OXFORD

Briefings in Bioinformatics, preprint, pp. 1–8

doi: preprint

EnGens: a computational framework for generation and analysis of representative protein conformational ensembles

Anja Conev [1], Mauricio Menagatti Rigo [1], Didier Devaurs [2], André Faustino Fonseca [3], Hussain Kalavadwala [3], Martiela Vaz de Freitas [3], Cecilia Clementi [4], Geancarlo Zanatta [5], Dinler Amaral Antunes [3,*] and Lydia Kavraki [1,*]

[1] Department of Computer Science, Rice University, Houston, 77005, TX, USA , [2] MRC Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, UK, [3]Department of Biology and Biochemistry, University of Houston, Houston, 77004, TX, USA, [4]Department of Physics, Freie Universität Berlin, Berlin, 14195, Germany and [5]Department of Biophysics, Institute of Biosciences, Federal University of Rio Grande do Sul, Porto Alegre, 91501-970, Brazil

*Corresponding author. antunesda@central.uh.edu*Corresponding author. kavraki@rice.edu

Abstract

Proteins are dynamic macromolecules that perform vital functions in cells. A protein structure determines its function, but this structure is not static, as proteins change their conformation to achieve various functions. Understanding the conformational landscapes of proteins is essential to understand their mechanism of action. Sets of carefully chosen conformations can summarize such complex landscapes and provide better insights into protein function than single conformations. We refer to these sets as representative conformational ensembles. Recent advances in computational methods have led to an increase in number of available structural datasets spanning conformational landscapes. However, extracting representative conformational ensembles from such datasets is not an easy task and many methods have been developed to tackle it. Our new approach, EnGens (short for ensemble generation), collects these methods into a unified framework for generating and analyzing protein conformational ensembles. In this work we: (1) provide an overview of existing methods and tools for protein structural ensemble generation and analysis; (2) unify existing approaches in an open-source Python package, and a portable Docker image, providing interactive visualizations within a Jupyter Notebook pipeline; (3) test our pipeline on a few canonical examples found in the literature. Representative ensembles produced by EnGens can be used for many downstream tasks such as protein-ligand ensemble docking, Markov state modeling of protein dynamics and analysis of the effect of single-point mutations.

Key words: proteins, conformational ensembles, clustering , dimensionality reduction, molecular dynamics (MD), crystal structure analysis

1. Introduction

Proteins are the main building blocks of cells, executing a variety of functions vital to life, such as signal transduction, immune defense, and DNA replication. These functions are driven by the three-dimensional arrangement (i.e., the structural conformation) of proteins (Kessel and Ben-Tal, 2018). However, proteins exist in a highly complex environment and are not static entities. The following examples demonstrate that a single protein conformation is not enough to characterize important protein dynamics driving diverse functions. First, allosteric modulations, driven by mutations or drug interactions far from the protein's active site, induce conformational changes within the active site (Nussinov and Tsai, 2013), which can modify the protein's activity (Todd *et al.*, 2002; Tsou, 1998; Weng *et al.*, 2011). Second, metamorphic proteins (Dishman and Volkman, 2018, 2022; Lella and Mahalakshmi, 2017) switch between drastically different folds of the same sequence, thereby performing different functions (Kim and Porter, 2021). Finally, intrinsically disordered proteins and intrinsically disordered protein regions constitute extreme examples of highly flexible structures. They exist as highly dynamic structural ensembles (Uversky, 2016) failing to form a globally stable three-dimensional shape in physiological solution, thereby performing different functions. All these examples demonstrate the importance of comprehensively characterizing a protein conformational landscape and identifying key conformational states to understand protein function (Henzler-Wildman and Kern, 2007).

The energy landscape theory (Frauenfelder *et al.*, 1991; Kumar *et al.*, 2000; Onuchic *et al.*, 1997) is one framework that provides an understanding of protein structure and dynamics by analyzing a protein's free energy landscape (or free energy

surface - FES) as a function of a few collective variables (CVs). However, the exact determination of the FES for large proteins is challenging, as it requires extensive sampling of the protein's conformational space. New methods for computational protein structure prediction (Jumper *et al.*, 2021; Baek *et al.*, 2021; Lin *et al.*, 2022) and simulation (Barhaghi *et al.*, 2022; Abraham *et al.*, 2015; Eastman *et al.*, 2013; Husic *et al.*, 2020; Hénin *et al.*, 2022) are emerging and there is an increased availability protein structure datasets. However, a full understanding of a protein's dynamics can be reached only when the dataset spans the FES sufficiently, allowing quantitative methods (such as Markov state modeling) to be applied. In this work, we do not tackle the sampling problem, as we rely on previously generated datasets. In other words, our approach focuses on structurally representative ensembles and not on thermodynamic ensembles.

There is a need to rapidly extract useful information from conformational datasets (Peng *et al.*, 2018) without directly modeling the dynamics. Subsets of conformations extracted to represent major conformational states contained within the data provide a useful conformational summary. We call such sets *representative conformational ensembles*. Note that, in this context, the term ensemble does not refer to a statistical ensemble. The task we address is that of extracting described representative conformational ensembles from datasets of protein structures. Extracted representative ensembles can be useful for many downstream tasks such as protein-ligand ensemble docking (Hall-Swan *et al.*, 2021), analysis of mutational effects (Kannan and Naganathan, 2022) and extensive Markov state modeling of protein dynamics (Abella *et al.*, 2020; Chan and Shukla, 2021). It is important to provide sufficient analysis of the extracted ensemble to summarize important properties of each protein state (e.g., the distance between protein domains or the distance between important residues in the active site) and help derive more intuitive insights (e.g., whether a member of the ensemble represents the protein in its active or inactive form). In this work, we develop EnGens - a computational pipeline for the generation and analysis of representative protein conformational ensembles.

Sources of protein structural datasets are now diverse. The Protein Data Bank (PDB) (wwPDB consortium, 2019; Burley *et al.*, 2021), first established in 1971, has experienced steady growth over the past decade. With more than 10,000 experimentally solved protein structures deposited annually, the total number of available entries to date is around 200,000. These data have allowed for new breakthroughs in the field of protein structure prediction, including machine learning techniques such as AlphaFold2 (Jumper *et al.*, 2021), RosettaFold (Baek *et al.*, 2021) or ESMFold (Lin *et al.*, 2022). The AlphaFold database (Varadi *et al.*, 2022) was released with over 200 million protein structure predictions. ESMFold has recently reported comparable performance to AlphaFold2 with the ESM Metagenomic Atlas, which contains 617 million predicted metagenomic structures. This vast amount of available data allows researchers to collect multiple conformations of the same protein (Takei and Ishida, 2022). Collected conformations make up datasets whose content can be summarized and analyzed with EnGens. We call such datasets "static" to highlight the fact that the conformations they contain are independent and not derived from simulating protein dynamics.

A more extensive analysis of protein dynamics can be performed using simulations that generate so-called "dynamic" datasets. Conformations within these datasets are not independent - they are time-ordered and form trajectories. Molecular dynamics (MD) simulations, first developed in the late 70s (Warshel and Levitt, 1976), have been established as a gold standard for exploring protein dynamics. Many computational packages have since been developed, including NAMD (Phillips *et al.*, 2020), GROMACS (BEKKER *et al.*, 1993; Berendsen *et al.*, 1995), AMBER (Salomon-Ferrer *et al.*, 2013), CHARMM (Brooks *et al.*, 2009), OpenMM (Eastman *et al.*, 2017). MD software is becoming more accessible with python plugins and graphical user interfaces (Barhaghi *et al.*, 2022). Markov State Modelling (MSM) approaches for interpreting MD simulations (Prinz *et al.*, 2011) have recently gained popularity. but constructing MSMs can be a lengthy process as this requires extensive sampling. On the other hand, EnGens can be used to gain insights into the content of MD datasets without fully modeling the dynamics.

