additional file with concatenated frames from a MD trajectory (standard output files from MD packages such as NetCDF, .gro, .trj and .trr), RMN models or any other source of protein conformational samples; (c) two additional files containing the set of collective coordinates $\{Q_i\}$on which the VGV is expressed, i.e., PCA or NMA modes (see eq. 1) and their corresponding frequencies. The outputs corresponding to each type of calculation are: (a) an optional PDB file with the coordinates of the protein cavity represented as the set of the void tetrahedrons obtained after Delaunay triangulation, which can be visualized using pymol or VMD, and a file with the value of the corresponding cavity volume; (b) an optional file containing the concatenated set of PDBs files with the coordinates of the protein cavity in the same format as the one provided in the *static* version, and an additional file with the list of corresponding cavity volumes; (c) a file with the list of displaced volumes in the directions of each collective coordinate provided as input, a second file with the VGVelements in the basis of these collective coordinates, and a third file with the value of the index of cavity flexibility.

In what follows, we show a few examples of ANA applications. In order to initially explore all the cavities present on a given protein structure, the user can use the following terminal command:

```
> ANA2 input_pdb.pdb -c configuration_file_1 -w cavity_residues -f output_prefix
```

with `input_pdb.pdb` being the input file with the standar pdb format, the `-c` flag indicates the configuration filename (`configuration_file_1`), `-w` the output file (`cavity_residues`) with the list of residues lining each cavity and `-f` the prefix for the PDB files, one per cavity, containing the void tetrahedrons obtained after Delaunay triangulation (i.e. output_prefix-i, with i=1, n, being n the total number of cavities found by ANA). In order to obtain in the output file the list of sequence numbers and types of the residues that line each cavity, the configuration file should contain the following command line:

```
list_wall = residue
```

Thereafter, the user can select one desired cavity whose IA can be subsequently recalculated according to a chosen precision (LP or HP) and/or a modified list of residues. In order to do that, a new configuration file with the following commands must be generated:

```
included_area_residues = 25 29 36 49 54 57 58 65 81 90 94 102 115 122 128

included_area_precision = 1
```

where the list of numbers after the command `included_area_residues =` indicates the sequence numbers of the residues that line the cavity, and `included_area_precision = 0 or 1` corresponds to LP and HP levels of calculation respectively.

In order track a cavity volume throughout a MD trajectory, the *dynamic* option of ANA can be invoke by using the terminal command:

```
> ANA2 input_pdb.pdb -c configuration_file_2 -d input_trajectory.nc -f cavity -o output_volume
```

where the flag `-d` indicates the input file `input_trajectory.nc` containing the concatenated frames from a MD trajectory (standard output files from MD packages such as NetCDF, .gro, .trj and .trr) and the output files `cavity.pdb`. and `output_volume` contain the concatenated set of cavities (defined by their corresponding set of void tetrahedrons) and their corresponding values of volume, respectively.

The change of the cavity volume along PCA or NMA modes, given by the elements of the VGV, and the index of relative flexibility of the cavity can be calculated by using the *VGV& flexibility* option of ANA. This is invoke with the following command added in the configuration file:

```
NDD_step = 3
```

and using terminal command:

```
> ANA2 input_pdb.pdb -c configuration_file_2 -M collective_coordinates -F frequencies
```

where the flags `-M` and `-F` point out to the plain text files `collective_coordinates` and `frequencies` with the corresponding PCA (or NMA) modes (eigenvectors) and frequencies (eigenvalues) respectively. ANA is also able to read PCA modes in the format provided by the Amber MD package [31,32] if the following line is added in the configuration file:

```
NDD_modes_format = amber
```

More details related to terminal and configurational commands can be found at ANA webpage https://ana.run/docs/config.html.

C. MD, NMA and apo-to-holo conformational changes

NMA has been performed using the Elastic Network Model (ENM) [16,33,34], which considers the protein as an elastic network with *N* nodes (α-carbons) linked by springs within a cutoff distance r_c . Within ENM, the interaction potential between residues is defined as $E = 1/2\, k_{ij}\left(\left|r_{ij}\right| - \left|r_{ij}^0\right|\right)^2$ being $r_{ij} \equiv r_i - r_j$ the vector connecting nodes i and j, and the zero superscript corresponds to the equilibrium position. The value of the force constant k_{ij} changes according to the type of interaction between nodes [33–35]. Normal modes are obtained by diagonalizing the second-order partial derivatives or Hessian matrix H of E. The resulted eigenvectors and eigenvalues are the set of normal modes $\{Q_i^{NMA}\}_{i=1,3N}$ and their corresponding frequencies. More details about our implementation of ENM and its parameters can be found elsewhere. [36]

MD simulation of the *RB* β-cylinder has been performed using AMBER16[31] software package. The porin (PDB[37] ID: 1PRN) is solvated with explicit TIP3P[38] water molecules in a truncated octahedric periodic box large enough to contain the protein and 10 Å of solvent on all sides and Amberff14SB[38–40] force field was used. Minimization of the molecular system is performed by 100-step of steepest-descent and 400-step conjugate gradient minimization applying constraints to the protein atoms. These are followed by a 400-step unconstrained conjugate gradient minimization.The system is then heated for 150 ps until it reaches the final temperature of 300 K. During heating, a harmonic restraint of 25.0 kcal/(mol·Å2) is applied to the protein atoms. A 2

fs time step with SHAKE[41] algorithm applied to all bonds involving hydrogen is used. Periodic boundary conditions and particle-mesh Ewald (PME) sums are used and a cutoff of 10 Å is applied to nonbonded interactions.The system is equilibrated at constant pressure using 26 steps of 100 ps and reducing the restraint on each step. After the last step with restraints, all restraints are lifted and final equilibration is achieved by a 300 ns simulation at the constant temperature of 300 K using Andersen barostat and Langevin thermostat with a γ collision frequency of 2 ps−1. Finally, 600 ns MD simulation is performed in order to collect equilibrated configurations at 10 ps intervals.

