MACHINE LEARNING IN BIOMOLECULAR SIMULATIONS

EDITED BY: Gennady Verkhivker, Vojtech Spiwok and Francesco L. Gervasio
PUBLISHED IN: Frontiers in Molecular Biosciences

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-136-0
DOI 10.3389/978-2-88963-136-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

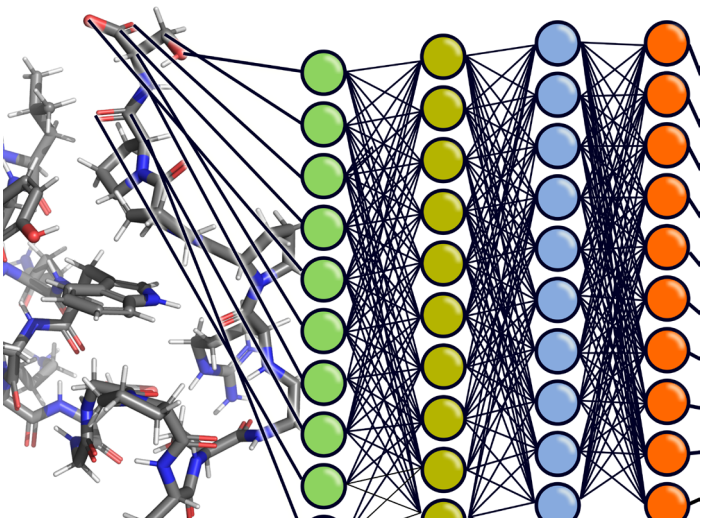
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

MACHINE LEARNING IN BIOMOLECULAR SIMULATIONS

Topic Editors:
Gennady Verkhivker, Chapman University, United States
Vojtech Spiwok, University of Chemistry and Technology, Czechia
Francesco L. Gervasio, University College London, United Kingdom



Machine learning methods such as neural networks, non-linear dimensionality reduction techniques, random forests and others meet in this research topic with biomolecular simulations. The authors of eight articles applied these methods to analyze simulation results, accelerate simulations or to make molecular mechanics force fields more accurate.

Citation: Verkhivker, G., Spiwok, V., Gervasio, F. L., eds. (2019). Machine Learning in Biomolecular Simulations. Lausanne: Frontiers Media. doi: 10.3389/978-2-88963-136-0

Table of Contents

frontiers in Molecular Biosciences

EDITORIAL
published: 29 August 2019
doi: 10.3389/fmolb.2019.00076

Editorial: Machine Learning in Biomolecular Simulations

Gennady Verkhivker[1,2], *Vojtech Spiwok*[3]* *and Francesco Luigi Gervasio*[4]

[1] Graduate Program in Computational and Data Sciences, Schmid College of Science and Technology, Chapman University, Orange, CA, United States, [2] Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine, CA, United States, [3] Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Czechia, [4] Protein Dynamics Research Group, Chemistry Department, University College London, London, United Kingdom

Keywords: machine learning, molecular modeling, molecular dynamics computer simulation, intrinsically disordered protein, sampling enhancement, non-linear dimension reduction

Editorial on the Research Topic

Machine Learning in Biomolecular Simulations

Interest in machine learning is growing in all fields of science, industry, and business. This interest was not primarily initiated by new theoretical findings. Interestingly, the theoretical basis of the majority of machine learning techniques, such as artificial neural networks, decision trees, or kernel methods, have been known for a relatively long time. Instead, there are other effects that triggered the recent boom of machine learning.

First, machine learning needs data to learn on. Huge data sets from Internet, Internet of Things, social networks, phones, wearable devices, and other sources are now available. Such datasets were not available a decade ago. Second, the recent wave of machine learning benefits from hardware advances, in particular from computing on graphical processing units and specialized hardware.

Biomolecular modeling and simulations are an ideal field for the application of machine learning approaches in the spirit of the recent boom of machine learning. Biomolecular simulations produce large amounts of data in the form of trajectories that can be used to train machine learning algorithms. At the same time, vast amounts of genomic data were critical in allowing AlphaFold in leading the field of *de novo* protein prediction in the most recent CASP protein prediction round. Moreover, GPUs are routinely used in biomolecular simulations for more than a decade to offload critical parts of calculation.

This Research Topic collects eight innovative works showcasing the application of machine learning in biomolecular simulations and related fields. It demonstrates major machine learning approaches such as artificial neural networks, random forests, and non-linear dimensionality reduction methods. These techniques are applied in analysis of trajectories, acceleration of biomolecular simulations, parametrization of force fields, and other tasks.

Helfrecht et al. present an alternative to classical definitions of structural motifs in proteins. Classical definitions of secondary and super-secondary structures are based on intuitive criteria, such as hydrogen bonds, dihedral angles, and others and have been widely used. However, they experience problems with borderline and partially disordered structures. This article presents an alternative based on machine learning, namely on Probabilistic Analysis of Molecular Motifs algorithm previously developed in the group.

The article from Trapl et al. presents a program Anncolvar. This tool makes it possible to approximate a collective variable using a simple neural network. The choice of optimal collective variables is crucial to the convergence of enhanced algorithms based on them. Anncolvar is

OPEN ACCESS

Edited and reviewed by:
Massimiliano Bonomi,
Institut Pasteur, France

***Correspondence:**
Vojtech Spiwok
spiwokv@vscht.cz

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *12 August 2019*
Accepted: *16 August 2019*
Published: *29 August 2019*

Citation:
Verkhivker G, Spiwok V and Gervasio FL (2019) Editorial: Machine Learning in Biomolecular Simulations. Front. Mol. Biosci. 6:76. doi: 10.3389/fmolb.2019.00076

shown to be very useful for collective variables that cannot be explicitly calculated on-the-fly or computationally expensive collective variables.

Wang et al. used classical as well as by unsupervised and supervised machine learning methods (principal component analysis, random forest) to analyze protein dynamics. They analyzed trajectories of an enzyme linked to antibiotic resistance β-lactamase, simulated in multiple conformational states.

Intrinsically disordered proteins (IDPs) are a hot topic given that about 10% of all proteins are disordered, and about 40% of eukaryotic proteins have at least one long disordered loop. It has been shown that proteins can have a function despite not having a stable conformation. This brings a new challenge in analysis of dynamics. Grazioli et al. use machine learning and network models on simulation trajectories of amyloid beta in its wild type and its medicinally relevant mutant. They show that machine learning analysis can explain the difference between protein variants. This was not possible by conventional trajectory analysis methods.

There is a growing number of works indicating that molecular mechanics potentials (force fields) developed for compactly folded proteins may fail in modeling of unfolded proteins and especially IDPs. This fact motivated Demerdash et al. to optimize force field for IDPs on the basis of data from small-angle X-ray/neutron scattering. This was done by iterative rounds of molecular dynamics simulations and comparison with experimental data. This approach was demonstrated on three IDPs.

The article of Agajanian et al. drives us more into the bioinformatics area. Recent applications of next-generation sequencing makes it possible to identify the role of mutations associated with cancer. The authors integrated multiple machine learning approaches to classify mutations an the basis of nucleotide sequence. The approach is further illustrated on biomolecular simulations of cancer associated protein kinases.

Tribello and Gasparotto use unsupervised machine learning methods to analyse simulation trajectories. Trajectory of the C-terminal fragment of the immunoglobulin binding domain B1 of protein G of *Streptococcus* was used as a model trajectory and analyzed by a range of mostly non-linear dimensionality reduction methods, namely principal component analysis, distance matching, Laplacian eigenmaps, Isomap, tSNE, and sketchmap. These methods are illustrated together with clustering methods. The article provides an overview of these methods and their advantages and disadvantages are discussed.

Kinetics of drug unbinding is recently becoming equivalently or even more important than binding thermodynamics in drug design as a parameter distinguishing between good and bad compounds. The article of Kokh et al. addresses this problem by machine learning. There are several trajectories of spontaneous drug binding available in literature. Drug unbinding is several orders of magnitude slower and today cannot be simulated without enhanced sampling. The authors analyzed a series of trajectories from enhanced sampling method Random Accelerated Molecular Dynamics, in particular its variant designed for simulation of drug unbinding kinetics. The approach has been tested on a series of heat shock protein 90 ligands differing by four orders of magnitude in their unbinding rates. Excellent agreement with experiment was obtained for most classes of compounds.

We believe that the papers included in this Research Topic demonstrate the great potential of machine learning in all fields pertaining biomolecular modeling and simulations, including in improving the accuracy of the models, in the analysis of molecular simulations and in providing effective variables to enhance the sampling. With this Research Topic *Frontiers in Molecular Biosciences* aspires to become a key forum for publishing of approaches combining machine learning with biomolecular simulations and further promote this multidisciplinary field.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to thank all authors, editors and reviewers of this Research Topic and Katie Powis and Justyna Lisowska from Frontiers for their help.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 April 2019
doi: 10.3389/fmolb.2019.00024

Atomic Motif Recognition in (Bio)Polymers: Benchmarks From the Protein Data Bank

*Benjamin A. Helfrecht, Piero Gasparotto, Federico Giberti and Michele Ceriotti**

Laboratory of Computational Science and Modeling, Institute of Materials, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Rationalizing the structure and structure–property relations for complex materials such as polymers or biomolecules relies heavily on the identification of local atomic motifs, e.g., hydrogen bonds and secondary structure patterns, that are seen as building blocks of more complex supramolecular and mesoscopic structures. Over the past few decades, several automated procedures have been developed to identify these motifs in proteins given the atomic structure. Being based on a very precise understanding of the specific interactions, these heuristic criteria formulate the question in a way that implies the answer, by defining a list of motifs based on those that are known to be naturally occurring. This makes them less likely to identify unexpected phenomena, such as the occurrence of recurrent motifs in disordered segments of proteins, and less suitable to be applied to different polymers whose structure is not driven by hydrogen bonds, or even to polypeptides when appearing in unusual, non-biological conditions. Here we discuss how unsupervised machine learning schemes can be used to recognize patterns based exclusively on the frequency with which different motifs occur, taking high-resolution structures from the Protein Data Bank as benchmarks. We first discuss the application of a density-based motif recognition scheme in combination with traditional representations of protein structure (namely, interatomic distances and backbone dihedrals). Then, we proceed one step further toward an entirely unbiased scheme by using as input a structural representation based on the atomic density and by employing supervised classification to objectively assess the role played by the representation in determining the nature of atomic-scale patterns.

Keywords: atomistic and molecular simulation, machine learning, biomolecules, molecular motifs, hydrogen bonds, secondary structure

OPEN ACCESS

Edited by:
Francesco Luigi Gervasio, University College London, United Kingdom

Reviewed by:
Carlo Camilloni, University of Milan, Italy
Fabio Pietrucci, Sorbonne Universités, France

****Correspondence:***
Michele Ceriotti
michele.ceriotti@gmail.com

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *03 December 2018*
Accepted: *26 March 2019*
Published: *18 April 2019*

Citation:
Helfrecht BA, Gasparotto P, Giberti F and Ceriotti M (2019) Atomic Motif Recognition in (Bio)Polymers: Benchmarks From the Protein Data Bank. Front. Mol. Biosci. 6:24. doi: 10.3389/fmolb.2019.00024

1. INTRODUCTION

Macromolecules are characterized by their capability of folding and assembling into hierarchical structures, which is a crucial element in their activity and stability. Understanding the structure of a macromolecule is thus a key step in discerning its function. Proteins are the archetypal example of complex molecular machines designed to perform unique and well-defined operations. Many polypeptides exhibit distinct secondary and tertiary structures in their native state, which are often used to explain their behavior. However, understanding and characterizing the structure of a macromolecule, even in the case of small proteins, can be rather difficult.

The structural description of a non-rigid molecule with many degrees of freedom relies on the identification of motifs, which can be used to classify their three-dimensional structure (e.g., an alpha-helix or beta-sheet in the case of a protein). The most common motifs that characterize these kinds of structures are intramolecular hydrogen bonds, such as those present in polysaccharides, as well as distinct dihedral angle patterns that are assumed by the backbone of a protein. Much work has been dedicated to understanding and classifying hydrogen bonds, ultimately producing several geometric criteria (e.g., distances and angles between donors and acceptors) as well as energetic criteria, to identify their presence or absence (Rahman and Stillinger, 1971; Brown, 1976; Mezei and Beveridge, 1981; Baker and Hubbard, 1984; Luzar and Chandler, 1993; McDonald and Thornton, 1994; Luzar and Chandler, 1996; Xu et al., 1997; Desiraju and Steiner, 2001; Arunan et al., 2011; Jeffrey and Saenger, 2012). Likewise, tabulating the different backbone dihedral angles exhibited by a macromolecule produces the so-called Ramachandran plot (Ramachandran et al., 1963), which finds widespread use in chemistry, biology, and biophysics to aid in the identification of protein secondary structure (Frishman and Argos, 1995).

There are several examples where this motif-based rationale was successfully employed to identify the secondary structure of proteins; the DSSP (Kabsch and Sander, 1983) and STRIDE (Frishman and Argos, 1995) algorithms are two notable examples. However, the identification of structural motifs in proteins is often based on a combination of human intuition and—sometimes generous—approximations, and may not be unique or readily applicable to different macromolecules. Moreover, the motif definitions are typically based on assessments of specific structures or, in the case of the hydrogen bond, focus solely on a single subset of the atomic species that may be involved.

In this context, a statistical framework capable of automatically identifying structural motifs that is free of energy approximations and relies on system-agnostic definitions would be advantageous. Having a purely data-driven definition of various motifs would be particularly useful in the field of bioinformatics, where they are used for structure prediction or the development of scoring functions for processes like protein-ligand docking. For example, Rosetta, one of the most well-known energy functions, has been developed to predict the structure of a protein given its amino acid sequence and local structural features such as dihedral angles (Simons et al., 1997, 1999).

Another example where purely data-driven definitions would be advantageous is in secondary structure classification. While several methods exist to classify protein secondary structure (Kabsch and Sander, 1983; Frishman and Argos, 1995, 1996; Jones, 1999; Cuff and Barton, 2000; Andersend et al., 2002; Martin et al., 2005; Nagy and Oostenbrink, 2014; Haghighi et al., 2016), these methods rely on amino acid sequences, hydrogen bonding energies, geometrical criteria, or some combination thereof. Machine learning techniques (Muggleton et al., 1992), and neural networks in particular (Holley and Karplus, 1989; Rost and Sander, 1993a,b; Jones, 1999; Cuff and Barton, 2000; Akkaladevi et al., 2004; Wood and Hirst, 2005; Rashid et al., 2016; Zhang et al., 2018) have also been used to classify the secondary structure of a protein based on a variety of features. Others have developed schemes to classify conformational patterns and secondary structure using dihedral angles alone (Hollingsworth et al., 2012; Nagy and Oostenbrink, 2014), but there remains a lack of a truly agnostic method for classifying (and predicting) secondary structures.

In this work, we illustrate how it is possible to use machine learning to obtain a statistical definition of atomic-scale motifs based on a data-driven analysis. Given a descriptor of the atomistic environments, we construct a probability density representing its occurrence in a given dataset. Then, using the Probabilistic Analysis of Molecular Motifs (PAMM) algorithm (Gasparotto and Ceriotti, 2014; Gasparotto et al., 2018), which casts the probability density into a Gaussian mixture model (GMM), we find the most probable motifs in the distribution. To create the density distribution we have used two different approaches: one using classical geometric descriptors such as interatomic distances and dihedral angles, and a more agnostic scheme that uses the Smooth Overlap of Atomic Positions (SOAP) framework (Bartók et al., 2013; Bartók and Csányi, 2015; De et al., 2016) as the input representation. The motif fingerprints obtained in this way have a general definition and are transferable between different systems. To illustrate this point, rather than selecting proteins of a given family or with small variations in the sequence, we have used entries from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Berman et al., 2000). The motifs obtained from PAMM were compared to a more "traditional" geometric definition of a hydrogen bond and to DSSP- and STRIDE-assigned secondary structures to assess their similarity. Furthermore, by comparing the fidelity of the unsupervised classification given by PAMM with that of a supervised scheme, we can assess whether classification errors stem from an incomplete representation or are a manifestation of the arbitrary nature of heuristic methods.

2. METHODS

The methods we used to represent structures and identify molecular motifs have been already discussed elsewhere. We used the PAMM scheme (Gasparotto and Ceriotti, 2014; Gasparotto et al., 2018) to identify modes in the probability distribution of atomic patterns. The PAMM algorithm takes as input a series of vectors representing local environments (distances, angles or more generic density-based representations such as SOAP feature vectors Bartók et al., 2013; De et al., 2016), performs a kernel density estimation on a sparse grid obtained by subsampling the input data, and performs a density-based clustering to identify local maxima in the estimate of the probability distribution. Finally, each cluster is represented as a Gaussian mode, which makes it possible to define probabilistic motifs identifiers (PMIs), structural indicators taking a value between zero and one that represent the degree of confidence by which a new local structure can be assigned to each of the clusters. In what follows we only summarize the aspects that are relevant to this specific

application, explaining in detail the preparation of the structures as well as how the pattern recognition has been performed for each descriptor. All the structures used in the definition of the structural motifs, regardless of the underlying descriptor used, were obtained from the RCSB PDB database on January 31, 2018. Note that the PDB contains redundant entries, i.e., protein structures with very similar sequences. These redundant structures were included in our analyses, and so the resulting models are biased according to the redundancies of the PDB.

2.1. Hydrogen Bond Definitions

As a first benchmark of the application of automatic pattern recognition schemes to (bio)polymers, we consider the case of the hydrogen bond (HB). While there is no shortage of alternative hydrogen-bond definitions based on structure, and PAMM has already been applied to the identification of HBs in water and ammonia (Gasparotto and Ceriotti, 2014; Gasparotto et al., 2016), proteins offer a test case that is more chemically diverse and one for which concrete definitions have been proposed. The latter makes it possible to establish a comparison between our automatic pattern recognition schemes and established categorical descriptions.

2.1.1. Hydrogen Bond Data Selection

The downside of using experimentally determined structures as the basis of our analysis is that the precision of the structural determination—particularly for hydrogen atoms—is limited and varies greatly between entries in the PDB. Given that hydrogen positions are obviously central to the definition of a hydrogen bond motif, only protein crystal structures obtained by X-ray diffraction with a resolution better than 1.2 Å and that included hydrogen atom positions were considered viable. Only 872 structures in the PDB met these requirements and could be properly parsed. Given that each structure contains hundreds of hydrogen bonds, this amount of data was sufficient for our statistical analysis.

From each protein structure, we considered only N, O, and H atoms with occupancy ≥ 0.95. Any oxygen and hydrogen atoms belonging to water or other small molecules were excluded. Four different hydrogen bond flavors were examined, depending on the nature of donor and acceptor: (1) $\mathrm{N-H\cdots N}$; (2) $\mathrm{N-H\cdots O}$; (3) $\mathrm{O-H\cdots O}$; (4) $\mathrm{O-H\cdots N}$.

2.1.2. Geometry Descriptors

For the determination of hydrogen bonding motifs, we examined all triplets of atoms, where one (O or N) is considered as the putative donor, one (O or N) is considered as the putative acceptor, and one is the H atom taking part in the bond. We considered separately the cases in which O and N act as either the donor or the acceptor, i.e., $\mathrm{N-H\cdots N}$, $\mathrm{N-H\cdots O}$, $\mathrm{O-H\cdots O}$, $\mathrm{O-H\cdots N}$. We did not use any additional criterion to identify which atoms could be part of a hydrogen bond, which means that the analysis considers as putative hydrogen bonds also triplets in which the three atoms are chemically bound or adjacent to one another in the backbone or in a side chain. Most of the traditional definitions of hydrogen bonds would implicitly discard these configurations and not consider them altogether. While it would be straightforward to eliminate such configurations as a preliminary step to our analysis, we retained them to serve as a demonstration of the robustness of using PAMM for identifying distinct structural patterns.

Even in protein structures obtained from high-resolution X-ray diffraction, hydrogen positions are often "refined." In other words, each hydrogen atom is often fixed at a predetermined distance from the atom to which it is covalently bound (Watkin, 2008; Cooper et al., 2010). To ensure that this artificial feature would not further bias the clustering, only the donor–acceptor and acceptor–hydrogen distances were chosen as geometrical descriptors for each hydrogen bond. Ignoring the donor–hydrogen distance does not limit the resolving power of a PAMM analysis, but makes it impossible to automatically eliminate some configurations with a very large donor–hydrogen distance. For this reason, before proceeding with the clustering, we further filtered the hydrogen bonds using the same geometric criteria that has been used in earlier studies of hydrogen bonding in water (Gasparotto and Ceriotti, 2014; Gasparotto et al., 2016), which relies on all of the donor–acceptor, donor–hydrogen, and acceptor–hydrogen distances (d_{DA}, d_{DH}, and d_{AH}, respectively). Those triplets in which the sum of d_{DH} and d_{AH} was greater than 4.5 Å were discarded in addition to those in which d_{DH} was greater than d_{AH}. The latter refinement reduces redundancies when examining different hydrogen bond flavors, as a given triplet with $d_{DH} > d_{AH}$ in $\mathrm{N-H\cdots O}$ is equivalent to that same triplet with $d_{DH} < d_{AH}$ in $\mathrm{O-H\cdots N}$; the donor and acceptor labels have just been interchanged. With these conditions, we identified several hundred thousand potential $\mathrm{N-H\cdots N}$ and $\mathrm{N-H\cdots O}$ triplets and 40–60 thousand $\mathrm{O-H\cdots O}$ and $\mathrm{O-H\cdots N}$ triplets that we retained for further analysis.

2.1.3. Clustering Parameters

To reduce the computational cost of the procedure while sampling all relevant values of the d_{DA} and d_{AH} distances we selected a sparse grid of 2000 configurations on which we computed a kernel density estimation of the probability distribution of different motifs. An approximately uniform distribution of grid points is achieved using a well-established farthest-point sampling (FPS) scheme (Ceriotti et al., 2013). The kernel bandwidth and local scale factors were determined automatically as discussed in Gasparotto et al. (2018). The automatically determined bandwidth was scaled by a factor of 0.3 to account for the strong multi-modality of the distribution, while we found the automatic choice of quick-shift distance to be appropriate. Clusters with weights less than 10^{-5} in the resulting mixture model were discarded, as they were sparsely populated and did not meaningfully contribute to the overall probability distribution and could be considered outliers.

2.1.4. Probabalistic Motif Indentifiers (PMIs)

For each hydrogen bond flavor, the PMI $f(\mathbf{x})$ at a point $\mathbf{x} = (d_{AH}, d_{DA})$ is calculated as in Gasparotto and Ceriotti (2014) and Gasparotto et al. (2018),

$$f(\mathbf{x}) = \frac{p_{HB} G(\mathbf{x}|\boldsymbol{\mu}_{HB}, \boldsymbol{\Sigma}_{HB})}{P(\mathbf{x}) + \zeta}, \tag{1}$$

where p_{HB} is the weight of the Gaussian G with mean $\boldsymbol{\mu}_{HB}$ and covariance $\boldsymbol{\Sigma}_{HB}$ describing the hydrogen bond, ζ is the background parameter, set to 10^{-5} for our purposes, and $P(\mathbf{x})$ is the total probability density of the GMM,

$$P(\mathbf{x}) = \sum_{k}^{N} p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where N is the total number of clusters in the model.

The PMI for a distance–angle geometry-based definition of the hydrogen bond is:

$$f(\mathbf{x}) = \begin{cases} 1, & d_{DA} < 3.5 \text{ Å}, d_{AH} < 2.5 \text{ Å}, d_{DH} < 1.5 \text{ Å}, \angle ADH < 30.0^{\circ} \\ 0, & \text{else} \end{cases} \tag{3}$$

As another example, the DSSP (Kabsch and Sander, 1983) definition of an N − H···O hydrogen bond, which is based on the distances d between the atoms participating in the C=O bond of one residue and the N − H bond of another residue, can also be used to construct a PMI.

To construct the DSSP-based PMI, we computed the required DSSP distances for all {N, H, C, O} quadruplets in each protein for which all four atoms have occupancy ≥ 0.95, and map the quadruplet to (d_{AH}, d_{DA}) space simply by taking d_{AH} as the oxygen–hydrogen distance and d_{DA} as the nitrogen–oxygen distance. Then for each $\mathbf{x} = (d_{AH}, d_{DA})$, we computed the joint probability distribution

$$P_{HB}(\mathbf{x}) = P(\mathbf{x}, E_{DSSP} < -0.5 \text{ kcal/mol}), \tag{4}$$

where E_{DSSP} is the DSSP electrostatic energy as defined in Kabsch and Sander (1983). The DSSP-based PMI can then be constructed following Equations 1, 2 by replacing $G(\mathbf{x}|\boldsymbol{\mu}_{HB}, \boldsymbol{\Sigma}_{HB})$ with the joint probability density $P_{HB}(\mathbf{x})$ and by defining the total probability density as

$$P(\mathbf{x}) = p_{HB} P_{HB}(\mathbf{x}) + (1 - p_{HB}) P(\mathbf{x}, E \geq -0.5 \text{ kcal/mol}). \tag{5}$$

where the weight p_{HB} is the fraction of C = O, N − H pairs that have $E < -0.5$ kcal/mol. It should be noted that the DSSP PMI is based on only a subset of the data used to define the PAMM PMIs and contains approximately 550,000 N − H···O triplets. As stated in 2.1.1 in the Methods, we discarded atoms from the analysis that had an occupation less than 0.95 in order to train PAMM on unambiguous atomic geometries. This, combined with the fact that the DSSP definition requires the positions of C atoms, means that the DSSP PMI was built considering only (C = O, N − H) pairs in the protein backbone for which each of the C, O, N, and H atoms had an occupation < 0.95, narrowing the dataset.

In order to compare different HB definitions and to quantify how often they disagree in identifying a local motif in $\mathbf{x} = (d_{AH}, d_{DA})$ space as an HB, we introduce the quantity

$$\delta_{AB} = \frac{1}{\lambda} \frac{\int P_{total}(\mathbf{x}) f_A(\mathbf{x}) f_B(\mathbf{x}) d\mathbf{x}}{\int P_{total}(\mathbf{x}) \left[f_A(\mathbf{x}) + f_B(\mathbf{x}) - f_A(\mathbf{x}) f_B(\mathbf{x}) \right] d\mathbf{x}}, \tag{6}$$

which is the probability that the PMIs A and B both identify point $\mathbf{x}$ as an HB relative to the probability that either one or the other identify an HB. $P_{total}(\mathbf{x})$ is the total probability distribution of observing (d_{AH}, d_{DA}) in the PDB dataset across all hydrogen bond flavors. The normalization factor λ is included to account for the fact that the PMIs f are posterior probabilities rather than true probability distributions. Thus, λ is chosen such that Equation 6 is equal to one when $f_A(\mathbf{x}) = f_B(\mathbf{x})$:

$$\lambda = \sqrt{\frac{\int P_{total}(\mathbf{x}) f_A^2(\mathbf{x}) d\mathbf{x}}{\int P_{total}(\mathbf{x}) \left[2 f_A(\mathbf{x}) - f_A^2(\mathbf{x}) \right] d\mathbf{x}} \cdot \frac{\int P_{total}(\mathbf{x}) f_B^2(\mathbf{x}) d\mathbf{x}}{\int P_{total}(\mathbf{x}) \left[2 f_B(\mathbf{x}) - f_B^2(\mathbf{x}) \right] d\mathbf{x}}} \tag{7}$$

2.2. Dihedral Angles for Secondary Structure Recognition

Secondary-structure patterns play a central role in rationalizing the structure and behavior of proteins. Well-established definitions exist based on the identification of HBs along the protein backbone, such as STRIDE (Frishman and Argos, 1995) and DSSP (Kabsch and Sander, 1983). There is, however, a need for definitions of secondary structure that are based on continuous structural coordinates, for instance, to bias atomistic simulations or to perform structure searches (Pietropaolo et al., 2008; Pietrucci and Laio, 2009). As an example of how one can use PAMM to provide a definition of secondary structure motifs that is based on a simple, local representation of the backbone, we used the Ramachandran dihedrals (Ramachandran et al., 1963), whose strong correlation to secondary structure has been long appreciated (Hollingsworth et al., 2012; Wood and Hirst, 2005; Kountouris and Hirst, 2009).

2.2.1. Dihedral Angle Data Selection

Because the calculation of dihedral angles is not sensitive to hydrogen atomic positions, the PAMM analysis of dihedral angles included all experimental protein crystal structures from the RCSB PDB (as of January 31, 2018) obtained from X-ray diffraction with a resolution better than 1.5 Å, totaling 12,708 structures and 4,275,677 residues from which dihedral angles could be extracted. Note again that no measures were taken to discard redundant structures from the PAMM analysis, hence the resulting mixture model is biased according to the redundancies of the PDB.

2.2.2. Clustering and Secondary Structure Classification

Using PAMM, a GMM of the backbone dihedral angles (ϕ and ψ) calculated with BioPython (Cock et al., 2009) was constructed. We performed a kernel density estimation on 4000 FPS grid points. In this case, we used a bandwidth scaling factor of 0.15, and a scaling of the quick-shift threshold of 0.20, compared to the values determined automatically based on the heuristics discussed in Gasparotto et al. (2018). We found that the automatic parameters were smoothing excessively the distribution, resulting in a loss of resolving power. We determined the optimal parameters by monitoring the number of clusters and their robustness as assessed by a bootstrapping

analysis. We also constructed PAMM GMMs based on higher dimensional feature spaces based on chains of ϕ and ψ angles in consecutive residues. Here we again used 4000 grid points but selected a bandwidth scaling factor of 0.30 and set the quick-shift scaling to 0.80. Similar to the case of the HB, we discarded clusters with weights $< 10^{-5}$.

2.2.3. Comparison of Secondary-Structure Definitions

Given that each point $\mathbf{x} = (\phi, \psi)$ corresponding to a single amino acid residue is associated with a secondary structure classification y from DSSP/STRIDE and a cluster assignment A with probability $p^{(A)}(\mathbf{x})$ from PAMM, a joint probability distribution $P(A, y)$ can be constructed by summing the cluster probabilities over all points $\mathbf{x}_y$ with secondary structure y,

$$P(A, y) = \frac{1}{N} \sum_{\mathbf{x}_y} p^{(A)}(\mathbf{x}_y), \tag{8}$$

where N is the total number of residues considered. $P(A, y)$ characterizes completely the relationship between the two definitions. Based on the joint probability we can compute the marginals $P(A)$ and $P(y)$ and the conditional probabilities $P(A \mid y)$ and $P(y \mid A)$, which provide equivalent information and make it easy to identify the correspondence—if any—between the PAMM-based PMI and the conventional definitions. For reference, the DSSP and STRIDE secondary structure classifications are as follows: B, isolated β-bridge; E, extended strand; G, 3_{10}-helix; H, α-helix; I, π-helix; T, turn; S, bend (DSSP only); C, loop, irregular element, or none of the above ("coil"). We use an "X" to signify an amino acid residue for which no secondary structure was assigned.

One can summarize the ability of the automatic definition to reproduce the classification given by STRIDE or DSSP by viewing the joint probability $P(A, y)$ in the framework of the Q3 (or Q8) accuracy score (Rost and Sander, 1993a). Given a particular clustering arrangement, one or more clusters can be selected that individually correspond to strands (B, E), helices (G, H, I) or coils (C, S, T) by assigning each cluster A the secondary structure that maximizes $P(y \mid A)$.

Thus, for sets of clusters $\mathcal{E}, \mathcal{H}, \mathcal{C}$ corresponding to strands, helices, and coils, the Q3 score is the sum $Q_\mathcal{E} + Q_\mathcal{H} + Q_\mathcal{C}$, where

$$Q_\mathcal{E} = \sum_{i \in \mathcal{E}} \big(P(i, B) + P(i, E)\big) \tag{9a}$$

$$Q_\mathcal{H} = \sum_{j \in \mathcal{H}} \big(P(j, G) + P(j, H) + P(j, I)\big) \tag{9b}$$

$$Q_\mathcal{C} = \sum_{k \in \mathcal{C}} \big(P(k, C) + P(k, S) + P(k, T)\big), \tag{9c}$$

and the secondary structure assignments B, E, G, H, I, C, S, and T are those determined by DSSP or STRIDE.

2.3. Smooth Overlap of Atomic Positions Representation

The analysis protocols that we have discussed above identify the presence of significant motifs based exclusively on how often a given local atomistic environment occurs in a reference dataset. While this procedure makes it possible to rely on simple and rather generic descriptors of local structure, it still requires a dose of chemical intuition, i.e., it is necessary to know the basis of hydrogen bonding and that dihedral angles can be used to identify the secondary structure of a protein. To fulfill our goal of creating a completely agnostic framework, one would need to use a more abstract, generally applicable measure of the atomistic environment that does not require any chemical intuition. To this end, we have employed SOAP, a method that can represent each chemical environment in a complete way and that can be applied seamlessly to any system, from biomolecules to materials.

2.3.1. Brief Introduction to SOAP

Before explaining the clustering procedure and parameters used with SOAP, we briefly introduce the representation. This is not meant to be a complete introduction, and we redirect the interested reader to more detailed papers previously published on the topic (Bartók et al., 2013; Bartók and Csányi, 2015; De et al., 2016). The SOAP vector is a recently introduced, atom-centered, density-based representation that has been used in many applications, from solids to molecular systems. It has been proven useful in describing and predicting many atomic and molecular properties such as structure and energy (De et al., 2016). The SOAP framework represents the atomic density around an atom j as a sum of Gaussians centered on each surrounding atom of species α. The sum can be cast into a smooth, local probability amplitude $\psi^\alpha_{\mathcal{X}_j}(\mathbf{r})$ by employing a cutoff function f_c that determines the extent of the local environment:

$$\langle \alpha \mathbf{r} | \mathcal{X}_j \rangle \equiv \psi^\alpha_{\mathcal{X}_j}(\mathbf{r}) = \sum_{i \in \alpha} f_c(\mathbf{r}_{ij}) g(\mathbf{r} - \mathbf{r}_{ij}). \tag{10}$$

The main parameters determining the behavior of the SOAP features are the cutoff distance—which defines the range of structural correlations that are deemed to be relevant—and the width of the Gaussian functions—which determines the sensitivity to atomic displacements.

In the original formulation of SOAP (Bartók et al., 2013), the atom density is expressed by expanding the environmental density in a basis of orthogonal radial basis functions $R_n(r)$ and spherical harmonics $Y^l_m(\hat{\mathbf{r}})$,

$$\langle \alpha n l m | \mathcal{X}_j \rangle = \int \mathrm{d}\mathbf{r} R_n(r) Y^l_m(\hat{\mathbf{r}}) \langle \alpha \mathbf{r} | \mathcal{X}_j \rangle. \tag{11}$$

This amplitude is invariant to translations in addition to permutations of atoms within each species α, but it is not invariant to rotations. Rotation invariance can be achieved by integrating the overlap between two atomic environments $\mathcal{X}$ over all relative rotations $\hat{R}$, yielding the kernel,

$$K^{(\nu)}(\mathcal{X}_j, \mathcal{X}_k) = \int \mathrm{d}\hat{R} \langle \mathcal{X}_j | R | \mathcal{X}_k \rangle^\nu. \tag{12}$$

For $\nu = 2$, the kernel is equivalent to the scalar product between the power spectra of environments j and k,

$$K^{(2)}(\mathcal{X}_j, \mathcal{X}_k) = \sum_{\alpha n \alpha' n' l} \langle \mathcal{X}_j | \alpha n \alpha' n' l \rangle \langle \alpha n \alpha' n' l | \mathcal{X}_k \rangle. \tag{13}$$

The power spectrum vectors $\langle \alpha n \alpha' n' l | \mathcal{X}_k \rangle$ can be used as an explicit, general, and complete representation of chemical environments.

2.3.2. SOAP Data Selection

Although SOAP is a powerful descriptor, the high dimensionality of the SOAP vectors $\langle \alpha n \alpha' n' l | \mathcal{X}_k \rangle$ makes PAMM pattern recognition based on these descriptors computationally intractable for large datasets. Therefore, we first performed a Principal Component Analysis (PCA) of the SOAP vectors with the aim of reducing the dimension of the input space for PAMM while maintaining the most discriminating SOAP features of the individual proteins. To accelerate the process, we used an FPS subset of SOAP components to reduce the input space for the PCA while maintaining its span. In particular, we selected 100 random structures from the same set used in the dihedral angle clustering and computed the SOAP vectors for all of the C_α atoms in the selected structures, taking into consideration all C, N, and O atoms within a cutoff radius of 6.0 Å as part of the local environment, which is large enough to incorporate information on several neighboring residues. From this collection of SOAP vectors, we selected 200 SOAP components via FPS, using the squared Euclidean distance between the SOAP vectors as the measure of separation (Imbalzano et al., 2018).

The SOAP vectors centered around all C_α atoms were then computed for all structures just as they were for the random subset, but only the FPS-selected components were kept and used to build the PCA representation; all other components of the SOAP vector were discarded. The full parameters used to generate the SOAP vectors are given in the **Supplemental Material**.

2.3.3. Clustering and Classification

The first 2, 6 and 10 PCA components of the reduced SOAP vectors were clustered by PAMM using 4000 grid points and a quick shift parameter of 1.0. The Kernel Density Estimation bandwidth scaling factor was chosen to be 0.20, 0.50, and 0.80 for the 2, 6, and 10 PCA component representations respectively. Clusters with weights $< 10^{-5}$ were discarded.

2.3.4. Probability Distribution

Because each individual reduced SOAP vector is based on an expansion around the C_α atoms, each vector corresponds to a single residue and therefore can be associated with a DSSP- or STRIDE-assigned secondary structure. The joint and conditional probability distributions for the reduced SOAP vectors clustered by PAMM were computed in the same manner as those for the dihedral angles, as were the Q3 and Q8 scores relative to DSSP and STRIDE.

2.4. Supervised Classification

Given that the SOAP representation can be tuned to encompass environments of different sizes and provide a complete description of the correlation between atomic positions, it gives us an opportunity to verify whether any discrepancy between the PAMM classification and the reference heuristics is due to the fact that the truncated representations that we use are incomplete, or due to the fact that the reference heuristics are not reflected in the probability distribution of motifs in the PDB. We can assess the completeness of the representations by training a *supervised* model to recognize DSSP or STRIDE motifs; that is, we can associate the SOAP description of the atomic environment $\mathcal{X}_i$ of each C_α atom with the label y_i assigned to it by DSSP or STRIDE. To perform this classification task we used a support vector machine (SVM) (Cortes and Vapnik, 1995) as implemented in the `scikit-learn` Python package (Pedregosa et al., 2011) to perform multiclass classification of a PCA of SOAP environments $\mathcal{X}_i$ according the labels y_i. For comparison, SVMs using backbone dihedral angles were also constructed. The SVMs employed a "one vs. one" classification scheme (Knerr et al., 1990) with a Gaussian kernel with width $\gamma = 1/N_f$, where N_f is the number of features, and regularization parameter $C = 1.0$. Furthermore, the SOAP PCA and dihedral angle data were scaled to have zero mean and unit variance before building the SVM. Of the approximately 4.3 million residues present in our dataset, we selected 200,000 residues at random (excluding those that were not assigned a secondary structure by DSSP or STRIDE) to train and evaluate the SVM. Of these 200,000 residues, 50,000 were randomly selected to serve as the training set, and the remaining 150,000 served as the test set. The asymptotic (large train set size) classification accuracy of the supervised model indicates the limit that can be achieved with a given environment representation. Learning curves of the Q3 and Q8 scores for the SVM are provided in the **Supplemental Material**.

3. RESULTS AND DISCUSSION

3.1. Hydrogen Bonds

Let us start by discussing the definition of HBs based on a traditional distance–angle criterion. **Figure 1** shows the probability distribution of (d_{AH}, d_{DA}) computed by accumulating simultaneously all four kinds of HBs. The PMI associated with the conventional definition of the hydrogen bond is highlighted. This definition encompasses a large peak in $P(\mathbf{x})$ that indeed corresponds to hydrogen-bonded configurations, but it also includes several additional peaks. By inspection, we found that these additional modes of the distribution are associated with motifs in which the putative donor and acceptor atoms are part of the same amino acid residue or where the H atom is not chemically bound to the donor. In practice, these geometries would be discarded a priori because most codes for analyzing biomolecular data take covalent bonding information into account. **Figure 1**, however, underscores the complex heuristics that are necessary to apply well-established definitions of atomic-scale motifs, and serves as a warning of the risks one could incur when blindly following these prescriptions in a different context than the usual forcefield simulations in which the chemical connectivity is fixed.