Our approach recognizes and addresses the need for a unified computational framework to help researchers summarize the vast amount of newly available structural data in an effort to understand the conformational landscape driving protein function. EnGens builds on several existing tools that have proven useful for protein structure analysis. For the computational representation of protein structure, EnGens utilizes the PDB module of BioPython (Cock *et al.*, 2009) as well as the rich featurization module of PyEmma (Scherer *et al.*, 2015), powered by MDTraj (McGibbon *et al.*, 2015). For dimensionality reduction and clustering steps EnGens provides a diverse set of algorithms implemented across deeptime (Hoffmann *et al.*, 2021), scikit-learn (Pedregosa *et al.*, 2011), UMAP (Trozzi *et al.*, 2021) and SRV (Chen *et al.*, 2019). EnGens brings all these tools closer to the community by providing an open-source pipeline wrapped into a portable Docker image and accompanied by extensive example workflows written in Jupyter Notebooks. Additionally, EnGens implements a set of customizable interactive visualizations that provide users with detailed insight into the generated conformational ensembles.

Other similar tools complement EnGens (Table S1). CoNSEnsX (Ángyán *et al.*, 2010) generates ensembles based on available NMR data. PENSA (Vögele *et al.*, 2022b,a) provides different metrics (Jensen-Shannon Distance, Kolmogorov-Smirnov Statistic, Overall Ensemble Similarity) for the comparison of generated ensembles. ProDy (Bakan *et al.*, 2011; Zhang *et al.*, 2021) provides a set of algorithms for studying protein dynamics, which includes normal mode analysis. The specificity of EnGens lies in that: (1) it provides customizable PyEmma featurization for both static and dynamic datasets; (2) it contains both linear and nonlinear dimensionality reduction techniques (linear PCA (Pearson, 1901) and TICA (Pérez-Hernández *et al.*, 2013; Schwantes and Pande, 2015); nonlinear UMAP and SRV); (3) it provides different clustering methods (hierarchical, K-means and Gaussian Mixture Models); (4) it is wrapped in an accessible Docker image and includes interactive Jupyter Notebook Workflows with rich ensemble visualizations. With these unique properties, EnGens enables users to automate the generation and analysis of protein conformational ensembles. We envision EnGens as an important and useful resource for data analysis of protein structure to support researchers in the era of big data.

In the following sections, we describe methods involved in the EnGens pipeline. Note that these methods have been previously published and extensively validated (Scherer *et al.*, 2015; Chen *et al.*, 2019; Trozzi *et al.*, 2021; Pérez-Hernández *et al.*, 2013; Schwantes and Pande, 2015). Hence, validating these methods is outside of the scope of our work. Instead, we showcase the use of the full EnGens pipeline on a set of

example molecules from the literature for which static or dynamic datasets are available. This includes molecules of different scales: a large protein complex (PI3K kinase), a peptide drug (Compstatin) and a small molecule (Nelfinavir).

2. Methods

We have developed EnGens, an automated pipeline for generating and analyzing protein conformational ensembles, given a dataset of protein structures as input. Note that EnGens pipeline has two distinct use-cases: i) processing static protein datasets (e.g., experimental structures); ii) processing dynamic protein datasets (e.g., MD simulations).

A static structural dataset could be experimentally derived and collected from the PDB or modeled computationally (e.g., using AlphaFold or Modeller). For a dataset extracted from the PDB, EnGens can be used to reveal different conformational states and extract a representative ensemble summarizing the dataset. For a dataset compiled by computationally modeling a protein and its common mutants, EnGens can describe the conformational landscape of mutants and help point out the impact of mutations.

A dynamic structural dataset is generally a trajectory derived from an MD simulation. If the simulation involves a protein with a ligand in its active site, EnGens can point out conformational changes that occur upon binding. It is important to note that for the analysis of MD-derived data much work has been done in the field of Markov state modeling (Abella *et al.*, 2020; Husic and Pande, 2018; Bernetti *et al.*, 2019). Modeling the dynamics of a system is outside of the scope of EnGens pipeline as its goal is only to generate and analyze the representative conformational ensemble. However, the dynamic use-case is largely inspired by the insights from Markov modeling approaches. For example, one important insight is that resolving slow processes can help identify biologically relevant conformational changes. Thus, using methods related to Markov modeling helps EnGens uncover conformational states and ensembles with biological relevance.

Both static and dynamic datasets can potentially include large numbers of structures that are difficult to systematically inspect visually. To address this issue, EnGens partitions the structural dataset into clusters and extracts a representative conformation from each cluster to form a structurally diverse conformational ensemble. The EnGens pipeline is divided into four workflows that are summarized in Figure 1. Below we give an overview of the workflows and their respective goals. A detailed description of each workflow is provided in the supplementary text.

2.1. *Workflow 1:* Extracting featurized representations from raw data

The first important step in computational analysis of protein structure is finding an appropriate representation for the data. We use the term raw data to refer to the form in which protein structure is stored in databases such as the PDB. Raw data is extracted from a comperhensive experimental study or simulation, and usually contains the three dimensional coordinates of all atoms. However, these coordinates are often redundant and noisy. To benefit from the downstream computational pipeline and avoid the effects of noise, extracting useful components from the raw data and generating a featurized representation is essential. Some features commonly used to describe the input proteins include: dihedral angles of the backbone, pairwise distances between residues and RMSD distances to a reference structure. The featurized representation of a protein structure is a numerical vector, which standard data science methods (such as dimensionality reduction and clustering) rely on.

2.2. *Workflow 2:* Projecting the featurized representation into an embedding in low dimensional space

Numerical vectors extracted from the first workflow often have very high dimensionality. Depending on the size of the protein and the type of featurization, this vector could contain thousands of elements to represent one structure. High dimensional data presents unique challenges for clustering algorithms, as metrics lose their utility in high dimensional spaces. It is thus important to embed the information into a lower dimensional space before clustering. In Workflow2 we provide implementations of four widely used algorithms for dimensionality reduction. For the static use-case we provide two standard methods: principal components analysis (PCA) and uniform manifold approximation and projection (UMAP). For the dynamic use-case we provide two additional methods that make use of the time ordered nature of the data: time-lagged independent components analysis (TICA) and state-free reversible VAMPnet (SRV). TICA and SRV are not suitable for static datasets, which lack the time component that is exploited by these methods. TICA and PCA are linear methods, while UMAP and SRV are non linear and can thus identify non linear relationships between features. The result of Workflow2 is an embedding of the data in a lower dimensional space, in which the data can be more efficiently partitioned into clusters to identify a representative ensemble.

2.3. *Workflow 3:* Clustering embeddings and extracting the ensemble

Low dimensional embeddings represent each conformation in the dataset. Various distance metrics can be used to calculate similarity between two conformations. This allows us to identify clusters of similar datapoints. In Workflow 3 we provide implementations of three widely used clustering algorithms: hierarchical clustering (Murtagh and Contreras, 2012), K-means (Hartigan and Wong, 1979) and gaussian mixture models (GMM) (Lindsay *et al.*, 1989). Hierarchical clustering provides a dendrogram of the data, allowing users to visually inspect the clusters and their relationships. The lower computational complexity of K-means makes it more suitable for large datasets. While K-means assumes a spherical data distribution, GMM can handle more complex distributions and provide a probabilistic model. Whatever the method, resulting clusters correspond to groups of structurally similar conformations. Then, we find the hub of each cluster as the point with the most neighbors and call it a cluster representative. Finally, we generate the resulting ensemble by extracting cluster representatives.