The set of protein structures that encompasses the guided pathway associated to apo-to-holo conformational transition has been obtained using our previously reported iterative procedure to explore conformational spaces of proteins based on their equilibrium dynamics at room temperature. The method has been used in its targeted version, that allows to find a conformational path that connects an initial chosen structure (e.g., the apo state) and a final target one (e.g., the holo state). The method makes use of the ENM model to iteratively introduce structural distortions according to the distribution function for normal modes in thermal equilibrium. More details about this procedure, variants and parameters can be found elsewhere.[20]

III. Results and discussion

A. Cavity volume, VGV and k_{VGV}: EGFR kinase

The epidermal growth factor receptor (EGFR) is a tyrosine kinase receptor that participates in a number of simultaneous cellular processes associated with signaling, and regulation of cell proliferation, differentiation, and migration.[42,43] Its kinase domain acts as a tumor marker in many cancer types [42,43]. It presents a main active site pocket limited by an N-terminal and C-terminal lobes (N and C lobes) that are connected by a hinge region.[44] Its native state is mainly clustered into active and inactive conformers, whose main pockets can be structurally and dynamically differentiated [45,46]. **Figure 2** illustrates the plasticity of ANA for the calculation of the IA corresponding to the main pocket of EGFR kinase active (PDB ID: 1M14) and inactive (PDB ID: 1XKK) conformers. Their definition using a common CH stresses the differential structural distortions between residues lining the cavities of both conformers. The cavity not only changes in size between both conformers (1350 Å and 1180 Å) (**Figures 2 (a** and **c)**) , but its VGV (see **Figures 2(b** and **d)**) and rigidity (4.9 and 6.0 kj/(molÅ) for active and inactive conformers respectively) also experience significant changes. These results show that active and inactive conformers can be differentiated not only by structural and dynamics aspects related to the protein backbone, but also for structural and dynamical features of their cavities. The larger flexibility of the pocket in the active EGFR conformer with respect to the inactive conformer can be directly related to the required functional hinge and shear motions between the C- and N-lobe associated to its collective low-frequency dynamics. This is consistent with previous studies that identify the two lowest frequency normal modes as dynamics fingerprints of the active conformers that allow to differentiate them from inactive conformers[46] The relative displacements in the direction of these modes lead to larger changes in cavity volumes than motions in the direction of the corresponding modes of inactive conformers, that is, a more flexible cavity in the active conformers with respect to inactive conformers.

B. Benchmarks of performance and accuracy: RB β-cylinder porin

ANA has been developed to calculate cavity volumes in an accurate and efficient way. This is a particularly necessary requirement when it is used in its *dynamic* version. In order to benchmark ANA against other available softwares, the membrane channel porin from the phototrophic bacteria Rhodopseudomonas blastica (RB) β-cylinder, consisting of 16 antiparallel β-strands forming a large barrel that lines the pore eyelet, has been considered[47]. **Figures 3 (a)** and **(b)** display the protein structure (PDB ID: 1PRN) with the corresponding IA and the cavity volume obtained by the Convex Hull and Delaunay Triangulation algorithms respectively (see Section **ANA Overview**).

The volume of the pore eyelet of RB-β-cylinder porin has been monitored using different available software packages. **Figure 3(c)** summarizes the results of benchmark ANA against MDPocket, POVME, and Epock when they are applied to 600 snapshots collected during the MD simulation. ANA, applied either at the LP or HP level (see Section **ANA Overview**), results faster than the other softwares. As it is expected due to the different strategies for cavity definitions, the calculated absolute values for cavity volumes differ among the different tested softwares (see Supplementary Information **Figure S2**). Nevertheless, their changes throughout the MD simulations present statistically significant correlations (**Table S1**).

Let us now test the robustness of ANA and other available softwares for the evaluation of partial derivatives of a cavity volume respect to any protein coordinate Q_i like cartesian coordinates, PCA or NMA modes, $\left(\frac{\partial V}{\partial Q_i}\right)$. Partial derivatives are calculated using finite differences, therefore, their robustness can be evaluated by analyzing the stability in the calculation of cavity volume changes with respect to variations in the incremental values of the coordinate Q_i. **Figure 4** shows results obtained in the calculation of volume changes for the pore eyelet of RB-□-cylinder porin under differential displacements along the first PCA mode using ANA and different available softwares (MDPocket[48], POVME[49], and Epock[23]). We can observe that, while other methods lead to rather unstable values of partial derivatives with respect to different finite coordinate displacements, ANA provides numerically stable results.

C. Cavity volume and flexibility changes associated to protein conformational changes: AK and LAO binding protein

In order to validate ANA for the analysis of changes in cavity volume and flexibility during protein conformational transitions, the apo-to-holo conformational change of the Adenylate Kinase (AK) and the Lysine-Arginine-Ornithine (LAO) binding protein have been considered. Both represent conformational changes with significant structural distortions upon ligand binding: AK (PDB IDs, 4AKE and 2ECK for apo and holo conformations, respectively) root-mean-square deviation (RMSD) = 6.9 Å between apo and holo conformations, and LAO binding protein (PDB IDs, 2LAO and 1LST for apo and conformations, respectively) with RMSD = 4.7 Å.

AK is a phosphotransferase that contributes to maintain the ATP/ADP balance in cells by catalyzing their interconversion and it has been extensively studied as a test system to study the mechanisms of large functional conformational transitions. [50] Its conformational change between apo and holo conformations corresponds to two open-close hinge bending motions involving displacements of its LID and NMP domains relative to the CORE domain. [51–55]. LAO binding protein is a permease of the periplasmic transport system in charge of the transport of

different kinds of substrates.[56,57] As AK, its apo-to-holo conformational change also involves global and hinge-bending motions of two domains denominated as 1 and 2.[58]

For both cases, AK and LAO binding protein, our previously reported iterative procedure to explore protein conformational spaces using NMA calculations (see Section II.C) was able to generate sets of validated structures that connect the apo and holo conformations. We calculate the cavity volumes and flexibility for each of these intermediate structures and the result is shown in **Figure 5**, where these values are depicted as a function of the RMSD for the corresponding structures with respect to holo structure. The effect that ligand binding has on the cavity volume flexibility is remarkable. Gradual changes of both quantities are observed while moving from the apo (open) to homo (close) conformations. In both cases, the cavity volumes decrease and their rigidities increase during the apo-to-holo transition. The decrease in the sizes of the cavities agrees with the expected volume changes for cavities localized at the center of the hinge of the open-to-close conformational transitions. Besides, the increase of cavity rigidities are the consequence of the increase of inter-domain interactions in the holo(close) with respect to the apo(open) conformer.

D. Cavity volume and flexibility related to protein conformational diversity: Dynein light chain

Protein fluctuations−cavity changes relationships can be further explored by analyzing the changes of cavity volume and flexibility within an ensemble of protein structures obtained by NMR experiments. For this purpose, we have considered the ensemble of 20 NMR structures of the dimer subunit of dynein light chain (DLC8) (PDB ID: 1F96). This subunit not only acts as an essential component of the dynein motor complex, but also regulates a large variety of biological events by binding to numerous proteins and enzymes [59]. DLC8 presents two conformational flexible peptide-binding channels to achieve binding specificity (see **Figure 6(a)**).