Similar considerations apply to the DSSP definition, whose corresponding PMI is shown in **Figure 2**. The DSSP definition follows more closely the main HB peak of the distribution, as one would expect given that it is heavily fine-tuned for one specific flavor of bond, N − H···O, between peptide groups. At the same time, DSSP also requires further heuristics to discard spurious

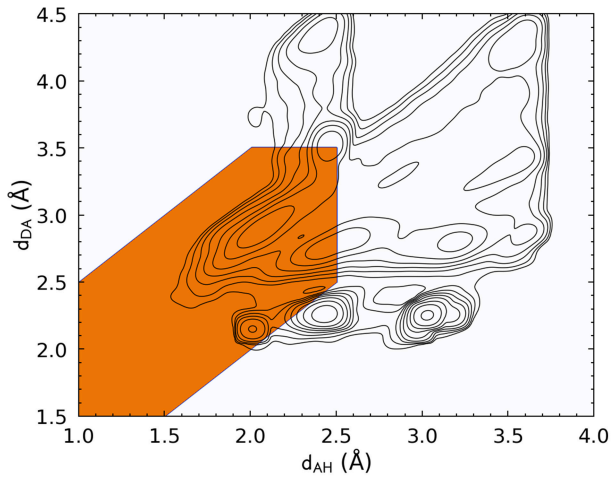


FIGURE 1 | Histogram of the acceptor–hydrogen and donor–acceptor distances across all hydrogen bond flavors, plotted with log-spaced contours. The maximum at ($d_{AH} \approx 2.1$Å, $d_{DA} \approx 2.8$Å) corresponds to the typical H-bond range. Other maxima are associated with other structural features, such as covalently bound groups on the side chains, geometries in which the two electronegative atoms are in the same residue, or configurations in which the hydrogen atom is not bound to the donor. The orange-shaded area corresponds to the distance-angle PMI as defined in Equation 3.

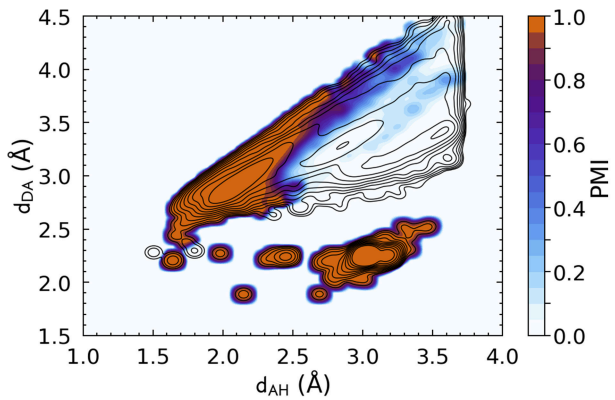


FIGURE 2 | Density plot of the PMI constructed using the DSSP hydrogen bond definition with $\zeta = 10^{-5}$. The PMI is plotted on top of a histogram of the distance features for N − H···O hydrogen bonds (discarding non-backbone groups, and any triplet for which it is not possible to define a DSSP H-bond energy, e.g., due to partial occupations), with log-spaced contours. DSSP identifies very clearly the H-bond peak, but also picks up spurious correlations corresponding to immediately adjacent residues [peak at ($d_{AH} \approx 3.0, d_{DA} \approx 2.25$)].

correlations corresponding to N − H and C=O in immediately adjacent residues, where ($d_{AH} \approx 3.0, d_{DA} \approx 2.25$).

Contrast these figures with the top row of **Figure 3**, which shows the PAMM PMIs for each cluster in the GMMs, computed separately for the four hydrogen bond flavors. The four distributions differ substantially from each other, and from the overall $P(\mathbf{x})$, while exhibiting multiple modes that are correctly identified by PAMM and assigned different cluster indices. Some of these modes correspond to correlations between covalently bound atoms, while others correspond to longer-range intermolecular correlations. For each flavor, the cluster that corresponds to the hydrogen bond is that with its center (mode) nearest to ($d_{AH} = 1.82$ Å, $d_{DA} = 2.74$ Å) (Gasparotto and Ceriotti, 2014). The corresponding PMIs, which are plotted in the bottom row of **Figure 3**, identify with great precision the region in the probability distribution that corresponds to the HB, and eliminate automatically the spurious configurations due to adjacent residues or covalently bound groups without the need for additional heuristics.

Figure 3 also shows that different kinds of hydrogen bonds correspond to noticeably different portions of (d_{AH}, d_{DA}) pattern space (a figure comparing different definitions is shown in the **Supplemental Material**). This means that a substantial fraction of molecular patterns would be misclassified if one tried to transfer the definition between different kinds of HB. As shown in **Table 1**, the probability that two definitions yield the same classification, as measured by Equation 6, can be as low at 50%. The agreement between the data-driven PMIs and the conventional distance–angle definition is even poorer, as shown in **Table 2** and in **Figure S2** in the **Supplemental Material**. It should be stressed, however, that this is largely due to the inclusion of correlations that are usually discarded by additional heuristics: if one computes the PMI similarity using a probability distribution $P_{total}(\mathbf{x})$ that discards atoms in the same or nearby residues, the probability increases substantially, particularly for N − H···N and N − H···O, as these are the flavors are responsible for the majority of spurious hydrogen bond geometries (e.g., intra-arginine or intra-histidine N − H···N triplets and backbone N − H···O triplets with donor and acceptor atoms in directly adjacent residues). The increase in PMI similarity is generally less pronounced when comparing two different hydrogen bond flavors because these PMIs are derived from a PAMM GMM, which automatically recognizes the spurious geometries as separate motifs. This example, although simple, demonstrates how one can use data-analytic techniques to extract definitions of molecular motifs based on experimental structural data. It also serves as a reminder of how heuristic definitions can lack transferability, and how their apparent simplicity is often contingent on a considerable amount of prior knowledge and the enforcement of additional conditions.

3.2. Dihedral Angles and Protein Secondary Structure

As another example of using simple geometric descriptors to find and evaluate atomic-scale motifs, we used PAMM to automatically detect dihedral angle motifs in proteins. Backbone dihedrals are central to our understanding of protein structure (consider, for example, the widespread use of the Ramachandran plot), and provide a rather unbiased description of a polymer chain that could be easily applied to other classes of polymers, whose structure is determined by different kinds of interactions.

The PMIs for each of the Gaussians in a PAMM GMM of the dihedral angles ϕ and ψ are shown in **Figure 4**. The PAMM dihedral angle clustering agrees well with those obtained

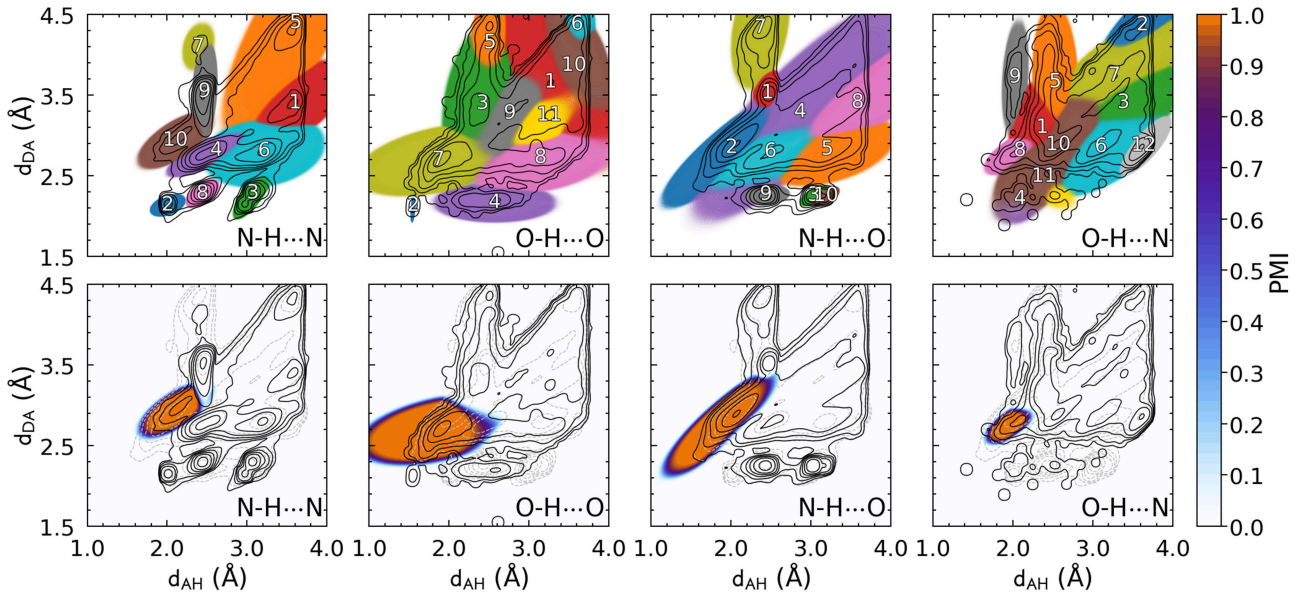


FIGURE 3 | The top panels represent all the clusters identified by PAMM. Clusters are numbered in an arbitrary order, and the colors reflect the cluster that is dominant in each region, as determined by its corresponding PMI [as defined in Equation (1), computed with $\zeta = 10^{-5}$]. Bottom panels highlights the PMI of the cluster associated with the hydrogen bond.

TABLE 1 | Probabilities that two PMIs corresponding to different hydrogen bond flavors agree that a point is a hydrogen bond (Equation 6).

PMI *A*	PMI *B*	δ_{AB}	$\delta_{AB}^{(i)}$	$\delta_{AB}^{(i+1)}$
N – H···N	N – H···O	0.92	0.93	0.94
N – H···N	O – H···O	0.57	0.63	0.74
N – H···N	O – H···N	0.60	0.59	0.60
O – H···O	N – H···O	0.55	0.61	0.71
O – H···O	O – H···N	0.60	0.68	0.85
N – H···O	O – H···N	0.57	0.57	0.58

The superscripts (i) and (i + 1) correspond to probabilities δ_{AB} where $P_{total}(\mathbf{x})$ excludes donor–hydrogen–acceptor triplets in which the donor and acceptor atoms are in the same residue (i), or additionally in adjacent residues (i + 1).

TABLE 2 | Probabilities that the hydrogen bond PMI and the distance–angle definition agree that a point is a hydrogen bond (Equation 6).

Bond type	δ_{AB}	$\delta_{AB}^{(i)}$	$\delta_{AB}^{(i+1)}$
N – H···N	0.56	0.65	0.89
N – H···O	0.60	0.71	0.93
O – H···O	0.63	0.65	0.68
O – H···N	0.33	0.39	0.53

The superscripts (i) and (i + 1) correspond to probabilities δ_{AB} where $P_{total}(\mathbf{x})$ excludes donor–hydrogen–acceptor triplets in which the donor and acceptor atoms are in the same residue (i), or additionally in directly adjacent residues (i + 1).

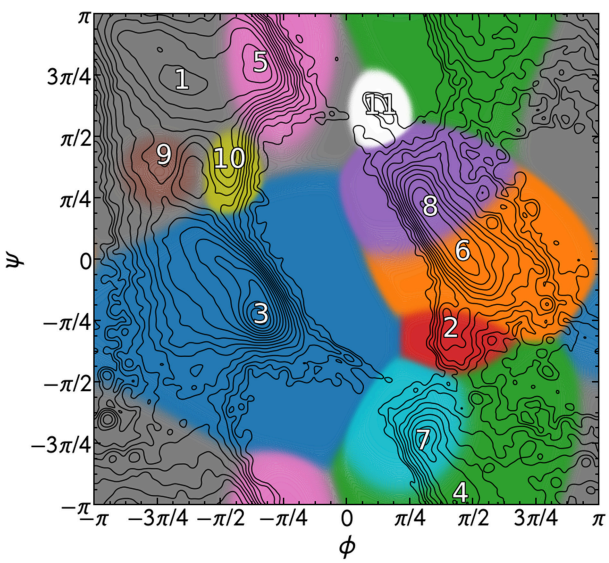


FIGURE 4 | PAMM clustering of all calculated dihedral angles with $\zeta = 0$. Cluster numbers are placed at the mode of the cluster, and each cluster has been colored differently. The isocontours of the total distribution are equally spaced on a logarithmic scale.

by Hollingsworth et al. (2012) and Nagy and Oostenbrink (Nagy and Oostenbrink, 2014), who have previously developed classification schemes based solely on dihedral angles. However, we observe like Hollingsworth et al. that dihedral angle patterns do not necessarily correspond to established secondary structure definitions, which is made clear upon comparison of **Figure 5**, which shows 100,000 randomly selected dihedral angle pairs colored according to their DSSP and STRIDE secondary structure assignments, and the clusters presented in **Figure 4**. As we

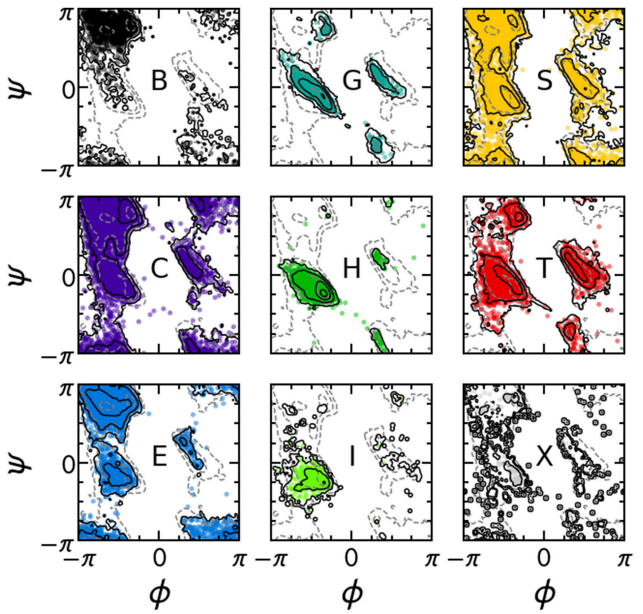


FIGURE 5 | Collection of 100,000 randomly selected (ϕ, ψ) pairs, separated according to the DSSP secondary structure classification of each pair. Solid contours correspond to the distribution of the secondary structure of interest; dashed contours correspond to the total distribution of all ϕ, ψ angles. Contours are equally spaced on a logarithmic scale.

will discuss further down, failure of dihedral angles to match established secondary-structure classifications is not due to an intrinsic lack of resolving power, but to the fact that dihedrals emphasize different kinds of structural correlations, so that secondary structure motifs are not associated with separate modes in feature space.

In order to quantify the correspondence between the PAMM cluster assignment and the secondary structure assignment, the joint and conditional probability distributions as outlined in section 2.2.3 were computed. **Figure 6** gives the joint and conditional probability distributions of the PAMM cluster assignment and the DSSP secondary structure assignment. (The probability distributions using the STRIDE secondary structure assignment are very similar to those using the DSSP assignment, and can be found in the **Supplemental Material**.)

Figure 6 Shows that there is a strong correlation between the most populated PAMM clusters (labeled by $A \in \{1, \ldots, 11\}$) and DSSP motifs (labeled by $y \in \{B, C, E, G, H, I, S, T, X\}$), with $A = 1, y = E$ and $A = 3, y = H$ being by large the most probable mutual assignments. The joint probability distribution, however, is not easy to interpret because of the widely varying populations of the different clusters. For this reason, the figure also shows the conditional probabilities, which normalize the joint assignments based on the DSSP $[P(A \mid y)]$ and PAMM $[P(y \mid A)]$ marginals. This analysis shows that the PAMM Cluster 1 encompasses most of the strand-like motifs (B, E) and Cluster 3 encompasses most of the helices (G, H, I). The distribution conditional on DSSP assignments is also insightful, showing that a large fraction of E and H motifs are assigned to PAMM Clusters 1 and 3, while the distribution conditional on PAMM cluster shows that disordered motifs are more evenly spread across all of the clusters. This comparison suggests that conventional heuristics are consistent with the actual distribution of structures in well-characterized proteins when it comes to well-defined sheet and helical motifs. On the other hand—at least when seen through the lens of the Ramachandran angles—DSSP bends, turns and coils are not clearly identifiable with separate peaks in the observed probability distribution. There are nevertheless clusters that are associated with clear peaks, and that are not associated with helices or strands. This suggests that "disordered" sections of proteins exhibit substantial order on the scale of the conformation of individual residues, and that looking at the statistics and correlations of these local motifs might be a better approach to characterize disordered polypeptides than trying to fit them within the existing categories.

One can further contextualize the probability distributions with the framework of the Q3 or Q8 score. Assigning Cluster 1 (see **Figure 4**) to the "strand" classification, Cluster 3 to the "helix" classification, and associating all other clusters with the "coil" designation yields a Q3 score of 0.70 relative to DSSP and 0.72 relative to STRIDE.

The rather low value of the Q3 score is comparable to the reported match scores of DISICL (Nagy and Oostenbrink, 2014) (with our PAMM PMI-based method performing better relative to DSSP but more poorly relative to STRIDE), which is also based solely on backbone dihedral angles. However, the Q3 score of our cluster-based secondary structure assignments is substantially lower than other methods that rely on dihedral angles in addition to amino acid sequences (Wood and Hirst, 2005; Kountouris and Hirst, 2009), or C_α distances (Martin et al., 2005). In this context, the underperformance of our method in classifying secondary structure could be given two different justifications. One is that the traditional secondary structure motifs are based on rather arbitrary thresholds, that recognize configurations as separate modes even when there are no clearly distinct maxima in the distribution of atomic configurations, regardless of the (reasonable) choice of input representation. Another is that our specific choice of representation, i.e., pairs of backbone dihedrals, is insufficient to distinguish between different motifs because of its excessive locality. The latter hypothesis is supported by the large overlap of different DSSP motifs in dihedral space (**Figure 5**), and can be tested by using different representations of the atomic motifs as the input to a PAMM analysis.

As a means of including more non-local information into the model while relying on a representation based purely on dihedrals, we also performed a PAMM clustering on the dihedral angles of consecutive residues, comparing the cluster assignment to the DSSP and STRIDE secondary structure classifications of the middle residue in the sequence. Just as in the two-dimensional case, in six dimensions (three consecutive residues) and ten dimensions (five consecutive residues) the helices and strands are localized to one or two clusters, while the other secondary structures are distributed across several clusters

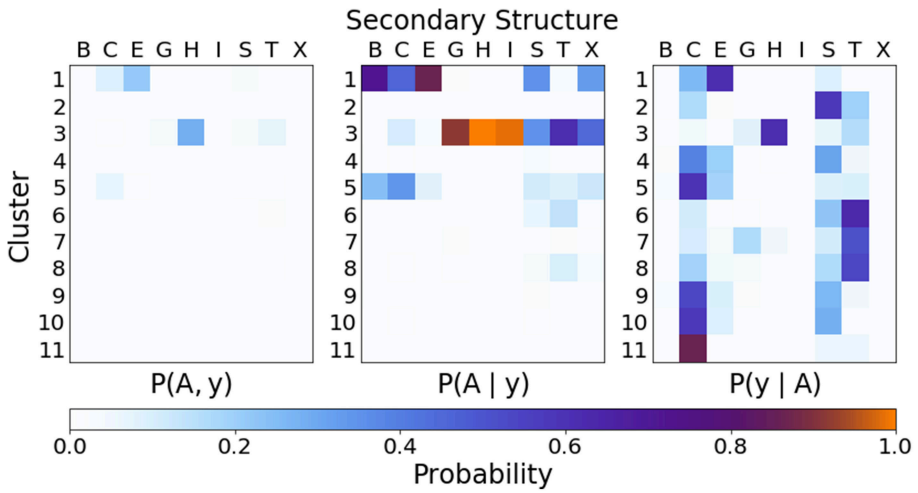


FIGURE 6 | Joint and conditional probabilities for the secondary structures obtained from DSSP and the clustering of dihedral angles from PAMM, where *A* is the cluster assignment and *y* the secondary structure classification.

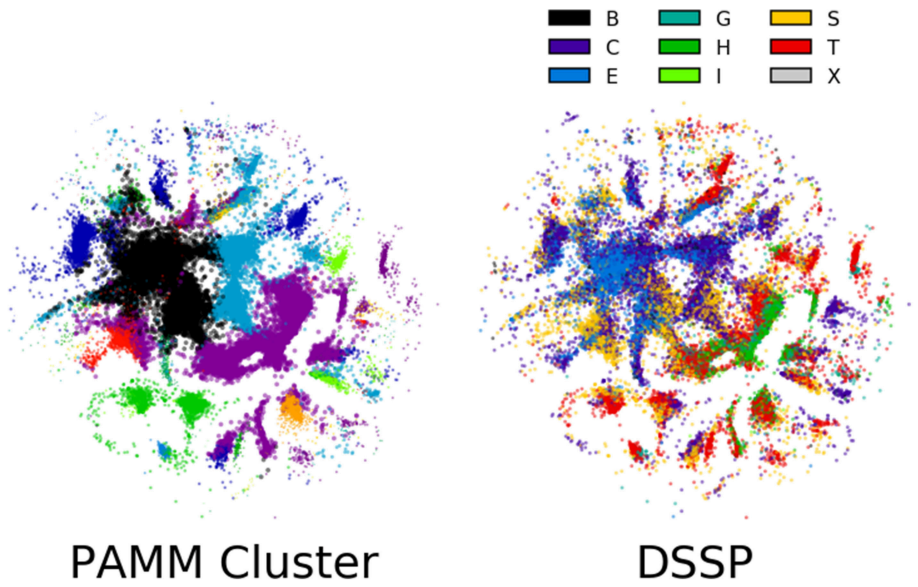


FIGURE 7 | Sketch-map representations of 100,000 randomly selected points in the six-dimensional ϕ, ψ space. Each point is colored according to its PAMM cluster assignment and middle residue DSSP secondary structure assignment. The lack of clear grouping observed among secondary structures suggests that secondary structure cannot be assigned based on dihedral angles alone. The points that are colored by their PAMM cluster are also sized based on the cluster weight; points belonging to a cluster with higher weight are larger.

(The probability distributions for the six- and ten-dimensional clusterings are given in the **Supplemental Material**). As a consequence, the Q3 score is largely the same among the two-, six-, and ten-dimensional representations (see **Table 3**). Moreover, we observe that the Q3 score can be sensitive to the choice of clustering parameters; relatively small changes to the parameters can change the resulting GMM such that the Q3 score increases or decreases by ≈ 0.05–0.10. For example, reducing the quick shift parameter from 0.90 to 0.80 in the ten-dimensional case roughly doubles the number of clusters and the Q3 score increases from approximately 0.68 to 0.73 for both DSSP and STRIDE.

The sensitivity of the classification to the parameters of the method is a general issue with unsupervised schemes, for which it is difficult to define a quantitative measure of the quality of the classification, based on which the performance of the algorithm can be automatically optimized. One possible solution would be to couple the unsupervised classification to a supervised learning task, as we discuss below. Another possibility involves the direct inspection of the cluster structure, which requires, in the case

of high-dimensional data, the application of another class of unsupervised learning algorithms that is aimed at obtaining a simplified low-dimensional representation. To this end, we have applied in **Figure 7** the Sketch-map dimensionality reduction method (Ceriotti et al., 2011; Tribello et al., 2012; Ceriotti et al., 2013) to the six-dimensional dihedral data.

The guiding principle of Sketch-map is to project high-dimensional data into a lower dimension such that points that are close to one another in the high-dimensional space are also close to one another in reduced dimension, and similarly for points that are far apart. Each point in the Sketch-map projection of the six-dimensional ϕ, ψ space is colored by its PAMM cluster assignment and its DSSP secondary structure assignment (**Figure 7**; the Sketch-map projection colored by STRIDE secondary structure assignment is given in the **Supplemental Material**). The Sketch-map projection corroborates our earlier observations that, with the exception of the helices and strands, any given secondary structure is distributed widely across the high-dimensional space. However, one can observe that there is considerably less overlap between regions associated with different DSSP motifs, and it appears that the failure of recognizing these regions as separate clusters is more a consequence of the scattered distribution of points rather than a lack of resolving power.

3.3. SOAP Environments

While it appears that established secondary structure definitions are not associated with well-separated modes in the PDB data, we cannot exclude that this is due to an incomplete description, and that a structure representation encoding more information than the sequence of backbone dihedrals would show greater correspondence between data-driven motifs and established structural definitions. For this reason, we turn to a radically different approach to represent local motifs. We use a SOAP-based representation (whose details are discussed above and in the SI) of the protein backbone for comparison with established secondary structure definitions. A PAMM GMM based on reduced SOAP vectors forms the basis for a truly agnostic method for identifying structural motifs and classifying secondary structure in proteins, as the only required information is the positions of the atoms in the protein backbone. The joint and conditional probability distributions of the clusterized SOAP vectors and DSSP secondary structure assignment are given in **Figure 8** (the probability distributions relative to the STRIDE assignment can be found in the **Supplemental Material**).

Compared to the dihedral angle probability distributions, the distributions based on a clustering of the SOAP vectors are more diffuse. Instead of the helices and strands being confined to one or two clusters as with the dihedral angles, in the SOAP clustering the helices and strands are divided among several clusters. However, from the perspective of the Q3 score, the SOAP representation performs as well as the dihedral angle representations, with scores in the range of 0.70–0.74 for two-, six- and ten-dimensional representations based on the principal components of the SOAP vectors.

3.4. Supervised Classification

The fact that increasing the complexity of the environment descriptors does not improve the match between PAMM PMIs and conventional secondary structure motifs suggests that the discrepancy is not due to lack of descriptive power, but to the fact that conventional motifs are not reflected in the environment distributions observed in the PDB. To substantiate this observation, we also use the dihedral angle and SOAP PCA representations to train an SVM to perform multiclass classification for the purpose of predicting secondary structures. The Q3 and Q8 scores resulting from SVMs built on the reduced SOAP representation and the dihedral angle representation at various dimensionalities are given in **Table 3** and are seen to improve systematically when the dimensionality of the

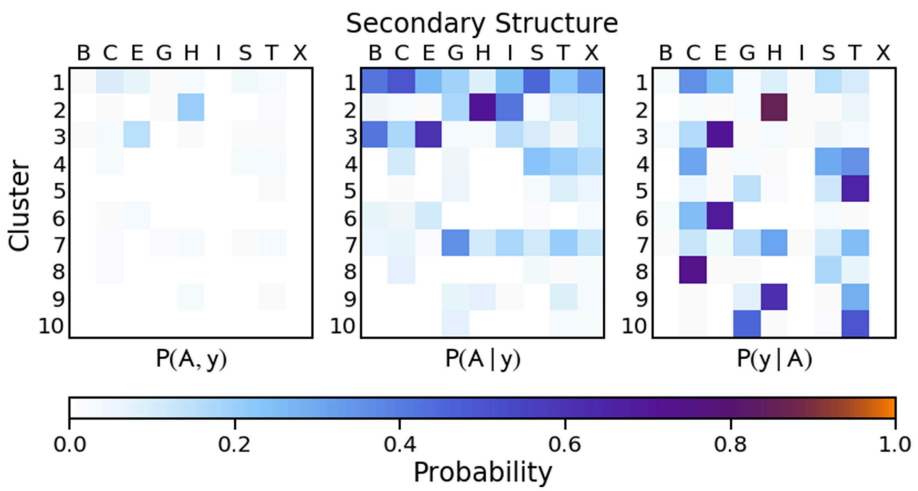


FIGURE 8 | Joint and conditional probabilities for the PAMM clustering of the first two principal components of the reduced SOAP vectors describing each residue of the protein backbone, where *A* is the PAMM cluster assignment and *y* is the DSSP secondary structure classification.

TABLE 3 | Q3 and Q8 scores relative to DSSP for PAMM PMI and SVM predictions of secondary structure based on a PCA of SOAP vectors and dihedral angles at various dimensionalities.

	PAMM PMI		SVM	
Representation	**Q3**	**Q8**	**Q3**	**Q8**
ϕ, ψ (2D)	0.71	0.61	0.78	0.67
ϕ, ψ (6D)	0.74	0.63	0.87	0.80
ϕ, ψ (10D)	0.73	0.61	0.88	0.82
SOAP PCA (2D)	0.73	0.58	0.75	0.61
SOAP PCA (6D)	0.72	0.58	0.84	0.73
SOAP PCA (10D)	0.71	0.55	0.90	0.79
SOAP PCA (100D)	—	—	0.95	0.89

The reported SVM scores are an average over five separate constructions of the SVM, each time using a new random subset of 200,000 residues, with 50,000 of these serving as the training set.

representation is increased—contrary to what observed with a PAMM analysis.

The improving Q3 and Q8 scores for the dihedral angles and reduced SOAP representations in the SVM coupled with the lack of obvious improvement in the cluster-based Q scores confirms that the limiting factor in the association between motifs is intrinsic to unsupervised learning. The reference heuristics—the DSSP and STRIDE secondary structure definitions—are simply not well-represented in the probability distribution of the data in the feature space that we use.

This simple example highlights both the difference in unsupervised and supervised learning methods while also emphasizing the importance of the choice of feature representation. A supervised learning scheme is well-suited to adapt an existing motif definition to a different representation of atomic environments, and—in the limit of a sufficiently large train set—serves as proof of whether the chosen representation is sufficiently complete to achieve an accurate classification. An unsupervised clustering model, on the other hand, is useful for finding new patterns in feature space. Provided that the representation is complete, it also can serve as validation for established pattern recognition heuristics, showing whether the presence of well separate motifs is robust to the choice of structural representation.

By comparing chains of dihedrals and backbone SOAP principal components, we have shown that the two representations possess a similar resolving power for a given size, and yield SOAP motifs that compare roughly in the same way to the DSSP/STRIDE classifications of secondary structure. While dihedral angles are certainly simpler and more straightforward to incorporate into existing analysis schemes, the general-purpose nature of SOAP makes the latter more suitable to be extended to different classes of supramolecular structures, and provides a somewhat less biased starting point for subsequent machine learning analyses.

4. CONCLUSIONS

In this work we have applied data-driven analysis techniques to experimental atomistic structure data of polypeptides extracted from the Protein Data Bank. Our objective has been to demonstrate that a generally applicable analysis protocol, that relies on little specific information for the system at hand can be used to re-discover some of the fundamental atomic-scale motifs that underlie the formation of complex supramolecular structures—specifically the hydrogen bond and secondary structure patterns. For this purpose, we used PAMM, a density-based algorithm that recognizes and associates local maxima in atomic feature space with particularly stable, frequently occurring configurations to highlight some of the shortcomings of more traditional definitions. For instance, we showed how conventional bond–angle criteria to recognize hydrogen bonds rely on multiple additional heuristics to avoid incorrectly classifying other recurring motifs that are associated to covalently bound groups. Furthermore, we quantified the substantial differences between various hydrogen-bond "flavors," underscoring the advantages of an adaptive, automatic definition.

The case of secondary structure patterns gave us the opportunity to compare the use of conventional representations of local atomic structure (backbone dihedrals) with an even more generally applicable strategy based on the principal components of the SOAP power spectrum. Despite being very different in spirit, the two representations yield very similar results; there is a good match between PAMM-based patterns and traditional heuristics for what concerns helices and strands, but rather poor agreement for other, less common motifs. By comparing representations of different complexity, and the outcome of both supervised and unsupervised classification schemes, we have shown that the conventional secondary structure recognition methods reflect only in part the intrinsic distribution of data of protein structures in the PDB.

While conventional secondary structure motifs have the advantage of being linked to structure–property relations and important design principles and have survived the test of time, data-driven definitions such as PAMM-based PMIs can be more easily adapted to specific simulations or, as in the present case, experimental data sets. Their robustness is highlighted by clustering outcomes that are rather insensitive to the choice of the structure representation. The possibility of using generic representations, such as the list of backbone dihedrals, or even more abstract feature vectors such as the SOAP power spectrum, makes a PAMM analysis well-suited for application to different classes of supramolecular and self-assembly problems, where less prior knowledge is available to define heuristic criteria. Finally, given that PMIs are smooth, differentiable functions that depend exclusively on atom coordinates, they show great promise for use in combination with automatic collective variable determination and in accelerated sampling schemes to probe structural transitions and rare events.

AUTHOR CONTRIBUTIONS

BH performed simulations, analyzed them, wrote pre- and post-processing code, and prepared figures. PG ran preliminary tests

and implemented the PAMM analysis software. All authors contributed to the design of the simulations and the writing of the text.

ACKNOWLEDGMENTS

BH, FG, and MC were supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 677013-HBMAP).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00024/full#supplementary-material

REFERENCES

Akkaladevi, S., Katangur, A. K., Belkasim, S., and Pan, Y. (2004). "Protein secondary structure prediction using neural network and simulated annealing algorithm," in *Proceedings of the 26th Annual International Conference of the IEEE EMBS* (San Francisco, CA), 2987–2990.

Andersend, C. A. F., Palmer, A. G., Brunak, S., and Rost, B. (2002). Continuum secondary structure captures protein flexibilty. *Structure* 10, 175–184. doi: 10.1016/S0969-2126(02)00700-1

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., et al. (2011). Defining the hydrogen bond: an account (IUPAC Technical Report). *Pure Appl. Chem.* 83, 1619–1636. doi: 10.1351/PAC-REP-10-01-01

Baker, E. N., and Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Progr. Biophys. Mol. Biol.* 44, 97–179. doi: 10.1016/0079-6107(84)90007-5

Bartók, A. P., and Csányi, G. (2015). Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quantum Chem.* 115, 1051–1057. doi: 10.1002/qua.24927

Bartók, A. P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. *Phys. Rev. B* 87, 184115. doi: 10.1103/PhysRevB.87.184115

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Brown, I. D. (1976). On the geometry of O − H···O hydrogen bonds. *Acta Crystallogr.* A32, 24–31.

Ceriotti, M., Tribello, G. A., and Parrinello, M. (2011). Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13023–13028. doi: 10.1073/pnas.1108486108

Ceriotti, M., Tribello, G. A., and Parrinello, M. (2013). Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.* 9, 1521–1532. doi: 10.1021/ct3010563

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Cooper, R. I., Thompson, A. L., and Watkin, D. J. (2010). CRYSTALS Enhancements: dealing with hydrogen atoms in refinement. *J. Appl. Crystallogr.* 43, 1100–1107. doi: 10.1107/S0021889810025598

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cuff, J. A., and Barton, G. J. (2000). Application of mutliple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40, 502–511. doi: 10.1002/1097-0134(20000815)40:3<502::AID-PROT170>3.0.CO;2-Q

De, S., Bartók, A. P., Csányi, G., and Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 18, 13754. doi: 10.1039/C6CP00415F

Desiraju, G. R., and Steiner, T. (2001). *The Weak Hydrogen Bond: In Structural Chemistry and Biology*, Vol. 9. Oxford: International Union of Crystal.

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23, 566–579. doi: 10.1002/prot.340230412

Frishman, D., and Argos, P. (1996). Incorporation of non-Local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* 9, 133–142. doi: 10.1093/protein/9.2.133

Gasparotto, P., and Ceriotti, M. (2014). Recognizing molecular patterns by machine learning: an agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* 141, 174110. doi: 10.1063/1.4900655

Gasparotto, P., Hassanali, A. A., and Ceriotti, M. (2016). Probing defects and correlations in the hydrogen-bond network of ab initio water. *J. Chem. Theory Comput.* 12, 1953–1964. doi: 10.1021/acs.jctc.5b01138

Gasparotto, P., Meißner, R. H., and Ceriotti, M. (2018). Recognizing local and global structural motifs at the atomic scale. *J. Chem. Theory Comput.* 14, 486–498. doi: 10.1021/acs.jctc.7b00993

Haghighi, H., Higham, J., and Henchman, R. H. (2016). Parameter-free hydrogen-bond definition to classify protein secondary structure. *J. Phys. Chem. B* 120, 8566–8570. doi: 10.1021/acs.jpcb.6b02571

Holley, L. H., and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* 86, 152–156. doi: 10.1073/pnas.86.1.152

Hollingsworth, S. A., Lewis, M. C., Berkholz, D. S., Wong, W.-K., and Karplus, P. A. (2012). $(\phi, \psi)_2$ motifs: a purely conformation-based fine-grained enumeration of protein parts at the two-residue level. *J. Mol. Biol.* 416, 78–93. doi: 10.1016/j.jmb.2011.12.022

Imbalzano, G., Anelli, A., Giofré, D., Klees, S., Behler, J., and Ceriotti, M. (2018). Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* 148, 241730. doi: 10.1063/1.5024611

Jeffrey, G. A., and Saenger, W. (2012). *Hydrogen Bonding in Biological Structures*. Berlin: Springer Science & Business Media.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202. doi: 10.1006/jmbi.1999.3091

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Neurocomputing* 1990, 41–50. doi: 10.1007/978-3-642-76153-9_5

Kountouris, P., and Hirst, J. D. (2009). Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinformatics* 10:437. doi: 10.1186/1471-2105-10-437

Luzar, A., and Chandler, D. (1993). Structure and hydrogen bond dynamics of water-dimethyl sulfoxide mixtures by computer simulations. *J. Chem. Phys.* 98, 8160–8173. doi: 10.1063/1.464521

Luzar, A., and Chandler, D. (1996). Effect of environment on hydrogen bond dynamics in liquid water. *Phys. Rev. Lett.* 76, 928–931. doi: 10.1103/PhysRevLett.76.928

Martin, J., Letellier, G., Marin, A., Taly, J.-F., de Brevern, A. G., and Gibrat, J.-F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BioMed Central Struct. Biol.* 5, 17. doi: 10.1186/1472-6807-5-17

McDonald, I. K., and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238, 777–793. doi: 10.1006/jmbi.1994.1334

Mezei, M., and Beveridge, D. L. (1984). Theoretical studies of hydrogen bonding in liquid water and dilute aqueous solutions. *J. Chem. Phys.* 74, 622–632. doi: 10.1063/1.440819

Muggleton, S., King, R. D., and Sternberg, M. J. E. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Eng.* 5, 647–657. doi: 10.1093/protein/5.7.647

Nagy, G., AND Oostenbrink, C. (2014). Bihedral-bsed segment identification and classification of biopolymers I: proteins. *Jo. Chem. Inf. Model.* 54, 266–277. doi: 10.1021/ci400541d

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Édouard Duchesnay, Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pietropaolo, A., Muccioli, L., Bernardi, R., and Zannoni, C. (2008). A chirality index for investigating protein secondary structures and their time evolution. *Proteins* 70, 667–677. doi: 10.1002/prot.21578

Pietrucci, F., and Laio, A. (2009). A collective variable for the efficient exploration of protein beta-sheet structures: application to SH3 and GB1. *J. Chem. Theory Comput.* 5, 2197–2201. doi: 10.1021/ct900202f

Rahman, A., and Stillinger, F. H. (1971). Molcular dynamics study of liquid water. *J. Chem. Phys.* 55, 3336–3359. doi: 10.1063/1.1676585

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7, 95–99. doi: 10.1016/S0022-2836(63)80023-6

Rashid, S., Saraswathi, S., Kloczkowski, A., Sundaram, S., and Kolinski, A. (2016). Protein secondary structure prediction using a small training set (Compact Model) combined with a complex-valued neural network approach. *BioMed Central Bioinf.* 17, 362. doi: 10.1186/s12859-016-1209-0

Rost, B., and Sander, C. (1993a). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599. doi: 10.1006/jmbi.1993.1413

Rost, B., and Sander, C. (1993b). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 90, 7558–7562. doi: 10.1073/pnas.90.16.7558

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* 268, 209–225. doi: 10.1006/jmbi.1997.0959

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. 34, 82–95. doi: 10.1002/(SICI)1097-0134(19990101)34:1<82::AID-PROT7>3.0.CO;2-A

Tribello, G. A., Ceriotti, M., and Parrinello, M. (2012). Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5196–5201. doi: 10.1073/pnas.1201152109

Watkin, D. (2008). Structure refinement: some background theory and practical strategies. *J. Appl. Crystallogr.* 41, 491–522. doi: 10.1107/S0021889808007279

Wood, M. J., and Hirst, J. D. (2005). Protein secondary structure prediction with dihedral angles. *Proteins* 59, 476–481. doi: 10.1002/prot.20435

Xu, D., Tsai, C.-J., and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* 10, 999–1012. doi: 10.1093/protein/10.9.999

Zhang, B., Li, J., and Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* 19:293. doi: 10.1186/s12859-018-2280-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 April 2019
doi: 10.3389/fmolb.2019.00025

Anncolvar: Approximation of Complex Collective Variables by Artificial Neural Networks for Analysis and Biasing of Molecular Simulations

*Dalibor Trapl[1], Izabela Horvacanin[1,2], Vaclav Mareska[1], Furkan Ozcelik[3], Gozde Unal[3] and Vojtech Spiwok[1]**

[1] Department of Biochemistry and Microbiology, University of Chemistry and Technology in Prague, Prague, Czechia, [2] Faculty of Science, University of Zagreb, Zagreb, Croatia, [3] Computer Engineering Department, Istanbul Technical University, Istanbul, Turkey

The state of a molecular system can be described in terms of collective variables. These low-dimensional descriptors of molecular structure can be used to monitor the state of the simulation, to calculate free energy profiles or to accelerate rare events by a bias potential or a bias force. Frequent calculation of some complex collective variables may slow down the simulation or analysis of trajectories. Moreover, many collective variables cannot be explicitly calculated for newly sampled structures. In order to address this problem, we developed a new package called *anncolvar*. This package makes it possible to build and train an artificial neural network model that approximates a collective variable. It can be used to generate an input for the open-source enhanced sampling simulation PLUMED package, so the collective variable can be monitored and biased by methods available in this program. The computational efficiency and the accuracy of *anncolvar* are demonstrated on selected molecular systems (cyclooctane derivative, Trp-cage miniprotein) and selected collective variables (Isomap, molecular surface area).