2.4. *Workflow 4:* Visualizing the data and analyzing the ensemble

In the final workflow we provide a set of customizable interactive plots to analyze the generated ensemble. Users can visualize and inspect the 2D embeddings and their clustering. The extracted representatives are highlighted and their position within the 2D embedding space can be identified. Additionally, users can visualize the 3D atomic-resolution conformations of the extracted representatives. The ensemble can be further

analyzed by generating a scatterplot of interesting features (e.g., the distance between important residues or RMSD to a template conformation) for each conformation. The same information can be summarized per cluster as a box plot. These visualizations are meant to help users interpret the ensemble (e.g., understand if the active and inactive states of a protein are represented within the ensemble).

3. Results

The algorithms gathered under the umbrella of the EnGens pipeline have been validated in past literature (Scherer *et al.*, 2015; Chen *et al.*, 2019; Trozzi *et al.*, 2021; Pérez-Hernández *et al.*, 2013; Schwantes and Pande, 2015). The validation of these methods being therefore outside of the scope of this work, in this section we showcase the use of the full EnGens pipeline. To this end, we have selected proteins for which structural data had been analyzed manually via often laborious processes to extract a conformational ensemble. We process the data entirely within the EnGens pipeline, and show that we can generate the same conformational ensemble as reported in previous studies. The examples we picked cover three systems of varying complexities. First, we process a large PI3K protein complex within both use-cases: a crystal structure dataset and an MD trajectory. Second, we apply EnGens to an MD trajectory of the peptide ligand Compstatin. Finally, we use the same methodology to process an MD trajectory of the small drug Nelfinavir.

3.1. Class I PI3K (PI3K-I) experiments

PI3K-I is a family of lipid kinase proteins that phosphorylate a lipid found on the plasma membrane, regulating cell growth and proliferation (Martini *et al.*, 2014). Increased activity of PI3K-I has been associated with oncogenesis and its the structural aspects have been widely studied. Members of PI3K-IA subfamily contain a regulatory (p85) and a catalytic (p110) subunit (Figure S1). Kinase activity is autoinhibited by the interaction between the nSH2 domain of the regulatory unit and the C2 domain of the catalytic unit (Yu *et al.*, 1998; Miller *et al.*, 2014). For instance, it has been shown that the nSH2 domain moves away from the catalytic unit upon contact with a phosphorylated tyrosine pY of the receptor tyrosine kinase (RTK). This movement leads to the activation of PI3K-IA (Buckles *et al.*, 2017; Nolte *et al.*, 1996; Vadas *et al.*, 2011). Two recent works performed further structural analysis of the PI3K, one using available PI3K crystal structure data (Zhang *et al.*, 2020) and another performing and analyzing MD simulations of a mutant (Galdadas *et al.*, 2020).

3.1.1. PI3K-IA: crystal structure dataset

We base this experiment on a study by Zhang *et al.* (2020) that extracted from the PDB a dataset of 49 dimer structures corresponding to alpha, beta and delta isoforms of PI3K-IA (Table S4). While all structures are dimers (containing both catalytic and regulatory units), they differ in the portion of the regulatory unit that is crystalized, namely the nSH2, iSH2 and cSH2 domains. 10 structures were crystalized without the nSH2 domain (PI3KΔnSH2), while the rest contain the nSH2 domain (PI3K+nSH2). The analysis by Zhang et al. was performed by manually engineering the feature of interest as distance between the C2 domain and the kinase domain of the PI3K catalytic unit. With a manually set threshold they divide the 49 structures in two groups: active/open (12) and inactive/closed (37). Zhang et al. conclude that all 10 PI3KΔnSH2 structures have nSH2 released and are active/open. Additionally, two of the PI3K+nSH2 structures have a mutation that leads to the activation. The other 37 structures are considered autoinhibited and are labeled inactive/closed.

When processing this dataset with the EnGens pipeline, our goal was to test the ability of EnGens to generate a diverse ensemble of structures that would include representative structures of the active and inactive states. We use PDB codes of the dataset as input (Table S4). EnGens extracts the maximum common substructure (MCS) for each structure (see Supplementary Material: Workflow 1S-2). The MCS includes the catalytic unit and the iSH2 domain of the regulatory unit (Figures S2, S3, S4). We featurize each structure by using the pairwise distances between the centers of mass of the MCS chains. We choose the PCA option for dimensionality reduction step and K-means for clustering. Results of the analysis as provided by the EnGens dashboard are shown in Figure 2.

The dataset is clustered into five clusters. Cluster #0 contains active/open conformations of the PI3K alpha isoform (with pdb codes: 3HHM, 3HIZ and 5DXH). Cluster #1 contains eight active/open conformations of the delta isoform. Cluster #3 contains a single active/open conformation of the beta isoform (2Y3A). Clusters #2 and #4 contain the remaining 37 inactive/closed conformations of the PI3K alpha isoform. The ensemble generated by EnGens contains the following representatives: 3HHM (cluster 0), 5VLR (cluster 1), 4L23 (cluster 2), 2Y3A (cluster 3), 5SXD (cluster 4). This ensemble is structurally diverse and contains both active (3HHM, 5VLR, 2Y3A) and inactive (4L23, 5SXD) conformations. Additionally, the clusters separate the isoforms present in the dataset, namely the alpha (3HHM, 4L23, 5SXD), beta (2Y3A) and delta (5VLR) isoforms.

3.1.2. PI3K-IA: MD trajectory

This experiment is based on a study by Galdadas *et al.* (2020) involving MD simulations of a PI3K-IA (with a hotspot E545K mutation leading to its increased activity), based on multiple walkers metadynamics simulations. Galdadas et al. manually defined two collective variables: CV1 - distance between the nSH2 domain of the regulatory unit and the helical domain of the catalytic unit; CV2 - distance to a reference state where nSH2 is detached. After inspecting the free energy surface landscape as a function of CV1 and CV2, they uncovered two energy basins: one containing a conformational ensemble corresponding to an active state with the nSH2 domain detached; the other containing two distinct conformational ensembles corresponding to an alternative activation path involving nSH2 sliding around the helical domain.

We process the MD performed by Galdadas et al. with EnGens to uncover the same conformational ensembles. To featurize the trajectory we select: (1) the RMSD distance of each frame to the reference structure (first frame of the trajectory) and (2) the Cartesian coordinates of the center of mass of the helical and nSH2 domains. Next, we select SRV with a lag time of 50 to reduce the dimensionality of our input to the top 3 SRV components. We select the GMM clustering, which produces three clusters. Finally, three representative conformations are extracted. The resulting EnGens dashboard is presented in Figure 3.

Clusters #0 and #1 contain conformations of the broad energy basin where the nSH2 domain is attached to the catalytic unit. Cluster #2 contains conformations in which PI3K is

active and the nSH2 domain is detached. This is verified by plotting the minimum distance between residues of the helical domain (catalytic unit) and residues of the nSH2 domain (Figure 3D) for all cluster members. This distance stands out for members of cluster #2 and is higher than the 2 Å threshold. Clusters #0 and #1 contain the two conformational ensembles located in the same energy basin, as identified by the original paper. These clusters differ in the distance between the residues Lys545 of the helical domain and Arg421 of the nSH2 unit (Figure 3E). In conclusion, the three described clusters identified by EnGens are consistent with the three described states reported by Galdadas et al.