Figure 6(b and **c)** shows the distributions of RMSD and cavity volumes obtained for the ensemble of NMR structures of DLC8, respectively. Besides, NMA has been performed for each of the NMR structures and the resulting distribution of values of rigidity is displayed in **Figure 6 (d)**. In order to compare the extent of variability in the RMSD, cavity volume and flexibility distributions, the coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, has been calculated. Values of 0.06, 0.15, and 0.13 were obtained for RMSD, cavity volume and flexibility distributions respectively. That is, structural variations in the protein backbone is accompanied by larger relative variations in the cavity volume and flexibility. This result points out that the extent of protein conformational diversity should not be directly associated with equivalent variations in cavity volumes and flexibilities.

Finally, we have performed PCA for the ensemble of NMR structures. The value of rigidity obtained using the corresponding PCA modes was 0.039 kj/(molÅ) (shown as vertical blue line in **Figure 6(c)**), in good agreement with the distribution of values of rigidity calculated for the ensemble of structures. This result points out the validity of NMA used in the calculation of cavity rigidities with respect to its direct calculation using the structural fluctuations obtained by PCA performed for the ensemble of experimental structures.

IV. Conclusions

We have presented ANA, an accurate and efficient software suitable for the analysis of changes of cavity volumes within ensembles of protein conformations obtained either experimentally (NMR, X-ray crystallography) or theoretically (MD simulations, and NMA among others). It has been developed to be robust enough for the quantification of cavity changes due to small structural distortions or numerical differentiations on specific collective coordinate displacements. Therefore, it can be used to calculate numerically stable partial derivatives of the cavity volumes with respect to predefined directions of motion (e.g. PCA and NMA modes or any other direction of motion associated to a specific conformational change).

The robustness of ANA for cavity volumes and numerical calculation of volume partial derivatives with respect to protein coordinates allows further explorations of new dynamical features of cavities, like their volume gradient vector and relative flexibility. We have shown that, in the case of EGFR kinase active and inactive conformers, these dynamical features of cavities can be used as conformer-specific features that can contribute to elucidate the functional role of the different conformers in the native state of a protein. Furthermore, by using collective coordinates like PCA and NMA modes, they can be directly associated with dynamical features of the protein backbone. The larger flexibility of the pocket in the active EGFR conformer with respect to the inactive conformer can be directly related to its functional collective motions, not present in the inactive conformer, that lead to larger changes in cavity volumes. This is an example that highlights the use of ANA for the exploration of protein fluctuations−cavity changes relationships.

In order to test ANA as a useful tool to analyze the behaviour of cavity volumes during protein conformational transitions, changes in cavity volume and flexibility were analyzed for the cases of apo-to-holo conformational changes of AK and LAO binding protein. In both cases, the cavity gradually becomes smaller and more rigid while moving from the apo (open) to homo (close) conformations, in agreement with expected volume changes for cavities localized at the center of the hinge of open-to-close conformational transitions. These results encourage the use of ANA for systematic studies of changes in the cavity volumes and flexibility upon ligand binding involving different types of conformational changes.

Finally, we have analyzed and compared the extent of variability in the RMSD, cavity volume and flexibility within an ensemble of NMR structures of the dimer subunit of dynein light chain (DLC8). We find that the extent of protein conformational diversity does not imply equivalent variations in cavity volumes and flexibilities. Nevertheless, more systematic studies performed on large data sets of NMR structures for different proteins are needed.

Source code, installation instructions, a quickstart tutorial, and further theoretical and practical details can be found at ANA's webpage https://ana.run/.

Acknowledgement

The authors would like to thank Guillaume Fraux for helping with the Chemfiles library.

Figures

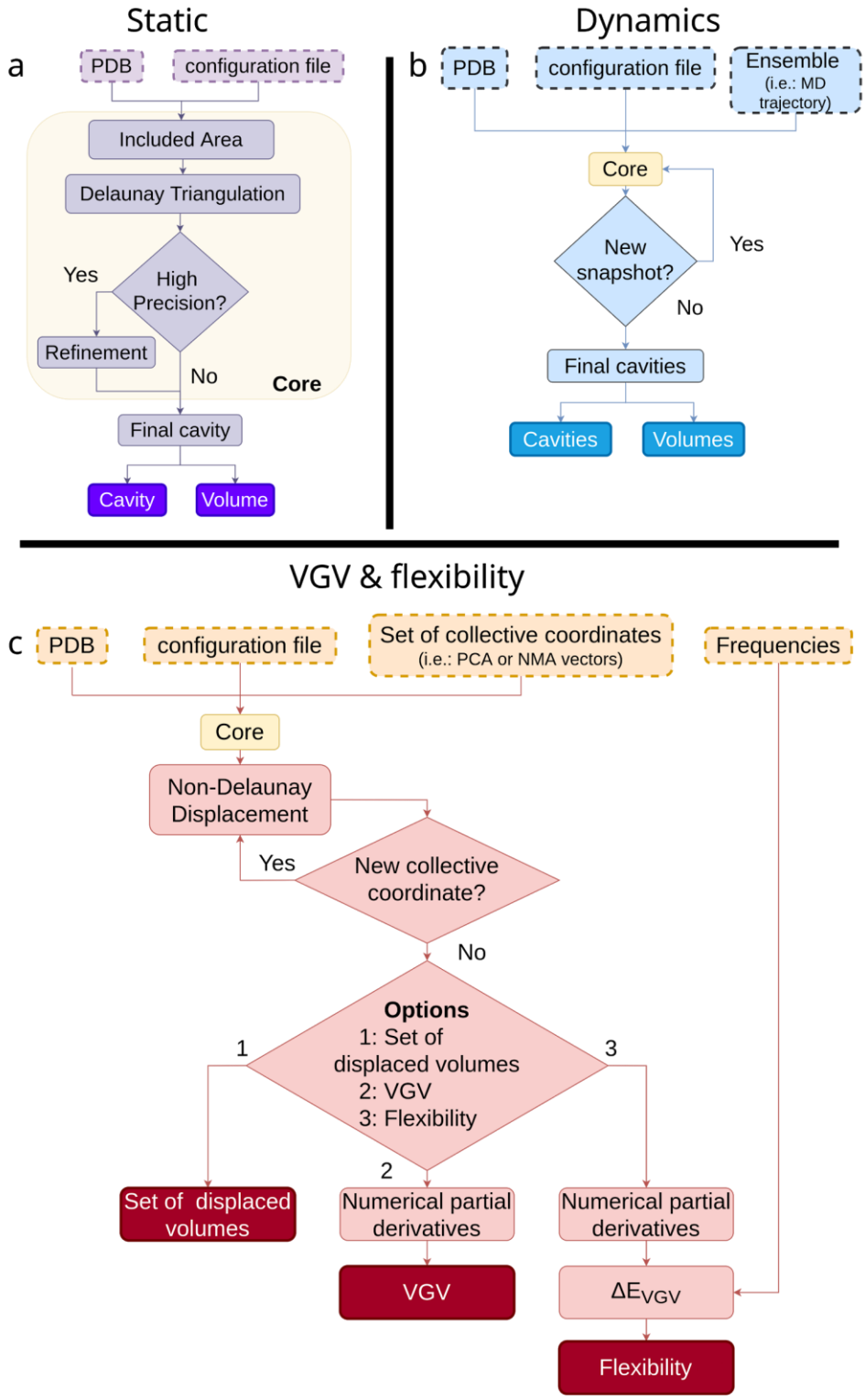

Figure 1. ANA workflows corresponding to its three options of use: *static*, *dynamic* and $VGV\&$ *flexibility*.