Keywords: metadynamics, neural networks, molecular dynamics simulation, collective variables, free energy simulations

OPEN ACCESS

Edited by:
Elena Papaleo, Danish Cancer Society Research Center, Denmark

Reviewed by:
Carlo Camilloni, University of Milan, Italy
Massimiliano Bonomi, Institut Pasteur, France

***Correspondence:**
Vojtech Spiwok
spiwokv@vscht.cz

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 January 2019*
Accepted: *01 April 2019*
Published: *18 April 2019*

Citation:
Trapl D, Horvacanin I, Mareska V, Ozcelik F, Unal G and Spiwok V (2019) Anncolvar: Approximation of Complex Collective Variables by Artificial Neural Networks for Analysis and Biasing of Molecular Simulations. Front. Mol. Biosci. 6:25. doi: 10.3389/fmolb.2019.00025

INTRODUCTION

Molecular dynamics simulation makes it possible to simulate any molecular process at the atomic level. In principle, structural and thermodynamical properties of a protein can be predicted by simulation of its folding and unfolding. Similarly, structure and stability of a protein-ligand complex can be predicted by simulation of binding and unbinding. Unfortunately, many molecular processes either cannot be simulated or their simulation is far from routine due to enormous computational costs of the molecular dynamics simulation method.

Several enhanced sampling methods have been developed in order to address this problem (Spiwok et al., 2015a). Some of these methods, such as umbrella sampling (Torrie and Valleau, 1977) or metadynamics (Laio and Parrinello, 2002), use a bias potential or a bias force to destabilize frequently sampled states and to enhance sampling of poorly sampled states. Tempering methods enhance sampling by means of elevated temperature (Abrams and Bussi, 2014). There are methods combining tempering and biasing as well as methods based on completely different principles.

Biased simulations usually require one or more preselected degrees of freedom on which the bias force or potential is applied. These degrees of freedom are referred to as collective variables (CVs). There are two technical prerequisites for CVs to be applicable in biased simulations. Firstly, a CV must be a function of atomic coordinates of the molecular system, i.e., it must be possible to calculate the value of a CV at every step of the simulation solely from atomic coordinates. Secondly, it must be possible to convert the force acting on the CV into forces acting on individual atoms, i.e., it must be possible to calculate the first derivative of the CV with respect to atomic Cartesian coordinates. Beside these technical prerequisites, in order to efficiently enhance sampling it is necessary to cover all slow motions in the molecular systems by few CVs.

There are many promising CVs that do not fulfill these requirements and therefore cannot be directly used in biased simulations. These include, for example, the results of non-linear dimensionality reduction methods (Das et al., 2006). There are examples of other CVs that fulfill these requirements; however, their calculation is computationally expensive. In order to make biased simulation with these CVs possible, we and others introduced approximations tailored for biased simulations (Branduardi et al., 2007; Spiwok and Králová, 2011; Spiwok et al., 2015b; Pazúriková et al., 2017).

Recent development of neural network algorithms allows the usage of artificial neural networks for the purpose of CV approximation. The advantage of neural networks is the fact that many of them are trained by the backpropagation algorithm (Goodfellow et al., 2016), which requires easy calculation of the derivatives of the output with respect to the input. This is exactly what is needed to convert forces acting on a CV into forces acting on atoms. Application of neural network models may also benefit from the current development of neural networks, which has lead to a number of new toolkits and programs.

Multiple recent studies have tested machine learning approaches to design collective variables for biased simulation to study thermodynamics and kinetics of molecular transitions (Galvelis and Sugita, 2017; Chen and Ferguson, 2018; Guo et al., 2018; Mardt et al., 2018; Pérez et al., 2018; Seo et al., 2018; Sultan and Pande, 2018; Wehmeyer and Noé, 2018). In this work we describe a new tool *anncolvar* for approximation of an arbitrary CV. Its function is outlined in **Figure 1**. This tool requires a set of structures, either a simulation trajectory or any other set of structures. For the sake of simplicity we will call this set a training trajectory. It must be accompanied with precomputed values of CVs. These data are used to train a simple neural network to approximate the value of CVs for other out-of-sample structures. It generates an input to a popular enhanced sampling program PLUMED (Bonomi et al., 2009; Tribello et al., 2014). The CV approximated by *anncolvar* can be calculated *a posteriori* for any 3D structure or trajectory. Furthermore, it can be used in metadynamics or other enhanced sampling methods available in PLUMED. This approach was tested on conformational changes of a cyclooctane derivative and Trp-cage mini-protein folding. Isomap (Tenenbaum et al., 2000) low-dimensional embeddings used as CVs in the metadynamics simulation of the former system represent CVs that cannot be calculated explicitly from Cartesian coordinates. Solvent-accessible surface area (SASA) used as a CV in simulations of the later system represents a CV that can be calculated explicitly from Cartesian coordinates, but such calculation is slow.

The program can be accessed for free at https://github.com/spiwokv/anncolvar or via PyPI.

METHODS

Use of Anncolvar

The program *anncolvar* is written in *Python* and uses packages *mdtraj* (McGibbon et al., 2015), *numpy* (Oliphant, 2006) and *keras* (Cholet, 2018)[1]. The machine learning package *keras* runs on top of one of three machine learning backends, namely *TensorFlow*, *Theano* or *CNTK*. Before installation of *anncolvar* it is necessary to install one of these backends. The package *anncolvar* was tested with *TensorFlow* on a laptop, personal computer and HPC cluster, with *Theano* on HPC cluster and with all three backends in continuous integration environment Travis-CI. Installation of other libraries may be required in order to enable use of GPU acceleration on GPU-equipped computers. Additionally, one needs to install *Python* (*Python* 2.7 and *Python* 3.6 were tested) and package management library *PyPI*.

Once the backend is installed, *anncolvar* can be installed by typing:

```
pip install numpy cython
pip install anncolvar
```

(or with *sudo*, depending on user rights and type of installation). *PyPI* installs all required python libraries. Successful installation can be checked by typing:

```
anncolvar -h
```

to print help. *Anncolvar* can be also installed from Anaconda Cloud (https://anaconda.org/spiwokv/anncolvar).

The program *anncolvar* is written in a way so that it requires a preprepared reference structure and a training trajectory. The reference structure is a single structure of the molecular system in PDB format. It is used as a template for RMSD fitting in order to remove translational and rotational motions. Furthermore, input data for artificial neural networks are typically scaled to lie between 0 and 1. The reference structure is used in this process. It must be prepared to fulfill following requirements:

1. It may contain only atoms intended for the analysis. Atoms not intended for the analysis, such as hydrogen atoms, must be deleted. The program *anncolvar* does not ask which atoms are to be analyzed and which are not. Numbering of atoms should not be changed by deletion of unwanted atoms, e.g., if atoms 2, 3, 5, 6, 8, etc. are deleted, the remaining atoms must be numbered 1, 4, 7, etc., not 1, 2, 3, etc.
2. It must be centered in a reasonably large box with coordinates of one corner set to [0,0,0] and the diagonal corner set to [l_x, l_y, l_z] (cubic boxes were used in this work). The size of the box must be sufficient to accommodate the analyzed molecule in all snapshots of the simulation (the program returns an error message if this fails). In the preprocessing step done by

[1] Cholet, F., and co-workers, https://keras.io, 2019.

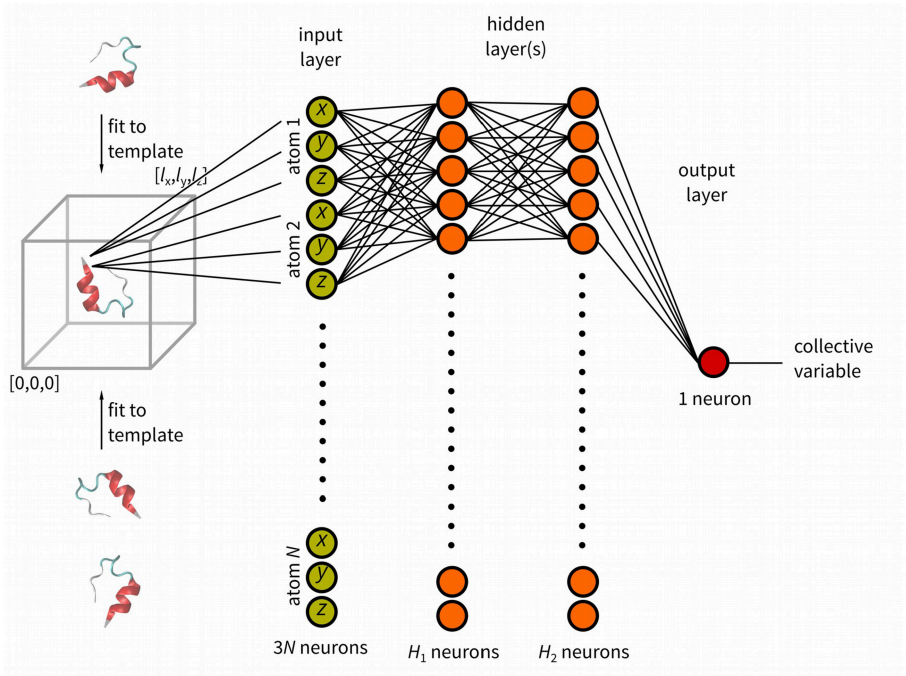


FIGURE 1 | Schematic representation of *anncolvar* function. Three input files are needed for training: (i). reference structure (in PDB) of the molecule located in the center of box with one corner with coordinates [0, 0, 0] and size of $[l_x, l_y, l_z]$, (ii) training trajectory (without molecules broken by periodic boundary condition) and, (iii) file containing precomputed values of the CV for each snapshot of the training trajectory. The program generates the input file for PLUMED. In PLUMED the molecule is fit to the template (reference structure) and the CV is calculated by neural network.

anncolvar the coordinates are fitted to the reference structure and then divided by l_x, l_y and l_z to lie between 0 and 1. The reference structure can be generated, for example, in Gromacs by a command:

```
gmx editconf -f input.pdb -o reference.pdb
-box 6 6 6 -c
```

for a box with $l_x = l_y = l_z = 6$ nm. The values of l_x, l_y and l_z must be specified by options `-boxx`, `-boxy` and `-boxz`.

The training trajectory must be prepared to fulfill following requirements:

1. It may contain only atoms intended for the analysis, i.e., the same atoms as in the reference structure.
2. The molecule must not be broken due to periodic boundary condition.

Fitting to a template is done by *mdtraj* library in *anncolvar*. For special fitting protocols it is possible to fit the training trajectory before running *anncolvar* and switch off fitting in *anncolvar* by `-nofit` option.

Finally, the program requires a set of precalculated values of collective variables for each snapshot of the training trajectory (option **-c**). This must be a space-separated file with a column containing values of the CV in the order of snapshots in the training trajectory. The index of the column can be specified by `-col` (e.g., `-col 2` for the second column).

The program makes it possible to modify the design of the neural network, namely the number of hidden layers (1, 2, or 3 is supported), activation functions in each layer (keras activation functions are supported), and the details of optimization (loss function, batch size and optimization algorithm). The results are written to a text output file for easy visualization of the correlation between original and predicted CV values. This output file controlled by **-o** option contains predicted and original values in the first and the second column, respectively. The third column indicates whether the value was used in the training (TR) or test (TE) set. Stratification of data into the training and test sets is controlled by `-test` (size of test set) and `-shuffle` (whether snapshots of the trajectory are or are not shuffled before the stratification).

Input file for the PLUMED open-source library for analyzing and biasing molecular dynamics simulations (Tribello et al., 2014) is also provided (`-plumed` option). This file (default name *plumed.dat*) makes it possible without much changes to calculate the CV for a trajectory (by PLUMED driver) or to monitor the value of the CV during a simulation. Application of the output PLUMED file in metadynamics or other enhanced sampling method supported by PLUMED requires minor changes easy for an experienced PLUMED user. In case the training trajectory and the biased simulation use a different atom numbering, it is necessary to renumber atoms in the PLUMED input file. The reference file is used as a template for fitting of the molecule in order to remove rotational and translational degrees of freedom. It may be

necessary to modify the PDB format to fulfill the requirements of PLUMED.

Proper function of *anncolvar* can be checked by recalculation of the CV in the training trajectory using *plumed driver* utility followed by comparison with the text output of *anncolvar*.

A sample training may be executed by:

```
anncolvar -i traj.xtc -p reference.pdb -c
results_isomap -col 2 \
  -boxx 1 -boxy 1 -boxz 1 -layers 1 -layer1
  64 -epochs 2000 \
  -o corr1.txt -plumed plumed1.dat
```

This carries out 2,000 epochs of training on an artificial neural network with the training trajectory in *traj.xtc* (Gromacs format), reference structure in *reference.pdb* and precalculated CV values in *results_isomap* (in the second column). The artificial neural network was composed of one hidden layer with 64 neurons with sigmoid (default) activation function.

Simulation Details

All simulations were carried out in Gromacs 5.1.1 (Abraham et al., 2015) with PLUMED 2.4 (Tribello et al., 2014).

Cyclooctane derivative (*trans*,*trans*-1,2,4-trifluorocyclooctane) was simulated as described elsewhere (Spiwok and Králová, 2011). Briefly, it was simulated in General AMBER force field (Wang et al., 2004) in vacuum using stochastic dynamics integrator with 1 fs step and without constraints. Temperature was kept constant at 300 K using Parrinello-Bussi thermostat (Bussi et al., 2007). Electrostatics was modeled without cut-off. The set of 8,375 reference structures was kindly provided by Brown and co-workers (Brown et al., 2008). They were generated by Brown and co-workers using a systematic generation algorithm as described in their work (Brown et al., 2008).

Trp-cage was modeled using Amber99SB-ILDN (Lindorff-Larsen et al., 2010) force field. The protein was placed in a periodic box of size $7 \times 7 \times 7$ nm (metadynamics, MTD) or $3.548 \times 3.896 \times 3.389$ nm (parallel tempering metadynamics, PT-MTD) containing 11,128 (MTD) (Laio and Parrinello, 2002) or 1,366 (PT-METAD) (Bussi et al., 2006) water molecules and one chloride anion. Step of molecular dynamics simulation was set to 2 fs. All bonds were constrained. Electrostatics was modeled by Particle-mesh Ewald method (Darden et al., 1993). Temperature was kept constant using Parrinello-Bussi thermostat (Bussi et al., 2007).

For MTD, the system was minimized by steepest descent algorithm. This was followed by 100 ps simulation in NVT and 100 ps simulation in NVT ensemble. This was followed by 100 ns well tempered metadynamics (Barducci et al., 2008) at 300 K.

For PT-MTD, the system was minimized by steepest descent algorithm. This was followed by 100 ps simulation in NVT and 100 ps simulation in NVT ensemble. The system was preequilibrated by 500 ps NVT simulations at 32 temperatures: 278.0, 287.0, 295.0, 303.0, 312.0, 321.0, 329.0, 338.0, 346.0, 355.0, 365.0, 375.0, 385.0, 396.0, 406.0, 416.0, 427.0, 437.0, 448.0, 459.0, 470.0, 482.0, 493.0, 505.0, 517.0, 528.0, 539.0, 551.0, 562.0, 573.0, 584.0, and 595.0 K. After that PT-METAD was performed at same temperatures. Replica exchange attempts were made every picosecond.

Trajectory of 208 μs simulation of Trp-cage folding/unfolding was kindly provided by D. E. Shaw Research (Darden et al., 1993). It was converted to Gromacs format and prepared by Gromacs tools for analysis in *anncolvar*.

RESULTS AND DISCUSSIONS

Cyclooctane Derivative Conformational Transitions

Cyclooctane non-symmetric derivative (*trans*,*trans*-1,2,4-trifluorocyclooctane) was introduced as a model molecular system by Brown and co-workers (Brown et al., 2008; Martin et al., 2010). They generated more than one million of conformations of this molecule by a systematic geometry-based algorithm. Then they filtered this set to obtain a set of 8,375 non-redundant structures. These structures were analyzed by a non-linear dimensionality method Isomap (Tenenbaum et al., 2000). Brown and co-workers demonstrated that it is possible to describe conformation of the model molecule using just three low-dimensional Isomap embeddings (see **Figure 2A** for the reproduction of the results of Brown and co-workers) (Brown et al., 2008).

It is very challenging to use low-dimensional embeddings as CVs in biased simulations. For this, it is necessary to calculate a low-dimensional embedding for a new out-of-sample structure. Furthermore, in order to apply biasing forces on a molecular structure it is necessary to calculate derivatives of the low-dimensional embedding with respect to the Cartesian coordinates. Unfortunately, using Isomap and most other non-linear methods it is not possible to directly calculate neither low dimensional embeddings for a new out-of-sample structure, nor their derivatives. For this purpose we have tested the Property Map Collective Variables (Spiwok and Králová, 2011), an extension of Path Collective Variables (Branduardi et al., 2007). An interesting alternative is application of autoencoders recently used by Chen and Ferguson (2018).

Here we test an artificial neural network performed by *anncolvar* to approximate Isomap embeddings. The set of 8,375 structures provided by Brown et al. (2008) was analyzed by Isomap to obtain three low-dimensional embeddings (**Figure 2A**). Next we use them to train a neural network to approximate these embeddings. Briefly, we used the command:

```
anncolvar -i traj_fit.xtc -p
reference.pdb \
  -c results_isomap -col 2 \
  -boxx 1 -boxy 1 -boxz 1 \
  -layers 3 -layer1 8 -layer2 8 -layer3 8 \
  -actfun1 sigmoid -actfun2 sigmoid
   -actfun3 sigmoid \
  -optim adam -loss mean_squared_error \
  -epochs 1000 -batch 256 \
  -o low1.txt -plumed plumed1.dat
```

The set of 8,375 structures was stored in Gromacs format in *traj_fit.xtc*. A reference structure was stored in the file

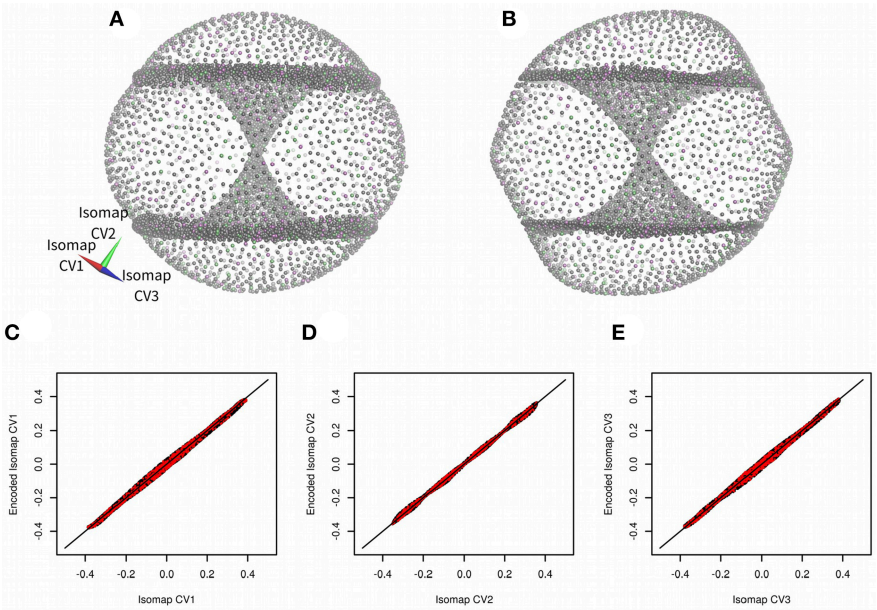


FIGURE 2 | Three-dimensional Isomap embeddings of *trans*,*trans*-1,2,4-trifluorocyclooctane **(A)** and its approximation using *anncolvar* **(B)**. Isomap embeddings in **(A)** and **(B)** were rotated by angle [0.00 rad, 2.40 rad, −0.55 rad] for better clarity. Training set points are in gray, test set points are in different colors depending on whether they were used to train Isomap embedding 1, 2, or 3. Comparison of Isomap embeddings 1 **(C)**, 2 **(D),** and 3 **(E)** original (horizontal) vs. approximated by *anncolvar* (vertical). Training set points are in black, test set points are in red. Distribution of differences between original and predicted CVs can be found in **Figure S1**.

reference.pdb. It was centered in the cubic box of size 1 nm with the corners at [0,0,0], [0,0,1], … [1,1,1] (in nm). Isomap low-dimensional embeddings were stored in the file *results_isomap* (space-separated, with structure ID and Isomap embedding 1, 2, and 3 in each column). This carried out 1,000 epochs of training (ADAM optimizer, mean square error loss function) of a network composed of an input layer with 72 neurons (for Cartesian coordinates of 24 atoms) and three hidden layers, each with eight neurons with the sigmoid activation function. By default, 10% of randomly selected structures are used as the test set and remaining as the training set.

This was repeated for the second and third Isomap coordinates (with `-col 3` and `4`, respectively). The resulting PLUMED input files were combined manually to one PLUMED input file. It was also necessary to renumber atoms due to a different numbering in the original data set and used force field.

There were visible differences between original Isomap embeddings and values approximated by *anncolvar* (**Figure 2**), nevertheless, these differences do not affect the functionality of embeddings. Pearson correlations of original and anncolvar-predicted Isomap low-dimensional embeddings were higher than 0.997. There was no significant difference between correlations in the training and test sets.

Next, the PLUMED input file was edited to enable metadynamics (Laio and Parrinello, 2002) with all three Isomap embeddings used as CVs. Hills were added every 1 ps with constant height of 0.2 kJ·mol^{-1} and width 0.02 (for all three Isomap CVs). The results of 100 ns metadynamics are depicted in **Figure 3**. The simulation started from one of boat-chair conformation located in the central "hourglass." After ∼20 ns all eight boat-chair conformations were flooded and the system started to explore one of boat conformations at the "equator." After ∼30 ns it started to explore the crown conformation at the "south pole." At time ∼50 ns also the inverted crown at the "north pole" was sampled. The convergence was assessed as the evolution of free energy difference between crown and boat-chair (see **Figure S2**). The free energy surface was visualized by Mayavi (Ramachandran and Varoquaux, 2011) and PoVRay (Persistence of Vision, 2018)[2]. The resulting free energy surface (**Figure 3B**) is in good agreement with the results of our previous studies (Spiwok and Králová, 2011; Pazúriková et al., 2017).

Trp-Cage Folding

Intuitively solvent-accessible surface area (SASA) of a protein is likely to be an interesting CV for protein folding simulation, because SASA of a protein in the folded state is likely to be smaller than for the unfolded state, which is one of requirements for a CV to be successful. For this purpose we used a 208-μs trajectory of Trp-cage miniprotein kindly provided by D. E. Shaw Research (Lindorff-Larsen et al., 2011). We admit that this is not solution to the "chicken-and-egg problem" [as discussed by (Chen and Ferguson, 2018)], because we cannot train the neural network without a long simulation trajectory with folding and unfolding events. Reinforcement learning (Nandy and Biswas, 2018) may be solution to this problem, but it is out of scope of this manuscript.

The trajectory provided by D. E. Shaw Research was converted to Gromacs format and SASA was calculated for 1,044,000 frames using *gmx sasa* tool from the Gromacs package (Abraham

[2]Persistence of Vision Pty. Ltd., http://www.povray.org, 2018.

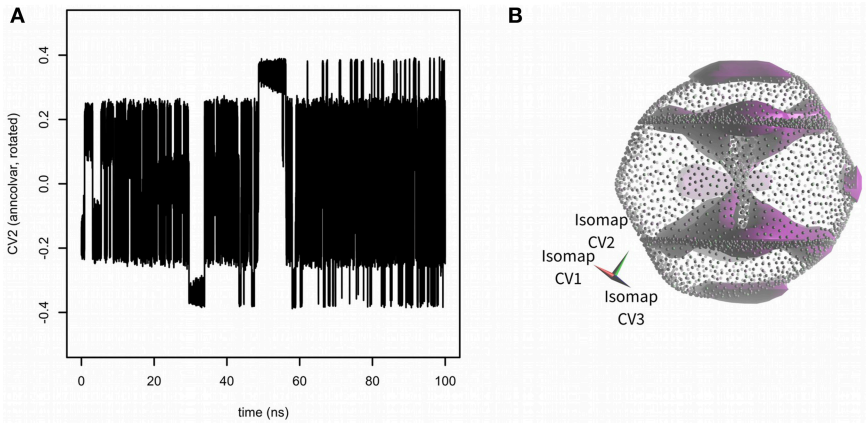


FIGURE 3 | Sampling of CVs in 100 ns metadynamics with Isomap low-dimensional embeddings calculated by *anncolvar* **(A)**. Free energy surface depicted as an isosurface (in violet) at + 30 kJ·mol^{-1} (relative to the global free energy minimum) **(B)**. Isomap embeddings were rotated by angle [0.00 rad, 2.40 rad, −0.55 rad] for better clarity.

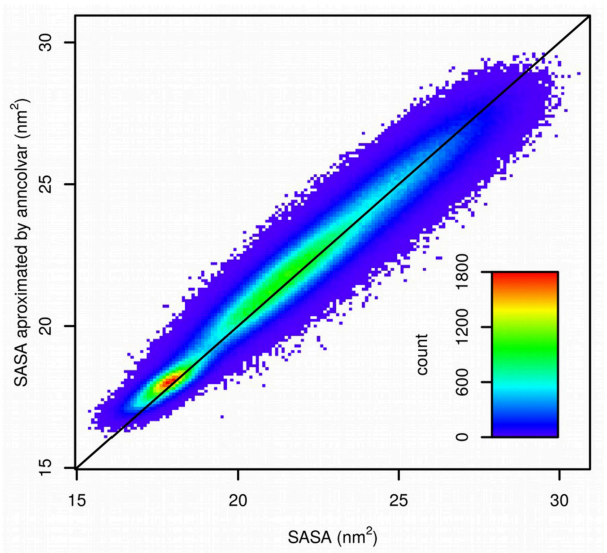


FIGURE 4 | Comparison of SASA calculated by *gmx sasa* (horizontal) and predicted by *anncolvar* (vertical). The line shows the diagonal ($y = x$).

et al., 2015). Next, a neural network was trained in *anncolvar* to approximate SASA. It contained 432 neurons in the input layer (for coordinates of 144 atoms placed in a cubic box of size 6 nm) and one hidden layer with 32 sigmoid neurons. The set of 10% of randomly selected structures was used as the test set and remaining as the training set. This provided a good 0.976 correlation (Pearson) between SASA calculated by *gmx sasa* and predicted by *anncolvar* (**Figure 4**).

We also examined the effect of training set size on *anncolvar* performance. The observed effect was small. The Pearson correlation coefficient for reference and predicted values ranged from 0.9750 (50% of trajectory frames used) to 0.9756 (90% of trajectory frames used), both using 1,000 epochs. We also tested training using a sub-optimal training set. Unfolded structures (RMSD form NMR structure >0.25 nm on all atoms, 879,759 structures) were selected from the trajectory and used as a training set. The resulting neural network predicts SASA with relatively good accuracy (Pearson correlation coefficient 0.96 for all structures and 0.77 for folded structures, see **Figure S7**). We plan to test *anncolvar* trained on sub-optimal training sets in future.

In order to evaluate performance of *anncolvar* we decided to estimate costs of SASA calculation by conventional program (*gmx sasa* from Gromacs package) and to compare it with *anncolvar*. The program *gmx sasa* calculates SASA of Trp-cage in approximately one millisecond. This corresponds to reasonably good performance of ∼0.6 s/ps or 10 min/ns. However, for biasing it is necessary to calculate not only SASA, but also its derivatives $d\text{SASA}/dx$. Methods for calculation of analytical surface derivatives have been reported in literature (Sridharan et al., 1995), but their implementation into available simulation packages would require intensive coding. In order to use numerical derivatives it would be necessary to evaluate delta SASA for incremental changes Δr of all coordinates of all atoms. This would downgrade performance to ∼days/ns. There are approaches that can be applied to address this problem, such as evaluation of CVs in multiple time steps (Ferrarotti et al., 2015), parallelization or GPU offloading. However, all these approaches either require intensive changes in a code or they may have other disadvantages.

The PLUMED input file was used to drive metadynamics (Laio and Parrinello, 2002) and parallel tempering metadynamics (PT-METAD) (Bussi et al., 2006) with SASA as a collective variable. Similarly to cyclooctane derivative it was necessary to manually edit *plumed.dat* file because of different atom numbering in the D. E. Shaw Research data set and the force field we used. Since formation of secondary structure is very important and potentially the slow step of Trp-cage folding, another CV was used to enhance formation of secondary structure. We selected an alpha helical content of a protein structure (ALPHARMSD)

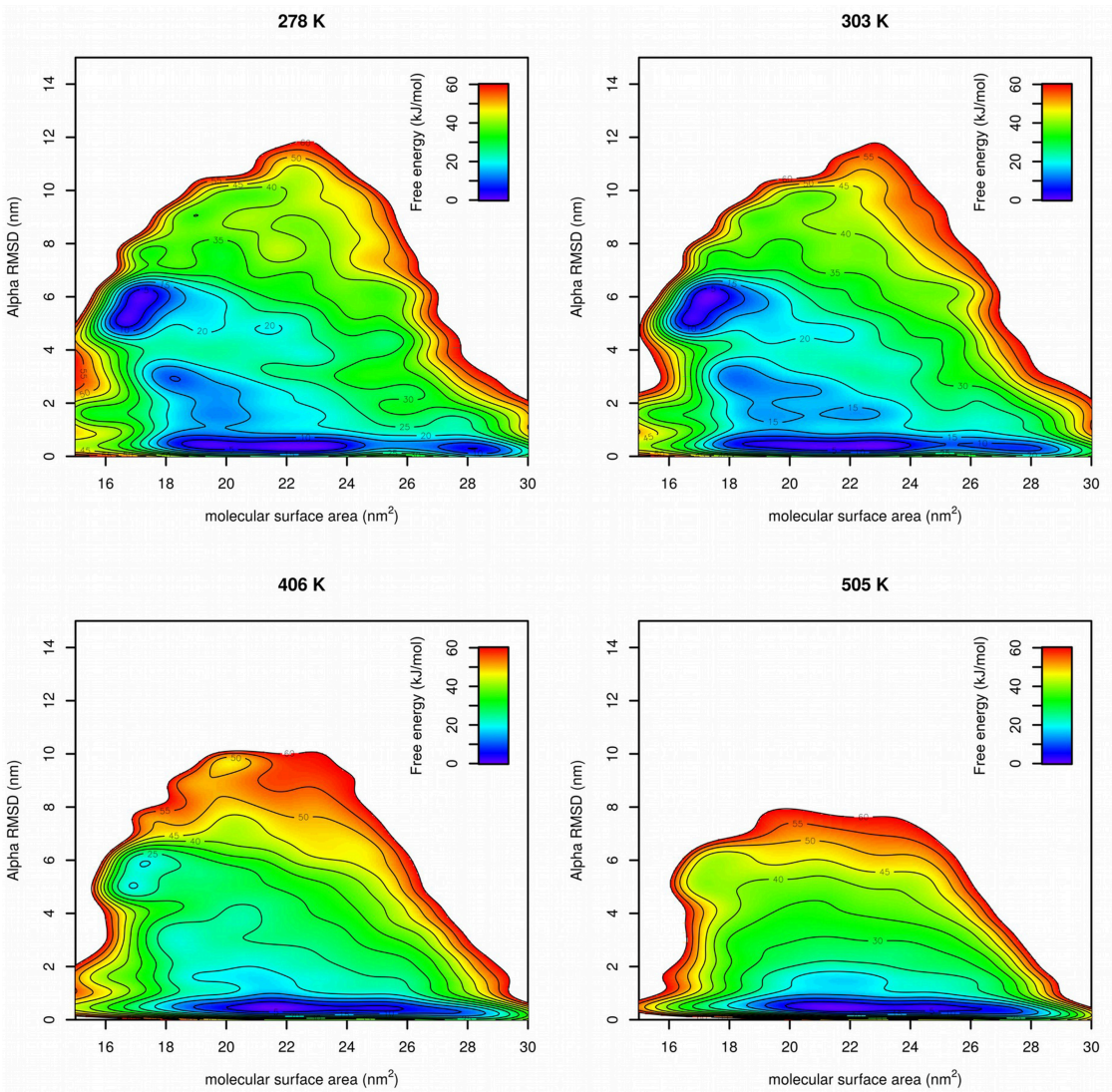


FIGURE 5 | Free energy surfaces of Trp-cage calculated by PT-METAD with SASA and Alpha RMSD collective variables at four selected temperatures.

(Pietrucci and Laio, 2009) collective variable with parameters set to default in PLUMED. Well-tempered metadynamics was performed with hills of height 1 kJ·mol^{-1} added every 1 ps with hill widths 1 nm^2 for SASA and 1 for ALPHARMSD, respectively. Bias factor of well-tempered metadynamics was set to 15 (Barducci et al., 2008). Unfortunately, 100 ns metadynamics starting from the folded structure lead to quick unfolding but not to folding (see Supporting Information, **Figure S3**).

Therefore, in order to enhance sampling in degrees of freedom that cannot be addressed by the applied CVs we replaced metadynamics by PT-METAD (Bussi et al., 2006). The system was simulated at 32 temperatures ranging from 278.0 to 595.0 K. Metadynamics parameters were not changed. The plot in **Figure S4** demonstrated significant overlap of potential energy histograms, which is a prerequisite for a reasonable replica exchange rate. During a PT-METAD (50 ns in each replica) we observed eight folding events (recognized by visual inspection of "demuxed" trajectories, see Supporting Information, **Figure S5**). This is in contrast to a parallel tempering molecular dynamics simulation with otherwise same parameters (without metadynamics), where no folding events were observed.

The size of box in PT-METAD was small to increase replica exchange probability and thus to reduce required number of replicas. We admit that this increases risk of self-interaction artifacts in folding simulations. We visually examined folding simulation trajectories and discovered examples of self-interactions (see **Figure S6**). These interactions were relatively short-living. Moreover, we believe that self-interactions complicate, not facilitate, folding. Therefore, neural network approximated SASA can be seen as a successful CV.

Free energy surfaces were calculated from Gaussian hills accumulated at each temperature in PT-METAD (**Figure 5**). Free energy surfaces are in a good agreement with the

results from literature (Lindorff-Larsen et al., 2011). At low temperatures there were two free energy minima with approximately same value of free energy. One at CVs [∼17 nm^2, ∼6] corresponds to the folded structure. The second one at CVs [∼21 nm^2, ∼0.5] corresponds to the unfolded structure. The fact that both minima have approximately same free energy value is in agreement with the fact that in an unbiased simulation (Lindorff-Larsen et al., 2011) the system spends approximately same time in unfolded and folded state. At slightly elevated temperatures the minimum corresponding to the folded state becomes more shallow and at high temperature it becomes almost indistinguishable. Other states, such as those with higher helical content or low-SASA states with low helical content, were predicted as energetically unfavorable.

One of the motivations for development of *anncolvar* was the potential speed gain compared to Path Collective Variables and Property Map. These two approaches require multiple RMSD-fitting processes in each step. This problem has been addressed by Close Structure algorithm (Pazúriková et al., 2017), which reduces the number of RMSD-fitting processes, but still requires multiple RMSD-fitting processes in some steps of the simulation. The approach presented here requires only one RMSD-fitting in each step. RMSD-fitting free approaches (such as those using interatomic distances) are not supported by *anncolvar*, but can be used in future if it turns out to be a viable strategy.

For the cyclooctane derivative, metadynamics was significantly slower than unbiased simulation (∼40 ns·day^{-1} vs. ∼7 μs·day^{-1} on single CPU). However, this can be explained by the fact that not only CV calculation, but also calculation of the bias potential takes large proportion of CPU load in the system much smaller (24 atoms) than biomolecular systems with explicit solvents. On the other hand, the situation was much more favorable in biomolecular systems with an explicit solvent. Metadynamics (Trp-cage with 11,128 water and one chloride) was approximately twice slower than corresponding unbiased simulation (both on 8 CPU cores). Examination of one part (5 ns) of metadynamics simulation revealed that metadynamics force calculation accounts for 78% of total force calculations and 59% of total calculations. Similarly PT-METAD (Trp-cage with 1,366 waters and one chloride) was also approximately twice slower than corresponding unbiased parallel tempering simulation (both on 32 CPU cores).

Neural networks architectures used in this study were relatively small to avoid slowing down of simulations. They are not deep enough to be called deep learning. There are several options to improve the program in order to enable deeper neural network models. For example, we plan to enable loading of weights and biases into PLUMED as text files. This would also simplify file handling. There is also space for parallelization and GPU offloading. We plan to work on this in near future.

In this work we used two different data sets to train the neural network. The first was generated by a systematic conformer generation. The second was generated by a long molecular dynamics simulation. Both approaches require that the structure corresponding to the free energy minimum is present in the training data set. This leads to the "chicken-and-egg problem" discussed by Chen and Ferguson (Chen and Ferguson, 2018). We have to know the structure of folded protein (or at least it must be present in the training data set without knowing that it is the folded one) in order to simulate folding of the protein. Therefore, the approach outlined in this work is suitable to study protein folding mechanisms with known folded structure, but it is not suitable for *de novo* structure prediction. Generative machine learning models, which involve models that can make accurate prediction outside the training set as they learn a broad distribution of the training set, may be useful to address this problem. Reinforcement learning can be another answer to this problem.

AUTHOR CONTRIBUTIONS

All authors contributed to the code of anncolvar. DT, IH, and VM: tested the code; DT, IH, and VS: carried out simulations, all authors contributed to the manuscript.

FUNDING

This work was funded by COST action OpenMultiMed (CA15120, Ministry of Education, Youth and Sports of the Czech Republic LTC18074), specific university research (Ministry of Education, Youth and Sports of the Czech Republic 21-SVV/2019) and Czech National Infrastructure for Biological Data (ELIXIR CZ, Ministry of Education, Youth and Sports of the Czech Republic LM2015047). Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the program Projects of Large Research, Development, and Innovations Infrastructures.

ACKNOWLEDGMENTS

Authors would like to thank Brown et al. and D. E. Shaw Research for data used in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00025/full#supplementary-material

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., et al. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001

Abrams, C., and Bussi, G. (2014). Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and

temperature-acceleration. *Entropy* 16, 163–199. doi: 10.3390/e16010163

Barducci, A., Bussi, G., and Parrinello, M. (2008). Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* 100:020603. doi: 10.1103/PhysRevLett.100.020603

Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., et al. (2009). PLUMED: a portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* 180, 1961–1972. doi: 10.1016/j.cpc.2009.05.011

Branduardi, D., Gervasio, F. L., and Parrinello, M. (2007). From A to B in free energy space. *J. Chem. Phys.* 126:054103. doi: 10.1063/1.2432340

Brown, W. M., Martin, S., Pollock, S. N., Coutsias, E. A., and Watson, J.-P. (2008). Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.* 129:064118. doi: 10.1063/1.2968610

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101. doi: 10.1063/1.2408420

Bussi, G., Gervasio, F. L., Laio, A., and Parrinello, M. (2006). Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* 128, 13435–13441. doi: 10.1021/ja062463w

Chen, W., and Ferguson, A. L. (2018). Molecular enhanced sampling with autoencoders: on-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* 39, 2079–2102. doi: 10.1002/jcc.25520

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089. doi: 10.1063/1.464397

Das, P., Moll, M., Stamati, H., Kavraki, L. E., and Clementi, C. (2006). Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9885–9890. doi: 10.1073/pnas.0603553103

Ferrarotti, M. J., Bottaro, S., Pérez-Villa, A., and Bussi, G. (2015). Accurate multiple time step in biased molecular simulations. *J. Chem. Theory Comput.* 11, 139–146. doi: 10.1021/ct5007086

Galvelis, R., and Sugita, Y. (2017). Neural network and nearest neighbor algorithms for enhancing sampling of molecular dynamics. *J. Chem. Theory Comput.* 13, 2489–2500. doi: 10.1021/acs.jctc.7b00188

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Guo, A. Z., Sevgen, E., Sidky, H., Whitmer, J. K., Hubbell, J. A., and de Pablo, J. J. (2018). Adaptive enhanced sampling by force-biasing using neural networks. *J. Chem. Phys.* 148, 134108–134109. doi: 10.1063/1.5020733

Laio, A., and Parrinello, M. (2002). Escaping free-energy minima. *Proc Natl. Acad. Sci. U.S.A.* 99, 12562–12566. doi: 10.1073/pnas.202427399

Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science*. 334, 517–520. doi: 10.1126/science.1208351

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., O'Dror, R., et al. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*. 78, 1950–1958. doi: 10.1002/prot.22711

Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* 9:5. doi: 10.1038/s41467-017-02388-1

Martin, S., Thompson, A., Coutsias, E. A., and Watson, J.-P. (2010). Topology of cyclo-octane energy landscape. *J. Chem. Phys.* 132:234115. doi: 10.1063/1.3445267

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109, 1528–1532. doi: 10.1016/j.bpj.2015.08.015

Nandy, A., and Biswas, M. (2018). *Reinforcement Learning*. New York, NY: Apress Media.