3.2. Compstatin experiment

Compstatin is a small, cyclic peptide that inhibits an immune surveillance mechanism associated with multiple diseases. Previously, we demonstrated that compstatin analogs (i.e., biochemical variants) adopt distinct conformations that ultimately affect binding affinity and inhibitor potential (Devaurs *et al.*, 2020).

We applied EnGens workflows to two compstatin analogs, using two MD simulations (Devaurs *et al.*, 2020). The 4MeW and Cp10 analogs were selected because of their conformational heterogeneity. To featurize conformations, we selected backbone torsions and carbon-alpha distances. Features are then summarized using UMAP and clustered using K-means.

As a result, we retrieve several representative conformations spanning different states of these analogs (Figure 4). In particular, EnGens could accurately retrieve conformational states associated with 4WeM, namely the open v-shaped state, the closed α-shaped, and three intermediate states. These states are identified as five distinct clusters and are structurally similar to our prior observations. This demonstrates again that EnGens can reproduce results obtained with distinct methodologies. The Cp10 analog showed intriguing results. We obraine three clusters corresponding to distinct conformations, including an intermediate state. In our previous study,we could assign only two states (the opened v-shaped and closed α-shaped ones) for the Cp10 analog. To our understanding, this discrepancy is due to the limitations of our previous analysis, which only relied on RMSD calculations and visual data interpretation.

These new findings suggest that EnGens has high sensitivity and can capture rapid transitions between conformational states.

3.3. Nelfinavir experiment

Nelfinavir is a potent HIV-1 protease inhibitor used in adults and children. Its action mechanism involves disabling the protease from cleaving gag-pol polyprotein. However, mutations of the protease might affect the impact of Nelfinavir on patients. Using MD simulations of Nelfinavir in solution, Antunes *et al.* (2014) inspected its conformational space and described three minimal energy Nelfinavir conformations.

We apply EnGens to these MD trajectories of Nelfinavir. We used the Cartesian coordinates of all the atoms of Nelfinavir as the featurization. Then, we apply SRV for dimensionality reduction and GMM for clustering (Figure 5A). As a result, EnGens identifies seven clusters (Figure 5B). One cluster representative conformation coincides with the conformation described as NF-i1 and other representatives are similar to the conformations described as NF-i2 and NF-i3 in the original paper (Figure 5C), considering an RMSD under 2 Å of difference (Table 1).

The first conformation, NF-i1, coincides with the representative of cluster #2 (RMSD = 0.413 Å). The NF-i2 structure matches cluster #1 and #6, with RMSD of 1.918 Å and 1.692 Å respectively. However, both clusters are at the end of the trajectory (Figure 5C), where they strongly overlap, indicating that EnGens slightly refined the state corresponding to the conformation presented in the original paper. The NF-i3 conformation is to EnGens' cluster #3, with an RMSD of 1.336 Å.

4. Discussion

Recent improvements in protein structure prediction tools are bringing the field of computational structural biology closer to the era of big data. One important "unsolved" task highlighted by the most recent CASP15 competition is modeling protein conformational ensembles. It is assumed that an ensemble of protein conformations will better represent the true state of a protein and will aid downstream tasks such as drug-target interaction prediction and molecular docking. Building a representative protein conformational ensemble from multiple input structures or an MD trajectory is not an easy task and many tools have been developed to tackle it. In this work, we recognized a need for a pipeline that we call EnGens.

We have evaluated the EnGens pipeline on systems of varying complexity. In each case, we recovered diverse ensembles that coincided with previously reported results. When analyzing a large protein complex such as PI3K, EnGens generated a representative ensemble containing both the active and inactive states. For the Compstatin peptide EnGens uncovered additional clusters of conformations, therefore enriching a previous study. In addition, EnGens also generated a relevant ensemble for the small drug Nelfinavir.

There are still big challenges for a pipeline of this sort. First, there are no clear guidelines on which method would perform best for a given molecular system. A number of alternative methods, each bearing its own set of hyper-parameters (Table S3), can be used to perform steps of the pipeline. We provide default values and some theoretical guidelines. For example, SRV and UMAP perform nonlinear dimensionality reduction, while TICA and PCA are linear. We thus suggest using SRV and UMAP for more complex systems where nonlinearity of features is expected. In addition, as TICA and SRV are techniques that are suitable for time-series data, they are expected to be less prone to noise resulting from fast fluctuations in the structure and should be suitable for the dynamic use-case. However, they can not be applied to the static use-case. Further theoretical analysis of some of these methods can be found in the literature (Glielmo *et al.*, 2021). Hyper-parameter optimization of the pipeline could be tackled with Bayesian optimization or other machine-learning approaches (Lee *et al.*, 2022). However, a wider benchmarking of these methods is necessary to evaluate the practical implications of the theory and provide good guidelines.

Second, expert knowledge of the analyzed system is still recommended for the featurization step. Some featurizations are generic, such as the pairwise residue distances that we applied to Compstatin. Others, such as the distance between the nSH2 domain and the helical domain of PI3K stem from a good understanding of the underlying system. Efforts have been made to automate this step. For the dynamic use-case new breakthroughs such as VAC (Variational Approach to Conformational dynamics) (Wu and Noé, 2020; Lorpaiboon *et al.*, 2020) and VAMP (Variational Approach for Learning Markov

Processes) (Wu and Noé, 2020) provide metrics to quantify the quality of featurization. Such metrics can be optimized using machine learning approaches to determine the most suitable featurization. However, these methods are highly dependent on the quality of the provided input MD data and are sensitive to different hyperparameters. Engineering features manually is still a widely used practice.

Third, we lack large standardized benchmarks and metrics for generating conformational ensembles. To avoid the hurdles we faced in this work, the community would greatly benefit from a public database collecting both static and dynamic datasets of protein conformations for which the representative conformational ensembles are known. Another problem is the lack of standardized metrics to compare the uncovered conformational states. Although RMSD is widely used to compare protein conformations, there are currently no equivalent standardized metrics for comparing two conformational ensembles. That is why our evaluation of EnGens is mostly qualitative and descriptive (e.g. determining if EnGens uncovered the active and inactive conformational states of PI3K).

These challenges will become more pressing as the field moves towards big data analysis to study protein flexibility. EnGens provides easy access to existing algorithms and can serve as a platform for the rapid development of new algorithms addressing these challenges.

5. Conclusion

EnGens is a novel tool for the end-to-end processing of large protein structural datasets with the aim of generating and analyzing representative protein conformational ensembles. EnGens unifies widely used Python libraries (PyEmma, deeptime, mdtraj, UMAP, sklearn, plotly, etc.) under one Docker image and provides interactive visualizations along with extensive examples of the pipeline in Jupyter Notebook workflows. For advanced users, we provide a Python package. Our code is open source and accessible through a github repository (https://github.com/anon528/supreme-couscous). We showcased how EnGens can be used to automate ensemble generation using examples from the literature. EnGens ensembles can be useful for many downstream tasks related to drug discovery such as molecular docking and drug-target interaction prediction. Additionally, EnGens can serve as a platform for further algorithmic development. Overall, we see the EnGens pipeline becoming part of many new efforts to utilize the structural data generated by novel structure prediction tools.

6. Competing interests

No competing interest is declared.