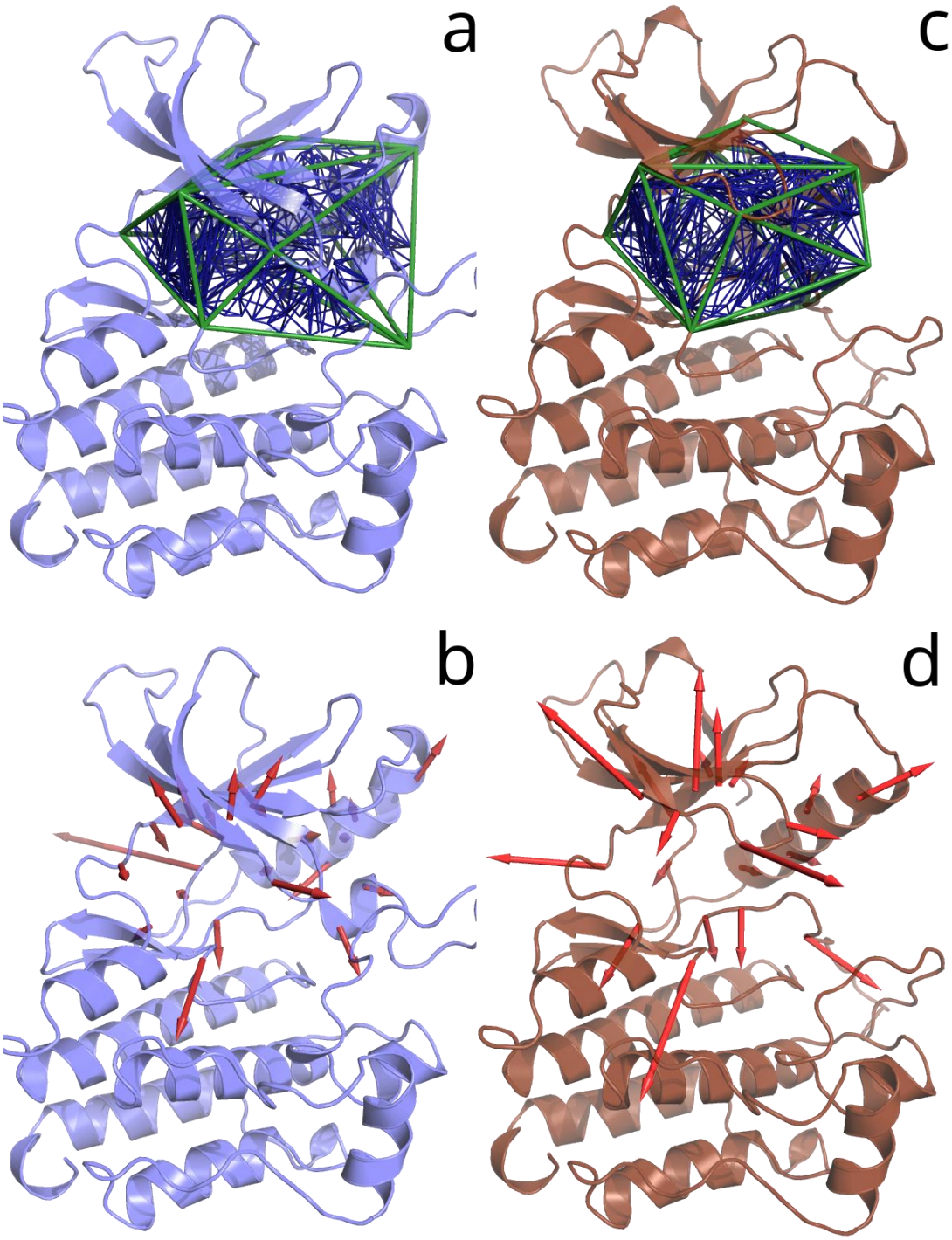

Figure 2. Calculated volume (a, c) and VGV (b, d) of the main pocket of the active site for active (c, d) and inactive (a, b) conformers of the kinase domain of the Epidermal Growth Factor Receptor (EGFR). Included Areas are shown in green, cavity tetrahedrons in blue and VGV elements as red arrows in the porcupine plot.