Oliphant, T. E. (2006). *A Guide to NumPy*. Spanish Fork, UT: Trelgol Publishing.

Pazúriková, J., Krenek, A., and Spiwok, V., Šimková, M. (2017). Reducing the number of mean-square deviation calculations with floating close structure in metadynamics. *J. Chem. Phys.* 146:115101. doi: 10.1063/1.4978296

Pérez, A., Martínez-Rosell, G., and De Fabritiis, G. (2018). Simulations meet machine learning in structural biology. *Curr. Opin. Struct. Biol.* 49, 139–144. doi: 10.1016/j.sbi.2018.02.004

Pietrucci, F., and Laio, A. (2009). A collective variable for the efficient exploration of protein beta-structures with metadynamics: application to sh3 and gb1. *J. Chem. Theory Comput.* 5, 2197–2201. doi: 10.1021/ct900202f

Ramachandran, P., and Varoquaux, G. (2011). Mayavi: 3D visualization of scientific data. *IEEE. Comput. Sci. Eng.* 13, 40–51. doi: 10.1109/MCSE.2011.35

Seo, B., Kim, S., Lee, M., Lee, Y.-W., and Lee, W. B. (2018). Driving conformational transitions in the feature space of autoencoder neural network. *J. Phys. Chem. C* 122, 23224–23229. doi: 10.1021/acs.jpcc.8b08496

Spiwok, V., Hošek, P., and Šućur, Z. (2015a). Enhanced sampling techniques in biomolecular simulations. *Biotech. Adv.* 6(pt 2), 1130–1140. doi: 10.1016/j.biotechadv.2014.11.011

Spiwok, V., and Králová, B. (2011). Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *J. Chem. Phys.* 135:224504. doi: 10.1063/1.3660208

Spiwok, V., Oborský, P., Pazúriková, J., Krenek, A., and Králová, B. (2015b). Nonlinear vs. linear biasing in Trp-cage folding simulations. *J. Chem. Phys.* 142:115101. doi: 10.1063/1.4914828

Sridharan, S., Nicholls, A., and Sharp, K. A. (1995). A rapid method for calculating derivatives of solvent accessible surface areas of molecules. *J. Comput. Chem.* 16, 1038–1044. doi: 10.1002/jcc.540160810

Sultan, M. M., and Pande, V. S. (2018). Automated design of collective variables using supervised machine learning. *J. Chem. Phys.* 149:094106. doi: 10.1063/1.5029972

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Torrie, G. M., and Valleau, J. P. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* 23, 187–199. doi: 10.1016/0021-9991(77)90121-8

Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). PLUMED 2: new feathers for an old bird. *Comput. Phys. Commun.* 185, 604–613. doi: 10.1016/j.cpc.2013.09.018

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Wehmeyer, C., and Noé, F. (2018). Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* 148:241703. doi: 10.1063/1.5011399

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 09 July 2019
doi: 10.3389/fmolb.2019.00047

Machine Learning Classification Model for Functional Binding Modes of TEM-1 β-Lactamase

Feng Wang[1], *Li Shen*[1], *Hongyu Zhou*[1], *Shouyi Wang*[2], *Xinlei Wang*[3] *and Peng Tao*[1]*

[1] *Department of Chemistry, Center for Scientific Computation, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, TX, United States,* [2] *Department of Industrial, Manufacturing, and Systems Engineering, University of Texas at Arlington, Arlington, TX, United States,* [3] *Department of Statistical Science, Southern Methodist University, Dallas, TX, United States*

TEM family of enzymes is one of the most commonly encountered β-lactamases groups with different catalytic capabilities against various antibiotics. Despite the studies investigating the catalytic mechanism of TEM β-lactamases, the binding modes of these enzymes against ligands in different functional catalytic states have been largely overlooked. But the binding modes may play a critical role in the function and even the evolution of these proteins. In this work, a newly developed machine learning analysis approach to the recognition of protein dynamics states was applied to compare the binding modes of TEM-1 β-lactamase with regard to penicillin in different catalytic states. While conventional analysis methods, including principal components analysis (PCA), could not differentiate TEM-1 in different binding modes, the application of a machine learning method led to excellent classification models differentiating these states. It was also revealed that both reactant/product states and apo/product states are more differentiable than the apo/reactant states. The feature importance generated by the training procedure of the machine learning model was utilized to evaluate the contribution from residues at active sites and in different secondary structures. Key active site residues, Ser70 and Ser130, play a critical role in differentiating reactant/product states, while other active site residues are more important for differentiating apo/product states. Overall, this study provides new insights into the different dynamical function states of TEM-1 and may open a new venue for β-lactamases functional and evolutional studies in general.

Keywords: TEM-1 β-lactamase, functional binding modes, structural analysis, random forest classification, machine learning, molecular dynamics

OPEN ACCESS

Edited by:
Gennady Verkhivker, Chapman University, United States

Reviewed by:
Elif Ozkirimli, Bogaziçi University, Turkey
Pavel Srb, Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Peng Tao
ptao@smu.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *11 February 2019*
Accepted: *11 June 2019*
Published: *09 July 2019*

Citation:
Wang F, Shen L, Zhou H, Wang S, Wang X and Tao P (2019) Machine Learning Classification Model for Functional Binding Modes of TEM-1 β-Lactamase. Front. Mol. Biosci. 6:47. doi: 10.3389/fmolb.2019.00047

INTRODUCTION

Antibiotic resistance against almost all the existing antibiotics presents a major risk to global health. Among many other factors, β-lactamases as a group of proteins that hydrolyze antibiotics play a key role in antibiotic resistance. The serine β-lactamases, which utilize a serine residue to hydrolyze the β-lactam ring-based antibiotics, and zinc based β-lactamases, are the two main groups of β-lactamases in general. Class A β-lactamases are one dominant subgroup in serine β-lactamases and are highly diversified. TEM-1, the most commonly encountered β-lactamase in Gram-negative bacteria, belongs to the Class A β-lactamases (Bradford, 2001). The structure and potential catalytic

mechanisms of TEM-1 have been studied extensively as a model system of Class A β-lactamases (Lamotte-Brasseur et al., 1991, 1999; Jelsch et al., 1992; Fonzé et al., 1995; Maveyraud et al., 1998; Petrosino et al., 1998; Minasov et al., 2002; Díaz et al., 2003; Hermann et al., 2003; Golemi-Kotra et al., 2004; Roccatano et al., 2005; Savard and Gagné, 2006; Doucet et al., 2007). The catalytic mechanism of TEM-1 can be divided into acylation and deacylation steps using penicillin as an example. The acylation step leads to an acylenzyme Michaelis-complex intermediate with a covalent bond formed between the Ser70 residue and ring opening product of penicillin β-lactam ring. This covalent bond in the acylenzyme intermediate is further hydrolyzed during the deacylation step, leading to an ineffective β-lactam ring-opening product detached from the enzyme. Catalytic functions of key residues at and surrounding an active site have been investigated extensively with some ongoing controversy (Oefner et al., 1990; Herzberg and Moult, 1991; Lamotte-Brasseur et al., 1991, 1992, 1994; Strynadka et al., 1992, 1996; Matagne et al., 1998). The active site of TEM-1 contains several conserved residues that are important for catalysis: Ser70, Lys73, Lys234, Glu166, and Ser130 (Fisette et al., 2010). Here and in the rest of the article, the sequence numbering of Ambler et al. (1991) is used to be consistent with the general literature about TEM-1 (Savard and Gagné, 2006; Doucet et al., 2007; Fisette et al., 2010). It is also believed that some residues, including Asn170, Ala237, Ser235, and Arg244, help to stabilize the acylenzyme intermediate. Although not fully determined, the contribution of these residues to TEM-1 catalytic mechanisms have been investigated extensively (Zafaralla et al., 1992; Stec et al., 2005; Marciano et al., 2009; Stojanoski et al., 2015; Palzkill, 2018). In addition, an allosteric site consisted of helixes 11 (residue 219–226) and 12 (residues 271–289) of TEM-1 were proposed (Horn and Shoichet, 2004). Two novel inhibitors were reported to destabilize the TEM-1 at high temperature. The two inhibitors can bind to the allosteric site in TEM-1, which locates in between helices 11 and 12. The allosteric site is 16 Å away from the active site. It was proposed that TEM-1 conformational changes were transmitted by a key catalytic residue, Arg244 (Horn and Shoichet, 2004). In another study, the allosteric site of TEM-1 was further detected through binding with a β-lactamase inhibitor protein (BLIP). It was suggested that the connections between active site and allosteric site may be modulated by the helix 10 region (residues 218–230) and Tpr229 in TEM-1 (Meneksedag et al., 2013). The allosteric site helixes 11 and 12 were also proposed as a cryptic pocket formation of TEM-1 (Oleinikovas et al., 2016). In addition, the residues P226-W229-P252 were identified as a PWP triad to stabilize the helix 10 region (Avci et al., 2016, 2018).

One important aspect of TEM-1 for its function is dynamics. Therefore, the molecular dynamics (MD) simulations were carried out to characterize dynamical properties of TEM-1 binding with benzyl penicillin molecule. A so-called Ω loop spans residues 163 through to 180 (including the key Glu166 residue for catalysis), and forms one edge of the active site (Dideberg et al., 1987; Herzberg and Moult, 1987; Moews et al., 1990; Jelsch et al., 1993; Vanwetswinkel et al., 2000). Some earlier MD simulations showed that the Ω loop was rather stable even with the absence of the ligand (Díaz et al., 2003). The whole TEM-1 has also been shown to be unusually rigid with limited motions on the picosecond-to-nanosecond time scale through a nuclear magnetic resonance (NMR) spectroscopy study (Savard and Gagné, 2006). Through more extended simulations and NMR studies, a variety of motions displayed by Ω loop are revealed to be potentially important for catalysis (Fisette et al., 2010). Another simulation study of TEM-1 binding with benzylpenicillin suggested that a substrate binding led to increased flexibility of Ω loop while making TEM-1 globally more rigid (Fisette et al., 2012). In addition to benzylpenicillin as a substrate, simulations were also carried out for TEM-1 bound with another two antibiotics, amoxicillin and ampicillin, to reveal that even the subtle differences in chemical structures of ligands could also regulate the substrate recognition (Pimenta et al., 2013).

One overlooked aspect of TEM-1's function is the binding with antibiotics and their hydrolysis product. Penicillin, for example, could bind with TEM-1 as favorable substrate, while the hydrolysis product of penicillin needs to leave the binding pocket for the turnover of this enzyme. Given the rigidity and sensitivity of the TEM-1 structure to the ligand, the response of protein dynamics to the ligand, in different chemical states through catalysis, could be significant and important for its function, however, this remains under-appreciated. One of the reasons for this is probably due to the fast turnover rate, which does not allow for a reliable experimental probe of the protein binding with ligands during its quick catalytic cycles. MD simulations provide an alternative way to scrutinize the difference between the binding modes of protein with similar ligands. However, due to the rigidity of TEM-1 and the similarity between two ligands of interest, some special analysis tools would be necessary for the purpose of comparison.

Machine learning methods are computational tools that construct data-driven prediction models based on training data. In recent years, machine learning methods have been successfully applied in computational chemistry (Husic and Pande, 2018), including pharmaceutical data analysis (Burbidge et al., 2001), protein–ligand binding affinity prediction (Ballester and Mitchell, 2010; Decherchi et al., 2015) and MD simulations based on machine learning analysis of quantum-mechanical forces (Li et al., 2015; Cortina and Kasson, 2018; Shcherbinin and Veselovsky, 2019). Recently, we have introduced two widely applied machine learning algorithms, a decision tree and an artificial neural network, to build classification models to differentiate two allosteric states of the second PDZ domain (PDZ2) in the human PTP1E protein as a dynamics-driven allosteric protein (Zhou et al., 2018). Despite the lack of a significant conformational change between two states of PDZ2, it was demonstrated that both algorithms could build effective prediction models and provide reliable quantitative evaluation of the contributions from individual residues to overall difference between the two states.

In this study, we applied another machine learning algorithm, random forest, to build models. Random Forest (Breiman, 2001) is a supervised learning algorithm that relies on an ensemble method to create an entire forest of random uncorrelated

decision trees, in order to achieve a more accurate and stable prediction. It has been found to be very useful in a wide scope of applications, due to its superior performance in classification and regression problems, as well as its ease of use and flexibility. The recognition of TEM-1 against ligands in different states is interrogated through simulations studies. The random forest method as an effective machine learning technique has been applied to analyze the simulations of TEM-1 in different binding states and evaluate the contribution from every residue and related secondary structures to the recognition of ligands in different states of TEM-1. Potential key residues could be identified based on their feature importance generated from the machine learning model of the simulation data of TEM-1 in different states. The TEM-1 hydrolysis mechanism is of great interest and has been subjected to extensive computational studies focusing on the TEM-1 active site or nearby residues (Díaz et al., 2001; Meroueh et al., 2005; Roccatano et al., 2005; Sgrignani et al., 2014). However, the potential contribution from protein dynamics in different states to catalysis has been largely overlooked. We hypothesize that TEM-1 in different catalytic states, including binding states with reactant and product, are differentiable and could provide further mechanistic details if subjected to appropriate analyses.

Therefore, the current study focuses on the development of classification models to differentiate dynamics of TEM-1 in different functional states and on obtaining information to correlate protein dynamics with individual residues regardless their positions relate to the active site. The dynamics of different states are compared with each other in the training process, governed by the random forest method. In the random forest method, the contribution from each residue to the overall classification model was measured as importance of features (Zhou et al., 2019). A higher importance value of a feature represented a higher contribution in classifying different functional states. Using the feature importance, important structures and residues identified by this computational study are also in agreement with previous studies of this enzyme. The analysis about active and allosteric sites of TEM-1 also sheds new light on the allosteric component of TEM-1 functions. The remainder of the paper is organized in four parts: computational methods, results, discussion, and conclusion.

COMPUTATIONAL METHODS

Molecular Dynamics (MD) Simulations

Three states of TEM-1 were subject to molecular dynamics (MD) simulations. TEM-1 bound with benzyl penicillin (**Figure 1A**) is referred to as the reactant state; TEM-1 bound with product of hydrolyzing benzyl penicillin (**Figure 1B**) is referred to as the product state, and TEM-1 alone without a ligand is referred to as the apo state. No crystal structure is available for TEM-1 binding with penicillin either as a reactant or product. The complex structure related to TEM-1 catalysis against penicillin with the best quality is an intermediate structure (PDB ID: 1fqg), which has been used for various computational studies. Therefore, this crystal structure was used to generate all three states of TEM-1, based on a hypothesis that equilibrium simulations could lead to sufficient sampling in these functional states. CHARMM molecular simulation program suite, version 40b1, was used to prepare and set up the systems (Halgren, 1992). Hydrogen atoms were added to the crystal structure of TEM-1 bound with benzyl penicillin using the hydrogen position construction facility (HBUILD) of the CHARMM. The benzyl penicillin ligand was removed to create the apo state of TEM-1. The benzyl penicillin structure was also modified using CHARMM internal coordinate editing functions to produce the benzyl penicillin hydrolysis product. CHARMM36 force field was used for TEM-1(Best et al., 2012). The CHARMM General Force Field (CGenFF) was generated for the benzyl penicillin and the benzyl penicillin hydrolysis product using online server ParamChem (https://cgenff.paramchem.org/). All systems are solvated in a water box using a TIP3P model with the addition of sodium and chloride ions to balance the charge and reproduce typical physiological ion concentrations.

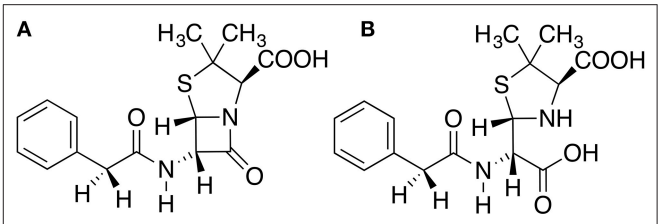


FIGURE 1 | Chemical structures of **(A)** Benzyl penicillin, **(B)** the hydrolysis product of benzyl penicillin.

The simulation boxes were subjected to 5,000 steps of the steepest descent energy minimization and further energy minimization using the adopted basis Newton-Raphson (ABNR) method until the total gradient of the system was lower than 0.02 kcal/mol•Å. Subsequently, the minimized simulation systems were subjected to 24 picoseconds (ps) isothermal-isobaric (NPT) ensemble equilibrium, gradually raising the temperature from 100 to 300 K. The system was then equilibrated via NVT ensemble MD simulations at 300 K. The time step for MD simulations is 2 fs, with all the bonds associated with hydrogen being fixed during the simulation using SHAKE method (Ryckaert et al., 1977). Periodic boundary condition was used in all simulations, and electrostatic interactions were calculated using the particle mesh Ewald method (Darden et al., 1993). For each state, five independent 100 ns NVT ensemble MD simulations were carried out as the production runs after 10 ns of equilibration. OpenMM simulation package was used to carry out the production MD simulations (Friedrichs et al., 2009; Eastman and Pande, 2015; Eastman et al., 2017).

Analysis of MD Simulations

Root-Mean-Square Deviation (RMSD)

RMSD is used to measure the difference in conformation for each snapshot of the MD simulations from a reference structure. For a molecular structure represented by Cartesian coordinate

vector r_i ($i = 1$ *to* N) of N atoms, the RMSD is calculated as the following:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(r_i^0 - Ur_i\right)^2}{N}}, \quad (1)$$

Where r_i^0 is the Cartesian coordinate vector of the i^{th} atom in the reference structure. The transformation matrix U is defined as the best-fit alignment between the TEM-1 structure along trajectories with respect to the reference structure.

Root-Mean-Square Fluctuation (RMSF)

RMSF is used to measure the fluctuation of conformation for each frame of the trajectories from the averaged structure.

$$RMSF_i = \left[\frac{1}{T}\sum_{t=1}^{T}\left|r_i(t) - r_i\right|^2\right]^{\frac{1}{2}}, \quad (2)$$

Where T is the time period and $\overline{r_i}$ is the averaged position of atom i over the whole time period.

Principal Component Analysis (PCA)

For each state, PCA was performed by projecting each of the extracted 25,000 frames from five independent trajectories on the principal normal modes. The analysis was carried out using mdtraj package (McGibbon et al., 2015) and scikit-learn library in python (Pedregosa et al., 2011). PCA is a method to reduce the dimensionality of the motion of molecules. It can extract the dominant modes of the motion from a trajectory of molecular dynamic simulation. The normal modes for PCA (Jolliffe, 2011) were obtained through diagonalizing the correlation matrix of the atomic position in one trajectory. The correlation matrix element is calculated by

$$C_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}} = \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\sqrt{\left[\left(\langle r_i^2\rangle - \langle r_i\rangle^2\right)\left(\langle r_j^2\rangle - \langle r_j\rangle^2\right)\right]}}, \quad (3)$$

Where C_{ij} is the Pearson correlation coefficient between atoms i and j.

The distributions of three TEM-1 states simulations in the PCA projection space are normalized and plotted as a density contour graph. The distribution density function was estimated by the Gaussian kernels (Scipy 1.2.1) (Turlach, 1993; Bashtannyk and Hyndman, 2001; Scott, 2015; Silverman, 2018).

Random Forest Model

The random forest classification was used in this study to develop classification models for the three states of TEM-1. The python package scikit-learn v0.20.3 was used to carry out the training and testing using this model. For each independent 100 ns simulation of all states, 5,000 frames were evenly extracted as the training and testing data. For each state, four simulations among five production runs were randomly selected as the training set with the remaining simulation used as the testing set. For each selected frame from the simulation, all the pairwise distances among the α carbons (Cα) of TEM-1 backbone are extracted as the features for training purpose. A total of 263 TEM-1 amino acid residues result in 34,453 pairwise distances as the training features. As a pre-step before the classification, the feature selection is carried out using the random forest classification model. Following a previous study to build feature selection using machine learning methods (Zhou et al., 2018), all features are pre-screened to select features accounting for 98.0% as total importance. The apo/product model has 901 features out of the total of 34,453 features. Similarly, after the feature selection, the reactant/product model has 1,170 features, the non-product/product model has 964 features and the apo/reactant model has 1,923 features for their classification models. The final classification models were developed using these preselected features. The number of preselected features for four training models with all preselected features are provided in the **Supplementary Material**.

A random forest algorithm was built on the decision tree models. First, training data was randomly divided into numerous sets and decision tree models were built based on each set. Then all the decision tree models were combined to generate final random forest classification model (Breiman, 2001; Geurts et al., 2006; Louppe, 2014). The random forest algorithm implemented in scikit-learn v0.20.3 (ensemble.RandomForestClassifier) was employed in this study. The number of decision trees generated in the random forest model (referred to as n_estimator) was varied for the best performance with the highest training and validation accuracy (**Supplementary Figure 1**). For each model, the number of decision trees to obtain the highest accuracy of validation was selected for the final classification model.

The random forest method was employed for two purposes in this study, including feature prescreening and classification model developing. In feature prescreening, the feature importance generated from preliminary random forest training process is assigned to each feature. All features are sorted based on their feature importance. The features with the sum of their importance accounting for 98% are selected for the final classification model. These pre-screened features of each classification model present in this study are listed in the **Supplementary Material**. The final classification models were trained using the pre-screened features and with new set of feature importance generated from the training process. The new set of feature importance is used for further analyses presented in this study.

Scores

In this study, the scores including accuracy, precision, recall, and F1 score were used to evaluate the performance of each classification model. The python package v0.20.3 (Pedregosa et al., 2011; Buitinck et al., 2013) was employed to generate these four scores. The accuracy score is defined as

$$accuracy = \frac{1}{N}\sum_{i=0}^{N-1} 1(\hat{y}_i = y_i), \quad (4)$$

where N is the number of samples, $\hat{y}_i$ is the predicted label and y_i is the true label for the i_{th} sample.

In a binary classification task, such as the classification models in this study with two labels, the predictions of the model are evaluated as the following. Positive/negative labels are used to reflect the prediction made by the model. True/false are used to represent whether the predicted labels correspond to the observed labels (real labels). Accordingly, precision, recall and F1 scores are defined as the following.

$$precision = \frac{tp}{tp + fp}, \tag{5}$$

$$recall = \frac{tp}{tp + fn}, \tag{6}$$

$$F1 = 2\frac{precision^{*}recall}{precision + recall}, \tag{7}$$

Term *tp* (true positive) represents the situation that the model gives positive prediction and the observed label is indeed positive. Term *fp* (false positive) represents that the model gives positive prediction, but the observed label is negative. Term *fn* (false negative) represents that the model gives negative prediction, but the observed label is actually positive. F1 score is a weighted mean of the precision and recall.

Feature Importance

The importance of each feature is generated by random forest algorithm based on Gini impurity (Equation 8). A higher importance represents a more important feature in distinguishing different states. The Gini importance implemented in python package scikit-learn v0.20.3 was used in this study and briefly introduced in the Equations (8–12) as the following.

The feature importance was calculated as Gini impurity:

$$Gini\ impurity = \sum_{i=1}^{C} -f_i(1 - f_i), \tag{8}$$

where f_i is the frequency of a label at a node, and C is the number of labels.

In the random forest models, many decision trees are constructed for training purpose. All the predictions from these individual trees are collected to make the final random forest classification model. The importance (n_j) of a node *j* in each decision tree was represented by Gini impurity:

$$n_j = w_j C_j - \sum_{1}^{m} w_{m(j)} C_{m(j)}, \tag{9}$$

where w_j is the weighted number of samples reaching node *j*, C_j is the impurity value of node *j*, and m is the number of child nodes of the tree.

The feature importance of feature *i* on decision tree is calculated as:

$$f_i = \frac{\sum_{1}^{s} n_j}{\sum_{k \in all\ nodes} n_k}, \tag{10}$$

where *s* is the times of node *j* split on feature *i*.

The normalized feature importance in a decision tree is calculated through:

$$norm\, f_i = \frac{f_i}{\sum_{j\ \in all\ features\ in\ a\ tree} f_j}, \tag{11}$$

The final feature importance in random forest classification is calculated as:

$$F_i = \frac{\sum_{j \in all\ trees} norm\, f_i}{N}, \tag{12}$$

where $norm\ f_i$ is the normalized feature importance values of a decision tree, *N* is the total number of trees (Breiman, 2001; Geurts et al., 2006; Pedregosa et al., 2011; Louppe, 2014).

In our classification models, the features are pairwise Cα distances. To evaluate the importance of each amino acid residue, all the feature importance of the pairwise distances relating to each residue are summed up and divided by two to generate the importance of a residue. Then the total importance of 263 residues were accumulated and the importance percentage of each residue could be calculated based on the total importance. The value of importance percentage represents the ability of a residue to differentiate three states. In other words, the importance could help to evaluate the contribution from a residue to differentiate three states in dynamic motions.

RESULTS

TEM-1 Three States Simulations Analysis

The time evolution of the RMSD of TEM-1 in five independent simulation sets in apo, reactant, and product states are plotted in **Figure 2**. All RMSD values were calculated with reference to the TEM-1 crystal structure. The averaged RMSD values are 1.5, 1.3, and 1.1 Å for the apo, reactant, and product states, respectively. The plots suggest that the TEM-1 is rather stable with low RMSD fluctuations in all three states. Among three states, the apo state displays the highest TEM-1 fluctuation, and the product state displays the lowest TEM-1 fluctuation. To address the concern of the simulation convergence, we also calculated the accumulative entropy of TEM-1 in each state along each independent simulation (**Supplementary Figure 2**). All three states display clear convergence tendency in each simulation.

RMSF of individual residues was calculated for each state using all five simulations and plotted in **Figure 3**. In agreement with the RMSD results, TEM-1 in the apo state has the highest fluctuation for most part of the protein (blue dashed line in **Figure 3**). However, TEM-1 in both the reactant and product states also displays higher fluctuation than the apo state in certain part, revealing that the binding with ligands and the type of ligand do exert a subtle impact on protein dynamics.

Then, we carried out PCA using all 15 simulations from three states as an attempt to develop a model differentiating three states of TEM-1. The simulations of each state are projected onto the surface as contour plots with normalization using the first principal component (PC1) and second principal component (PC2) (**Figure 4**). Overall, all three states largely overlap with

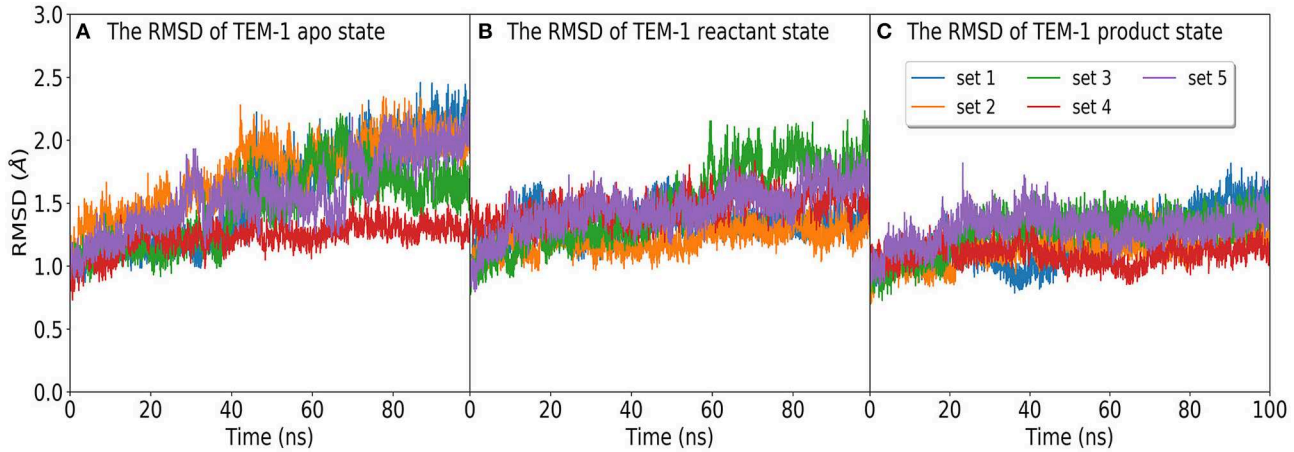


FIGURE 2 | The RMSD distribution of molecular dynamics simulations of TEM-1 in **(A)** apo state, and binding with benzyl penicillin in **(B)** reactant and **(C)** product states. In each state, the RMSD are calculated in five independent 100 ns simulations labeled as set 1 to set 5.

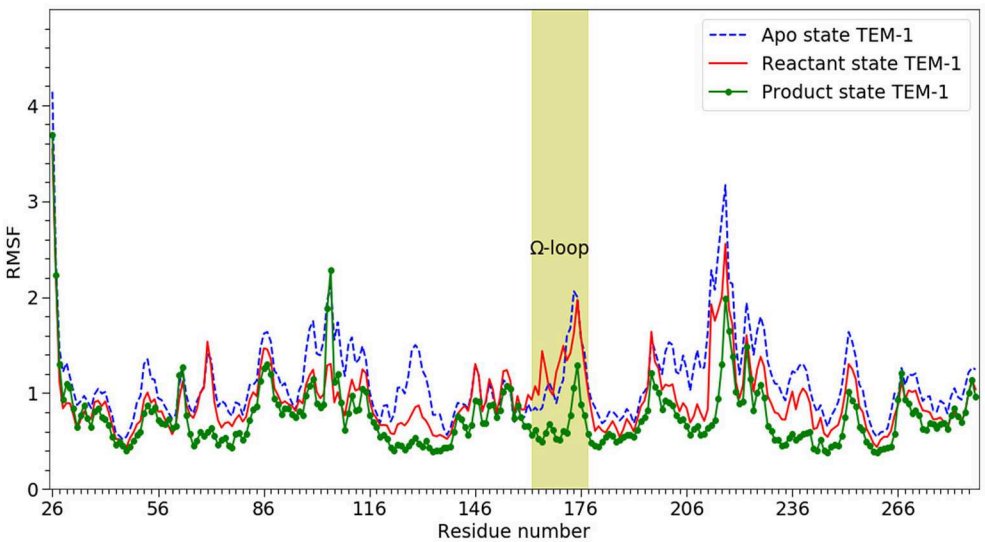


FIGURE 3 | The RMSF of α-carbons (Cα) from 26 to 288 on TEM-1 β-lactamase in apo (blue dash line), reactant (red line) and product states (dot line), Ω loop (residue 163–180) highlighted. All three states have overall similar distribution but with significant difference. The product shows the lowest overall RMSF. The apo state show the highest overall RMSF.

each other on the PC1/PC2 surface, and each state has two or three minima, which are referred to as attraction basins. The reactant and product states cover similar areas and largely overlap with each other, with their attraction basins close to each other. The apo state has different attraction basins and has much narrower distribution than the other two states. The PCA results reflect that the TEM-1 structure is generally rigid without significant global conformational change. However, the subtle differences among the distributions of TEM-1 in different states in the PCA space do indicate the shift in population of TEM-1 in different binding states. The following analysis using the random forest model provides more insight into these subtle differences.

Random Forest Model

The training and testing results of the random forest model for all three states, including accuracy, precision, recall, and F1 scores, are plotted in **Figure 5**. Classification models were developed to differentiate between apo and product states, reactant and product states, non-product (combining the apo and reactant states) and product states, as well as between apo and reactant states. For the classification model to differentiate the reactant and product states, the training with cross-validation provides high performance, and testing provides better than 87% accuracy in all categories (**Figure 5A**), suggesting that the TEM-1 reactant and product states are highly differentiable using the Cα pairwise distances as protein structural information. Slightly better scores

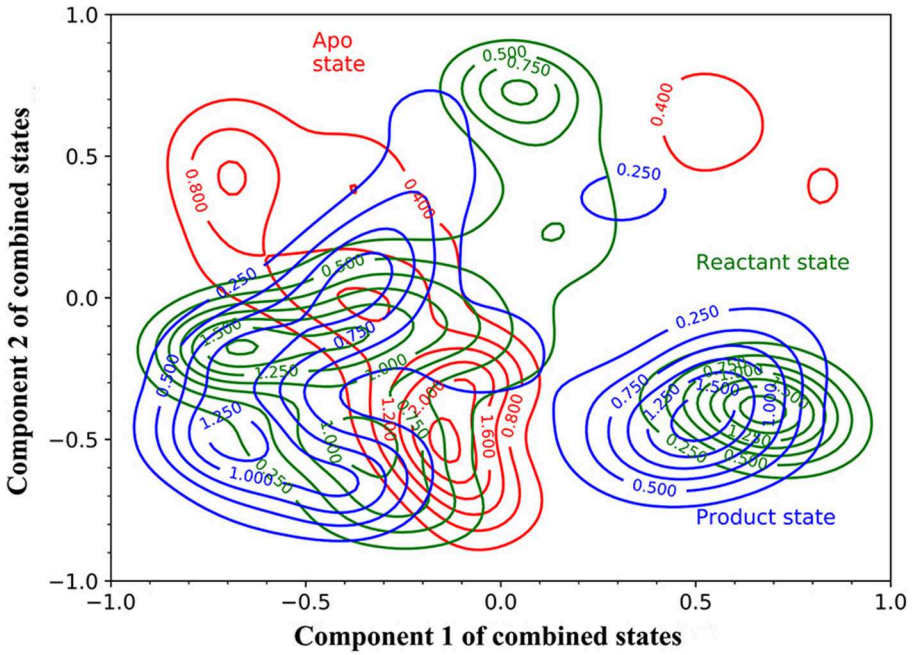


FIGURE 4 | The projection of the simulations of TEM-1 in apo (red), reactant (green) and, product (blue) states onto Component 1 and Component 2 of combined states. Components 1 and 2 are the first and second components from the principal component analysis (PCA) based on the simulations of all three states. The projection on to components 1 and 2 are normalized.

are obtained for the classification model to differentiate the apo and product states (**Figure 5B**). These results show that the TEM-1 in the product state is clearly distinguishable from TEM-1 in the apo and reactant states. However, distinguishability between the apo and reactant states of TEM-1 is significantly lower than the first two pairs (**Figure 5C**), suggesting that these two states share significant similarity in terms of protein backbone structural distributions represented as Cα pairwise distances. To further test this, both apo and reactant states are combined together to be considered as non-product state vs. product state. A classification model differentiating the non-product and the product states is built with cross-validation performance measures close to 100% and testing performance measures ranging between 82 and 99% (**Figure 5D**), similar to the models for apo/product and reactant/product pairs.

As part of preliminary study, two other widely applied machine learning methods, artificial neural network and support vector machine methods, were also applied to develop classification models for TEM-1. Both methods produced models with performance worse than random forest model (**Supplementary Figures 3, 4**). In addition, the random forest method provides importance numerical value for each feature, which could be used to search for key residues and functional groups in protein structure. Therefore, the remainder of the study focuses on random forest model result.

Secondary Structures Contribution

In random forest classification models, each Cα pair is given an importance value reflecting its contribution for the classification model. These values could be used to evaluate, to some extent, the importance of individual amino acid residues. We first used these values to evaluate the contribution of secondary structures in TEM-1, with regard to the differences among different states. For each secondary structure, all the importance values associated with residues in that structure are summed together and divided by two as the overall importance. Three well performing classification models, apo/product, reactant/product, and non-product/product, are used for this comparison purpose. The TEM-1 structure is divided into β-sheets, α-helices, coils and turns as secondary structures and the residues inclusive in these structures. The β-sheet and α-helices of TEM-1 are defined in a previous study (Savard and Gagné, 2006), and are commonly used in general literatures of TEM-1 (Simm et al., 2007; Fisette et al., 2010, 2012). The definition of coils and turns in the database of secondary structure assignments (DSSP) are used in this study (Kabsch and Sander, 1983). There are some coils and turns with just one or two residues. Some of them have small importance values. For simplification, when such a short coil or turn is adjacent to another coil or turn, they are combined as a new coil or turn structure for analysis. However, if a short coil or turn is between β-sheets or α-helices, it was kept by itself.

We further calculate the importance of individual secondary structures and plot it in **Figure 6**. All five β-sheets in TEM-1 have importance values lower than 5% (**Figure 6A**), indicating that the β sheets may not play an important role, with regard to ligand binding. There are 11 helices with varying lengths in TEM-1. Most helices have low importance (**Figure 6B**). The only exception is helix (69–85), which has overall importance close

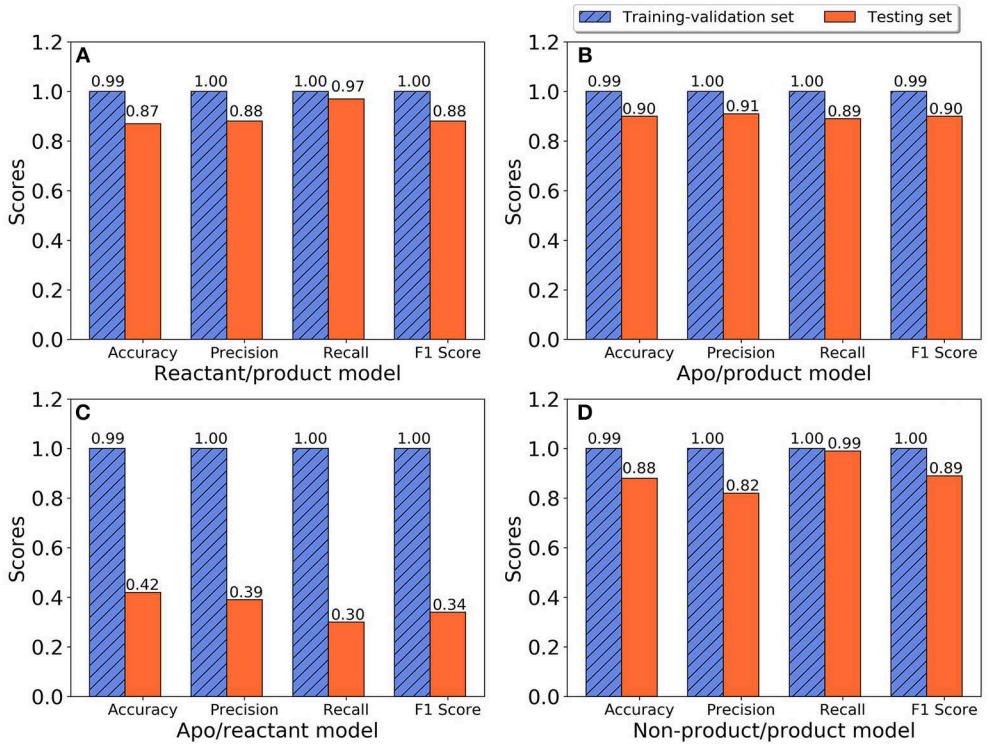


FIGURE 5 | The performance of random forest classification models in accuracy, precision, recall, and F1 scores for training-validation set (blue shadow) and testing set (red). **(A)** Reactant and product states model; **(B)** Apo and product states model; **(C)** Apo and reactant states model; **(D)** Non-product and product states model.

to 16% in the reactant/product model (**Figure 6B**), and also one of the helices around the active site of TEM-1 (**Figure 7** green transparent surface).

There are 10 short fragments being considered as random coils in TEM-1. Among this, residues 213–215 coil shows the highest importance in all three models (**Figure 6C**), which is illustrated and highlighted as cyan structure in **Figure 7**. The second important coil is residues 129–131, with three residues accounting for more than 8% importance in the non-product/product model and around 5% in the other two models. Both 213–215 and 129–131 (highlighted as red structure in **Figure 7**) coils are adjacent to the active site.

There is a total of 15 turn structures in TEM-1, some with significant difference among three classification models. The importance of the residue 216–220 turn (highlighted as yellow structure in **Figure 7**) is the highest on average among all turn structures, followed by residues 102–108 turn (highlighted as green structure in **Figure 7**). Both turns are positioned as gate to cap the TEM-1 active site.

For a better understanding of each residue, the mapping of importance percentage of each residue in TEM-1 obtained from the machine learning training process is plotted in **Figure 8** (divided into three parts A, B, and C). The serial numbers of residues from the PDB file that start from 26 to 111 are used in **Figure 8A**, from 112 to 198 are used in **Figure 8B** and from 199 to the end 288 are used in **Figure 8C**. The overall distributions of TEM-1 individual residue importance based on different classification models resemble each other. Residue 213 has the highest percentage (9.3%) in the apo/product model (**Figure 8C**), which is also the highest percentage for a single residue among all three models. In reactant/product model, residue 70 has the highest percentage as 8.4% (**Figure 8A**). In all three models, residues 67–73, 103–107, 127–135, 162–171, 176–182, and 210–220 have relative high importance percentages in all three models. Interestingly, these residue regions were proposed to undergo conformational changes in a previous NMR study (Savard and Gagné, 2006).