7. Data availability

The data underlying this article are available in the github repository available at https://github.com/anon528/supreme-couscous. The data used for the PI3K MD analysis generated by Galdadas *et al.* (2020) is available from the EBRAINS repository at https://search.kg.ebrains.eu/ instances/Model/7f44abeb3068cdc74506bac6e72a8802.

8. Author contributions statement

A.C., M.M.R., D.D., A.F.F., M.V.F., C.C., G.Z., and D.A.A. conceived the experiments. A.C., A.F.F., H.K., and M.V.F. conducted the experiments. A.C., A.F.F., and M.V.F. wrote the manuscript, A.C., M.M.R., D.D., A.F.F., M.V.F., C.C., G.Z., D.A.A. and L.K. analyzed the results. All authors reviewed the manuscript.

9. Acknowledgments

The authors would like to thank colleagues from Kavraki Lab for many helpful discussions.

10. Funding

Work on this project by A.C. and L.K. have been supported in part by National Institutes of Health NIH [U01CA258512]. Other support included: University of Edinburgh and Medical Research Council [MC_UU_00009/2 to D.D.]; Computational Cancer Biology Training Program fellowship [RP170593 to M.M.R.]; The Brazilian National Council for Scientific and Technological Development [CNPq no. 440412/2022-6 to G.Z]; University of Houston Funds and Rice University Funds.

References

Abella, J. R. *et al.* (2020). Markov state modeling reveals alternative unbinding pathways for peptide–MHC complexes. *Proceedings of the National Academy of Sciences*, **117**(48), 30610–30618. Publisher: Proceedings of the National Academy of Sciences.

Abraham, M. J. *et al.* (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1-2**, 19–25.

Antunes, D. A. *et al.* (2014). New Insights into the In Silico Prediction of HIV Protease Resistance to Nelfinavir. *PLOS ONE*, **9**(1), e87520. Publisher: Public Library of Science.

Baek, M. *et al.* (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**(6557), 871–876. Publisher: American Association for the Advancement of Science.

Bakan, A. *et al.* (2011). ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics*, **27**(11), 1575–1577.

Barhaghi, M. S. *et al.* (2022). py-MCMD: Python Software for Performing Hybrid Monte Carlo/Molecular Dynamics Simulations with GOMC and NAMD. *Journal of Chemical Theory and Computation*, **18**(8), 4983–4994. Publisher: American Chemical Society.

BEKKER, H. *et al.* (1993). GROMACS - A PARALLEL COMPUTER FOR MOLECULAR-DYNAMICS SIMULATIONS: 4th International Conference on Computational Physics (PC 92). *PHYSICS COMPUTING '92*, pages 252–256. Place: SINGAPORE Publisher: World Scientific Publishing.

Berendsen, H. J. C. *et al.* (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, **91**(1), 43–56.

Bernetti, M. *et al.* (2019). An Integrated Markov State Model and Path Metadynamics Approach To Characterize Drug Binding Processes. *Journal of Chemical Theory and Computation*, **15**(10), 5689–5702. Publisher: American Chemical Society.

Brooks, B. *et al.* (2009). CHARMM: The Biomolecular Simulation Program. *Journal of computational chemistry*, **30**(10), 1545–1614.

Buckles, T. C. *et al.* (2017). Single-Molecule Study Reveals How Receptor and Ras Synergistically Activate PI3K and PIP3 Signaling. *Biophysical Journal*, **113**(11), 2396–2405.

Burley, S. K. *et al.* (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology,

biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, **49**(D1), D437–D451.

Chan, M. C. and Shukla, D. (2021). Markov state modeling of membrane transport proteins. *Journal of Structural Biology*, **213**(4), 107800.

Chen, W. *et al.* (2019). Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets. *The Journal of Chemical Physics*, **150**(21), 214114. Publisher: American Institute of Physics.

Cock, P. J. A. *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Devaurs, D. *et al.* (2020). Computational analysis of complement inhibitor compstatin using molecular dynamics. *Journal of Molecular Modeling*, **26**(9), 231.

Dishman, A. F. and Volkman, B. F. (2018). Unfolding the Mysteries of Protein Metamorphosis. *ACS Chemical Biology*, **13**(6), 1438–1446.

Dishman, A. F. and Volkman, B. F. (2022). Design and discovery of metamorphic proteins. *Current Opinion in Structural Biology*, **74**, 102380.

Eastman, P. *et al.* (2013). OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *Journal of Chemical Theory and Computation*, **9**(1), 461–469. Publisher: American Chemical Society.

Eastman, P. *et al.* (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, **13**(7), e1005659.

Frauenfelder, H. *et al.* (1991). The Energy Landscapes and Motions of Proteins. *Science*, **254**(5038), 1598–1603. Publisher: American Association for the Advancement of Science.

Galdadas, I. *et al.* (2020). Unravelling the effect of the E545K mutation on PI3K kinase. *Chemical Science*, **11**(13), 3511–3515. Publisher: The Royal Society of Chemistry.

Glielmo, A. *et al.* (2021). Unsupervised Learning Methods for Molecular Simulation Data. *Chemical Reviews*, **121**(16), 9722–9758.

Hall-Swan, S. *et al.* (2021). DINC-COVID: A webserver for ensemble docking with flexible SARS-CoV-2 proteins. *Computers in Biology and Medicine*, **139**, 104943.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1), 100–108. Publisher: [Wiley, Royal Statistical Society].

Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, **450**(7172), 964–972. Number: 7172 Publisher: Nature Publishing Group.

Hoffmann, M. *et al.* (2021). Deeptime: a Python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, **3**(1), 015009. Publisher: IOP Publishing.

Husic, B. E. and Pande, V. S. (2018). Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, **140**(7), 2386–2396. Publisher: American Chemical Society.

Husic, B. E. *et al.* (2020). Coarse graining molecular dynamics with graph neural networks. *The Journal of Chemical Physics*, **153**(19), 194101. Publisher: American Institute of Physics.

Hénin, J. *et al.* (2022). Enhanced sampling methods for molecular dynamics simulations. *Living Journal of Computational Molecular Science*, **4**(1). arXiv:2202.04164 [cond-mat, physics:physics].

Jumper, J. *et al.* (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873), 583–589. Number: 7873 Publisher: Nature Publishing Group.

Kannan, A. and Naganathan, A. N. (2022). Ensemble origins and distance-dependence of long-range mutational effects in proteins. *iScience*, **25**(10), 105181.

Kessel, A. and Ben-Tal, N. (2018). *: Structure, Function, and Motion, Second Edition*. Chapman and Hall/CRC, New York, 2 edition.

Kim, A. K. and Porter, L. L. (2021). Functional and Regulatory Roles of Fold-Switching Proteins. *Structure*, **29**(1), 6–14.

Kumar, S. *et al.* (2000). Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science*, **9**(1), 10–19. Publisher: Cambridge University Press.

Lee, Y. *et al.* (2022). Adaptive Experience Sampling for Motion Planning Using the Generator-Critic Framework. *IEEE Robotics and Automation Letters*, **7**(4), 9437–9444. Conference Name: IEEE Robotics and Automation Letters.

Lella, M. and Mahalakshmi, R. (2017). Metamorphic Proteins: Emergence of Dual Protein Folds from One Primary Sequence. *Biochemistry*, **56**(24), 2971–2984. Publisher: American Chemical Society.

Lin, Z. *et al.* (2022). Evolutionary-scale prediction of atomic level protein structure with a language model. Pages: 2022.07.20.500902 Section: New Results.