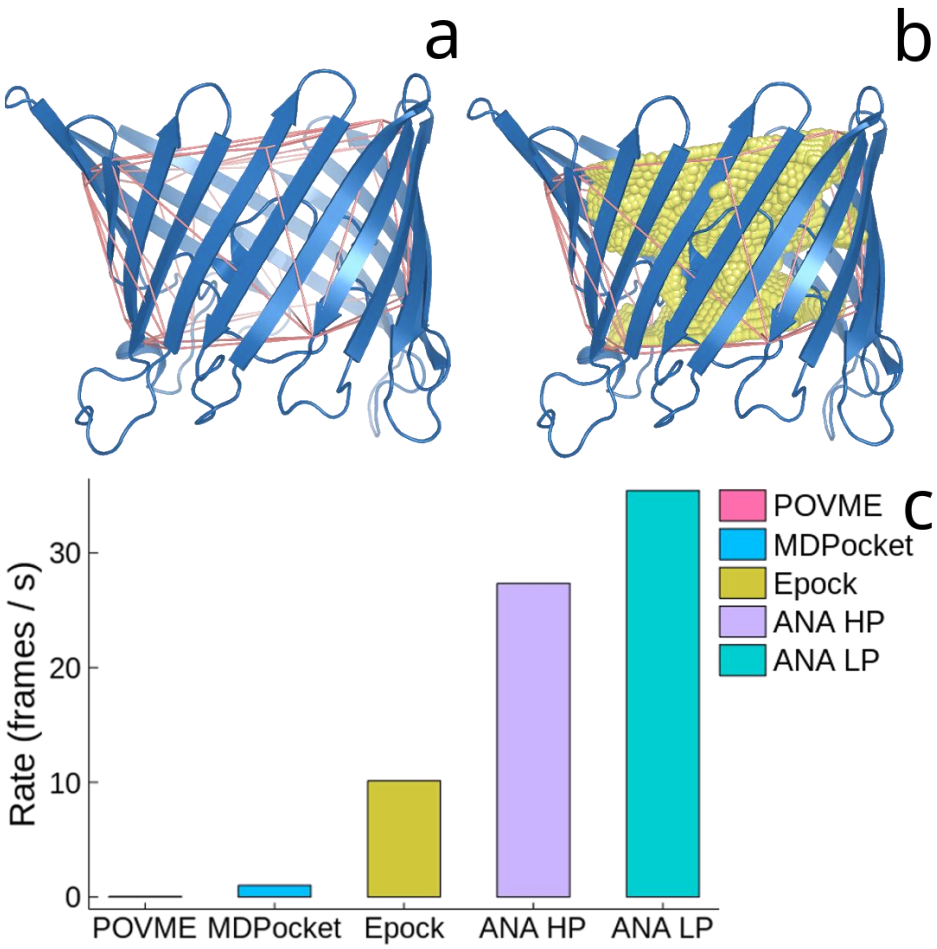

Figure 3. Summary of applying ANA to the analysis of the calculation of the volume for the main tunnel of Rhodopseudomonas Blastica (RB) β-cylinder porin: protein structure including (a) the IA and (b) the final cavity volume; (c) benchmark ANA against MDPocket, POVME, and Epock when they are applied to 600 snapshots collected during a MD simulation.

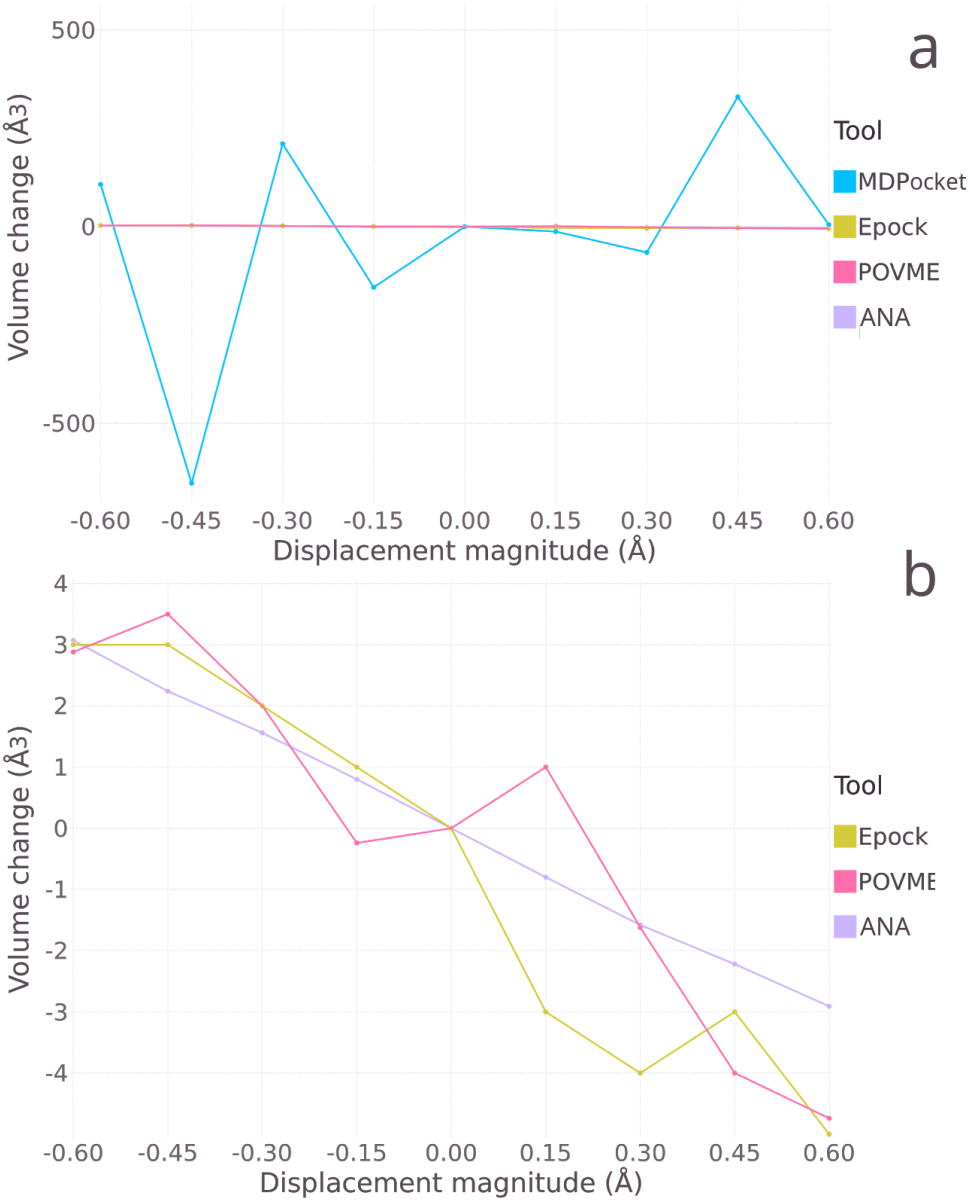

Figure 4. Comparison of robustness in the calculation of volume changes for the main tunnel of *Rhodopseudomonas Blastica* β-cylinder porin under differential displacements along the first PCA mode using different available softwares. Two different scales have been used in order to enhance the comparison with **(a)** MDPocket, and **(b)** Epock and POVME.