For each model, the top 10 residues with the highest percentages are listed and illustrated with the TEM-1 structure in **Figures 9A–F**. Most of the key residues identified through the classification model are not on either helices or strands secondary structures. However, few active site residues are among the top 10 residues (illustrated in green in **Figures 9D–F**). The percentages of active site residues are significantly different, which is plotted for all three models (**Supplementary Figure 5**). Ser70 from the TEM-1 active site has significantly high importance in the reactant/product model. Ser70 in the other two models, and all other active site residues, only display importance lower than 3%. These are in the agreement that the TEM-1 active site is generally rigid for the purpose of catalysis.

We further investigate the distribution of residues importance with reference to the active site. The importance of residues lying within a certain distance range (i.e., between 4 and 5 Å) from the active site residues are accumulated and

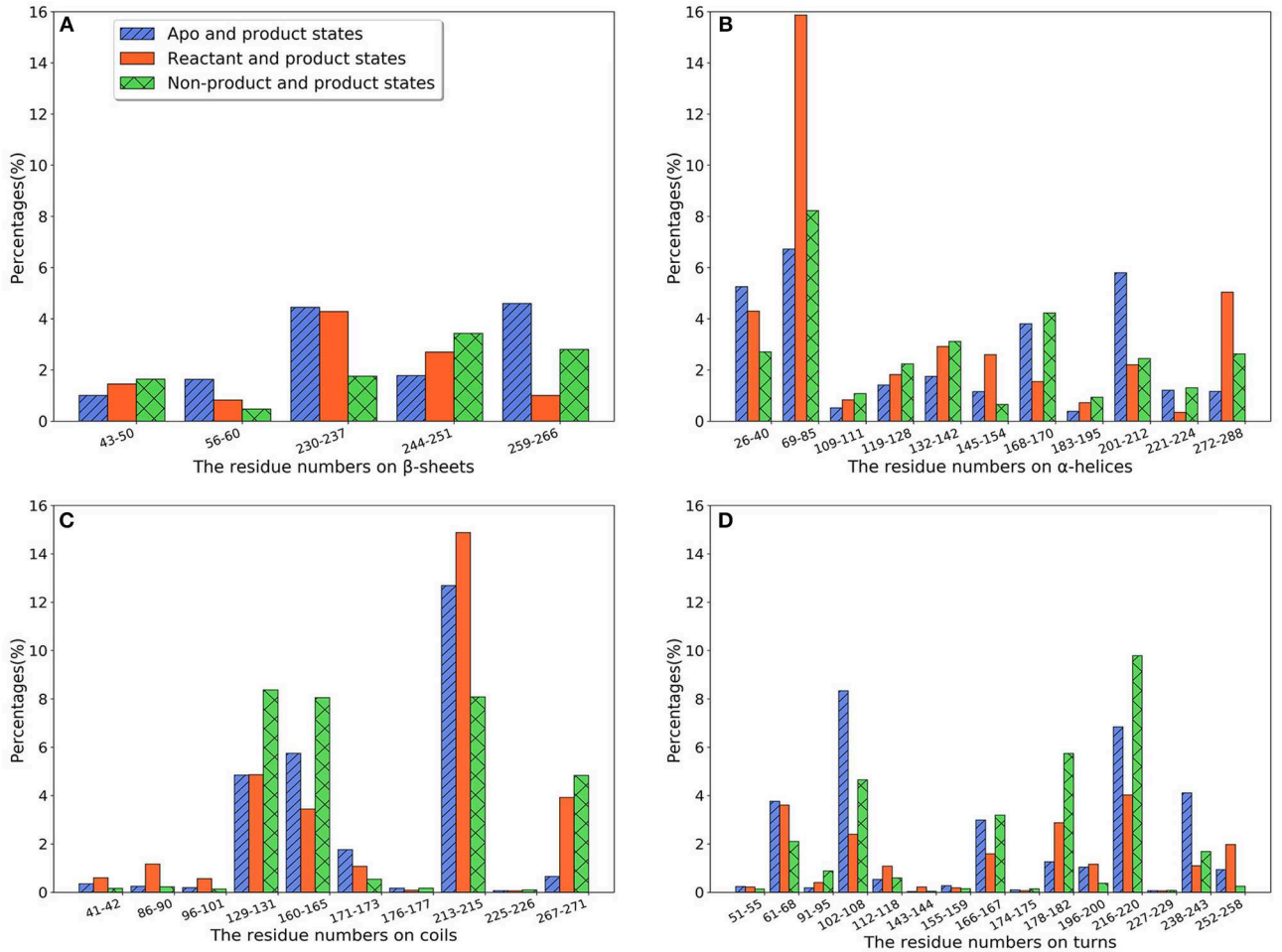


FIGURE 6 | The total feature importance of individual secondary structure in TEM-1: **(A)** β-sheets; **(B)** α-helix; **(C)** Random coils; **(D)** Turns. Each secondary structure is labeled by residue number. There are five β-sheets and 11 α-helices with varying sizes in TEM-1 structure.

normalized by the number of residues within a distance range, which is shown in **Figure 10A**. There are clearly three peaks of importance for the shells around 4, 7, and 10 Å away from the active site. The sums of importance of residues away from the active site region in the three models are plotted in **Figure 10B**. The accumulative importance of residues surrounding the active site is smoothly increasing along the distance.

The Conformational Analysis

In three states classifications, the key residues are identified based on the feature importance obtained from the classification models. However, the conformational changes in three states are very important for detecting the catalytic mechanism of TEM-1 bound with penicillin G complex. Therefore, further conformational analysis is carried out based on the selected key residues with top feature importance. Among the top 10 residues based on their accumulative feature importance, Tyr105 as a gatekeeper of the active site could stabilize the ligand binding (Doucet and Pelletier, 2007; Doucet et al., 2007). However, the interaction between Asn132 and Tyr105 may perturb the stabilizations (Wang et al., 2002). And a mutant of Asn105 has been proposed to create disruptive steric clashes with Asn132 and destabilize the ligand binding (Doucet and Pelletier, 2007). Asn132 is also a special residue, which was proposed to provide additional space for active site (Swarén et al., 1998). Therefore, the distance between Cα atoms of Tyr105 and Asn132 was selected for further analysis to reveal detailed conformational change relevant to functional states. In addition, the interaction between Lys73 and Asn132 was reported as important residues for TEM-1's catalytic function (Swarén et al., 1998). Accordingly, the Cα atoms distance between Lys73 and Asn132 is subjected to further analysis in this study. Two residues Gln39 and Thr269 among the top 10 residues are distal from the active site. Thr269 is really close to the allosteric site Helices 12 (Residue 272–288) identified in previous study (Horn and Shoichet, 2004). To reveal potential correlation between the active site and Gln39 as well as Thr269, the Cα atoms distance from Ser70 as the center of active site to these two residues are also subjected to further analysis.

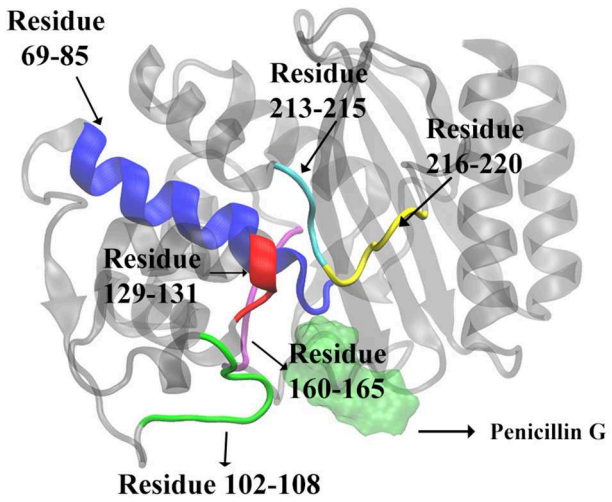


FIGURE 7 | The secondary structures of TEM-1 with significant total feature importance. Residues 69 through 85 as α-helix (blue), residues 213 through 215 as random coil (cyan); residues 216 through 220 as turn (yellow); Residues 129 through 131 as random coil (red), residues 102 through 108 as turn (green), residues 160 through 165 as random coil (purple). The ligand penicillin-G molecule is also illustrated as green transparent surface. The residue 160–165 is behind of the residue 69–85 in this view.

The density distributions of Cα atom distances of Tyr105-Asn132, Lys73-Asn132, Ser70-Gln39, Ser70-Thr269, and residue pairs for all three TEM-1 states are plotted in **Figure 11**. The Cα atom distance distribution of Tyr105-Asn132 has only one main peak close to 6 Å for reactant state (**Figure 11A**). However, the conversion from reactant to product leads to a second peak between 8 and 9 Å. Interestingly, the apo state without a ligand shows a similar distance distribution to the product state of this pair with two peaks between 6–7 Å and 8–9 Å. The density distribution of Lys73-Asn132 Cα atom distance has two peaks in the reactant state, one close to 9 Å and one between 10 and 11 Å (**Figure 11B**). The conversion to the product leads to only one peak around 9.2 Å of this distribution. In apo state, this distribution has a peak around 9.3 Å and a small shoulder about 10.3 Å.

For Ser70-Gln39 pair, the distributions of their Cα atom distance in all three states have only one peak (**Figure 11C**), which are located at 23.8, 24, and 24.5 Å for the apo, reactant and product states, respectively. Similarly, the density distributions of Ser70-Thr269 Cα atom distance also have only one peak for all three states, all between 19 and 20 Å (**Figure 11D**). These analyses demonstrated that the key residues with high feature importance do behave significantly in different functional states of protein. The residues Lys73, Asn132, Gln39, Ser70, and Thr269 are illustrated in the TEM-1 apo, reactant and product aligned structures with green transparent surface representing the ligand penicillin G binding pocket (**Figure 12**).

We further investigated four groups including Ω loop (residues 163–180), residues 213–220 including a turn and random coil structure and residues 102–108 as a turn structure, which are related to structures with high importance percentages illustrated in **Figure 7**. The helix 12 (residues 272–288) with high importance (>5%) in reactant/product model is also included. To reveal a potentially significant conformational change of these groups, the RMSD of these groups with the TEM-1 (1fqg) crystal structure as a reference are calculated and plotted in **Figure 13**. In TEM-1 bound with inhibitors, helix 11 (residues 219–226) and helix 12 (residues 272–288) were identified as an allosteric site (Horn and Shoichet, 2004). In the classification models generated in this study, helix 11 has a low feature importance and residues 213–220 have high importance. The RMSD distributions of residues 213–220 and helix 12 as potential allosteric sites are plotted in **Figures 13B,C**. The RMSD of residues 102–108 as a turn structure containing key residue Tyr105 is plotted in **Figure 13D**. The positions of the four residues group in TEM-1 are also illustrated in **Figure 12**. Interestingly, although Ω loop has high importance percentage, the RMSD distributions of Ω loop in three states are similar with each other displaying one main peak around 0.7 Å (**Figure 13A**). It indicated that Ω loop is not very flexible, agreeing with some NMR studies (Roccatano et al., 2005; Bös and Pleiss, 2009; Fisette et al., 2010). On the contrary, the RMSD distributions of 213–220 turn are significantly different among three states. In the reactant state, there are two main peaks around 1.2 and 2 Å and one small peak around 2.5 Å. In the product state, the RMSD distribution shift toward lower values with three peaks around 0.8, 1.3, and 2.5 Å. In the apo state, there is a dominant peak around 1.3 Å with a smaller peak around 2.6 Å. This clearly revealed significant conformational changes of this turn structure. The RMSD densities of helix 12 (residues 272–288) are similar in all three states with only one peak around 0.4 Å (**Figure 13C**), suggesting little conformational change of this secondary structure. The RMSD densities of residues 102–108 turn have one dominant distribution in three states (**Figure 13D**). The reactant and product states have the peak smaller than 0.4 Å. The apo state has the peak larger than 0.4 Å. These analyses demonstrate that the conformational change may play important role only in a limited local structure to differentiate functional states.

DISCUSSIONS

The role of protein dynamics in catalysis is becoming essential in understanding enzyme's catalytic mechanisms. TEM-1 is one of the proteins that has been interrogated for the correlation between dynamics and functions both experimentally and computationally (Farmer et al., 2017; Modi and Ozkan, 2018). In a detailed study of TEM-1 using NMR, the backbone motion of several TEM-1 mutants at Tyr105 was characterized and linked to its enzymatic function, because the residue in TEM-1 plays a key role in substrate differentiation and stabilization (Doucet et al., 2007). Coincidently, Tyr105 is identified as the second most important residue to differentiate the apo and product states in the current study (**Figure 9A**). The NMR study of TEM-1 also revealed that the mutations at residue 105 led to the change

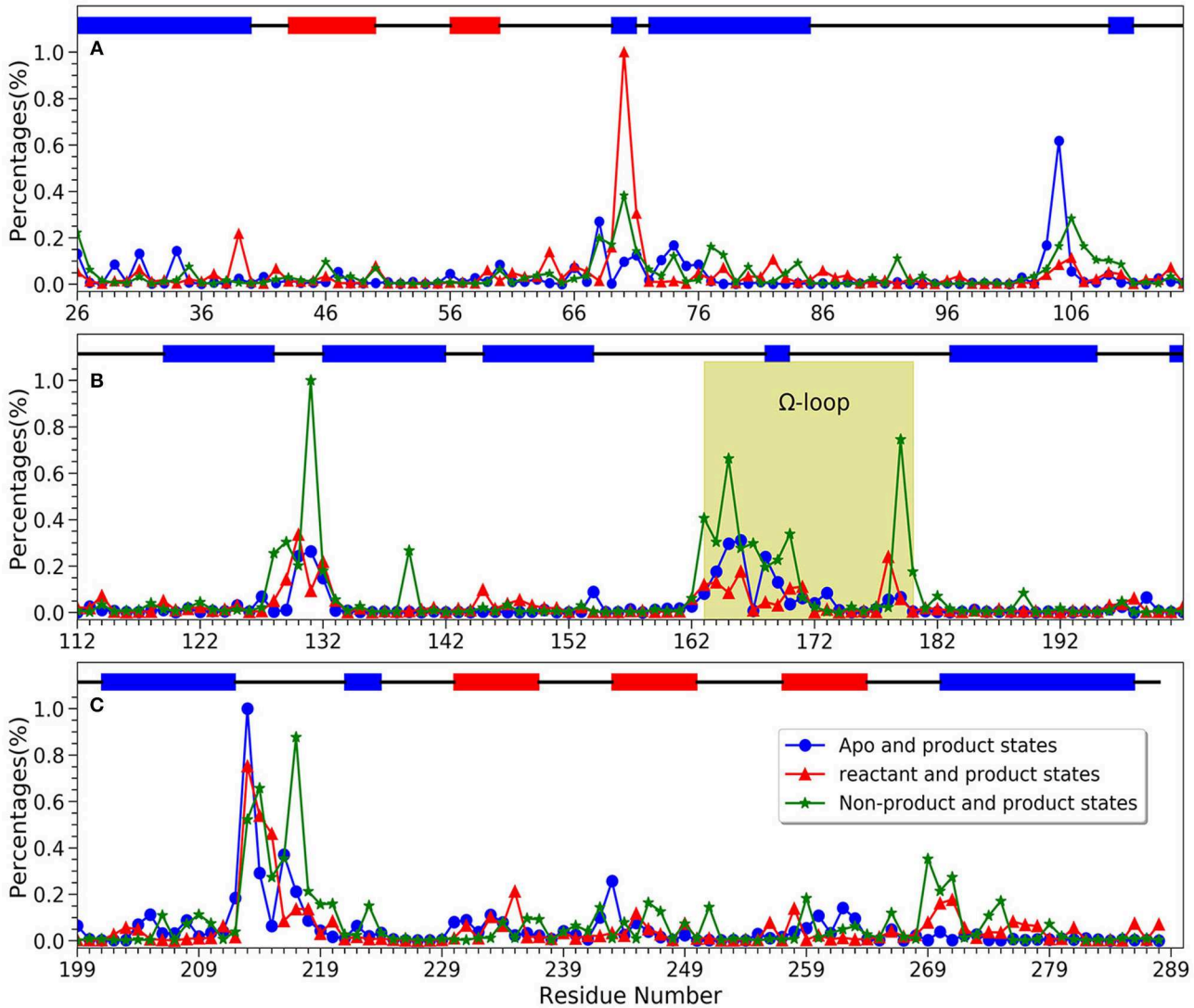


FIGURE 8 | The accumulative feature importance of each residue in TEM-1. The blue circles represent the apo and product states classification model, the red triangles represent the reactant and product states classification model, and the green stars represent the non-product and product states classification model. On the top of each sub-figure, the β-sheets and α-helices are labeled as red and blue rectangles, respectively. Ω-loop is highlighted in yellow. **(A)** Residues 26 through 115; **(B)** Residues 112 through 202; **(C)** Residues 199 through 288.

of backbone motion exceeding the TEM-1 active site and with a wide range of motion time scales. Interestingly, many key residues discovered in this study to be important for TEM-1 dynamical functional states are in a good agreement with the comprehensive NMR study.

The comparison among NMR spectroscopy of TEM-1 mutants showed that the most significant effect on backbone amide motion, marked as chemical shift differences, occur in the residues in 68–80, 100–115, 120–140, 163–170, 213–218, and 235–246 regions (Doucet et al., 2007). All these regions have significant feature importance from all classification modes developed in the current study (**Figure 8**). In general, the chemical shift differences observed in NMR spectroscopy have no direct connections with protein dynamics. But the backbone amide chemical shift is sensitive to the hydrogen bonding interactions in protein (Paramasivam et al., 2018). In another study, it was proposed that TEM-1 with mutant Tyr105 displayed effects on the backbone amide chemical shift of wild-type TEM-1 and can reduce the catalytic efficiency of TEM-1 binding with benzyl penicillin complex (Doucet et al., 2004). Although the backbone amide chemical shift difference is caused by the Tyr105 mutation of TEM-1 in the reference, there is indeed a relationship between the chemical shift difference and the catalytic efficiency for TEM-1 with benzyl penicillin complex. Therefore, the correlation between feature importance of key residues with the backbone amide chemical shift differences may help us to further understand the meaning of the machine learning based classification model. It is possible that the

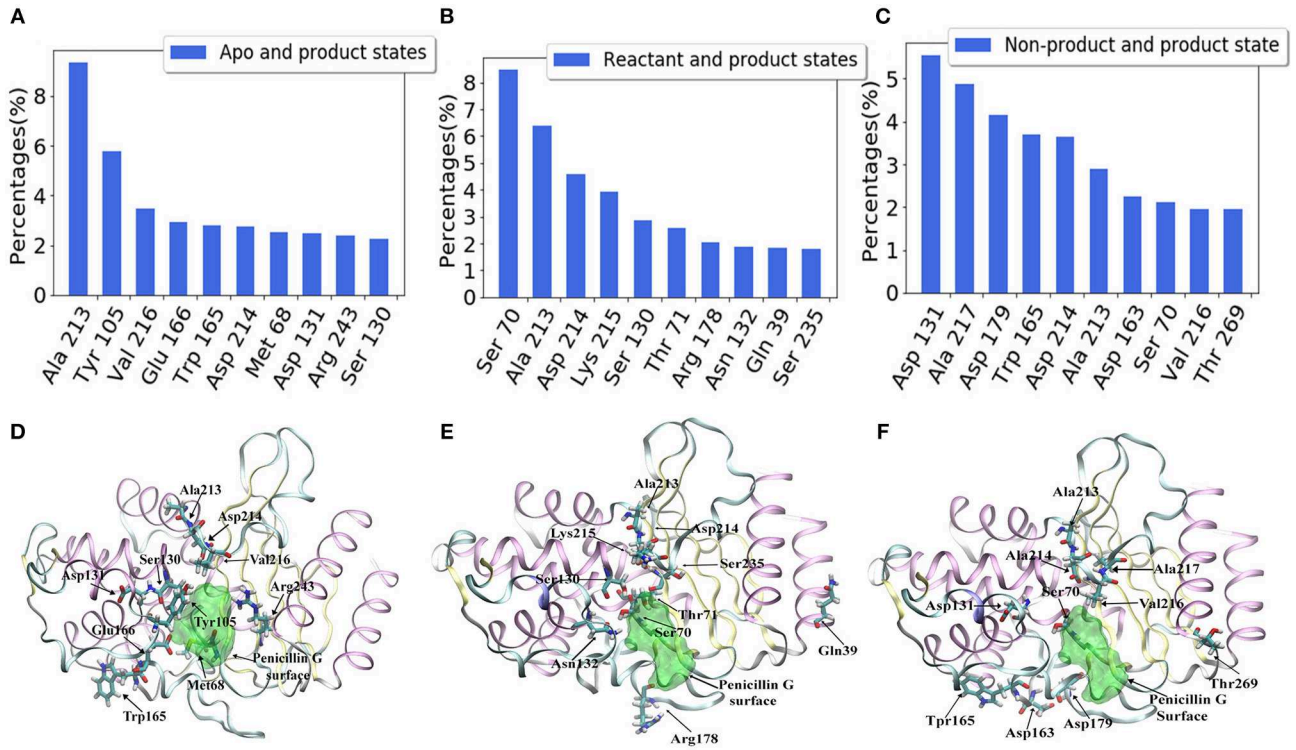


FIGURE 9 | Top 10 residues with the highest feature importance in TEM-1 β-lactamase based on classification models: **(A)** Apo/product model, **(B)** Reactant/product model, **(C)** Non-product/product model; The structures of top 10 residues with the highest feature importance and their positions comparing with the active site of TEM-1 in each model: **(D)** apo/product model, **(E)** reactant/product model, and **(F)** non-product/product model. The ligand penicillin-G molecule is also illustrated as green transparent surface. The active site of TEM-1 is the pocket holding the penicillin-G molecule.

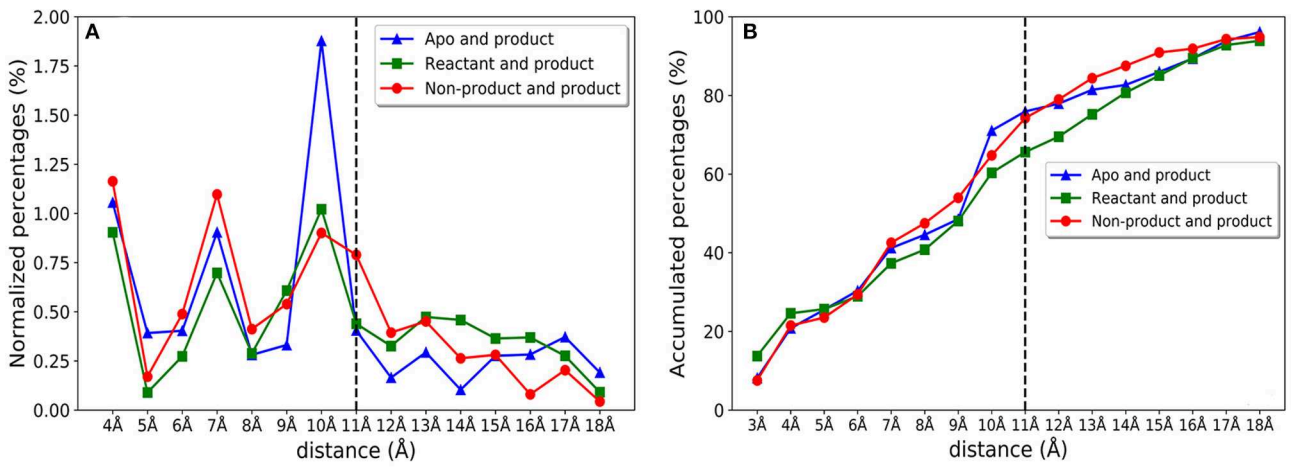


FIGURE 10 | The feature importance of residues with reference to the distance from active site for apo/product, reactant/product, and non-product/product classification models. **(A)** The normalized feature importance of residues within certain distance from active site (using 1 Å window). For example, the importance percentage 0.25% in 5 Å represents the importance percentage of all residues located in distance range 4–5 Å away from the active site; **(B)** Accumulative feature importance of residues with a certain distance from the active site. For example, the importance percentage of 5 Å represents the importance percentage including all the residues within 5 Å away from the active site.

backbone amide motion indicated by the NMR spectroscopy is well-coupled with the backbone Cα motion, which is used to construct features for the machine learning training models in this study. Further comparison also shows remarkable agreement at the individual residue level. Some conserved residues and residues at the so-called active site wall showed significant NMR

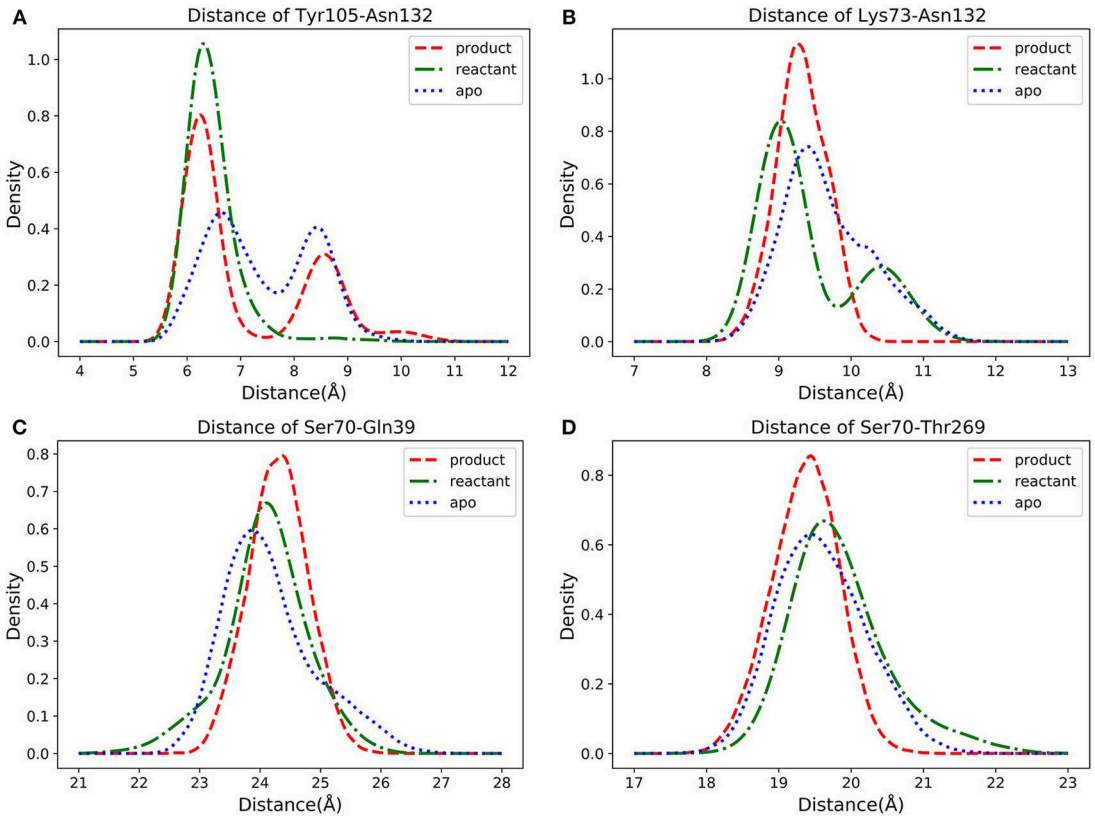


FIGURE 11 | The density distributions of pairwise α carbon atoms distance in apo (blue dot line), reactant (green dot dashed line) and product (red dashed line) states: **(A)** Tyr105 and Asn132, **(B)** Lys73 and Asn132, **(C)** Ser70 and Gln39, **(D)** Ser70 and Thr269.

TABLE 1 | The key residues from current study and a NMR study.

Residues with high feature importance[a]	Adjacent key NMR residues[b]
Met68, Ser70	Thr71
Ser130, Asp131	Met129 Asn132
Asp163, Trp165, Glu166	Arg164, Glu168
Arg178, Asp179	Thr181
Ala213, Asp214, Ala217	Lys215, Val216, Gly218
Ser235	Lys234
Thr269	Met270

[a] *Current study*
[b] *Table 4 in a NMR study (Doucet et al., 2007).*

relaxation parameter changes between the wild type and the most significantly different Y105D mutant (Doucet et al., 2007). Six residues (Asn132, Tyr105, Lys215, Val216, Thr71, and Arg243) among the 21 residues with the highest important features from the classification modes in this study (**Figure 9**) are among the key residues for the local dynamic effects identified in the NMR study. Many more residues (a total of 14) selected by the feature importance are also in the adjacent region within the key residues selected in the NMR study (**Table 1**).

Comparison between the NMR spectroscopy between wild type and Y105D mutant also revealed that significant local differences in the regions of residues 70–80, 124–135, and most importantly in 211–221. Our analysis shows that these regions display high accumulative feature importance as various secondary structures, such as residues 70–80 belonging to α-helix, residues 124–135 spreading across random coil and α-helix, and residues 211–221 containing both random coil and turn structures (**Figure 6**).

Ω loop (163–180) is a key secondary structure close to the ligand binding site in TEM-1 and important for its catalytic function. A previous NMR and MD simulation work showed that Ω loop displayed limited flexibility with the key translational component (Bös and Pleiss, 2009) It was proposed that the Ω loop is a key structural feature for substrate binding and recognition (Fisette et al., 2012). It was observed in the same study that the inter-Ω loop salt bridge between Arg164 and Asp179 is prone to be affected by the substrate binding, while the Arg164-Thr71 interaction is stabilized by the ligand binding. Accordingly, the Ω loop shows significant and various importance in our three classification models, with the most significance in the non-product/product model. Residues Asp163, Arg164, Trp165, and Asp179 are very important residues (>3% in **Figure 8B** Ω loop green highlighted part) for the non-product/product differentiation model. Residues Trp165, Glu166, and Glu168 are important residues (>2% in **Figure 8B** Ω loop green highlighted

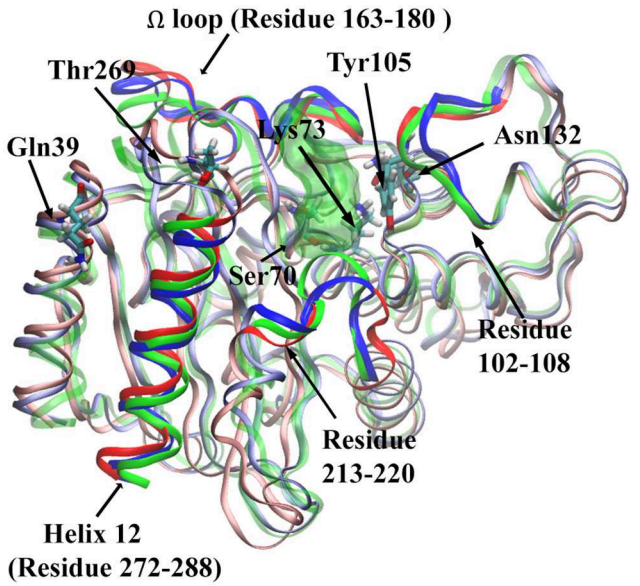


FIGURE 12 | The structure of TEM-1 in apo (blue), reactant (green) and product (red) states. Ω loop (residues 163–180), helix 12 (residues 272–288), residues 102–108, and residues 213–220 are highlighted in each state with same colors. Also key residues Gln39, Thr269, Ser70, Lys73, Tyr105, and Asn132 are labeled. The ligand penicillin-G molecule is represented as green transparent surface.

part) for the apo/product differentiation model. In comparison, the Ω-loop is somewhat less important in the reactant/product model than in the other two models, indicating the importance of differentiating the product from other states. In the non-product/product model, both Arg164 (close to 0.3% percentages of importance) and Asp179 (close to 0.8% percentages of importance) are emphasized as important residues. The Asp179 and Arg164 locate at the entrance of the active site and form the inter-Ω loop salt-bridge to stabilize the loop. In reactant/product and apo/product models, the importance of Arg164 and Asp179 are not obvious, the combination of apo and reactant magnify their importance in non-product/product model. We hypothesized that the interaction between Arg164 and Asp179 exist in all three states to stabilize the loop. Both hydrolyzed benzyl penicillin and benzyl penicillin molecules as substrates can strengthen the interaction. That may be the reason why the overall Ω loop does not carry high importance percentage in reactant/product model. The overall Ω loop is more stable in reactant and product states than in the apo state. In addition, Trp165 is highlighted in both non-product/product and apo/product models, which indicates that Tpr165 is a key residue to classify the apo/reactant and product states. Therefore, it is likely that Tpr165 plays an important role in de-acylation step of the catalytic mechanism, which is also mentioned in experimental study (Petrosino et al., 1998). Another key residue for acylation, Glu166, has a relative high importance in apo/product model. We proposed that Glu166 is not only as a general base in acylation (Minasov et al., 2002) but also very important in the de-acylation step. These detailed comparisons with experimental study provided further insight into the functions of the Ω loop of ligand binding in addition to enzyme catalysis.

The NMR study suggested the key Ω loop motion was in the microsecond (μs)-millisecond (ms) time scale, which was beyond the current simulation study. However, it was also pointed out that the Ω loop dynamics is more focused and less random than other secondary structures even at a large time scale. The good agreement and complimentary comparison between the classification models developed in this study and previous NMR studies of TEM-1 suggests the effectiveness of the machine learning method in the application of protein dynamics and functional analysis. The usage of Cα distance as training features from extensive MD simulations for training practically bridges among protein dynamics with inter-residue correlation, regardless the distance region within the framework of different functional states.

Asn132 was identified as a residue controlling the size of the TEM-1 active site cavity. Distance distribution analysis of Lys73 and Asn132 reveals that the binding with reactant effectively compresses the active site into a closed active site and creates a minor open state representing by two peaks of Cα distance distribution in reactant state (**Figure 11B**). However, the product binding state only has one main peak as a closed active site without a minor open state. This could be a key dynamical difference between reactant and product binding states. The interaction between Tyr105 and Asn132 also related to the active site. Opposite to the Lys73 and Asn132 Cα distance distribution, the Cα distance distribution of Tyr105 and Asn132 changes from single dominant peak in reactant state to double peaks in the product state (**Figure 11A**). The difference of the apo state distribution from both reactant and product states also sheds light on these TEM-1 functional states. Helix 11 (residues 219–226) and 12 (residues 272–288) were proposed as an allosteric site with 3–7 Å shift in helix 11 and 1–3 Å shift in helix 12 comparing to the apo structure (Horn and Shoichet, 2004; Avci et al., 2018). The significant conformational change of residues 213–220 as a turn and random coil structure adjacent to helix 11 could be coupled with the allosteric function residing in this region. The similarity of the helix 12 RMSD distributions shared by all three states warrants further study to elucidate the allosteric mechanism associated with this local structure (**Figures 13B,C**).

It could be a concern that the initial structures for apo, reactant and product state, generated from the same crystal structure (1FQG) in catalytic intermediate state, may not present three target states well. To address this concern, we collected a total of eight crystal structures of wild type TEM-1 in apo states and five crystal structures of wild type TEM-1 binding with various ligands from PDB, including the one with penicillin used as starting structure in this study, as reference structures for the simulations. The averaged RMSDs of each functional state simulations with reference to these crystal structures were calculated and plotted in **Supplementary Figure 6**. It is interesting that the product state simulations consistently have lower RMSDs with reference to all 13 crystal structures, including both apo and holo states of TEM-1 and the structure used in this

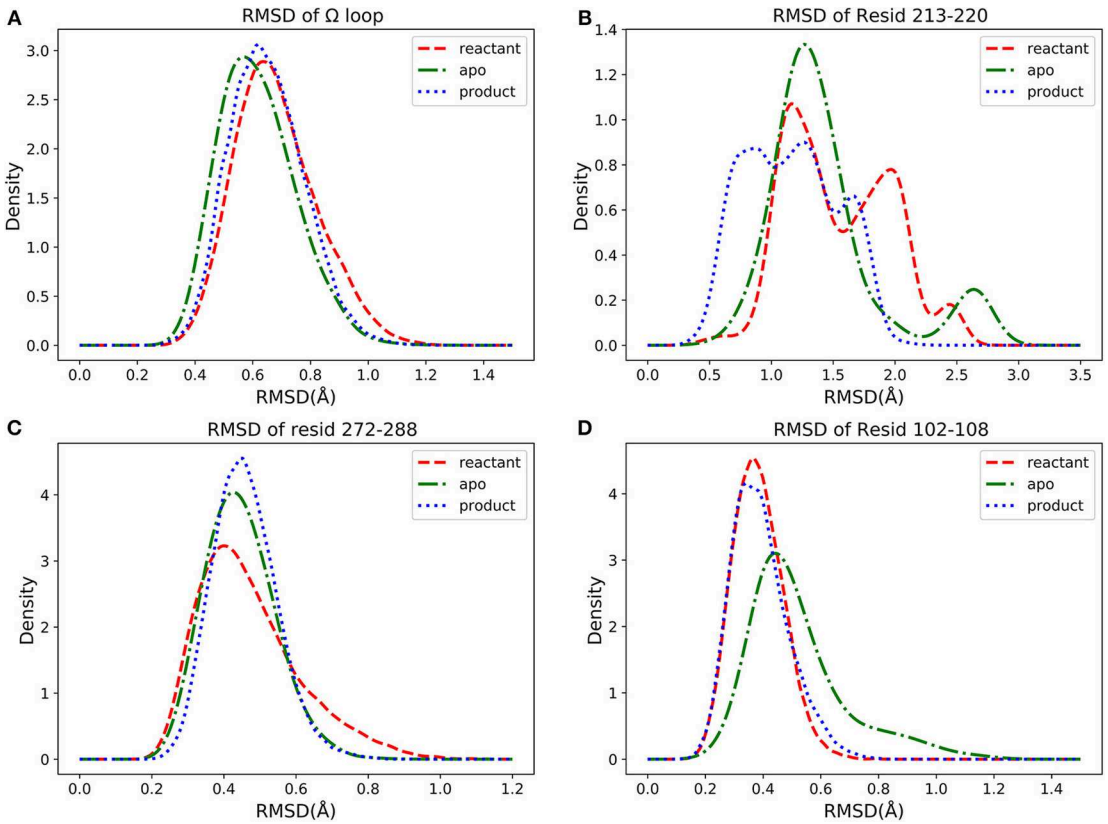


FIGURE 13 | The density distributions of residues groups' RMSDs in apo (green dot dashed line), reactant (red dashed line) and product (blue dot line) states: **(A)** RMSD of Ω loop (residues 163–180), **(B)** RMSD of residues 213–220, **(C)** RMSD of helix 12 (residues 272–288), **(D)** RMSD of residues 102–108.

study, than both apo and reactant state simulations. In addition, both apo and reactant states simulations consistently have similar RMSDs with reference to these TEM-1 crystal structures. Although these results could prove either the simulations are sufficient for the sampling of each state or not, these results are consistent with our results that the apo and reactant states are similar to each, and both are different from the product state. It may suggest that binding with the catalysis product is a dynamically stable state for TEM-1 and contributes to the catalytic activities of TEM-1 against antibiotics. This could lead to some intrinsic dynamical properties of TEM-1 in different functional states, which warrant further in-depth studies.

CONCLUSION

In this study, we developed classification models for TEM-1 β-lactamases in different binding modes against penicillin using a machine learning method called random forest. Using the backbone Cα distances of all residue pairs as the features for the model training purpose, the developed classification models effectively correlate the global protein dynamics with the individual residue correlation, with regard to the different binding modes. The feature importance generated from the classification model training process was used to evaluate the contribution from individual residues, as well as secondary structures in TEM-1, to each model. It is shown that the random coil structures carry the highest feature importance among secondary structures, including α-helix, β-strands, and turns. It may indicate that the motions of coils contribute significantly to the differences among three states, and lead to more flexibility of random coils than in other secondary structures. Accordingly, the protein flexibility is proposed to be a key factor in ligand recognition of TEM-1. Detailed comparison also revealed that the individual key residues identified from the machine learning models not only have a good agreement with the NMR study, but also provide unprecedented insight into the function of individual residues with regard to differentiating protein in different binding modes. Specifically, it is suggested that some catalytically important residues at the active site are also critical for recognizing the hydrolyzed product of antibiotics. Overall, this study demonstrates that machine learning methods provides effective tools to analyze protein dynamics in different binding modes and produce intriguing insight into the correlation between protein functional states and various structural levels.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

FW wrote the manuscript and carried out the four independent MD simulations for three states (1,200 ns) and performed all the analysis. LS carried out 1 MD simulation for three states (300 ns). HZ provided some scripts of machine learning. SW and XW authors contributed to the final version of the manuscript. PT contributed to the final version of the manuscript and supervised the project.

FUNDING

The work was supported by National Science Foundation under a CAREER Grant [1753167] and SMU Dean's Research Council research grant. Computational time was provided by Southern Methodist University's Center for Scientific Computation.