Lindsay, B. *et al.* (1989). Mixture Models: Inference and Applications to Clustering. In *Journal of the American Statistical Association*, volume 84, page 337. ISSN: 01621459 Issue: 405 Journal Abbreviation: Journal of the American Statistical Association.

Lorpaiboon, C. *et al.* (2020). Integrated Variational Approach to Conformational Dynamics: A Robust Strategy for Identifying Eigenfunctions of Dynamical Operators. *The Journal of Physical Chemistry B*, **124**(42), 9354–9364. Publisher: American Chemical Society.

Martini, M. *et al.* (2014). PI3K/AKT signaling pathway and cancer: an updated review. *Annals of Medicine*, **46**(6), 372–383. Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/07853890.2014.912836.

McGibbon, R. *et al.* (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, **109**(8), 1528–1532.

Miller, M. S. *et al.* (2014). Structural basis of nSH2 regulation and lipid binding in PI3K. *Oncotarget*, **5**(14), 5198–5208.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, **2**(1), 86–97. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.53.

Nolte, R. T. *et al.* (1996). Crystal structure of the PI 3-kinase p85 amino-terminal SH2 domain and its phosphopeptide complexes. *Nature Structural Biology*, **3**(4), 364–374. Number: 4 Publisher: Nature Publishing Group.

Nussinov, R. and Tsai, C.-J. (2013). Allostery in Disease and in Drug Discovery. *Cell*, **153**(2), 293–305.

Onuchic, J. N. *et al.* (1997). Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, **48**, 545–600.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**(85), 2825–2830.

Peng, J.-h. *et al.* (2018). Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chinese Journal of Chemical Physics*, **31**(4), 404–420. Publisher: American Institute of Physics.

Phillips, J. C. *et al.* (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, **153**(4), 044130. Publisher: American Institute of Physics.

Prinz, J.-H. *et al.* (2011). Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, **134**(17), 174105. Publisher: American Institute of Physics.

Pérez-Hernández, G. *et al.* (2013). Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, **139**(1), 015102. Publisher: American Institute of Physics.

Salomon-Ferrer, R. *et al.* (2013). An overview of the Amber biomolecular simulation package. *WIREs Computational Molecular Science*, **3**(2), 198–210. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1121.

Scherer, M. K. *et al.* (2015). PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, **11**(11), 5525–5542. Publisher: American Chemical Society.

Schwantes, C. R. and Pande, V. S. (2015). Modeling Molecular Kinetics with tICA and the Kernel Trick. *Journal of Chemical Theory and Computation*, **11**(2), 600–608. Publisher: American Chemical Society.

Takei, Y. and Ishida, T. (2022). How to select the best model from AlphaFold2 structures? preprint, Bioinformatics.

Todd, A. E. *et al.* (2002). Plasticity of enzyme active sites. *Trends in Biochemical Sciences*, **27**(8), 419–426. Publisher: Elsevier.

Trozzi, F. *et al.* (2021). UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *The Journal of Physical Chemistry B*, **125**(19), 5022–5034. Publisher: American Chemical Society.

Tsou, C.-L. (1998). Active Site Flexibility in Enzyme Catalysisa. *Annals of the New York Academy of Sciences*, **864**(1), 1–8. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1998.tb10282.x.

Uversky, V. N. (2016). p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure–Function Continuum Concept. *International Journal of Molecular Sciences*, **17**(11), 1874.

Vadas, O. *et al.* (2011). Structural Basis for Activation and Inhibition of Class I Phosphoinositide 3-Kinases. *Science Signaling*, **4**(195), re2–re2. Publisher: American Association for the Advancement of Science.

Varadi, M. *et al.* (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**(D1), D439–D444.

Vögele, M. *et al.* (2022a). drorlab/pensa: PENSA 0.2.8.

Vögele, M. *et al.* (2022b). Systematic Analysis of Biomolecular Conformational Ensembles with PENSA. arXiv:2212.02714 [physics, q-bio].

Warshel, A. and Levitt, M. (1976). Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, **103**(2), 227–249.

Weng, Y.-Z. *et al.* (2011). A study on the flexibility of enzyme active sites. *BMC Bioinformatics*, **12**(1), S32.

Wu, H. and Noé, F. (2020). Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science*, **30**(1), 23–66.

wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, **47**(D1), D520–D528.

Yu, J. *et al.* (1998). Regulation of the p85/p110 Phosphatidylinositol 3-Kinase: DISTINCT ROLES FOR THE N-TERMINAL AND C-TERMINAL SH2 DOMAINS*. *Journal of Biological Chemistry*, **273**(46), 30199–30203.

Zhang, M. *et al.* (2020). Structural Features that Distinguish Inactive and Active PI3K Lipid Kinases. *Journal of Molecular Biology*, **432**(22), 5849–5859.

Zhang, S. *et al.* (2021). ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics*, **37**(20), 3657–3659.

Ángyán, A. F. *et al.* (2010). CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data. *BMC Structural Biology*, **10**(1), 39.

Table 1. Root mean square deviation (RMSD in Å) of the EnGens cluster representatives (I) for Antunes *et al.* (2014) trajectory to the conformations from the original paper (II). Representatives are generated by EnGens using SRV for dimensionality reduction and GMM for clustering. RMSD values below 2Å are bolded.

II \ I	cluster #0	cluster #1	cluster #2	cluster #3	cluster #4	cluster #5	cluster #6
NF-i1	2.150	2.879	**0.413**	1.436	2.837	1.475	3.682
NF-i2	3.954	**1.918**	3.887	3.282	3.966	4.084	**1.692**
NF-i3	2.304	2.701	2.467	**1.336**	2.572	2.292	3.563