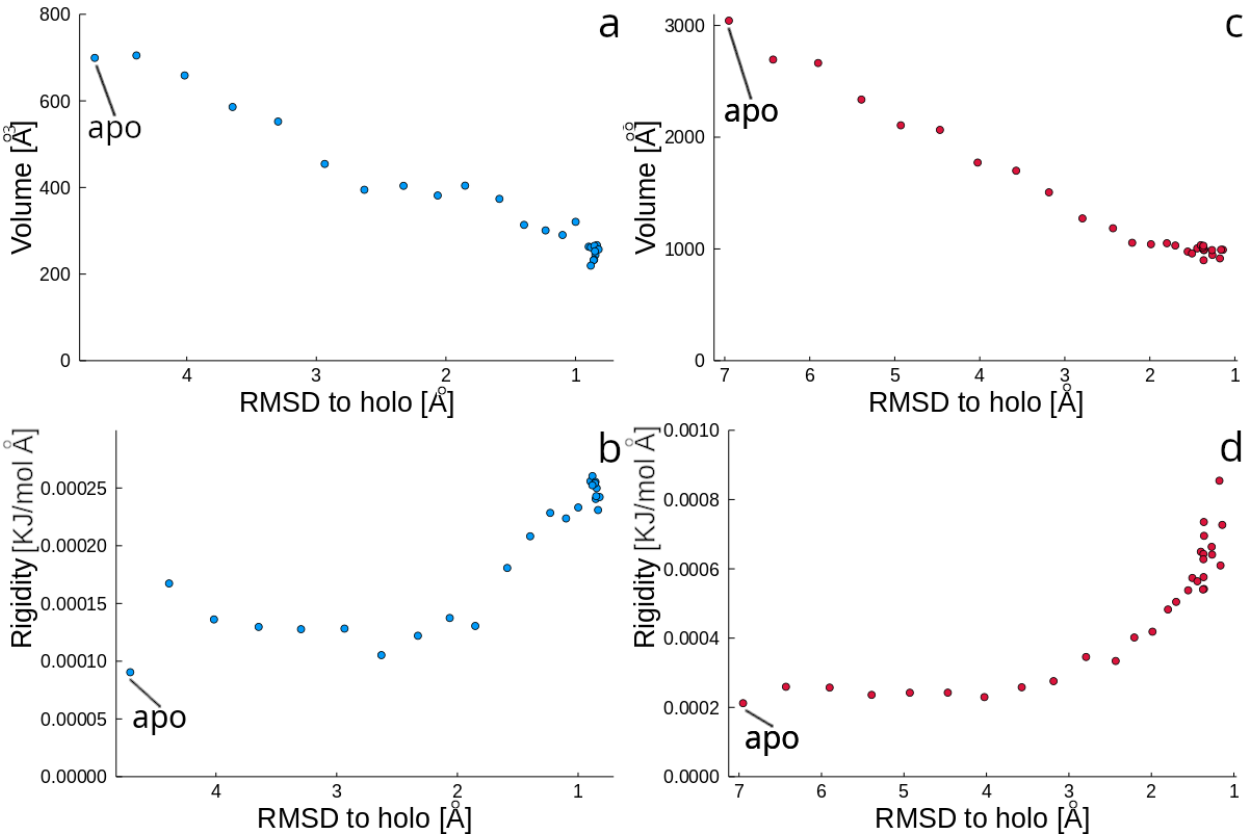

Figure 5. Variations of cavity volume and rigidity for Lysine-Arginine-Ornithine (LAO) binding protein (a, b), and Adenylate Kinase (c, d) for sets of intermediate structures connecting apo and holo conformations as a function of the RMSD for the corresponding structures with respect to holo structure.

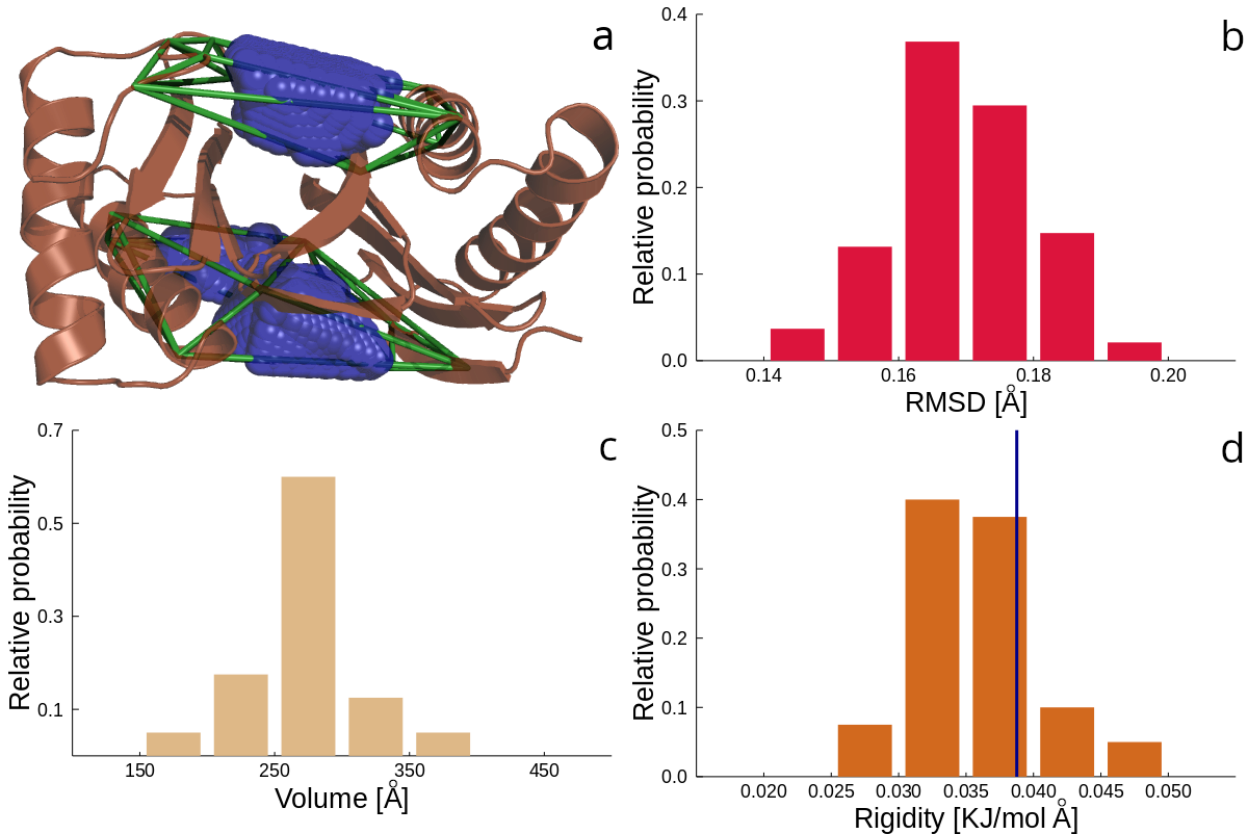

Figure 6. (a) Calculated volumes for the two peptide-binding channels of the dimer subunit of dynein light chain (DLC8). The Included Areas are shown in green. Distributions of (b) RMSD, (c) cavity volume, and (d) cavity rigidity obtained for the ensemble of NMR structures. The vertical blue line in (d) indicates the value of rigidity obtained by performing PCA for the ensemble of NMR structures.