ACKNOWLEDGMENTS

Computational time was provided by Southern Methodist University's Center for Scientific Computation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00047/full#supplementary-material

REFERENCES

Ambler, R. P., Coulson, A. F., Frère, J. M., Ghuysen, J. M., Joris, B., Forsman, M., et al. (1991). A standard numbering scheme for the class A beta-lactamases. *Biochem. J.* 276(Pt 1), 269–270. doi: 10.1042/bj2760269

Avci, F. G., Altinisik, F. E., Karacan, I., Senturk Karagoz, D., Ersahin, S., Eren, A., et al. (2018). Targeting a hidden site on class A beta-lactamases. *J. Mol. Graph. Model.* 84, 125–133. doi: 10.1016/j.jmgm.2018.06.007

Avci, F. G., Altinisik, F. E., Vardar Ulu, D., Ozkirimli Olmez, E., and Sariyar Akbulut, B. (2016). An evolutionarily conserved allosteric site modulates beta-lactamase activity. *J. Enzyme Inhib. Med. Chem.* 31, 33–40. doi: 10.1080/14756366.2016.1201813

Ballester, P. J., and Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26, 1169–1175. doi: 10.1093/bioinformatics/btq112

Bashtannyk, D. M., and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Stat. Data Anal.* 36, 279–298. doi: 10.1016/S0167-9473(00)00046-3

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.* 8, 3257–3273. doi: 10.1021/ct300400x

Bös, F., and Pleiss, J. (2009). Multiple molecular dynamics simulations of TEM beta-lactamase: dynamics and water binding of the omega-loop. *Biophys. J.* 97, 2550–2558. doi: 10.1016/j.bpj.2009.08.031

Bradford, P. A. (2001). Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. *Clin. Microbiol. Rev.* 14, 933–951. doi: 10.1128/CMR.14.4.933-951.2001

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint.* arXiv:1309.0238.

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* 26, 5–14. doi: 10.1016/S0097-8485(01)00094-8

Cortina, G. A., and Kasson, P. M. (2018). Predicting allostery and microbial drug resistance with molecular simulations. *Curr. Opin. Struct. Biol.* 52, 80–86. doi: 10.1016/j.sbi.2018.09.001

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., and Cavalli, A. (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun.* 6:6155. doi: 10.1038/ncomms7155

Díaz, N., Sordo, T. L., Merz, K. M., and Suárez, D. (2003). Insights into the acylation mechanism of class A β-lactamases from molecular dynamics simulations of the TEM-1 enzyme complexed with benzylpenicillin. *J. Am. Chem. Soc.* 125, 672–684. doi: 10.1021/ja027704o

Díaz, N., Suárez, D., Sordo, T. L., and Merz, K. M. (2001). Acylation of class A β-lactamases by penicillins: a theoretical examination of the role of serine 130 and the β-lactam carboxylate group. *J. Phys. Chem. B* 105, 11302–11313. doi: 10.1021/jp012881h

Dideberg, O., Charlier, P., Wéry, J. P., Dehottay, P., Dusart, J., Erpicum, T., et al. (1987). The crystal structure of the beta-lactamase of *Streptomyces albus* G at 0.3 nm resolution. *Biochem. J.* 245, 911–913. doi: 10.1042/bj2450911

Doucet, N., De Wals, P.-Y., and Pelletier, J. N. (2004). Site-saturation mutagenesis of Tyr-105 reveals its importance in substrate stabilization and discrimination in TEM-1 β-lactamase. *J. Biol. Chem.* 279, 46295–46303. doi: 10.1074/jbc.M407606200

Doucet, N., and Pelletier, J. N. (2007). Simulated annealing exploration of an active-site tyrosine in TEM-1 beta-lactamase suggests the existence of alternate conformations. *Proteins* 69, 340–348. doi: 10.1002/prot.21485

Doucet, N., Savard, P.-Y., Pelletier, J. N., and Gagné, S. M. (2007). NMR investigation of Tyr105 mutants in TEM-1 β-lactamase: dynamics are coorrelated with function. *J. Biol. Chem.* 282, 21448–21459. doi: 10.1074/jbc.M609777200

Eastman, P., and Pande, V. S. (2015). OpenMM: a hardware independent framework for molecular simulations. *Comput. Sci. Eng.* 12, 34–39. doi: 10.1109/MCSE.2010.27

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., et al. (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comp. Biol.* 13:e1005659. doi: 10.1371/journal.pcbi.1005659

Farmer, J., Kanwal, F., Nikulsin, N., Tsilimigras, M. C. B., and Jacobs, D. J. (2017). Statistical measures to quantify similarity between molecular dynamics simulation trajectories. *Entropy* 19:646. doi: 10.3390/e19120646

Fisette, O., Gagné, S., and Lagüe, P. (2012). Molecular dynamics of class A β-lactamases—effects of substrate binding. *Biophys. J.* 103, 1790–1801. doi: 10.1016/j.bpj.2012.09.009

Fisette, O., Morin, S., Savard, P.-Y., Lagüe, P., and Gagné, S. M. (2010). TEM-1 backbone dynamics—insights from combined molecular dynamics and nuclear magnetic resonance. *Biophys. J.* 98, 637–645. doi: 10.1016/j.bpj.2009.08.061

Fonzé, E., Charlier, P., To'Th, Y., Vermeire, M., Raquet, X., Dubus, A., et al. (1995). TEM1 β-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Crystallogr. D Biol. Crystallogr.* 51, 682–694. doi: 10.1107/S0907444994014496

Friedrichs, M. S., Eastman, P., Vaidyanathan, V., Houston, M., Legrand, S., Beberg, A. L., et al. (2009). Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* 30, 864–872. doi: 10.1002/jcc.21209

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Golemi-Kotra, D., Meroueh, S. O., Kim, C., Vakulenko, S. B., Bulychev, A., Stemmler, A. J., et al. (2004). The importance of a critical protonation state

and the fate of the catalytic steps in class A β-lactamases and penicillin-binding proteins. *J. Biol. Chem.* 279, 34665–34673. doi: 10.1074/jbc.M313143200

Halgren, T. A. (1992). The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* 114, 7827–7843. doi: 10.1021/ja00046a032

Hermann, J. C., Ridder, L., Mulholland, A. J., and Höltje, H.-D. (2003). Identification of Glu166 as the general base in the acylation reaction of class A β-lactamases through QM/MM modeling. *J. Am. Chem. Soc.* 125, 9590–9591. doi: 10.1021/ja034434g

Herzberg, O., and Moult, J. (1987). Bacterial resistance to beta-lactam antibiotics: crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.5 A resolution. *Science* 236, 694–701. doi: 10.1126/science.3107125

Herzberg, O., and Moult, J. (1991). Penicillin-binding and degrading enzymes. *Curr. Opin. Struct. Biol.* 1, 946–953. doi: 10.1016/0959-440X(91)90090-G

Horn, J. R., and Shoichet, B. K. (2004). Allosteric inhibition through core disruption. *J. Mol. Biol.* 336, 1283–1291. doi: 10.1016/j.jmb.2003.12.068

Husic, B. E., and Pande, V. S. (2018). Markov state models: from an art to a science. *J. Am. Chem. Soc.* 140, 2386–2396. doi: 10.1021/jacs.7b12191

Jelsch, C., Lenfant, F., Masson, J. M., and Samama, J. P. (1992). β-lactamase TEM1 of *E. coli* crystal structure determination at 2.5 Å resolution. *FEBS Lett.* 299, 135–142. doi: 10.1016/0014-5793(92)80232-6

Jelsch, C., Mourey, L., Masson, J.-M., and Samama, J.-P. (1993). Crystal structure of *Escherichia coli* TEM1 β-lactamase at 1.8 Å resolution. *Proteins* 16, 364–383. doi: 10.1002/prot.340160406

Jolliffe, I. (2011). "Principal component analysis," in *International Encyclopedia of Statistical Science,* ed M. Lovric (Berlin: Springer), 1094–1096.

Kabsch, W., and Sander, C. (1983). DSSP: definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Lamotte-Brasseur, J., Dive, G., Dideberg, O., Charlier, P., Frère, J. M., and Ghuysen, J. M. (1991). Mechanism of acyl transfer by the class A serine β-lactamase of *Streptomyces albus* G. *Biochem. J.* 279, 213–221. doi: 10.1042/bj2790213

Lamotte-Brasseur, J., Jacob-Dubuisson, F., Dive, G., Frère, J. M., and Ghuysen, J. M. (1992). *Streptomyces albus* G serine beta-lactamase. Probing of the catalytic mechanism via molecular modelling of mutant enzymes. *Biochem. J.* 282(Pt 1), 189–195. doi: 10.1042/bj2820189

Lamotte-Brasseur, J., Knox, J., Kelly, J. A., Charlier, P., Fonzé, E., Dideberg, O., et al. (1994). The structures and catalytic mechanisms of active-site serine β-lactamases. *Biotechnol. Genet. Eng. Rev.* 12, 189–230. doi: 10.1080/02648725.1994.10647912

Lamotte-Brasseur, J., Lounnas, V., Raquet, X., and Wade, R. C. (1999). pKa calculations for class A β-lactamases: influence of substrate binding. *Protein Sci.* 8, 404–409. doi: 10.1110/ps.8.2.404

Li, Z., Kermode, J. R., and De Vita, A. (2015). Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* 114:096405. doi: 10.1103/PhysRevLett.114.096405

Louppe, G. (2014). Understanding random forests: from theory to practice. *arXiv preprint*. arXiv:1407.7502.

Marciano, D. C., Brown, N. G., and Palzkill, T. (2009). Analysis of the plasticity of location of the Arg244 positive charge within the active site of the TEM-1 β-lactamase. *Protein Sci.* 18, 2080–2089. doi: 10.1002/pro.220

Matagne, A., Lamotte-Brasseur, J., and FrÈRe, J.-M. (1998). Catalytic properties of class A β-lactamases: efficiency and diversity. *Biochem. J.* 330, 581–598. doi: 10.1042/bj3300581

Maveyraud, L., Pratt, R. F., and Samama, J.-P. (1998). Crystal structure of an acylation transition-state analog of the TEM-1 β-lactamase. Mechanistic implications for class A β-lactamases. *Biochemistry* 37, 2622–2628. doi: 10.1021/bi972501b

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109, 1528–1532. doi: 10.1016/j.bpj.2015.08.015

Meneksedag, D., Dogan, A., Kanlikilicer, P., and Ozkirimli, E. (2013). Communication between the active site and the allosteric site in class A beta-lactamases. *Comput. Biol. Chem.* 43, 1–10. doi: 10.1016/j.compbiolchem.2012.12.002

Meroueh, S. O., Fisher, J. F., Schlegel, H. B., and Mobashery, S. (2005). Ab Initio QM/MM study of class A β-lactamase acylation: dual participation of Glu166 and Lys73 in a concerted base promotion of Ser70. *J. Am. Chem. Soc.* 127, 15397–15407. doi: 10.1021/ja051592u

Minasov, G., Wang, X., and Shoichet, B. K. (2002). An ultrahigh resolution structure of TEM-1 β-lactamase suggests a role for Glu166 as the general base in acylation. *J. Am. Chem. Soc.* 124, 5333–5340. doi: 10.1021/ja0259640

Modi, T., and Ozkan, B. S. (2018). Mutations utilize dynamic allostery to confer resistance in TEM-1 β-lactamase. *Int. J. Mol. Sci.* 19:E3808. doi: 10.3390/ijms19123808

Moews, P. C., Knox, J. R., Dideberg, O., Charlier, P., and Frère, J.-M. (1990). β-lactamase of *Bacillus licheniformis* 749/C at 2 Å resolution. *Proteins* 7, 156–171. doi: 10.1002/prot.340070205

Oefner, C., D'Arcy, A., Daly, J. J., Gubernator, K., Charnas, R. L., Heinze, I., et al. (1990). Refined crystal structure of β-lactamase from *Citrobacter freundiiindicates* a mechanism for β-lactam hydrolysis. *Nature* 343, 284–288. doi: 10.1038/343284a0

Oleinikovas, V., Saladino, G., Cossins, B. P., and Gervasio, F. L. (2016). Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc.* 138, 14257–14263. doi: 10.1021/jacs.6b05425

Palzkill, T. (2018). Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of TEM, CTX-M, and KPC β-lactamases. *Front. Mol. Biosci.* 5:16. doi: 10.3389/fmolb.2018.00016

Paramasivam, S., Gronenborn, A. M., and Polenova, T. (2018). Backbone amide 15N chemical shift tensors report on hydrogen bonding interactions in proteins: a magic angle spinning NMR study. *Solid State Nucl. Magn. Reson.* 92, 1–6. doi: 10.1016/j.ssnmr.2018.03.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Petrosino, J., Cantu, C., and Palzkill, T. (1998). β-lactamases: protein evolution in real time. *Trends Microbiol.* 6, 323–327. doi: 10.1016/S0966-842X(98)01317-1

Pimenta, A. C., Martins, J. M., Fernandes, R., and Moreira, I. S. (2013). Ligand-induced structural changes in TEM-1 probed by molecular dynamics and relative binding free energy calculations. *J. Chem. Inf. Model.* 53, 2648–2658. doi: 10.1021/ci400269d

Roccatano, D., Sbardella, G., Aschi, M., Amicosante, G., Bossa, C., Nola, A. D., et al. (2005). Dynamical aspects of TEM-1 β-lactamase probed by molecular dynamics. *J. Comput. Aided Mol. Des.* 19, 329–340. doi: 10.1007/s10822-005-7003-0

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi: 10.1016/0021-9991(77)90098-5

Savard, P.-Y., and Gagné, S. M. (2006). Backbone dynamics of TEM-1 determined by NMR: evidence for a highly ordered protein. *Biochemistry* 45, 11414–11424. doi: 10.1021/bi060414q

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York, NY: John Wiley & Sons.

Sgrignani, J., Grazioso, G., De Amici, M., and Colombo, G. (2014). Inactivation of TEM-1 by Avibactam (NXL-104): insights from quantum mechanics/molecular mechanics metadynamics simulations. *Biochemistry* 53, 5174–5185. doi: 10.1021/bi500589x

Shcherbinin, D., and Veselovsky, A. (2019). "Analysis of protein structures using residue interaction networks," in *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process,* ed C. G. Mohan (Cham: Springer International Publishing), 55–69.

Silverman, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. New York, NY: Routledge.

Simm, A. M., Baldwin, A. J., Busse, K., and Jones, D. D. (2007). Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. *FEBS Lett.* 581, 3904–3908. doi: 10.1016/j.febslet.2007.07.018

Stec, B., Holtz, K. M., Wojciechowski, C. L., and Kantrowitz, E. R. (2005). Structure of the wild-type TEM-1 β-lactamase at 1.55 Å and the mutant enzyme Ser70Ala at 2.1 Å suggest the mode of noncovalent catalysis for the mutant enzyme. *Acta Crystallogr. D Biol. Crystallogr.* 61, 1072–1079. doi: 10.1107/S0907444905014356

Stojanoski, V., Chow, D.-C., Hu, L., Sankaran, B., Gilbert, H. F., Prasad, B. V. V., et al. (2015). A triple mutant in the Ω-loop of TEM-1 β-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *J. Biol. Chem.* 290, 10382–10394. doi: 10.1074/jbc.M114.633438

Strynadka, N. C. J., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C., et al. (1992). Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature* 359, 700–705. doi: 10.1038/359700a0

Strynadka, N. C. J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., et al. (1996). Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nat. Struct. Biol.* 3, 233–239. doi: 10.1038/nsb0396-233

Swarén, P., Maveyraud, L., Raquet, X., Cabantous, S., Duez, C., Pédelacq, J.-D., et al. (1998). X-ray analysis of the NMC-A β-lactamase at 1.64-Å resolution, a class A carbapenemase with broad substrate specificity. *J. Biol. Chem.* 273, 26714–26721. doi: 10.1074/jbc.273.41.26714

Turlach, B. A. (1993). *Bandwidth Selection in Kernel Density Estimation: A Review*. Berlin: Citeseer.

Vanwetswinkel, S., Avalle, B., and Fastrez, J. (2000). Selection of β-lactamases and penicillin binding mutants from a library of phage displayed TEM-1 β-lactamase randomly mutated in the active site Ω-loop11Edited by A. *R. Fersht. J. Mol. Biol.* 295, 527–540. doi: 10.1006/jmbi.1999.3376

Wang, X., Minasov, G., and Shoichet, B. K. (2002). Noncovalent interaction energies in covalent complexes: TEM-1 β-lactamase and β-lactams. *Proteins* 47, 86–96. doi: 10.1002/prot.10058

Zafaralla, G., Manavathu, E. K., Lerner, S. A., and Mobashery, S. (1992). Elucidation of the role of arginine-224 in the turnover processes of class A beta-lactamases. *Biochemistry* 31, 3847–3852. doi: 10.1021/bi00130a016

Zhou, H., Dong, Z., and Tao, P. (2018). Recognition of protein allosteric states and residues: machine learning approaches. *J. Comput. Chem.* 39, 1481–1490. doi: 10.1002/jcc.25218

Zhou, H., Dong, Z., Verkhivker, G., Zoltowski, B. D., and Tao, P. (2019). Allosteric mechanism of the circadian protein vivid resolved through markov state model and machine learning analysis. *PLoS Comp. Biol.* 15:e1006801. doi: 10.1371/journal.pcbi.1006801

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 12 June 2019
doi: 10.3389/fmolb.2019.00042

Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods

Gianmarc Grazioli[1,2], *Rachel W. Martin*[2,3] *and Carter T. Butts*[1,4,5]*

[1] California Institute for Telecommunications and Information Technology (Calit2), University of California, Irvine, Irvine, CA, United States, [2] Department of Chemistry, University of California, Irvine, Irvine, CA, United States, [3] Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, United States, [4] Department of Computer Science, University of California, Irvine, Irvine, CA, United States, [5] Department of Sociology, Statistics, and Electrical Engineering and Computer Science, University of California, Irvine, Irvine, CA, United States

Simulations of intrinsically disordered proteins (IDPs) pose numerous challenges to comparative analysis, prominently including highly dynamic conformational states and a lack of well-defined secondary structure. Machine learning (ML) algorithms are especially effective at discriminating among high-dimensional inputs whose differences are extremely subtle, making them well suited to the study of IDPs. In this work, we apply various ML techniques, including support vector machines (SVM) and clustering, as well as related methods such as principal component analysis (PCA) and protein structure network (PSN) analysis, to the problem of uncovering differences between configurational data from molecular dynamics simulations of two variants of the same IDP. We examine molecular dynamics (MD) trajectories of wild-type amyloid beta ($A\beta_{1-40}$) and its "Arctic" variant (E22G), systems that play a central role in the etiology of Alzheimer's disease. Our analyses demonstrate ways in which ML and related approaches can be used to elucidate subtle differences between these proteins, including transient structure that is poorly captured by conventional metrics.

Keywords: machine learning, intrinsically disordered proteins, molecular dynamics, amyloid fibrils, amyloid beta, protein structure networks, support vector machines, clustering

OPEN ACCESS

Edited by:
Francesco Luigi Gervasio, University College London, United Kingdom

Reviewed by:
Elodie Laine, Université Pierre et Marie Curie, France
Ilpo Vattulainen, University of Helsinki, Finland

***Correspondence:**
Carter T. Butts
buttsc@uci.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *15 February 2019*
Accepted: *20 May 2019*
Published: *12 June 2019*

Citation:
Grazioli G, Martin RW and Butts CT (2019) Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. Front. Mol. Biosci. 6:42. doi: 10.3389/fmolb.2019.00042

1. INTRODUCTION

Molecular dynamics (MD) simulations, either alone or guided by experimental data, have greatly enhanced our ability to probe molecular motions at the atomic scale. Unfortunately, these advances can also lead to the creation of a map that is almost as complex as the territory it describes: as simulation methodology has improved, the need for approaches to analyze and make sense of increasingly information-rich simulated trajectories has grown. This is particularly true in the case of intrinsically disordered proteins (IDPs), where recent developments in the combined use of simulation methods with NMR (Dedmon et al., 2005; Salmon et al., 2010; Salvi et al., 2016) and small angle x-ray scattering data (Sibille and Bernadó, 2012) have led to a proliferation of configurational information. The dynamics of and transient conformations explored by IDPs are often extremely high dimensional and are not always well described by the standard vocabulary of structural biology. Machine learning and network analytic approaches offer potentially valuable

ways of addressing such problems by facilitating (respectively) the detection of systematic patterns in high-dimensional data and the representation and modeling of complex structures that do not follow simple, regular motifs (e.g., alpha helices or beta strands). In this paper, we show how tools drawn from both traditions can give purchase on the comparative exploratory analysis of molecular dynamics trajectories from protein variants, yielding insights that would be difficult to obtain using more conventional methods. We illustrate our approach using simulations of the wild type (WT) Aβ_{1-40}, a well-known intrinsically disordered protein and its E22G ("Arctic") variant, which is implicated in familial Alzheimer's disease (Nilsberth et al., 2001), and which has been a system of interest for many previous molecular dynamics studies (Cecchini et al., 2006; Lam et al., 2008; Urbanc et al., 2010).

The majority of proteins have a well-defined structure-function relationship, whereby the protein's biological role is contingent on it being correctly *folded* into its flexible, but locally stable, functional configuration. By contrast, intrinsically disordered proteins (along with proteins possessing a significantly large intrinsically disordered region) owe their function to not being confined to a small number of stable regions of configuration space. For example, many signaling proteins are able to bind a wide variety of targets due to their intrinsic disorder (Iakoucheva et al., 2002). The study of IDPs presents challenges inherent to both the molecular systems themselves and the standard conventions used by the scientists who study proteins. In addition to the difficulty of distilling down the complex motions of these "moving targets" of structural biology to some intuitable form, there are additional difficulties due to the standard descriptive and experimental toolkits used by structural biologists and chemists, from Ramachandran plots to X-ray crystallography, being tailored toward gaining insight about proteins within the paradigm of a small number of favored static configurations. Thus, if we wish to search for latent order characteristics of a particular IDP, we must establish methodologies for characterizing and interpreting IDP data. Such problems, where vast amounts of high-dimensional unstructured data is available for a set of known classes (e.g., WT class vs. E22G class) are the exact situations where machine learning algorithms excel. In fact, a great deal of progress has been made in the development of ML-based technologies for the interpretation of chemical and biochemical systems, such as automated optimal partitioning of configuration space for building kinetic models (Grazioli et al., 2017), clustering-based methods for building Markov models of protein folding (Husic and Pande, 2017), protein conformational space mapping with self-organizing maps (Bouvier et al., 2014), protein-ligand interaction scoring (Ragoza et al., 2017), automating the definition of atom types in molecular mechanics force fields (Zanette et al., 2018), and even the *inverse design* of materials, using ML to guide material design, given a set of desired material properties (Sanchez-Lengeling and Aspuru-Guzik, 2018).

A related problem is summarizing the transient structures of IDPs in a way that is reductive enough to provide useful simplification while still being flexible enough to accommodate a wide range of irregular structural configurations. Network representations, which have been extensively studied in the context of human social networks (Wasserman and Faust, 1994), provide a natural tool for this purpose. Most relevant to IDP behavior are protein structure networks (PSNs), which represent protein structures in terms of relationships (e.g., bonded or non-bonded interactions) among groups of atoms (e.g., moieties, residues, or whole secondary structure elements). PSNs are useful for coarse-graining protein structure while retaining topological information describing internal contacts, and have been employed to rapidly identify enzymes with distinct but non-obvious structural features (Butts et al., 2016), characterize local packing characteristics distinguishing closely related enzyme classes (Unhelkar et al., 2017), distinguish structural features particular to thermophilic vs. mesophilic proteins (Brinda and Vishveshwara, 2005), analyze simulation trajectories (Benson and Daggett, 2012), and predict differences in overall protein (Atilgan et al., 2001; Jacobs et al., 2001) and active site (Duong et al., 2018) flexibility, among other tasks (Csermely et al., 2012). PSNs can be modeled using statistical techniques adapted from social network analysis (Yaveroğlu et al., 2015), allowing for very flexible and computationally efficient identification of structural biases distinguishing groups of proteins, tests of hypotheses relating to protein topology, and simulation of PSN structure. Here we leverage these techniques to uncover differences in the respective energy landscapes of Aβ_{1-40} wild type and E22G.

In addition to providing broadly applicable methodology, we also present applications of this approach to the elucidation of the dynamic, and often subtle, characteristics of wild-type Aβ_{1-40} and its variant E22G that lead to their distinct behavior in solution, despite their being identical in all but one amino acid. Although the present discussion is focused on applying our methodologies to IDPs, it is noteworthy that there are also examples of well-folded proteins, like TEM-1 β-lactamase (Roccatano et al., 2005) or ZASP PDZ (Fratev et al., 2014), where the structural changes caused by point mutations can also be very difficult to discern in molecular simulations, despite the mutations having known physiological effects. Thus, the approaches discussed here may have applicability beyond the IDP case. The remainder of the paper is organized as follows: we begin by applying simple and well-established methods for comparing data generated by molecular dynamics simulations of both WT Aβ_{1-40} and the E22G variant (e.g., Ramachandran plots), highlighting their limitations in the context of intrinsically disordered proteins. Although the two proteins seem at first blush to exhibit nearly identical behavior, we show how support vector machines (SVMs) can be employed to construct a metric that readily distinguishes them. Projection of conformations obtained from structures of Aβ fibrils onto this metric can then be used to predict differences in fibrillization behavior. Moving from torsion angles to topology, we employ exponential family random graph models (ERGMs) to characterize the properties of favorable transient structures in Aβ_{1-40} residue-level PSNs, and use this to explore the structures most energetically favored by WT vs. E22G (and vice versa). We then close with a demonstration of how joint k-means clustering of conformations from long WT and E22G trajectories and network analysis of

the Markov transition graph on the resulting conformational states reveals substantial differences in dynamics that are not apparent on casual inspection. Additional technical details regarding our simulations and analysis are provided in the following section, and we conclude with a discussion of our findings and how approaches such as these can be used to select targets for further experimental biophysical characterization and structural biology.

2. RESULTS

2.1. Exploring the Torsion Angle Space of Energy Minima

Prior to applying more complex, ML-based techniques for identifying the characteristic differences between the configurational dynamics of the WT and E22G variants, it is reasonable to first apply more established approaches toward that same end. Thus, we begin by calculating a Ramachandran plot (**Figure 1**) from a large set of configurations generated by MD simulations from a highly dispersed set of seed conformations (details provided in section 4), as well as from conformations associated with large samples of local energy minima. It is clear from the data shown in **Figure 1** that WT and E22G cannot be distinguished by their distributions in Ramachandran space. This result illustrates the core problem of exploratory analysis of intrinsically disordered proteins: many of the simple and familiar tools of structural biology exploit the fact that folding confines typical proteins to a narrow range of conformations, and the lack of such confinement leaves them with little signal to leverage.

Given that the Ramachandran plot does not display any obvious differences that could be used to distinguish between WT and E22G conformations in torsion angle space, it is natural to ask whether these variants might still be distinguished by the distribution of their *angular velocities* in the same space. Employing a large number of trajectories initialized from a set of widely dispersed local minima (see section 4), we plot the distribution of local ψ and ϕ angular velocities in the equivalent of a Ramachandran space (**Figure 2**). As can be seen, the resulting velocity distribution is homogeneous both by residue index (left) and by variant (right), with the points colored for each attribute overlapping so completely that they appear to form a single undifferentiated distribution. Plainly, this property cannot differentiate between WT and E22G. Moreover, the similarity in velocity distributions between variants suggests that differences in the energy landscape associated with the E22G mutation are extremely subtle, despite its known differences in aggregation behavior relative to wild type (Lord et al., 2006; Norlin et al., 2012).

The lack of distinguishing features in either the Ramachandran space of conformations or the "differentiated Ramachandran" space of angular velocities highlights the subtle nature of differences in IDP behavior, and points to the need for more flexible—and high dimensional—techniques to identify differences. We now turn to a family of kernel learning methods that are well-suited to this purpose.

2.2. Finding Relatively Favored Conformations via SVM

The observation that WT and E22G Aβ_{1-40} differ by a single residue, yet exhibit differing propensities for fibrillization in experiments (Norlin et al., 2012), seems to imply that the conformations they sample in solution must originate from differing equilibrium distributions in configuration space. Further, we note that if a configuration is defined as a vector of all torsion angles for residues 1 through 40, the respective distributions for WT and E22G both "live" in the same coordinate space. Thus, we may posit some *characteristic axis*, onto which any configuration in the shared torsion angle space can be projected, where points at one extreme are most characteristic of WT (and least likely to be sampled by E22G) and points on the other extreme are most characteristic of E22G (and least likely to be sampled by WT). If we, for the sake of argument, were to imagine that the sets of conformations sampled by each variant were linearly separable—i.e., a separating hyperplane in torsion angle space could be placed between them with all WT points on one side and all E22G points on the other—such an axis would be trivial to define: it would be the vector normal to the separating hyperplane. Unfortunately, the condition of linear separability is an unrealistic assumption for two systems that are both highly similar and high dimensional, and the Ramachandran analysis of **Figure 1** suggests that it is inapplicable here. However, we could consider an alternative version of our construction, in which we nonlinearly map our torsion angle space into an alternative space (called a *feature space*) in which our conformations *are* linearly separable and then find the characteristic axis within this modified space. The resulting characteristic axis would no longer take a simple form in our original space (the *input space*), but we could nevertheless use it to "score" hypothetical conformations for similarity to WT vs. E22G by mapping them into the feature space and finding their projection onto the characteristic axis in that space.

Finding transformations of this type in high-dimensional data is a central problem of *kernel learning* (Scholkopf et al., 1999), and identifying a "characteristic axis" like the one envisioned above is a natural application of *support vector machines* (SVMs) (Vapnik, 2013). In a classification context, SVMs seek maximum-margin separating hyperplanes between sets of observations, with the characteristic axis corresponding to a quantity (often called the *decision value*) that is used to predict class membership. While "pure" SVMs are linear algorithms, kernelized SVMs (i.e., SVMs operating on kernel-transformed inputs) are powerful tools for finding complex separating surfaces (or, in the case of imperfect separability, approximate separating surfaces) in more general contexts.

A heuristic illustration of how SVMs can be used to extract a characteristic axis from linearly non-separable data classes is shown in **Figure 3**, as an aid to intuition. Note that in the input space $\{x, y\}$ (**Figure 3A**), no single plane can be defined that perfectly separates the blue class from the red class. By mapping the data to the higher-dimensional space of all polynomials in x and y (truncated to the subspace $\{x, y, x^2\}$ in **Figure 3B**, chosen for visualization purposes), this same data set is now linearly

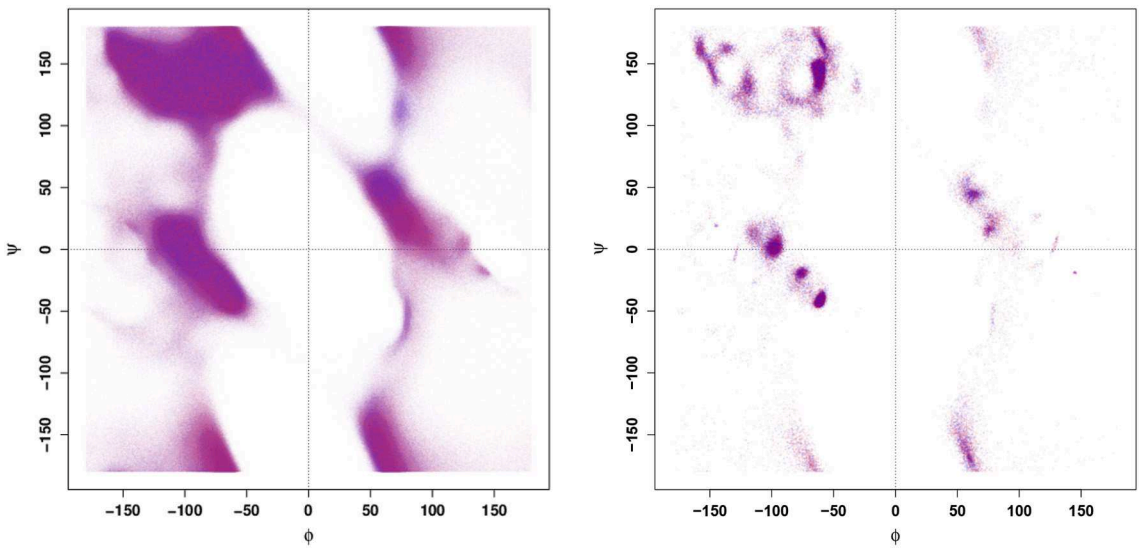


FIGURE 1 | Ramachandran plots for all conformations **(Left panel)** and conformations of local minima **(Right panel)**. Points are colored by variant (blue - WT, red - E22G); apparent purple color indicates near identical distribution of torsion angles.

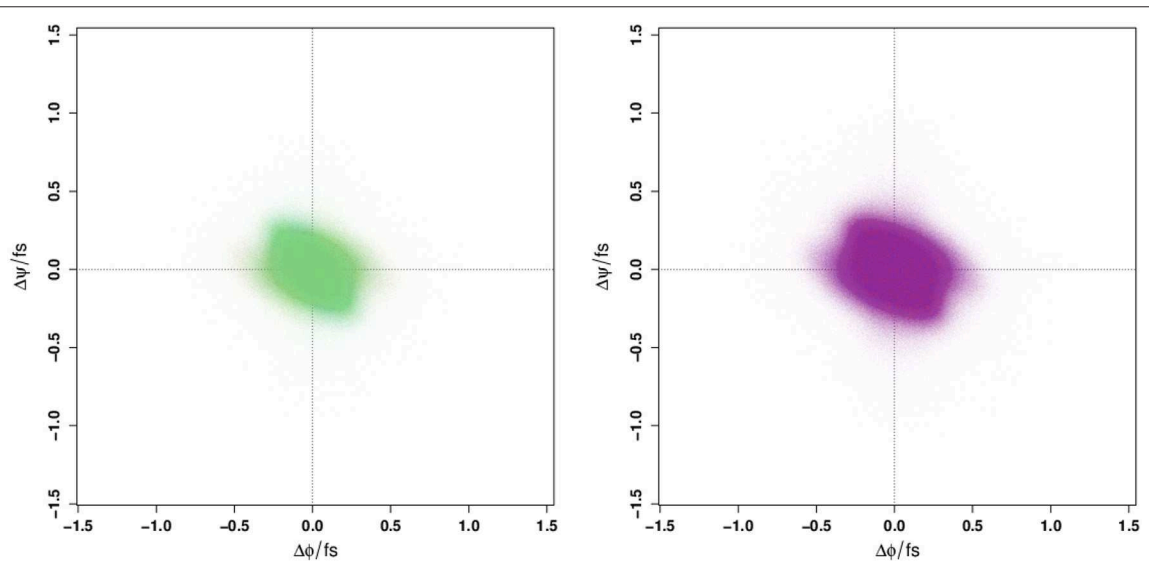


FIGURE 2 | Angular velocity distributions by dihedral angle colored by residue index **(Left panel)** and by protein **(Right panel)**. No systematic variation is visible in either case.

separable. Such a mapping onto quadratic functions of the inputs constitutes a polynomial kernel of order 2, with mapping into higher-order polynomials corresponding to higher-order kernels; mapping to polynomial functions of arbitrary order can be performed by selection of e.g., the Gaussian or radial basis function (RBF) kernel, whose basis set can be interpreted in terms of Taylor series expansions of exponential functions. Such an expansion can in principle find a separating hyperplane for any point set (subject to regularity conditions), making the RBF kernel a so-called "universal" kernel. With a separating plane now defined in the kernel-transformed feature space, the data points can be projected onto the vector normal to that plane (C). This vector is our characteristic axis, with the 0 point corresponding to the point of maximum margin when dividing the two classes.

To apply this idea to the case of our Aβ variants, we trained an SVM classifier under a RBF kernel to distinguish low-energy conformations of WT (obtained by independent annealing trajectories seeded with an overdispersed sample of conformations obtained via a high-temperature trajectory) from those of E22G (see section 4 for details). To gain insight into conformations that are relatively favorable for E22G vs. WT, we approximately linearize the decision surface (i.e., the pre-image

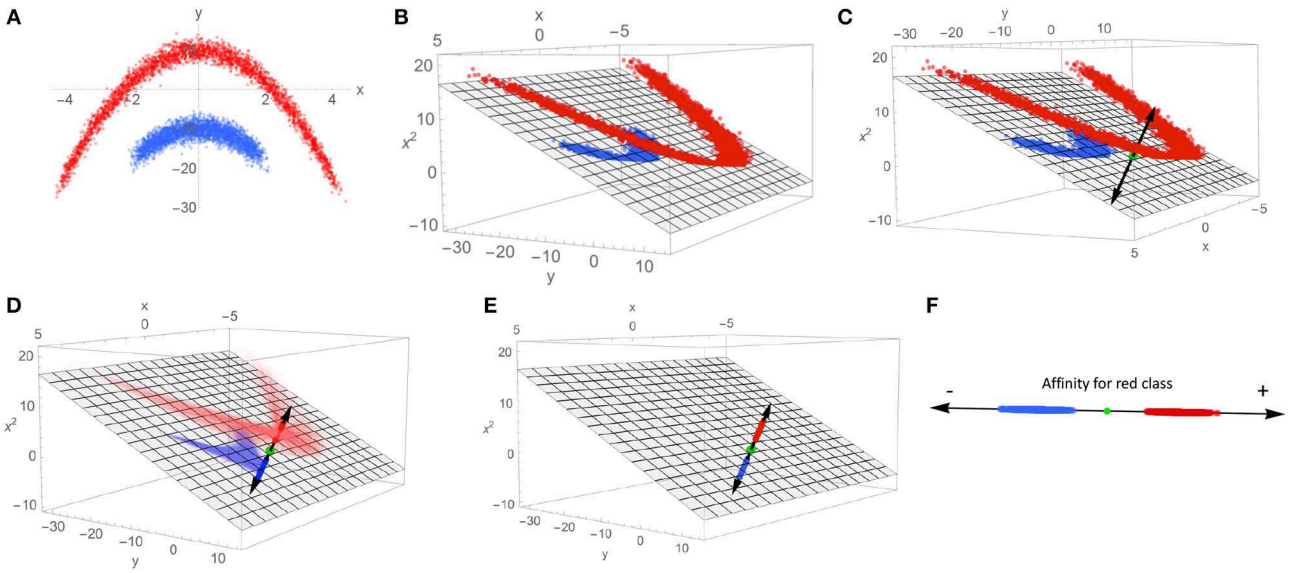


FIGURE 3 | This schematic demonstrates the process of mapping linearly non-separable data to a single coordinate that indicates membership to each class. **(A)** An example of linearly non-separable 2D data with two classes indicated by red and blue dots. **(B)** An exact mapping of the points shown in **(A)**. To a 3D feature space, where the data is linearly separable by the plane shown as gray mesh. **(C)** A vector normal to the separating plane (black double-headed arrow) is introduced, along with the intersection of the separating plane and normal vector (green dot). **(D)** The points in the 3D feature space are projected onto the normal vector. **(E)** The projected points on the characteristic axis are shown in the 3D space. **(F)** The characteristic axis now serves as the reference frame for a scalar scoring indicating a point's affinity for one class vs. the other.

of the separating hyperplane in the input space) and examine its characteristics averaged over the E22G/WT conformations that are closest to it. Specifically, we identify the *support vectors* from the SVM solution (i.e., the data cases with non-zero weight, from which the decision surface is defined), and identify pairs of WT and E22G support vectors that are as close as possible within the input space (as measured by Euclidean distance between inputs). Each of these pairs can be envisioned as straddling the decision surface, with no other pair being strictly closer to it (since, if so, at least one point in the pair would not be a support vector). Taking the difference of properties between one conformation in the pair and the other thus allows us to approximate the gradient of the decision surface with respect to those properties in the original (input) space, at some point between the conformation pair. Considering the distribution of such differences over all such pairs then gives us insight into the properties that typically do (or do not) typically distinguish E22G trajectories from those of WT.

Figure 4 shows the result of such a calculation performed for the (circular) mean differences in torsion angles between paired E22G and WT support vectors, for the low-energy conformation model. Although many angles show no significant differences—indicating that, on average, there is no *net* contribution of position on this angle to relative favorability—some show a clear and systematic difference across the decision boundary. Perhaps most notable are the torsion angles for ϕ_{22} and ψ_{22}, both of which show positive change when moving across the decision boundary from the WT to the E22G side. (Put another way, ψ_{22} tends to be turned approximately 0.35 radians to the right within E22G minima from its value in similar WT trajectories). In addition to confirming the intuition that the substantial loss of side chain steric hindrance brought about by the E22G mutation alters the local backbone curvature at the mutation site, our analysis allows us to focus on the torsion angle changes that best distinguish otherwise similar local minima. For instance, we also see significant increases in ϕ angles for residues 18, 20, 25, and 38, and decreases for residues 5, 23, and 37, showing systematic effects on several (but not all) sites along the backbone. Similarly, we see significant additional increases in ψ angles for residues 6 and 36, and decreases for residues 13, 21, 23, 26, and 39, showing that the two torsion angles are affected differently by the mutation but that those effects show signs of clustering (e.g., the relatively numerous angular differences near the mutation site, or for residues 37-39 at the C terminus).