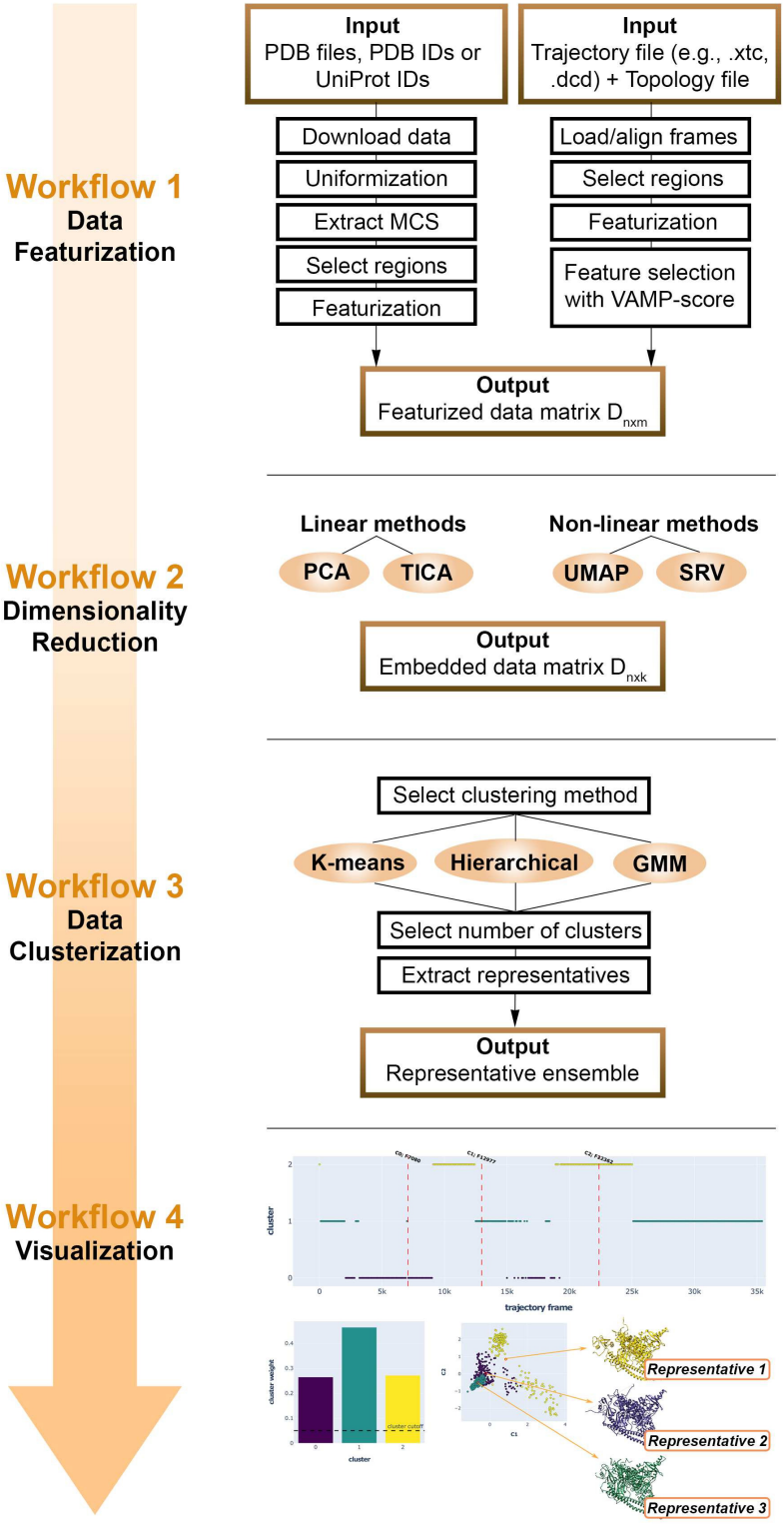

Fig. 1. ***Overview of the EnGens methodology.*** Workflows are listed across the vertical arrow on the left. Individual steps are listed in the diagram on the right.

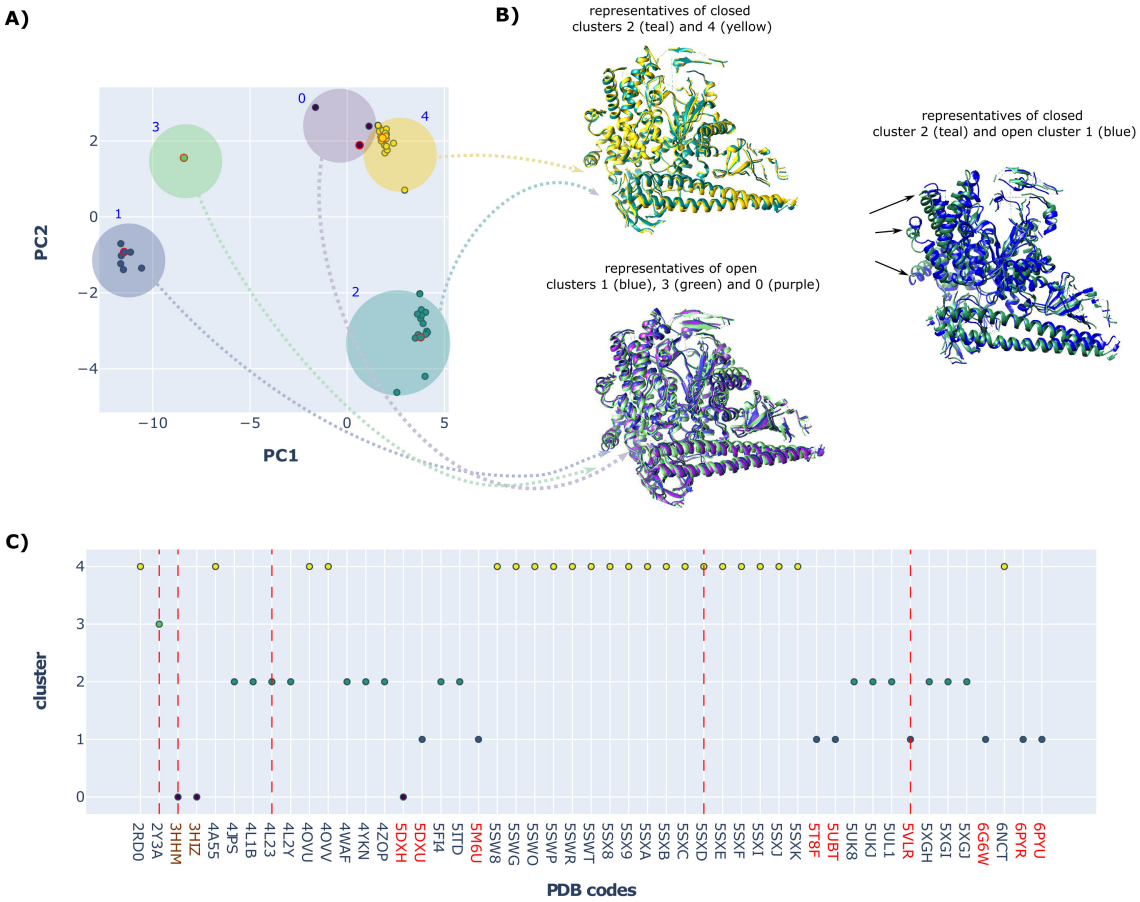

Fig. 2. ***EnGens processing of the Zhang et al. dataset of PI3K crystal structures.*** A) Each point corresponds to a structure from Zhang et al. dataset. Points are colored based on the cluster they were assigned to, and clusters are indicated as large circles on the plot. Points extracted as cluster representatives are highlighted in red. The x and y axis represent the first and second principal components of these data. B) 3D structural models of the EnGens representatives: upper left - representatives of clusters 2 and 4 (inactive/closed states); bottom left - representatives of clusters 0, 1 and 3 (active/open states); right - comparison between the representative of the active state cluster 1 and the representative of the inactive state cluster 2 (the arrows point to the regions showing the biggest differences). C) PDB codes of the crystal structures are listed on the x axis. These codes are colored based on the conformational state identified by Zhang et al. (black - inactive/closed states; red - active/open states; brown - states active due to mutation). The EnGens cluster assignment is shown on the y axis. Red vertical lines indicate cluster representatives that were selected by EnGens (codes: 2Y3A, 3HHM, 4L23, 5SXD, 5VLR).