References

(1) Kellis, J. T.; Nyberg, K.; S˘ail, D.; Fersht, A. R. Contribution of Hydrophobic Interactions to Protein Stability. *Nature*. 1988, pp 784–786. https://doi.org/10.1038/333784a0.

(2) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry*. 1990, pp 7133–7155. https://doi.org/10.1021/bi00483a001.

(3) Merski, M.; Fischer, M.; Balius, T. E.; Eidam, O.; Shoichet, B. K. Homologous Ligands Accommodated by Discrete Conformations of a Buried Cavity. *Proceedings of the National Academy of Sciences*. 2015, pp 5039–5044. https://doi.org/10.1073/pnas.1500806112.

(4) Maeno, A.; Sindhikara, D.; Hirata, F.; Otten, R.; Dahlquist, F. W.; Yokoyama, S.; Akasaka, K.; Mulder, F. A. A.; Kitahara, R. Cavity as a Source of Conformational Fluctuation and High-Energy State: High-Pressure NMR Study of a Cavity-Enlarged Mutant of T4 Lysozyme. *Biophys. J.* **2015**, *108* (1), 133–145.

(5) Lopez, C. J.; Yang, Z.; Altenbach, C.; Hubbell, W. L. Conformational Selection and Adaptation to Ligand Binding in T4 Lysozyme Cavity Mutants. *Proceedings of the National Academy of Sciences*. 2013, pp E4306–E4315. https://doi.org/10.1073/pnas.1318754110.

(6) Černý, J.; Biedermannová, L.; Mikulecký, P.; Zahradník, J.; Charnavets, T.; Šebo, P.; Schneider, B. Redesigning Protein Cavities as a Strategy for Increasing Affinity in Protein-Protein Interaction: Interferon-γReceptor 1 as a Model. *BioMed Research International*. 2015, pp 1–12. https://doi.org/10.1155/2015/716945.

(7) Hubbard, S. J.; Argos, P. Cavities and Packing at Protein Interfaces. *Protein Science*. 1994, pp 2194–2206. https://doi.org/10.1002/pro.5560031205.

(8) Pravda, L.; Berka, K.; Vařeková, R. S.; Sehnal, D.; Banáš, P.; Laskowski, R. A.; Koča, J.; Otyepka, M. Anatomy of Enzyme Channels. *BMC Bioinformatics*. 2014. https://doi.org/10.1186/s12859-014-0379-x.

(9) Monzon, A. M.; Rohr, C. O.; Fornasari, M. S.; Parisi, G. CoDNaS 2.0: A Comprehensive Database of Protein Conformational Diversity in the Native State. *Database* **2016**, *2016*. https://doi.org/10.1093/database/baw038.

(10) Hrabe, T.; Li, Z.; Sedova, M.; Rotkiewicz, P.; Jaroszewski, L.; Godzik, A. PDBFlex: Exploring Flexibility in Protein Structures. *Nucleic Acids Res.* **2016**, *44* (D1), D423–D428.

(11) Monzon, A. M.; Zea, D. J.; Fornasari, M. S.; Saldaño, T. E.; Fernandez-Alberti, S.; Tosatto, S. C. E.; Parisi, G. Conformational Diversity Analysis Reveals Three Functional Mechanisms in Proteins. *PLoS Comput. Biol.* **2017**, *13* (2), e1005398.

(12) Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the Barrier. *Front Mol Biosci* **2019**, *6*, 117.

(13) Bussi, G.; Laio, A.; Parrinello, M. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Phys. Rev. Lett.* **2006**, *96* (9), 090601.

(14) Zhou, R. Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Protein Folding Protocols*. pp 205–224. https://doi.org/10.1385/1-59745-189-4:205.

(15) Wu, X.; Wang, S. Self-Guided Molecular Dynamics Simulation for Efficient Conformational Search. *The Journal of Physical Chemistry B*. 1998, pp 7238–7250. https://doi.org/10.1021/jp9817372.

(16) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77* (9), 1905–1908.

(17) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. Exploring the Suitability of Coarse-Grained Techniques for the Representation of Protein Dynamics. *Biophys. J.* **2008**, *95* (5), 2127–2138.

(18) Kantarci-Carsibasi, N.; Haliloglu, T.; Doruker, P. Conformational Transition Pathways

Explored by Monte Carlo Simulation Integrated with Collective Modes. *Biophys. J.* **2008**, *95* (12), 5862–5873.
(19) Kurkcuoglu, Z.; Bahar, I.; Doruker, P. ClustENM: ENM-Based Sampling of Essential Conformational Space at Full Atomic Resolution. *J. Chem. Theory Comput.* **2016**, *12* (9), 4549–4562.
(20) Saldaño, T. E.; Freixas, V. M.; Tosatto, S. C. E.; Parisi, G.; Fernandez-Alberti, S. Exploring Conformational Space with Thermal Fluctuations Obtained by Normal-Mode Analysis. *J. Chem. Inf. Model.* **2020**. https://doi.org/10.1021/acs.jcim.9b01136.
(21) Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids. *J. Mol. Graph.* **1992**, *10* (4), 229–234.
(22) Berka, K.; Hanák, O.; Sehnal, D.; Banás, P.; Navrátilová, V.; Jaiswal, D.; Ionescu, C.-M.; Svobodová Vareková, R.; Koca, J.; Otyepka, M. MOLEonline 2.0: Interactive Web-Based Analysis of Biomacromolecular Channels. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W222–W227.
(23) Laurent, B.; Chavent, M.; Cragnolini, T.; Dahl, A. C. E.; Pasquali, S.; Derreumaux, P.; Sansom, M. S. P.; Baaden, M. Epock: Rapid Analysis of Protein Pocket Dynamics. *Bioinformatics* **2015**, *31* (9), 1478–1480.
(24) Brooks, B.; Karplus, M. Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **1983**, *80* (21), 6571–6575.
(25) Van Aalten, D. M. F.; De Groot, B. L.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. A Comparison of Techniques for Calculating Protein Essential Dynamics. *Journal of Computational Chemistry*. 1997, pp 169–181. https://doi.org/3.0.co;2-t.">10.1002/(sici)1096-987x(19970130)18:2<169::aid-jcc3>3.0.co;2-t.
(26) Hayward, S.; Groot, B. L. Normal Modes and Essential Dynamics. *Methods in Molecular Biology*. 2008, pp 89–106. https://doi.org/10.1007/978-1-59745-177-2_5.
(27) Barletta, G. P.; Fernandez-Alberti, S. Protein Fluctuations and Cavity Changes Relationship. *J. Chem. Theory Comput.* **2018**, *14* (2), 998–1008.
(28) Barber, C. B.; Bradford Barber, C.; Dobkin, D. P.; Huhdanpaa, H. The Quickhull Algorithm for Convex Hulls. *ACM Transactions on Mathematical Software (TOMS)*. 1996, pp 469–483. https://doi.org/10.1145/235815.235821.
(29) Boissonnat, J.-D.; Devillers, O.; Pion, S.; Teillaud, M.; Yvinec, M. Triangulations in CGAL. *Computational Geometry*. 2002, pp 5–19. https://doi.org/10.1016/s0925-7721(01)00054-2.
(30) Fabri, A.; Giezeman, G.-J.; Kettner, L.; Schirra, S.; Sch[illegible]nherr, S. On the Design of CGAL a Computational Geometry Algorithms Library. *Software: Practice and Experience*. 2000, pp 1167–1202. https://doi.org/3.0.co;2-b.">10.1002/1097-024x(200009)30:11<1167::aid-spe337>3.0.co;2-b.
(31) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.
(32) Lee, T.-S.; Cerutti, D. S.; Mermelstein, D.; Lin, C.; LeGrand, S.; Giese, T. J.; Roitberg, A.; Case, D. A.; Walker, R. C.; York, D. M. GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J. Chem. Inf. Model.* **2018**, *58* (10), 2043–2050.
(33) Bahar, I.; Erman, B.; Jernigan, R. L.; Atilgan, A. R.; Covell, D. G. Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function. *J. Mol. Biol.* **1999**, *285* (3), 1023–1037.
(34) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80* (1), 505–515.