Another method for determining which degrees of freedom contribute most substantially to the classification of a configuration as belonging to either WT or E22G is to combine SVMs with principal component analysis (PCA), as shown in **Figure 5**. In this treatment, the differences in torsion angles between WT and E22G minima across the SVM decision boundary are processed using PCA, resulting in a new reference frame in which the principal components are linear combinations of the original dimensions that begin with the direction of maximum variance and proceed in subsequent orthogonal directions in order of diminishing variance (Pearson, 1901). Thus, plots of the first two principal components, such as **Figure 5**, display the two directions through the space of torsion angle differences that best summarize (in a least squares sense) the total pattern of variation in torsion angle differences

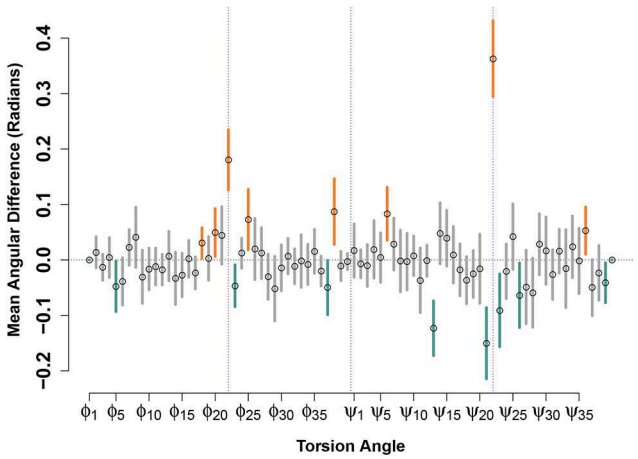


FIGURE 4 | Mean angular difference in torsion angles between WT and E22G minima, across the SVM decision boundary. For each torsion angle, *y* axis values show the average across all support vector pairs of the angular difference in WT and E22G values; 95% confidence intervals indicated by thick lines (orange values significantly positive, teal values significantly negative). Prominent deviations are seen at the mutation site (blue dotted lines) but can also be seen at many other locations along the backbone.

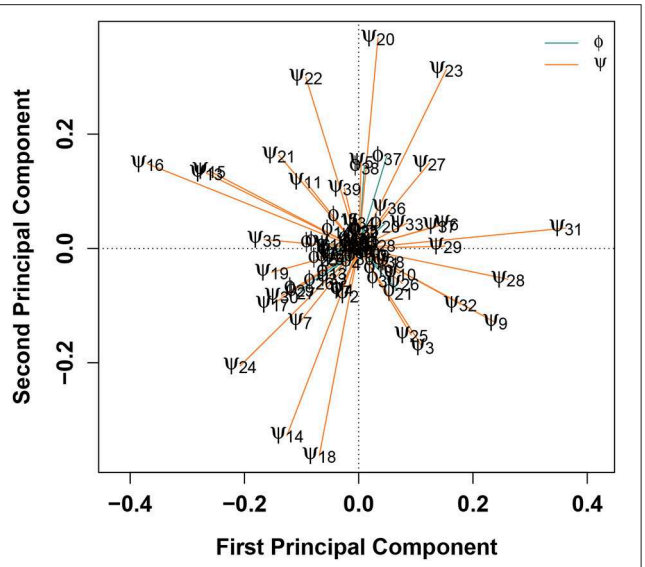


FIGURE 5 | PCA of angular differences in torsion angles between WT and E22G minima, across the SVM decision boundary. Vectors show ϕ (teal) and ϕ (orange) torsion angle loadings on the first two principal components; angles with similar loadings tend to show similar patterns of differences for E22G vs. WT conformations.

across the decision surface. The loadings on these components hence provide information on which angles contribute most to these directions, and on the sense of that contribution (i.e., positive or negative).

Figure 5 shows that, with the exceptions of ϕ_{37-38}, the first two components are strongly dominated by the ψ torsion angles. This result is consistent with the greater variance in ψ relative to ϕ in standard protein secondary structures, but it was not observable from the Ramachandran plot of Aβ conformations given in **Figure 1**. The strongest contrasts seen are between: ψ_{13} and ψ_{15-16} (left) and a group of angles including ψ_{31-32}, ψ_3, ψ_9, ψ_{25}, and ψ_{28} (right); and ψ_{20} and ψ_{22-23} (top) and ψ_{14}, ψ_{18}, and ψ_{24} (bottom). The first contrast involves a cluster of residues marking the N-terminal end of a stretch of residues forming a (transient) α-helix in a solution-state NMR structure (PDBID: 2LFM) (Vivekanandan et al., 2011) vs. a collection of several residues in the terminal regions of the protein. The second contrast, interestingly, pits a cluster of residues at the C-terminal end of the aforementioned helix-forming region with three residues spanning it (two at either end and one in the middle). This suggests one mode involving the extent of helical structure in range of residues 14-23, and another involving a broader pattern of curvature throughout the protein. By identifying such patterns, we can potentially focus attention on particular conformational features that are differentially favored by E22G vs. wild-type Aβ.

One obvious application for a score distinguishing WT and variant conformations is in screening for the potential to exhibit distinct patterns of fibrillization. Fibrillization is difficult to probe directly via MD trajectories, due to the long timescales and large atom counts involved, and fibrillization experiments with new systems are costly. In particular, structure determination efforts are time-consuming and often require technological innovations to achieve. Although amyloid fibrils by definition form a common cross-β structure, they often differ in detailed structural topology. Therefore, given a new variant with potential clinical significance, it is useful to be able to obtain some indication of whether or not it is likely to form fibrils with the same structural topology as the wild-type protein. While the SVM analysis conducted here cannot provide a definitive answer to this question, it can tell us (based on the sets of trajectories available) whether known fibril structures involve monomeric conformations that are *more characteristic of wild-type than the variant*. If WT and the variant (here E22G) have similar affinity for a particular set of fibrillar conformations, then this suggests that the variant will have a similar propensity to produce such fibrils in practice; however, if the affinity differs strongly between WT and the variant, then this may indicate a difference in the propensity to produce fibrils of this topology.

Such an approach is illustrated in **Figure 6**, where the relative similarity of fibrillar conformations to E22G vs. WT (as expressed by projection onto the characteristic axis) is shown for all conformations from 10 Aβ fibril structures found in the Protein Data Bank. While some individual configurations appear more favorable for E22G than WT (positive values), all fibril structures were overall significantly more typical of WT solution minima than the minima observed for E22G (hence all plot markers are blue in **Figure 6**), suggesting that the latter has a different fibrillization pattern. Interestingly, the two non-wild type fibrils included (PDBIDs 2LNQ and 2MPZ, both of the D23N or "Iowa" variant) show particularly strong relative affinity for WT vs. E22G, suggesting that E22G's fibrillization behavior differs from that of both variants. These results are compatible with

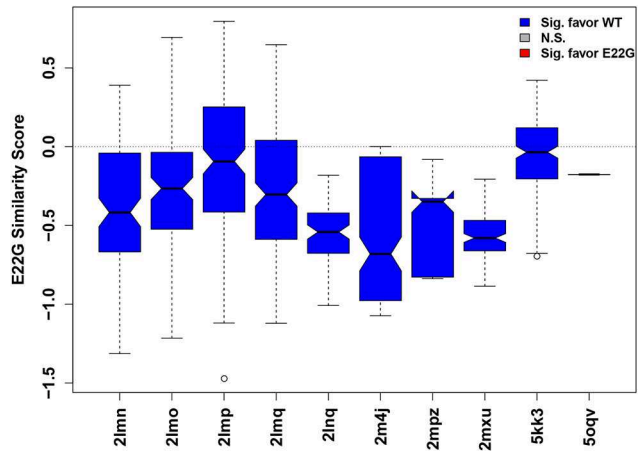


FIGURE 6 | Distribution of similarity scores for conformations from known fibril structures (Aβ residues 15–40). Higher values (above dotted mid-line) indicate greater net E22G affinity, while lower values indicate greater net WT affinity. The overwhelming majority of conformations are more characteristic of WT minima than E22G minima, suggesting that there is a different fibrillization pattern for E22G, making this variant a high priority for future structural studies.

experimental findings that have previously suggested that E22G may have a different fibrillization mode from WT, potentially proceeding through a different oligomeric precursor. A study employing a variety of biophysical techniques concluded that aggregation of this species proceeds via a characteristic type of on-pathway intermediates and then quickly progresses to a highly polymorphic variety of fibrils (Norlin et al., 2012), making high-resolution structure determination difficult. Given the time and expense necessary for solving atomic-resolution structures of even a single fibrillar conformation, measures of potential dissimilarity in fibrillization behavior are useful tools for choosing new structural targets. Disease-relevant variants, such as E22G, that are likely to occupy one or more novel fibril topologies can be considered high-priority targets for further structure determination efforts. It is important to reiterate here that the similarity scores for each fibril type represent how similar each fibril structure is to WT vs. E22G, thus two fibril structures whose similarity scores are close in value may or may not be similar to each other.

As with the decision surface, we can gain some additional insights regarding the local factors that tend to lead fibrillar conformations to be more favorable for E22G vs. WT by local linearization. In this case, we proceed by regressing the similarity score (projection onto the characteristic axis) for each fibrillar conformation onto the input features of each conformation (the real and imaginary components of its torsion angles). The coefficients from this regression represent the mean gradient of the score over the fibrillar conformations; to convert these into statements involving the original torsion angles, we express the gradient elements associated with each angle (i.e., $\hat{\beta}_i \sin(\theta_i) + \hat{\beta}_i' \cos(\theta_i)$, for angle θ_i with regression coefficients $\hat{\beta}_i, \hat{\beta}_i'$) in the periodic form $b_i \sin(\theta_i - y_i)$ [where $b_i = \sqrt{\hat{\beta}_i^{\,2} + \hat{\beta}_i'^{\,2}}$ and $y_i = \tan^{-1}(\hat{\beta}_i'/\hat{\beta}_i)$]. Intuitively, the modulus b_i scales the absolute magnitude of the contribution of local changes to the ith torsion angle to changes in the expected similarity score, while the argument y_i defines a *reference angle* or *angular offset* such that small increments above y_i increase similarity to E22G, while small decrements below y_i decrease it.

A schematic detailing how such an approach is implemented is shown in **Figure 7** using a single pair of ϕ and ψ torsion angles in a simplified, two-dimensional example. We consider two variants of a hypothetical protein (designated "blue" and "red") with two torsion angles of interest, ϕ and ψ. The blue and red dots on the angular plots for ϕ and ψ in **Figure 7A** represent the values for these angles for 1,000 different configurations sampled for each variant. From these conformations we may create an affinity score surface by training an SVM classifier to classify blue vs. red configurations using the real and imaginary components of both angles ϕ and ψ as the training data ($\{Re(\psi), Im(\psi), Re(\phi), Im(\phi)\}$). **Figure 7B** shows this affinity score surface in ϕ, ψ space (lighter values favor blue, while darker values favor red), together with the sampled red and blue configurations from panel **Figure 7A**. Now, consider a set of comparable torsion angles obtained from fibril structures; these may also be projected into our angular space, as shown in **Figure 7C** (cyan points). Each fibrillar conformation can be assigned an affinity score based on its location on the affinity score surface, indicating the extent to which it is more typical of the blue vs. the red variant. Regressing the affinity scores of the fibrillar conformations on the underlying torsion angles yields the mean gradient of the affinity score surface in angular space across the fibrillar conformations (orange arrow). From this we can equivalently construct a set of reference angles (green dot) that expresses the torsion angles that would provide the average greatest tendency to be more blue-like (vs. red-like) in the vicinity of the fibrillar conformations. Returning to an angular representation, **Figure 7D** shows both the mean vectors for the fibrillar conformations (cyan) and the reference angles (orange/green) in polar space. Local rotations toward the reference angle are here associated with increasing "blueness," while rotations away are associated with increasing "redness."

In applying this methodology at scale to the Aβ system, we display these regression coefficients in the form of what we call an *orrery plot* in **Figure 8**. Each y axis value in the orrery plot gives the reference point for the associated torsion angle, while moduli are shown by point radius. Higher moduli indicate greater local contributions to the affinity score. (Note that, due to unreported residues in the fibrillar PDB structures, we limit our examination to residues 15-40). At a glance, the orrery plot tells us that the dominant local contributors to E22G similarity are the torsion angles at the mutation site, as well as angles such as ϕ_{17}, ϕ_{27}, ϕ_{32}, ψ_{18}, ψ_{21}, and ψ_{34}. The offset values show that not all torsion angles of the same type are in phase with each other (in the sense of having a common reference such that values higher or lower than the reference have the same impact on the similarity score), although some sets of residues do have very similar offsets. This may suggest particular groups of residues whose local conformations play a similar role in initiating or stabilizing fibril structure in wild-type Aβ. We also see many residues whose

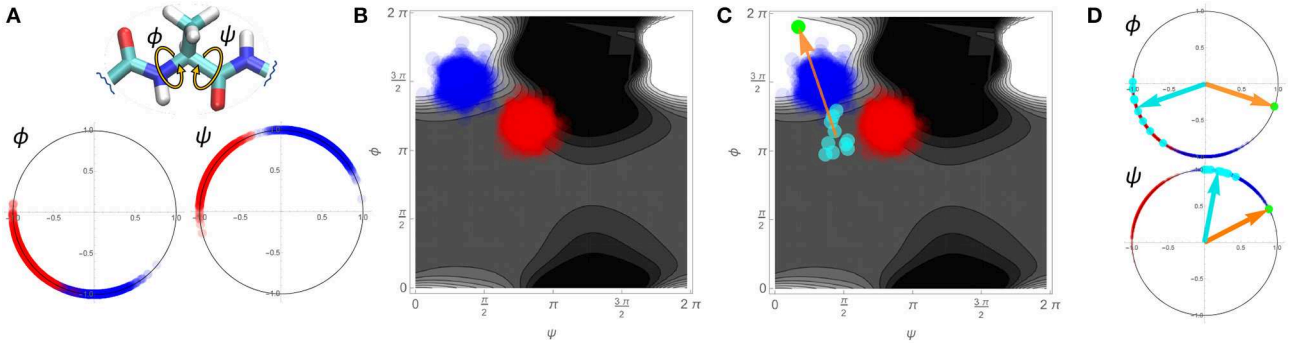


FIGURE 7 | Schematic demonstration of the linearized similarity score used in **Figure 8** using a simplified 2-dimensional example. **(A)** 1000 sampled conformations for two protein variants (blue and red dots) on two torsion angles. **(B)** Depiction of blue and red conformations in Ramachandran space, with an SVM-derived affinity score (background gradient). Lighter regions are more typical of the blue variant, while darker regions are more typical of the red variant. **(C)** Mapping a set of hypothetical fibrillar conformations into Ramachandran space (cyan) allows them to be associated with affinity scores. Regression of affinity scores onto the underlying angular components yields the mean gradient of the affinity score local to the fibrillar conformations (orange arrow), which can be re-expressed in terms of a reference angle (green dot). **(D)** Re-expression of conformations in angular space, with reference angles (orange/green) and mean fibrillar conformations (cyan arrows) indicated. Direction of rotation relative to the reference indicates whether local angular changes (from the mean fibrillar conformation) would make the structure more "blue-like" vs. "red-like".

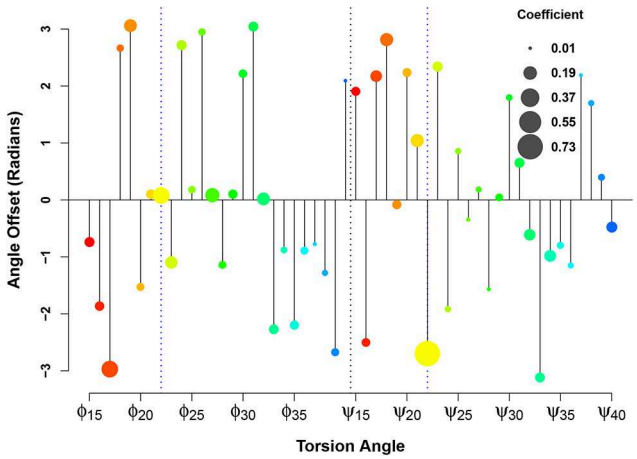


FIGURE 8 | Linearization of the similarity score for fibrillar conformations. Relative affinity for E22G (vs. WT) increases as $b\sin(\theta - y)$, where y is the angle offset (y axis), b is the weight coefficient (circle size), and θ is the torsion angle. The color spectrum from red to blue indicates the sequence position from N-terminal to C-terminal.

conformations do not seem to be strongly associated with relative affinity for wild-type vs. E22G (e.g., ϕ_{37} or ψ_{24}), which suggests that differences in fibrillization behavior between the two variants are not likely to depend on the local conformations of these residues. The orrery plot thus provides us with guidance on the angular degrees of freedom that are more or less likely to distinguish protein variants with respect to their propensity to adopt fibrillar conformations.

2.3. Identifying Differences in Transient Structure via Network Analysis

As noted, a central challenge in the analysis of IDPs is their lack of the characteristic secondary structure motifs that are the primary point of reference for describing and comparing the tertiary structures of folded proteins. Although IDPs by definition lack stable secondary structure, they nevertheless form other types of transient structures that can be characterized. Transient structural features have been observed in weakly structured proteins (Williamson and Miranker, 2007; Lee et al., 2014) or partially folded intermediates (Teilum et al., 2002; Bernard et al., 2005), often using the sensitivity of NMR chemical shifts to local backbone conformation (Spera and Bax, 1991); such features are often found to resemble more stable structural elements formed upon interaction with a binding partner (Song et al., 2008). A natural approach to characterizing transient structural elements is via the use of residue-level PSNs to characterize the pattern of interactions among residues within sampled conformations, giving rise to coarse-grained representations that are flexible enough to represent the wide range of conformational variation exhibited by IDPs. A residue-level PSN is a network structure (or, more formally, a *graph*) whose nodes or vertices correspond to individual residues, and whose edges correspond to inter-residue contacts. Here, we define two residues v_i, v_j to be in contact (*adjacent*) if there exists an atom a_i in residue v_i and atom a_j in residue v_j such that the inter-atomic distance between a_i and a_j is less than 1.2 times the sum of their respective van der Waals radii. We compute a PSN for each conformation in our set of respective WT and E22G energy minima, giving us an ensemble of PSNs (each a 40-node network) for each Aβ_{1-40} variant.

2.3.1. Where Is Transient Structure Formed in E22G and WT?

A natural first question to address is where transient structure is potentially formed in the wild-type and variant proteins. While there are many types of local network structure that might be considered, we follow (Unhelkar et al., 2017) in using the degree k-cores of the PSN to indicate areas of cohesive interaction among residues. A (degree) *k-core* (Wasserman and Faust, 1994)

is a maximum set of nodes such that every member of the set is adjacent to at least k other members of the set; the highest k such that vertex v belonging to the kth core of a graph is referred to as v's *core number*, and is an indication of v's embeddedness in locally cohesive structure. While k-cores need not be globally cohesive, high-numbered k-cores are composed of locally cohesive elements, and hence vertices with high core numbers represent residues belonging to regions of the protein connected by multiple redundant contacts. By contrast, vertices with low core numbers represent residues residing in regions that are at best very loosely connected.

To summarize global tendencies toward structure formation in the two variants, **Figure 9** shows the mean core numbers for each WT and E22G residue, averaged over all minimum energy conformations in each respective set. Observed mean core numbers range from just over 1 at the N-terminus to over 3 in the internal region of the protein, falling again near the C-terminus. The relatively low core numbers near the termini are reflective of the high flexibility of these regions, though we observe a substantial and significant difference between the N-terminal and C-terminal regions (with the former being far less structured, on average, than the latter). In general, WT and E22G show very similar patterns of core structure throughout the N-terminal region, although E22G shows significantly higher core numbers for the majority of residues. The largest differences in core numbers are observed for a band of residues extending roughly from G15 to M35. Within this region, E22G produces substantially more local cohesion, on average, than WT. The elevated level of structure within this band for both variants may stem in part from interactions among the numerous nonpolar residues located within it, but the cross-variant difference points to a major role for E22 in destabilizing possibly aggregation-inducing local interactions throughout the C-terminal region. Although comparative experimental results are not available for these proteins, this central region of higher connectivity is consistent with the observations of Rosenman et al. (2013) from NMR experiments on the wild-type protein at low temperature. Based on measured J-couplings and molecular dynamics simulations, several frequently populated structural elements were observed, including a transient salt bridge between E22 and K28 [also observed by Rosenman et al. (2013)], which was observed in the minima of our wild-type models.

To get a better sense of how these differences in structure arise, it is useful to distinguish the residue contacts that arise more often in E22G than WT (and vice versa). **Figure 10** shows, for both sets of PSNs, the edges that are found significantly more often in E22G (red) or in WT (blue). Mutation of the glutamic acid at position 22 to glycine clearly enhances a large complex of potential contacts, prominently including residues 7-8, 11-12, and 22-23 (among others); in addition, we see a weaker but more broad-based enhancement of contact rates throughout the protein, but particularly in the C-terminal region. By contrast, relatively few contacts are more prevalent in WT, among the exceptions being pairwise contacts between 1 and 22 and 3 and 11, as well as some relatively local contacts in the C-terminal region (appearing to involve interactions among nearby non-polar residues). Overall, the broad pattern suggests that in WT,

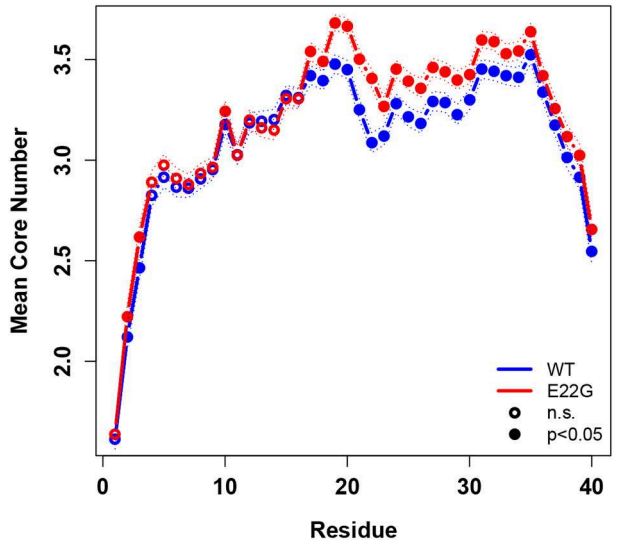


FIGURE 9 | Mean degree core number by residue, WT vs. E22G, with 95% confidence intervals. Significant differences in core numbers are indicated by dark circles. E22G shows a region of markedly higher average cohesion involving internal residues in the C-terminal portion of the protein, suggesting a major role for these residues in the formation of transient structural elements that may be involved in aggregate formation.

E22 both blocks interactions among residues in its immediate vicinity and limits the ability of the two large patches of nonpolar residues within the C-terminal region to interact (with some of these instead participating more often in ephemeral internal interactions). In E22G, the replacement of the bulky glutamic acid with the small and highly flexible glycine appears to allow these previously blocked groups to interact with much higher frequency, raising the average local cohesion.

It should be noted that all of the above contacts are transient, with typical conformations being quite sparsely connected (though some do have considerable self-interaction). Thus, these patterns reveal biases or general tendencies in a fluctuating system, rather than the stable structures characterizing more typical proteins. This raises the question of which particular structures are more strongly favored for WT vs. E22G, to which we now turn.

2.3.2. What Transient Structures Characterize the Difference Between WT and E22G?

The above give us some sense of where transient structure is being formed in WT and E22G, but they do not provide a strong holistic sense of which sorts of global structures are more characteristic of E22G vs. WT. For that purpose, we must consider the networks as a whole. To do this, we fit statistical models to the respective E22G and WT minima that identify the network features that are more or less enhanced for each variant. We do this by leveraging ERGMs (Hunter et al., 2008), parametric statistical models for graphs that allow direct representation of complex dependence among edges. Given a random graph G, defined on support $\mathbb{G}$,

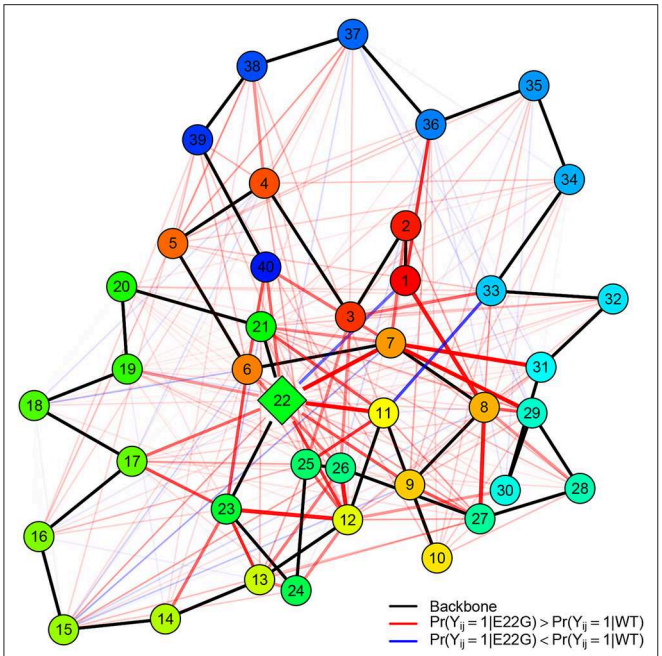


FIGURE 10 | Relative PSN edge frequencies, E22G vs. WT. Of edges that differ statistically between E22G and WT, more of them are characteristic of the E22G variant (red) than wild type Aβ(blue); backbone edges (black) are provided as a reference, with vertices colored by residue index (red to blue). Edge width and translucency reflect the magnitude of relative difference in appearance rates (heavier edges differ more). Although E22G forms structure more readily in most locations, a key complex of activity involves residues 7, 8, 11, 22, and 23 (with several other residues also participating).

we may write its probability mass function in ERGM form as

$$\Pr(G = g|\theta, t, X) = \frac{\exp(\theta^T t(g, X))h(g)}{\sum_{g' \in \mathbb{G}} \exp(\theta^T t(g', X))h(g')}, \qquad (1)$$

where $t : \mathbb{G}, X \mapsto \mathbb{R}^k$ is a vector of sufficient statistics, $\theta \in \mathbb{R}^k$ is a parameter vector, h is a *reference measure* satisfying $0 \leq h(g') \leq \infty$ for $g' \in \mathbb{G}$ and $h(g') = 0$ otherwise, and X is a set of covariates. In the case of residue-level PSNs, $\mathbb{G}$ is the set of all simple graphs on N vertices (where N is the length of the primary sequence), subject to the constraint that each vertex is tied to the vertices corresponding to its neighbors in the protein backbone. Here, we follow typical practice for unvalued, fixed-N networks and take h to be the counting measure on $\mathbb{G}$, implying that $h(g') \propto 1$ for $g' \in \mathbb{G}$ and 0 otherwise. Since h then cancels for graphs in the support, we henceforth omit it in our notation (it being tacitly assumed that the probability of graphs outside the support is 0).

An extensive statistical literature exists on ERGMs, and in particular on the problem of inferring an unknown θ from observations of G. Substantively, the model can be understood as describing *biases* in the distribution of G relative to the reference measure (in our case, the uniform distribution over possible 40-node PSNs), with the nature of each bias determined by the choice of statistics (t) and the direction and strength of each bias determined by θ. Here, we fit separate ERGMs to the sets of observed WT and E22G minima (respectively), inferring θ in each case by approximate Bayesian inference using Laplace parameter priors analogous to the L1 regularization employed in the well-known LASSO procedure (Tibshirani, 1996). **Table 1** shows the posterior mean estimates, posterior standard deviations, and 95% central posterior intervals for the parameters (i.e., θ) of each fitted model. The estimated effects (i.e., t) are described in greater detail in section 4, but may be summarized as follows: an *Edges* effect sets the baseline PSN density; *Backbone Dist* indicates the effect of the absolute distance through the backbone (in units of residues) on the propensity of each residue pair to be in contact; *Hydophobicity* indicates the effect of hydrophobicity (as measured by the scale of Kyte and Doolittle, 1982) on the overall propensity of each residue to form contacts; *Charge Mixing* indicates the effect of like or unlike charges to be respectively in contact or not in contact (for charged residues); *Polar/Nonpolar Mixing* indicates the propensity of polar residues to be in contact with nonpolar residues; *Polar/Polar* mixing indicates the propensity of polar residues to be in contact with other polar residues; *Volume* indicates the effect of residue van der Waals volume (in Å^3) on the propensity to form contacts; *Mass* indicates the effect of residue mass (in Da) on the propensity to form contacts; *Dist from Termini* indicates the effect of residue distance from the nearest terminus (ranging from 1 at the center to 0 at either terminus) on the propensity to form contacts; *GWESP(0.5)* indicates a geometrically weighted shared partner statistic with a decay parameter of 0.5, reflecting the tendency toward triadic clustering within the PSN; and *Prior Scale* refers to the scale of the Laplace parameter prior (which determines the strength of regularization).

Of the estimated effects, all except for hydrophobicity and mass have 95% credible intervals that do not contain 0, and posterior means for both models are quite similar. Broadly, we may interpret the parameter estimates as follows. The low baseline density (as determined by the edges parameter) is compatible with the general observation that both WT and E22G are generally unstructured, with most residues having few non-backbone contacts at any given time. We observe a mild tendency for residues that are far from each other in the primary sequence to interact; the high flexibility of Aβ implies relatively little inhibition of long-range contacts, however, and the effect is fairly small. As would be expected on physical grounds, electrostatic and nonpolar effects are fairly large (with pairs of nonpolar residues relatively more likely to form contacts than pairs of polar residues or polar/nonpolar pairs). Volume also has a small effect on contact formation, with larger residues being somewhat more likely to have more contacts. Perhaps more interestingly, distance from the nearest terminus (equivalently, placement in the middle of the primary sequence) is a strong positive predictor of the tendency to form contacts, and there is a strong overall tendency toward clustering (as might be expected on geometric grounds). Thus, there is a net bias toward structure formation for the interior of the protein, despite its overall high mobility and lack of persistent secondary structure.

Although these models are highly simplified, they can be thought of as expressing approximate "force fields" describing

TABLE 1 | Posterior estimates for the WT and E22G PSN ERGMs (respectively).

	Wild type				E22G			
Parameter	**Post mean**	**Post SD**	**Q2.5%**	**Q97.5%**	**Post mean**	**Post SD**	**Q2.5%**	**Q97.5%**
Edges	−6.137	0.0719	−6.286	−5.986	−6.356	0.0667	−6.484	−6.218
Backbone dist	−0.025	0.0017	−0.028	−0.021	−0.019	0.0014	−0.021	−0.016
Hydrophobicity	−0.003	0.0038	−0.010	0.005	0.002	0.0039	−0.006	0.009
Charge mix	−0.999	0.0449	−1.083	−0.909	−0.996	0.0533	−1.108	−0.901
Polar/Nonpolar mix	−0.347	0.0320	−0.411	−0.285	−0.365	0.0295	−0.419	−0.308
Polar/Polar mix	−0.512	0.0531	−0.614	−0.410	−0.478	0.0465	−0.571	−0.393
Volume	0.004	0.0007	0.003	0.006	0.003	0.0007	0.001	0.004
Mass	−0.001	0.0007	−0.002	0.001	0.001	0.0007	0.000	0.002
Dist from termini	0.140	0.0239	0.097	0.188	0.190	0.0247	0.145	0.241
GWESP(0.5)	2.137	0.0235	2.090	2.182	2.205	0.0221	2.159	2.246
Prior scale	0.941	0.0102	0.922	0.960	0.958	0.0080	0.941	0.974

the relative favorability of different PSN structures with respect to each variant. Drawing on this intuition, we may use the models to construct a log "favorability ratio" that, for a given PSN, measures the extent to which it is relatively favorable for E22G vs. WT. In particular, let $\hat{\theta}_{WT}$ be the estimated coefficients for the WT model, and $\hat{\theta}_{E22G}$ the corresponding coefficients for the E22G model. Then, for PSN G, the quantity

$$f_{WT}^{E22G}(G) = \hat{\theta}_{E22G} t_{E22G}(G) - \hat{\theta}_{WT} t_{WT}(G) \tag{2}$$

is the log favorability ratio for E22G vs. WT (where t_{E22G} and t_{WT} indicate the vectors of graph statistics for G evaluated for each respective sequence, the two having slightly different residue properties). It may be observed from Equation 1 that $f_{WT}^{E22G}(G)$ is equal to the log ratio of the probability of observing G under the two respective models, up to an additive constant that does not depend upon the PSN. Thus, while the absolute level of $f_{WT}^{E22G}(G)$ cannot be interpreted, differences in the log ratio for different choices of G are meaningful; in particular, if $f_{WT}^{E22G}(G) > f_{WT}^{E22G}(G')$, then PSN G is relatively favored by E22G vs. WT vis a vis G'.

The log favorability ratio provides considerable insight into the types of transient structures that are most heavily favored by E22G relative to WT. For instance, **Figure 11** shows the five PSN structures most favored by E22G and WT, respectively, out of all minima from both (pooled) sets. As can be seen, the minima most favored by E22G involve extensive, cohesive structures, while those favored by WT tend to be extremely sparse (with most structure being local with respect to the backbone). Interestingly, where the wild type-favored PSNs have more extensive structure, it tends to be near the termini (in contrast with E22G, which shows more extensive structure within the interior of the protein). As noted above, both models encourage structure formation within the interior of the primary sequence; however, wild type $A\beta_{1-40}$ appears to favor conformations with terminal structure more than the E22G variant (plausibly because the E22G places far more probability mass on globally cohesive structures that are destabilized in the wild-type protein). Examination of these extreme cases thus gives us an immediate intuition for the nature of the subtle differences in transient structure formation that distinguish the two variants.

2.4. Comparative Cluster Analysis of WT and E22G Dynamics

Cluster analysis is a useful tool for subdividing conformational spaces, having been successfully employed in applications such as refinement of protein structure homology models (Raval et al., 2012), building Markov models for protein folding (Husic and Pande, 2017), and probing the configurational and hydrogen bonding structure of solvating water molecules in confined regions of proteins (Young et al., 2007). Here, we show how cluster solutions calibrated for accurate treatment of conformational dynamics combined with comparative analysis of cluster-induced transition networks can be used to reveal differences in the behavior of the WT and E22G Aβ variants.

2.4.1. Can Differences in Physiological Temperature Dynamics for WT and E22G Be Detected?

It has been shown in the present study and elsewhere (Chebaro et al., 2015; Granata et al., 2015) that the thermodynamics of intrinsically disordered proteins are governed by vast potential energy surfaces with numerous or perhaps innumerable local minima corresponding to nearly isoenergetic microstates, rather than a single well-defined global minimum. This situation makes comparative analysis of thermodynamic distributions for similar IDPs extremely difficult compared to systems where only a few local minima exist. At the same time, experiments have confirmed that even subtly different IDPs, such as the WT and E22G proteins being studied in the present work, do exhibit a marked difference in their capacity to form amyloid fibrils (Lord et al., 2006; Norlin et al., 2012). This sharp contrast between the thermodynamic similarities of WT and E22G and the substantial difference in their behavior under solution conditions strongly suggests that there may be more easily discernible kinetic differences between them. In other words, although the configurations of both systems are distributed very similarly

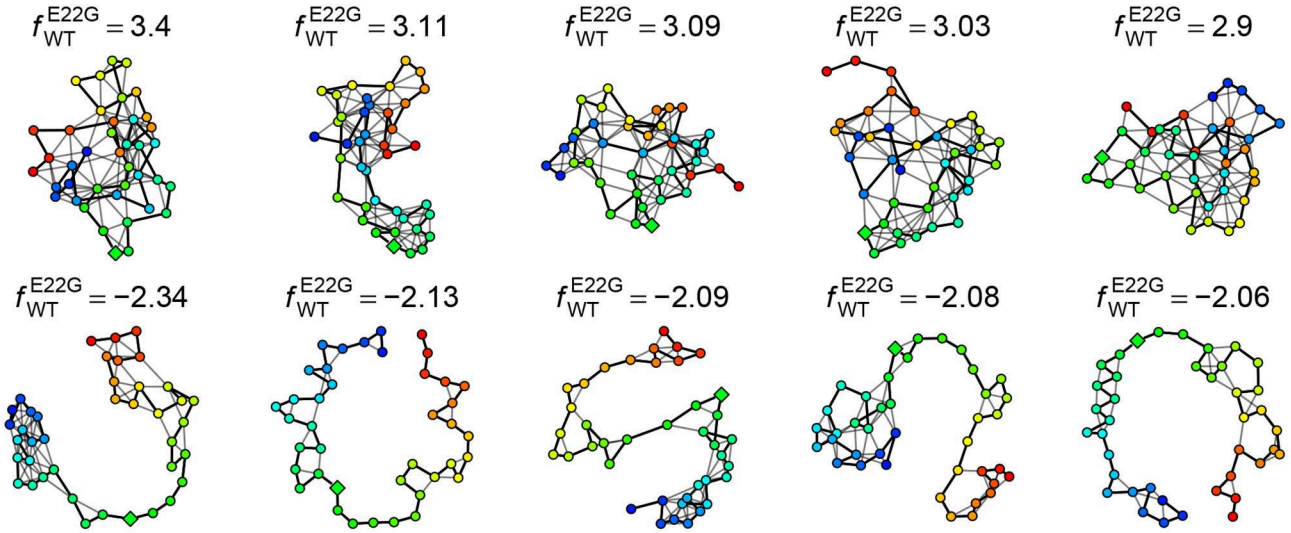


FIGURE 11 | Five PSN structures most heavily favored by E22G **(Top)** and WT **(Bottom)**, respectively. E22G-favored structures are highly cohesive, while those favored by WT have little structure (most of which is local).

when time-marginalized, the way the proteins transition between regions of the conformation space may be distinct.

While the conventional intuition motivating clustering or segmentation of conformational space in the context of protein dynamics is that the protein will restricted to a relatively small number of low free energy basins (with relatively rare transitions over free energy barriers between basins) (Bolhuis et al., 2002), this cannot be assumed for IDPs: while local minima exist, they are extremely numerous and widely dispersed across a relatively flat energy landscape (Granata et al., 2015). However, even without the assumption of well-defined basins, we can segment conformational space into a set of discrete regions and use this as the basis for a coarse-grained treatment of protein dynamics (estimating transition rates from observed simulation trajectories). While many approaches could be used for this purpose, *k*-means clustering (Hartigan and Wong, 1979) on input space of torsion angles is a natural choice: it is highly scalable, adaptively places boundaries around regions of high conformational density, and leads to cells that are both convex and relatively compact. Here, we apply *k*-means clustering (using the R implementation R Core Team, 2018) to trajectories in torsion angle space produced by 500 ns long molecular dynamics simulations (10×10^6 time steps each), jointly clustering WT and E22G to create a shared coarsening of their common conformational space. We then examine the dynamics on this coarsened space to reveal differences between the two systems.

2.4.2. Choosing the Number of Clusters to Optimize Dynamic Accuracy

An important parameter to determine in fitting any *k*-means clustering model is *k*, i.e., the number of clusters the algorithm will generate. One of the most common and straightforward metrics for determining the optimal choice of *k* is to plot the mean squared distance between the data points and their respective cluster centers, a.k.a. an *elbow plot*. For data sets with a strong characteristic number of clusters, a sharp decline in this distance will be observed when *k* is set to that characteristic number of clusters. As shown in **Figure 12A**, the configurations produced by the MD simulations of the WT and E22G variants of $A\beta_{1-40}$ showed no well-defined elbow, a pattern compatible with a widely dispersed range of conformations with no deep potential energy wells. Although somewhat diminishing gains are observed somewhere between $k = 5$ and $k = 10$, this result is by no means conclusive, thus additional metrics for selecting *k* are needed.

Another commonly used metric for finding an optimal value of *k* for k-means clustering is to plot mean silhouette width as a function of *k* and look for a well-defined maximum (Rousseeuw, 1987). The silhouette width of a given data point *i* is defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3}$$

where $a(i)$ is the mean distance between point *i* and all other points within its cluster, $b(i)$ is the mean distance between point *i* and all points in the cluster it is nearest to but to which it does not belong. This equation produces silhouette width values $-1 \geq s(i) \leq 1$, where, on the extremes, 1 indicates ideal cluster membership for point *i* and -1 indicates that *i* has been grouped into the wrong cluster. Silhouette analysis of our system is shown in **Figure 12B**. Although the optimal choice of *k* is clearly greater than 8, again, the standard metric provides evidence for the wide dispersal of conformations, and a need to choose a *k*-selection approach that is tailored for the case of IDP trajectories.

Given that our goal is to segment a continuous conformation space for the purpose of building a coarse-grained approximation to the underlying dynamics, an alternative approach is to estimate the accuracy of the dynamic model produced by a given choice of *k*, and to find the *k* that leads to the

lowest level of approximation error. Intuitively, the error involved in a Markov approximation of the true dynamics is dominated by two terms: the *coarsening loss* due to the approximation of each specific conformation within a voxel by the voxel centroid; and the *transition rate error* associated with imperfect estimation of the inter-voxel transition rates. Given a fixed set of trajectories, it is apparent that the coarsening loss is diminishing in k: the more finely we divide the space, the more accurately each observed conformation is represented. At the same time, however, larger choices of k also reduce the information available to estimate each inter-voxel transition rate, leading to errors that are increasing in k. Minimizing the total error is thus expected to lead to a k that optimizes the trade-off between coarsening and rate estimation errors.