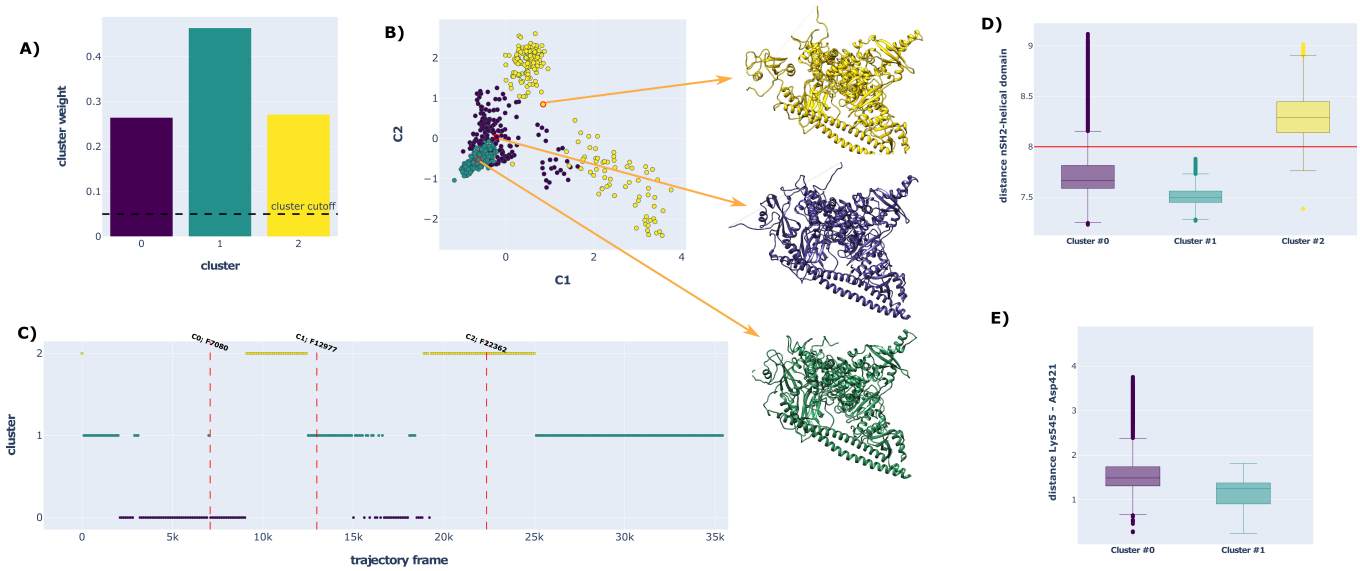

Fig. 3. ***EnGens processing of the Galdadas et al. MD trajectory of PI3K.*** A) The proportion belonging to each cluster is plotted on the y-axis (cluster weight). Cluster indexes are listed on the x-axis. B) Two-dimensional embedding based on the components identified by the SRV method. Datapoints represent frames and are colored based on their respective cluster (same colors as in A). The 3D structural models of the three cluster representatives are shown on the right of this plot. C) The timeline view of the trajectory, where the x-axis lists the index of each frame, and the y-axis lists the corresponding cluster index. Vertical red lines highlight the representative frames extracted in the generated ensemble. D) The distance between nSH2 and the helical domain of PI3K is plotted on the y-axis. The x-axis lists the clusters. The red horizontal line represents the threshold of 8Å. E) The distance between Lys545 and Asp421 of the PI3K regulatory unit is plotted on the y-axis. The x-axis lists the clusters.

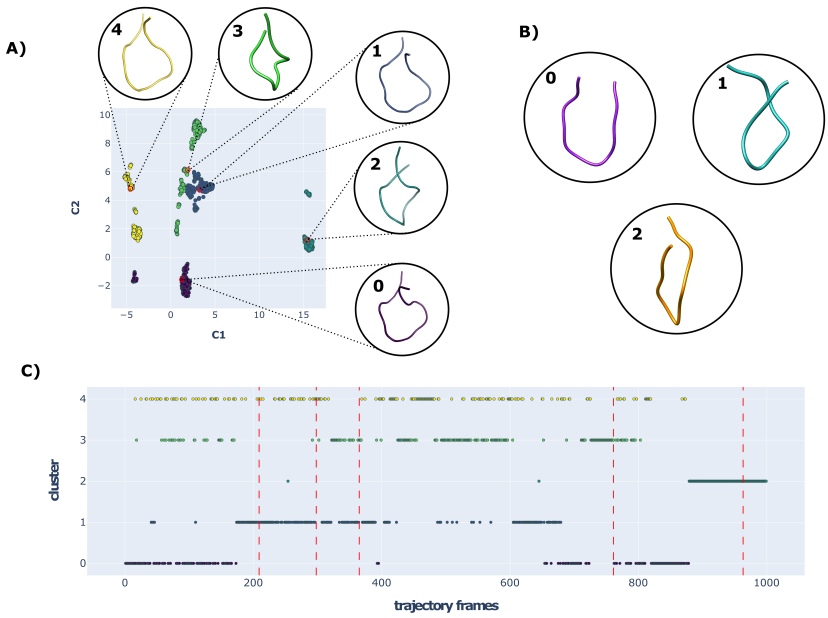

Fig. 4. ***EnGens analysis of Compstatin analogs.*** A) UMAP visualization of the MD trajectory of the 4MeW analog. Points represent MD frames, and are colored based on their respective cluster. Cluster representatives (points highlighted in red) were selected as the closest conformation from the k-mean centroid. The 4MeW cartoon backbone for each cluster representative is presented, including a single closed state (Cluster #2), a single open state (Cluster #1), and three intermediate states (Cluster #0, #3, #4). B) The Cp10 analog: cartoon backbone visualizations of the three cluster representatives, namely the open opened, closed, and intermediate states, respectively. C) Time-oriented plot displaying the association between each MD frame (x-axis) and the clusters.

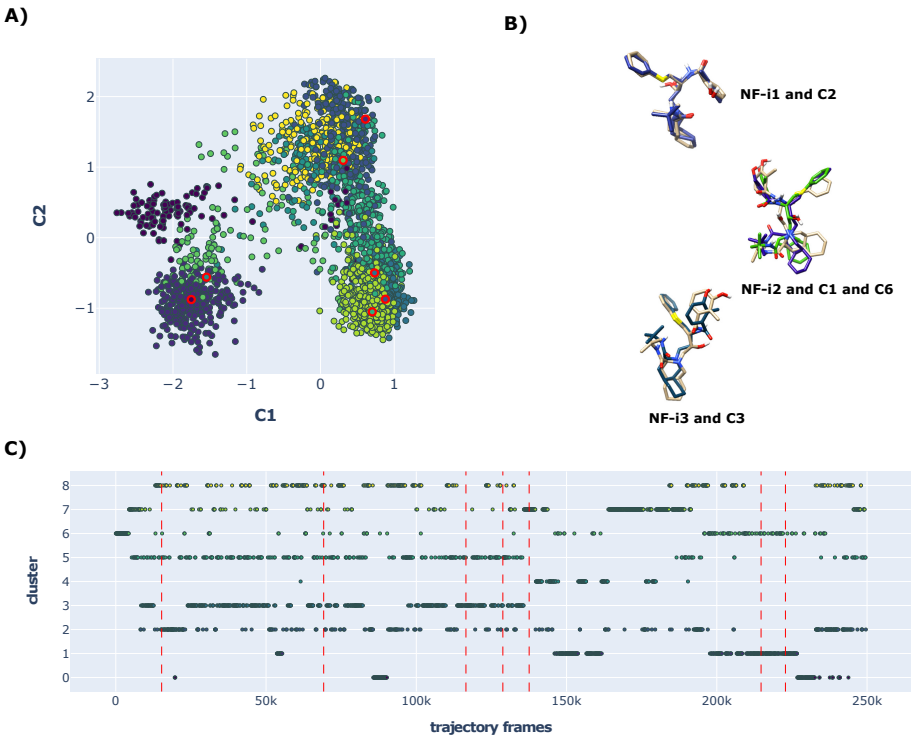

Fig. 5. ***EnGens analysis of the Nelfinavir trajectory.*** A) Projection of all MD frames into a 2D space produced by SRV. Points are colored based on the cluster to which frames were assigned. B) Three plots showing the representative of the clusters aligned with the conformations identified by Antunes *et al.* (2014) (NF-i1, NF-i2, NF-i3 in pale tan color). C) Timeline of the MD trajectory showing which frame (x-axis) belongs to which cluster (y-axis).