(35) Jeong, J. I.; Jang, Y.; Kim, M. K. A Connection Rule for Alpha-Carbon Coarse-Grained Elastic Network Models Using Chemical Bond Information. *J. Mol. Graph. Model.* **2006**, *24* (4), 296–306.
(36) Saldaño, T. E.; Monzon, A. M.; Parisi, G.; Fernandez-Alberti, S. Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLoS Comput. Biol.* **2016**, *12* (3), e1004775.
(37) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics*. 1983, pp 926–935. https://doi.org/10.1063/1.445869.
(39) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Protein Simulations*. 2003, pp 27–85. https://doi.org/10.1016/s0065-3233(03)66002-x.
(40) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*. 2015, pp 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.
(41) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *Journal of Computational Physics*. 1977, pp 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.
(42) Yarden, Y.; Sliwkowski, M. X. Untangling the ErbB Signalling Network. *Nat. Rev. Mol. Cell Biol.* **2001**, *2* (2), 127–137.
(43) Henry, N. L.; Hayes, D. F. Cancer Biomarkers. *Mol. Oncol.* **2012**, *6* (2), 140–146.
(44) Shan, Y.; Arkhipov, A.; Kim, E. T.; Pan, A. C.; Shaw, D. E. Transitions to Catalytically Inactive Conformations in EGFR Kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (18), 7270–7275.
(45) Hasenahuer, M. A.; Barletta, G. P.; Fernandez-Alberti, S.; Parisi, G.; Fornasari, M. S. Correction: Pockets as Structural Descriptors of EGFR Kinase Conformations. *PLoS One* **2018**, *13* (2), e0192815.
(46) Barletta, G. P.; Hasenahuer, M. A.; Fornasari, M. S.; Parisi, G.; Fernandez-Alberti, S. Dynamics Fingerprints of Active Conformers of Epidermal Growth Factor Receptor Kinase. *J. Comput. Chem.* **2018**, *39* (29), 2472–2480.
(47) Kreusch, A.; Schulz, G. E. Refined Structure of the Porin from Rhodopseudomonas Blastica. Comparison with the Porin from Rhodobacter Capsulatus. *J. Mol. Biol.* **1994**, *243* (5), 891–905.
(48) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: Open-Source Cavity Detection and Characterization on Molecular Dynamics Trajectories. *Bioinformatics* **2011**, *27* (23), 3276–3285.
(49) Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J. Chem. Theory Comput.* **2017**, *13* (9), 4584–4592.
(50) Yan, H.; Tsai, M. D. Nucleoside Monophosphate Kinases: Structure, Mechanism, and Substrate Specificity. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1999**, *73*, 103–134, x.
(51) Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hübner, C. G.; Kern, D. Intrinsic Motions along an Enzymatic Reaction Trajectory. *Nature* **2007**, *450* (7171), 838–844.
(52) Vonrhein, C.; Schlauderer, G. J.; Schulz, G. E. Movie of the Structural Changes during a Catalytic Cycle of Nucleoside Monophosphate Kinases. *Structure* **1995**, *3* (5), 483–490.
(53) Hanson, J. A.; Duderstadt, K.; Watkins, L. P.; Bhattacharyya, S.; Brokaw, J.; Chu, J.-W.; Yang, H. Illuminating the Mechanistic Roles of Enzyme Conformational Dynamics. *Proc.*

Natl. Acad. Sci. U. S. A. **2007**, *104* (46), 18055–18060.
(54) Seyler, S. L.; Beckstein, O. Sampling Large Conformational Transitions: Adenylate Kinase as a Testing Ground. *Molecular Simulation*. 2014, pp 855–877. https://doi.org/10.1080/08927022.2014.919497.
(55) Beckstein, O.; Denning, E. J.; Perilla, J. R.; Woolf, T. B. Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open↔Closed Transitions. *Journal of Molecular Biology*. 2009, pp 160–176. https://doi.org/10.1016/j.jmb.2009.09.009.
(56) Oh, B. H.; Pandit, J.; Kang, C. H.; Nikaido, K.; Gokcen, S.; Ames, G. F.; Kim, S. H. Three-Dimensional Structures of the Periplasmic Lysine/arginine/ornithine-Binding Protein with and without a Ligand. *J. Biol. Chem.* **1993**, *268* (15), 11348–11355.
(57) Kang, C. H.; Shin, W. C.; Yamagata, Y.; Gokcen, S.; Ames, G. F.; Kim, S. H. Crystal Structure of the Lysine-, Arginine-, Ornithine-Binding Protein (LAO) from Salmonella Typhimurium at 2.7-A Resolution. *J. Biol. Chem.* **1991**, *266* (35), 23893–23899.
(58) Echols, N.; Milburn, D.; Gerstein, M. MolMovDB: Analysis and Visualization of Conformational Change and Structural Flexibility. *Nucleic Acids Res.* **2003**, *31* (1), 478–482.
(59) Fan, J.; Zhang, Q.; Tochio, H.; Li, M.; Zhang, M. Structural Basis of Diverse Sequence-Dependent Target Recognition by the 8 kDa Dynein Light Chain. *J. Mol. Biol.* **2001**, *306* (1), 97–108.

For Table of Contents Only