To put these two error sources on an even footing, we unify them by defining a one-step *prediction error* for the coarsened Markov model. Specifically, given an observed conformation within a particular voxel, we predict the next conformation in the trajectory by (1) drawing the next voxel state from the Markov model, and (2) drawng a random conformation from the set of all observed conformations within the voxel. The distance between this drawn conformation and the observed next conformation is the one-step prediction error. Minimizing this error (summed over all observed transitions) automatically incorporates both the coarsening loss and the transition rate error, in a manner that is conceptually true to our end goal (approximating complex, high-dimensional conformational trajectories with a coarse-grained Markov model).

The one-step prediction error summed over all trajectories is referred to as the *total Markov error,* and is computed as follows. First, assume a set of observed trajectories, a clustering solution, and an estimated transition rate matrix. Next, begin with the first observed conformation, and proceed as follows:

1. Taking the current cluster ID as input to the Markov model, predict the cluster membership of the next time point.
2. Draw a configuration from the cluster into which the model predicted a transition.
3. Measure the distance between the predicted configuration and the actual configuration for that time step, and add that distance to the total Markov error.
4. Repeat steps 1 through 3 for the remainder of the trajectory, and either repeat with the next trajectory if any remain or else return the TME for that model.

The TME metric for k-means clustering was applied to 20 separate k-means model fitting calculations, varying k from 2 through 16 and averaged to produce the plot shown in **Figure 12C**. The metric shows a well-defined optimum at $k = 11$, where the total Markov error is at a minimum. The TME methodology implicitly strikes a balance in bias-variance tradeoff between the extremes of too few clusters, where transition frequencies are more likely to be well-sampled but the configuration draws from step 2 are drawn from higher variance clusters, and too many clusters, where smaller clusters have lower variance but under-sampling of transitions imparts a bias to the random walks in cluster space.

2.4.3. Transition Frequency Graphs From k-Means Clustering of WT and E22G Trajectories

Once the optimal number of clusters of $k = 11$ was identified using the total Markov error metric, the lowest TME of the 20 k-means models with $k = 11$ was selected for further analysis. The matrices of transition frequencies between clusters (see section 4) are ideally represented using graphs (**Figures 13**, **14**). A few key observations are immediately apparent when comparing **Figure 13** with **Figure 14**. The E22G graph displays a much higher degree of connectivity compared to WT, with more evenly distributed populations across the clusters visited along its trajectory. Notably, cluster number 6, the highest populated cluster in the E22G transition graph, is both highly connected and minimally populated in the WT graph. This is noteworthy because although transitions were observed between cluster 6 and 9 of the 10 other clusters present in the WT trajectories, the trajectories did not remain in cluster 6 long enough to produce a more substantial population in that cluster. This implies that while cluster 6 is highly accessible to both WT and E22G variants, E22G appears to exhibit substantially higher stability in this region of configuration space.

Given the sharp contrast between the transition frequency graphs in **Figures 13**, **14**, it is necessary to examine the possibility that the difference in configuration space sampling is due to the trajectories being too short. More specifically, since the configuration space of Aβ_{1-40} is believed to be expansive, it is necessary to demonstrate that the observed differences are not occuring because the two variants simply did not have time to cover the distance between the configuration subspaces favored by one vs. the other. As a way to address this, we generated the cluster proximity graph shown in **Figure 15**. It is immediately obvious that this is a very well-connected graph, with many of the strongest ties occurring between vertices whose populations are dominated by differing variants. For example, note that most of the strongest ties in the graph are between nodes of substantially different relative populations of WT vs. E22G. As a specific case, consider the three most WT-dominant nodes on the graph, nodes 4, 8, and 9: all exhibit some strong ties, yet none of their respective strong ties are shared between each other. The cluster center proximity graph provides strong evidence that the disparity between the clusters sampled in the WT and E22G simulations are indeed inherent to their respective dynamics, and not an artifact of under-sampling.

3. DISCUSSION

This comparative study of the wild type Aβ_{1-40} protein and its "Arctic" E22G variant identifies some key differences in the types of transient structures formed by monomers of the disease-related variant. Although the Ramachandran plots and angular velocity distributions of MD trajectories for these proteins are essentially identical, SVM analysis finds key sets of torsion angles that are indicative of conformations that are more characteristic of either wild-type or E22G. Combining this approach with PCA provides a more detailed view of the differences in transient structural motifs formed, namely

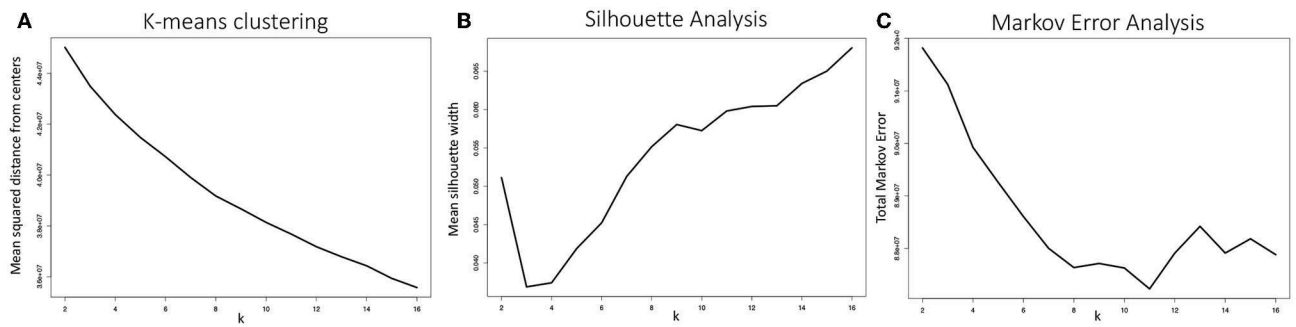


FIGURE 12 | (A) Mean distance from cluster centers as a function of the number of clusters generated k, often called an elbow plot, is a typical metric for determining the ideal number of clusters to use in k-means clustering. In strongly clustered systems, the point where this mean distance drops abruptly, or the "elbow," is the optimal value for k. Here we show the mean elbow plot for 20 separate clustering calculations using k values ranging from 2 through 16. **(B)** Mean silhouette width as a function of k is another common metric for choosing an optimal value of k. Although the metric shows that the optimal k value ought to be greater than 6, no clear optimum is demonstrated. The silhouette widths shown were generated by sampling 10 sets of 5,000 silhouette width values from all 20 clustering calculations and averaging. **(C)** Our metric for optimal k selection, whereby the quality of a given clustering was evaluated on the basis of how well a Markov model generated from a transition frequency matrix of transitions between clusters performed at predicting the actual trajectory's path through torsion angle configuration space. Here we see a clear optimum: $k = 11$, where the total Markov error metric is at a minimum.

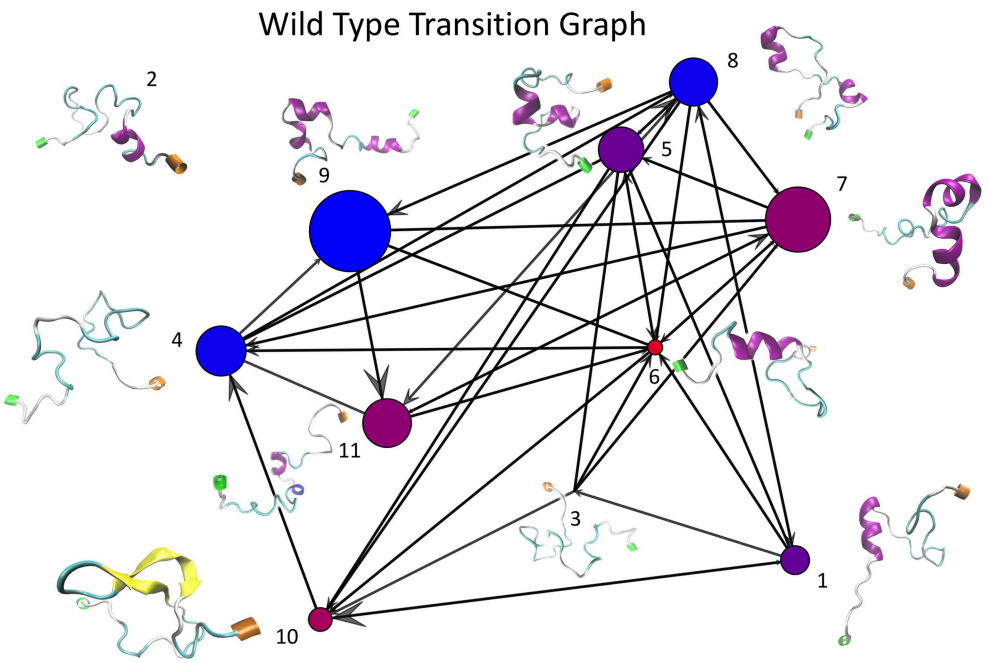


FIGURE 13 | This graph of transition frequencies between clusters 1 through 11 was generated by clustering the trajectory in torsion angle space from a 500 ns simulation of wild type $A\beta_{1-40}$. This particular clustering model was selected due to it having the minimum total Markov error (sum of both WT and E22G) of all 20 replicates having $k = 11$. Each vertex represents one cluster, edges indicate that transitions were observed between each respective pair of vertices, and molecular structures shown are the structures from the trajectories nearest each cluster center (N termini are labeled green and C termini are labeled orange). Note that this is a directed graph, thus the size of the arrowheads are proportional to the number of transitions that occurred in that direction. The size of each vertex is proportional to the number of configurations from the WT trajectory belonging to that cluster, and the color of each vertex represents the relative populations of WT vs. E22G in the aggregate of both trajectories (the bluer the more WT, the redder the more E22G).

the amount of helical character in the vicinity of residues 14-23 and the amount of contact between the C-terminal region and other parts of the protein. Comparisons of the similarity scores for the wild-type and E22G monomers investigated here with known fibril structures from the Protein Data Bank reveal that most of the known fibril structures occupy more wild-type like conformations, suggesting that E22G may fibrillize into a different topology, a hypothesis that is consistent with morphological differences in experimentally observed fibrils (Norlin et al., 2012), although high-resolution structures have

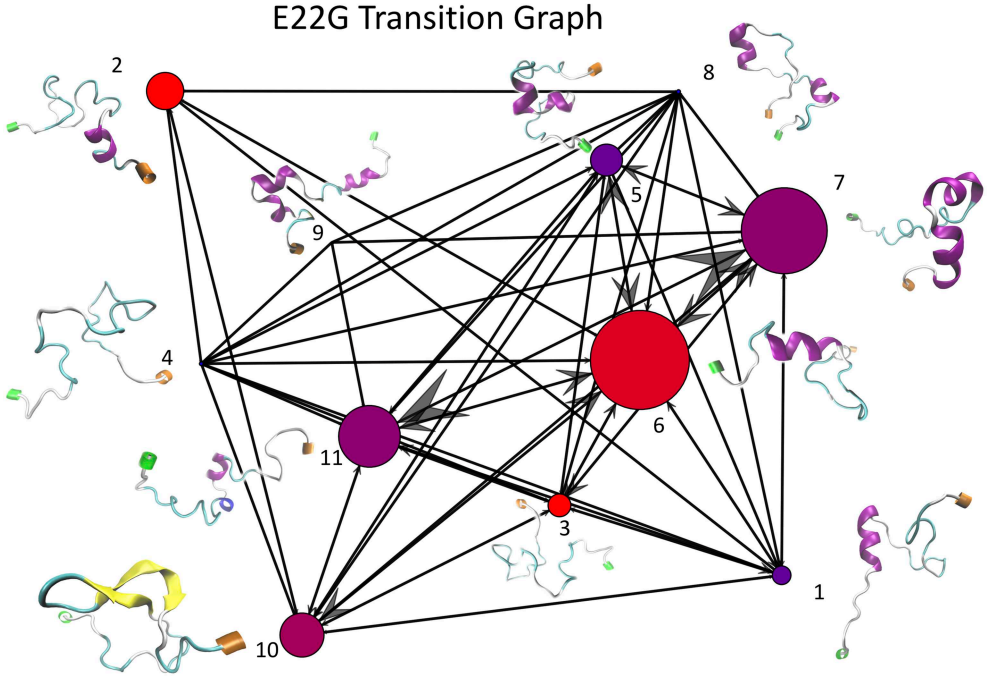


FIGURE 14 | This graph of transition frequencies between clusters 1 through 11 was generated by clustering the trajectory in torsion angle space from a 500 ns simulation of E22G mutants of Aβ_{1-40}. The clustering model used is the same as that shown in **Figure 13**, only applied to the E22G trajectory instead. The image convention used is also the same as that of **Figure 13**.

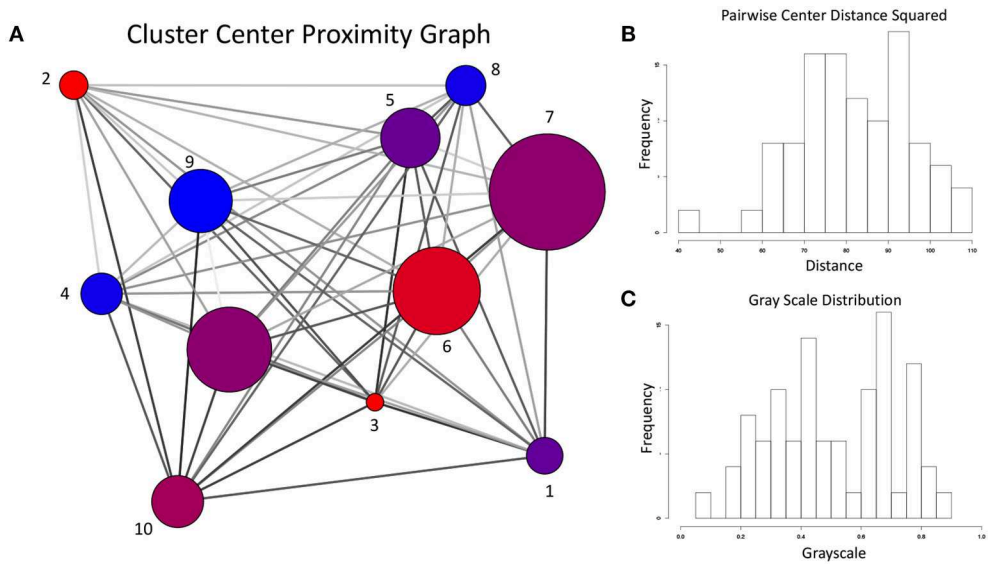


FIGURE 15 | (A) For the cluster center proximity graph, the vertex color follows the same convention as **Figures 13**, **14**, while vertex size is proportional to the aggregate population (both WT and E22G). The edges of this graph represent the proximity between their respective vertices in configuration space (the darker the edge, the closer the vertices are positioned in torsion angle space). **(B)** Distribution of the pairwise distances between cluster centers. **(C)** Distribution of grayscale values (0 is black, 1 is white) for the edges in **(A)** was produced by mapping the pairwise distances with a sigmoidal function. Note that the shape of the distribution is very close to that of **(B)**, indicating an accurate representation of the proximity data in the grayscale.

not yet been solved for this variant. The previously discussed results approach the structures from torsion angle space, which is a convenient representation of backbone conformations, but does not address intramolecular connectivity. Protein structure networks (PSNs) enable a parsimonious representation of local and long-range cohesion. We find that the mean degree core

number, a measure of each residue's connectivity, is significantly higher for most residues in E22G compared to wild-type, with particularly large differences observed in residues G15 to M35. This region of enhanced structural cohesion in the E22G variant may represent a nucleation site for the formation of pathological aggregates. PSN analysis of the five structures most favored by wild-type vs. E22G shows that the former prefers much sparser, extended structures, while the latter is prone to compact, densely connected conformations. Overall, this enhanced propensity of E22G to form denser patterns of inter-residue contacts, even if these species exist only transiently, is indicative of its increased susceptibility to aggregation. Our results not only provide insight into this protein system, but also illustrate a more general approach that can be applied to comparative analysis of intrinsically disordered proteins in other settings. While a strong precedent exists for applying frameworks devised for characterizing proteins with well-defined folded states, like DSSP (definition of Secondary Structure of Protein Joosten et al., 2010), toward characterizing the transient structure present in IDPs (Rosenman et al., 2013), we present a methodology that allows the latent structure of the data itself to define the metrics for similarity or difference between variants. Our approach does not risk the confirmation bias that can result from applying methods that search for a particular known type of order in an intrinsically disordered system. Rather, the ML-based methods shown herein search for the most predictive latent structure in the data and then maps that structure onto some intuitable paradigm. In most supervised machine learning applications, the goal is to train a classifier or regression model that can be used to make predictions on future data points after being trained on training data from the past. We have demonstrated that tools from the ML toolkit, such as SVMs and clustering algorithms, can be used in ways that go beyond traditional "black box" approaches, and instead be used to answer mechanistic questions about *how* and *why* subtle structural differences in complex systems like IDPs can lead to markedly different dynamics. Although fitting the models remains an important step in the present work, the utility of a well-trained ML model goes beyond being able to make accurate predictions. Using our approach, the fact that we are able to train a model to accurately classify or group structures as having WT or E22G character, given the training data, serves as an indication that the input data is indeed a set of sufficient statistics for discerning between the classes of interest. This is a key piece of information for molecular simulations in general, as one must always be wary that an inconclusive result is due to the inherent problems of molecular simulations, such as under-sampling or insufficiently detailed models. For example, in the case of the present work, wild-type Aβ and the E22G variant are known to exhibit radically different fibrillization dynamics on experimentally accessible size and time scales, yet standard approaches to analysis of MD simulations of these systems show little to no difference in their behavior (e.g., the Ramachandran plot if **Figure 1**). As is the case for MD-based study, when standard methods of analysis are inconclusive, a legitimate concern is that lack of detail in the MD forcefields and/or under-sampling could be to blame for the inability to differentiate between WT and E22G dynamical data with the standard methods. By using multiple ML approaches to first prove that indeed enough simulation data is present to reliably differentiate between variants, and subsequently probe the ML models themselves to determine which input characteristics and even which specific configuration data points were most informative, we have demonstrated that our ML-based methods can be used to simultaneously verify the adequacy of the sampling while providing a less biased interpretation of the dynamics of intrinsically disordered proteins.

While there is no one-size-fits-all approach for characterizing the transient structure of IDPs–different questions demand different representations–we would suggest that several methods shown here are likely to prove widely useful in practice. As noted, we find residue-level PSNs to provide a fairly simple way to represent transient structure that complements traditional, secondary structure-based methods while capturing features that are hard to express via the latter. Measures of local cohesion (like the core numbers used here) are easily computed, and provide immediate insight into which regions of the protein tend to occupy locally folded conformations; comparing these measures across variants allows the impact of mutations on transient structure to be assessed without requiring formation of recognizable secondary structure. Model-based analysis of PSN structure using ERGMs is more complex, but provides a powerful tool for identifying transient structures that are differentially favored across variants. Given the rich analytic toolkit developed for the study of social networks (Wasserman and Faust, 1994; Brandes and Erlebach, 2005; Butts, 2008a) (which are themselves characterized by irregular and often transient structure), this would seem to be an area with substantial potential for further development.

4. MATERIALS AND METHODS

4.1. Molecular Dynamics Simulation of AβMonomers

All MD simulations $A\beta_{1-40}$ monomers were carried out using the NAMD 2.10 molecular dynamics software package (Phillips et al., 2005) with the CHARMM36 force field (Best et al., 2012) in Generalized Born implicit solvent (Qiu et al., 1997) with an electrostatic interaction cutoff of 14Å, an alpha (i.e., descreening) cutoff of 12Å, a 2fs step size, and an ionic concentration of 0.1M; except as noted below, all simulations were performed at constant temperature using a Langevin thermostat with a damping coefficient of 1/ps. The seed structure for WT $A\beta_{1-40}$ was taken from the lowest energy conformation of the monomeric solution structure of (Paravastu et al., 2008) (PDB: 2LMN). The seed structure for the E22G variant was obtained via homology modeling using SWISS-MODEL (Schwede et al., 2003) (template PDB 2M4J Lu et al., 2013). Visualizations of the molecular structures were generated using the VMD software package (Humphrey et al., 1996), with additional processing performed using R (R Core Team, 2018).

4.1.1. Identification of Local Minima

To obtain an overdispersed set of seed conformations, 100 ns MD simulations at 450K were carried out for WT and E22G,

respectively using the above protocol; 1,000 conformations were collected in each case (1 per 100 ps), with the first being discarded and the rest being retained for subsequent analysis. Each conformation obtained from the above process was then used to seed a 1ns annealing trajectory in which temperature was systematically lowered from 310K to 0K by constant increments of 1K (i.e., with approximately 1,600 time steps between increments) using velocity reinitialization (no Langevin thermostat). The final conformation from each of 1998 annealing runs was retained as a local minimum for further analysis (resulting in 999 minima for each of WT and E22G, respectively).

4.1.2. Simulation of Conformations and Angular Velocities from Dispersed Starting Points

To sample Aβ_{1-40} conformations across a wide range of conformation space, we use the above-identified local minima as seeds for short secondary trajectories at physiological temperature. For each minimum, we simulated 10 independent trajectories at 310K, using our base protocol. Each trajectory was simulated for 50 intervals of 2 ps, separated by "bursts" in which conformations were recorded 10 times separated by intervals of 20 fs. This resulted in a total length per trajectory of approximately 110 ps. In total, 9,990 trajectories were simulated for each of WT and E22G, with approximately 500,000 10-configuration "bursts" recorded for analysis. Mean angular velocities were then estimated for each burst by taking the mean of the circular (angular) difference between frames on each torsion angle and dividing by the interval between frames.

4.1.3. Simulation of Dynamics at Physiological Temperature

To examine longer-range $A\beta_{1-40}$ dynamics at physiological temperature, independent trajectories using our base protocol were simulated for WT and E22G at 310K for 500 ns. 250,000 conformations (1/2ps) were retained from each trajectory for subsequent analysis.

4.2. Support Vector Machine Analysis of Low-Energy Conformations

Backbone dihedral angles were obtained for all local minima using a combination of R and VMD scripts; for subsequent analysis, each torsion angle was represented via its real and imaginary components (for a total of 160 input features per conformation). SVM analysis was performed using the e1071 package for R (Meyer et al., 2018), using a Gaussian (aka radial basis function) kernel. Hyperparameter tuning for the kernel bandwidth and cost parameters was performed via a grid search using 10-fold cross-validation. For local analysis of mean angular differences across the decision surface, the set of all support vectors for the SVM solution was obtained and sorted into matched E22G/WT pairs by Euclidean distance in the input space (with the closest pair being matched first, then the next closest, and so on until no pairs remained). Angular (i.e., minimum circular) differences were then computed for the torsion angles in each pair, expressed as the angular displacement needed to go from the WT angle to its E22G counterpart (in radians).

For analysis involving fibrillar conformations, all models were extracted from PDB Berman et al. (2000) entries 2LMN (Paravastu et al., 2008), 2LMO (Paravastu et al., 2008), 2LMP (Paravastu et al., 2008), 2LMQ (Paravastu et al., 2008), 2LNQ (Qiang et al., 2012), 2M4J (Lu et al., 2013), 2MPZ (Sgourakis et al., 2015), 2MXU (Xiao et al., 2015), 5KK3 (Colvin et al., 2016), and 5OQV (Gremer et al., 2017). The conformation of each monomer in each fibril structure was extracted and converted to torsion angle features as described above. Because many reported structures were missing most or all of the N-terminal residues, we limited analysis to residues 15-40. A second SVM solution was obtained from the minima using only these residues using the above protocol, which was employed for this analysis. The projection of each fibril onto the feature space vector normal to the separating hyperplane (the "affinity score") was performed by obtaining the decision value for the classification prediction (E22G vs. WT) for each fibrillar conformation. To obtain information on the mean gradient of the affinity score over the fibrillar conformations, scores were regressed on the input features of the conformations; the resulting coefficients estimate the mean gradient of the affinity score for the real and imaginary portions (respectively) of each torsion angle, averaged across conformations. For visualization, the two coefficients for each torsion angle were transformed into modulus/argument representation [i.e., for torsion angle θ_i, $\beta_i sin(\theta_i) + \beta_i' cos(\theta_i) = b_i sin(\theta_i - y_i)$ with $b_i = \sqrt{\beta_i{}^2 + \beta_i'^2}$ and $y_i = \tan^{-1}(\beta_i'/\beta_i)$]. All calculations were performed using R (R Core Team, 2018).

4.3. Protein Structure Network Analysis

Residue-level PSNs were obtained for each local minimum conformation by calculating distances among all atom pairs and forming an edge between residues r_i and r_j if there existed atoms $a_i \in r_i$, $a_j \in r_j$ such that the a_i, a_j distance was smaller than 1.2 times the sum of their van der Waals radii. All analysis and visualization was performed using R and statnet (Handcock et al., 2008; R Core Team, 2018); van der Waals radii were taken from Alvarez (2013). k-cores were calculated for all PSNs using the sna library for R (Butts, 2008b).

ERGM estimation was performed using an approximate Bayesian procedure building on the approach of Desmarais and Cranmer (2012). We independently estimate a model for each sample of PSNs, with the structure

$$\begin{aligned} \sigma &\sim \text{Inv} - \text{Gamma}(\kappa, \zeta) \\ \theta_1, \ldots, \theta_p &\sim \text{Laplace}(0, \sigma) \\ Y_1, \ldots, Y_n &\sim \text{ERGM}(\theta, X), \end{aligned}$$

where σ is the prior scale (with hyperparameters κ and ζ), $\theta = (\theta_1, \ldots, \theta_p)$ is the vector of ERGM coefficients, $Y = (Y_1, \ldots, Y_n)$ is a PSN sample, and X is a set of protein-specific covariates (e.g., residue properties). Draws at each level are taken to be conditionally independent. Intuitively, this model is a Bayesian analog to the LASSO procedure applied to a pooled ERGM, with the Laplace parameter priors inducing the equivalent of L1 regularization on the posterior mode. (To improve regularization performance, we rescale the

changescores associated with θ to unit variance during the estimation process, so that each coefficient is on the same scale; reported estimates have been returned to the original scale). Because direct posterior simulation for this model would be prohibitively computationally expensive on the large sample of networks used here, we instead employ an approximate inference strategy closely related to that of Schmid and Desmarais (2017) for single networks and Desmarais and Cranmer (2012) dynamic networks. Our approach proceeds as follows. For a specific sample, Y, we approximate the posterior mode $\theta|Y, X$ by numerically maximizing the quantity

$$\int_0^\infty p(\theta|\sigma)p(\sigma|\kappa,\zeta)\prod_{i=1}^{n}\mathcal{PL}(Y_i|\theta,X)d\sigma$$

where $\mathcal{PL}$ is the conditional *pseudo-likelihood* of Y_i (Strauss and Ikeda, 1990) given the constraint that all residues must be adjacent to their neighbors along the protein backbone. The pseudo-likelihood is an easily calculated approximation to the exact ERGM likelihood whose mode, for large conditionally independent samples, approaches that of the true likelihood (Strauss and Ikeda, 1990). To obtain approximate posterior quantities, we then perform Bayesian bootstrap (Rubin, 1981) simulation of $\theta|Y^{(j)}, X$ over replicates $Y^{(1)}, \ldots, Y^{(m)}$ of the original data set (with graphs as the independently resampled units). We report approximate posterior mean, standard deviations, and 95% credible intervals obtained through this procedure for θ and σ.

Model terms used for the PSN ERGM analysis were computed using a combination of R scripts and tools within the ergm statnet package (Hunter et al., 2008); descriptions for model terms used here follow e.g., Morris et al. (2008). A standard *edges* term was used as a density offset, with an *absdiff* term for distance along the backbone, and a *nodemix* for polar/nonpolar interaction (with nonpolar/nonpolar as the reference category). Electrostatics were implemented via an *edgecov* term with a covariate matrix Z such that $Z_{ij} = 1$ if r_i and r_j have the same nonzero charge, $Z_{ij} = -1$ if r_i and r_j have the different nonzero charge, and $Z_{ij} = 0$ if either r_i or r_j are uncharged. *nodecov* terms were included for hydrophobicity (using the scale of Kyte and Doolittle, 1982), residue volume (in Å^3), residue mass (in Da), and residue-wise distance from the nearest terminus (scaled from 0 to 1). Finally, we account for endogenous clustering using a fixed-decay geometrically weighted edgewise shared partner term (*GWESP(0.5)*). For the Laplace scale, we employ a minimally informative (i.e., diffuse) hyperprior ($\kappa = 0.1, \zeta = 1.1$).

Computation for the log relative favorability ratio was performed for each PSN by calculating the model statistics (i.e., terms) for the adjacency structure of the PSN under the respective residue properties of each variant and then multiplying by their respective parameter estimates per equation 2. f_{WT}^{E22G} was then calculated for all WT and E22G minima PSNs, with the highest and lowest scoring configurations (respectively) being chosen for visualization.

4.4. Comparative Cluster Analysis of WT and E22G Dynamics

All k-means clustering was carried out using the standard R implementation of k-means clustering (R Core Team, 2018). Torsion angle vectors used to define the configuration space were expanded into real and imaginary components, as outlined in section 4.2. The Markov models for the total Markov error metric were generated matrices of transition frequencies by defining a Jeffreys prior on each row, with the observed transitions for that row treated as multinomial data, leading to a posterior mean for the c_{ij} transition of $(Z_{ij} + 0.5)/(N_i + k/2)$, where N_i is the number of cluster pairs starting in c_i and Z_{ij} is the total number of transitions from cluster i to cluster j.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

Simulation, analysis, and method development were performed by GG and CB. GG, CB, and RM wrote the paper.

FUNDING

This work was supported by NSF award DMS-1361425.

REFERENCES

Alvarez, S. (2013). A cartography of the van der Waals territories. *Dalton Trans.* 42, 8617–8636. doi: 10.1039/c3dt50599e

Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80, 505–515. doi: 10.1016/S0006-3495(01)76033-X

Benson, N. C., and Daggett, V. (2012). A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *J. Bioinform. Comput. Biol.* 10:1250008. doi: 10.1142/S0219720012500084

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bernard, C., Houben, K., Derix, N., Marks, D., van der Horst, M., Hellingwerf, K., et al. (2005). The solution structure of a transient photoreceptor intermediate: δ25 photoactive yellow protein. *Structure* 13, 953–962. doi: 10.1016/j.str.2005.04.017

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ, and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* 8, 3257–3273. doi: 10.1021/ct300400x

Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. (2002). Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* 53, 291–318. doi: 10.1146/annurev.physchem.53.082301.113146

Bouvier, G., Desdouits, N., Ferber, M., Blondel, A., and Nilges, M. (2014). An automatic tool to analyze and cluster macromolecular

conformations based on self-organizing maps. *Bioinformatics* 31, 1490–1492. doi: 10.1093/bioinformatics/btu849

Brandes, U., and Erlebach, T., editors (2005). *Network Analysis: Methodological Foundations*. Berlin: Springer-Verlag.

Brinda, K., and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophys. J.* 89, 4159–4170. doi: 10.1529/biophysj.105.064485

Butts, C. T. (2008a). Social networks: a methodological introduction. *Asian J. Soc. Psychol.* 11, 13–41. doi: 10.1111/j.1467-839X.2007.00241.x

Butts, C. T. (2008b). Social network analysis with SNA. *J. Stat. Softw.* 24, 1–51. doi: 10.18637/jss.v024.i06

Butts, C. T., Zhang, X., Kelly, J. E., Roskamp, K. W., Unhelkar, M. H., Freites, J. A., et al. (2016). Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Comput. Struct. Biotechnol. J.* 14, 271–282. doi: 10.1016/j.csbj.2016.05.003

Cecchini, M., Curcio, R., Pappalardo, M., Melki, R., and Caflisch, A. (2006). A molecular dynamics approach to the structural characterization of amyloid aggregation. *J. Mol. Biol.* 357, 1306–1321. doi: 10.1016/j.jmb.2006.01.009

Chebaro, Y., Ballard, A. J., Chakraborty, D., and Wales, D. J. (2015). Intrinsically disordered energy landscapes. *Sci. Rep.* 5:10386. doi: 10.1038/srep10386

Colvin, M. T., Silvers, R., Ni, Q. Z., Can, T. V., Sergeyev, I., Rosay, M., et al. (2016). Atomic resolution structure of monomorphic Aβ42 amyloid fibrils. *J. Am. Chem. Soc.* 138, 9663–9674. doi: 10.1021/jacs.6b05129

Csermely, P., Singh Sandhu, K., Hazai, E., Hoksza, Z., Kiss, H. J., Miozzo, F., et al. (2012). Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function: hypotheses and a comprehensive review. *Curr. Prot. Peptide Sci.* 13, 19–33. doi: 10.2174/138920312799277992

Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., and Dobson, C. M. (2005). Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Amer. Chem. Soc.* 127, 476–477. doi: 10.1021/ja044834j

Desmarais, B. A., and Cranmer, S. J. (2012). Statistical mechanics of networks: estimation and uncertainty. *Physica A* 391, 1865–1876. doi: 10.1016/j.physa.2011.10.018

Duong, V. T., Unhelkar, M. H., Kelly, J. E., Kim, S. H., Butts, C. T., and Martin, R. W. (2018). Network analysis provides insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*. *Integr. Biol.* 10, 768–779. doi: 10.1039/C8IB00140E

Fratev, F., Mihaylova, E., and Pajeva, I. (2014). Combination of genetic screening and molecular dynamics as a useful tool for identification of disease-related mutations: zasp pdz domain g54s mutation case. *J. Chem. Inform. Model.* 54, 1524–1536. doi: 10.1021/ci5001136

Granata, D., Baftizadeh, F., Habchi, J., Galvagnion, C., De Simone, A., Camilloni, C., et al. (2015). The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Sci. Rep.* 5:15449. doi: 10.1038/srep15449

Grazioli, G., Butts, C. T., and Andricioaei, I. (2017). Automated placement of interfaces in conformational kinetics calculations using machine learning. *J. Chem. Phys.* 147:152727. doi: 10.1063/1.4989857

Gremer, L., Schölzel, D., Schenk, C., Reinartz, E., Labahn, J., Ravelli, R. B., et al. (2017). Fibril structure of amyloid-β (1–42) by cryo–electron microscopy. *Science* 358, 116–119. doi: 10.1126/science.aao2825

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). statnet: software tools for the representation, visualization, analysis and simulation of network data. *J. Stat. Softw.* 24:1548. doi: 10.18637/jss.v024.i01

Hartigan, J. A., and Wong, M. A. (1979). Algorithm as 136: a k-means clustering algorithm. *J. R. Stat. Soc.* 28, 100–108. doi: 10.2307/2346830

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). ergm: a package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* 24:nihpa54860. doi: 10.18637/jss.v024.i03

Husic, B. E., and Pande, V. S. (2017). Ward clustering improves cross-validated Markov state models of protein folding. *J. Chem. Theory Comput.* 13, 963–967. doi: 10.1021/acs.jctc.6b01238

Iakoucheva, L., Brown, C., Lawson, J., Obradović, Z., and Dunker, A. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 323, 573–584. doi: 10.1016/S0022-2836(02)00969-5

Jacobs, D. J., Rader, A., Kuhn, L. A., and Thorpe, M. (2001). Protein flexibility predictions using graph theory. *Proteins* 44, 150–165. doi: 10.1002/prot.1081

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., et al. (2010). A series of pdb related databases for everyday needs. *Nucl. Acids Res.* 39(Suppl. 1):D411–D419. doi: 10.1093/nar/gkq1105

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157, 105–132. doi: 10.1016/0022-2836(82)90515-0

Lam, A., Teplow, D., Stanley, H., and Urbanc, B. (2008). Effects of the arctic (e22$\rightarrow$g) mutation on amyloid β-protein folding: discrete molecular dynamics study. *J. Am. Chem. Soc.* 130, 17413–17422. doi: 10.1021/ja804984h

Lee, C., Kalmar, L., Xue, B., Tompa, P., Daughdrill, G., Uversky, V. N., et al. (2014). Contribution of proline to the pre-structuring tendency of transient helical secondary structure elements in intrinsically disordered proteins. *Biochim. Biophys. Acta* 1840, 993–1003. doi: 10.1016/j.bbagen.2013.10.042

Lord, A., Kalimo, Hannuand Eckman, C., Zhang, X.-Q., Lannfelt, L., and Nilsson, L. N. (2006). The Arctic Alzheimer mutation facilitates early intraneuronal Aβ aggregation and senile plaque formation in transgenic mice. *Neurobiol. Aging* 27, 67–77. doi: 10.1016/j.neurobiolaging.2004.12.007

Lu, J.-X., Qiang, W., Yau, W.-M., Schwieters, C. D., Meredith, S. C., and Tycko, R. (2013). Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* 154, 1257–1268. doi: 10.1016/j.cell.2013.08.035

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-0.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of exponential-family random graph models: terms and computational aspects. *J. Stat. Softw.* 24, 1–24. doi: 10.18637/jss.v024.i04

Nilsberth, C., Westlind-Danielsson, A., Eckman, C., Condron, M., Axelman, K., Forsell, C., et al. (2001). The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced A β protofibril formation. *Nat. Neurosci.* 4, 887–893. doi: 10.1038/nn0901-887

Norlin, N., Hellberg, M., Filippov, A., Sousa, A. A., Gröbner, G., Leapman, R. D., et al. (2012). Aggregation and fibril morphology of the Arctic mutation of Alzheimer's Aβ peptide by CD, TEM, STEM and *in situ* AFM. *J. Struct. Biol.* 180, 174–189. doi: 10.1016/j.jsb.2012.06.010

Paravastu, A. K., Leapman, R. D., Yau, W.-M., and Tycko, R. (2008). Molecular structural basis for polymorphism in Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18349–18354. doi: 10.1073/pnas.0806270105

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Mag. J. Sci.* 2, 559–572. doi: 10.1080/14786440109462720

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with namd. *J. Comput. Chem.* 26, 1781–1802. doi: 10.1002/jcc.20289

Qiang, W., Yau, W.-M., Luo, Y., Mattson, M. P., and Tycko, R. (2012). Antiparallel β-sheet architecture in Iowa-mutant β-amyloid fibrils. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4443–4448. doi: 10.1073/pnas.1111305109

Qiu, D., Shenkin, P. S., Hollinger, F. P., and Still, W. C. (1997). The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *J. Phys. Chem. A* 101, 3005–3014. doi: 10.1021/jp961992r

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Statistical Software Package.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *J. Chem. Inform. Model.* 57, 942–957. doi: 10.1021/acs.jcim.6b00740

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* 80, 2071–2079. doi: 10.1002/prot.24098

Roccatano, D., Sbardella, G., Aschi, M., Amicosante, G., Bossa, C., Di Nola, A., et al. (2005). Dynamical aspects of tem-1 β-lactamase probed

by molecular dynamics. *J. Comput. Aided Mol. Design* 19, 329–340. doi: 10.1007/s10822-005-7003-0

Rosenman, D. J., Connors, C. R., Chen, W., Wang, C., and García, A. E. (2013). Aβ monomers transiently sample oligomer and fibril-like configurations: ensemble characterization using a combined MD/NMR approach. *J. Mol. Biol.* 425, 3338–3359. doi: 10.1016/j.jmb.2013.06.021

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Rubin, D. B. (1981). The bayesian bootstrap. *Ann. Stat.* 9, 130–134. doi: 10.1214/aos/1176345338

Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M., Zweckstetter, M., et al. (2010). NMR characterization of long-range order in intrinsically disordered proteins. *J. Am. Chem. Soc.* 132, 8407–8418. doi: 10.1021/ja101645g

Salvi, N., Abyzov, A., and Blackledge, M. (2016). Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J. Phys. Chem. Lett.* 7, 2483–2489. doi: 10.1021/acs.jpclett.6b00885

Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365. doi: 10.1126/science.aat2663

Schmid, C. S., and Desmarais, B. A. (2017). "Exponential random graph models with big networks: maximum pseudolikelihood estimation and the parametric bootstrap," in *IEEE International Conference on Big Data* (Boston, MA), 116–121.

Scholkopf, B., Mika, S., Burges, C. J., Knirsch, P., Muller, K.-R., Ratsch, G., et al. (1999). Input space versus feature space in kernel-based methods. *IEEE Trans. Neural Netw.* 10, 1000–1017. doi: 10.1109/72.788641

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). Swiss-model: an automated protein homology-modeling server. *Nucl. Acids Res.* 31, 3381–3385. doi: 10.1093/nar/gkg520

Sgourakis, N. G., Yau, W.-M., and Qiang, W. (2015). Modeling an in-register, parallel "Iowa" Aβ fibril structure using solid-state NMR data from labeled samples with Rosetta. *Structure* 23, 216–227. doi: 10.1016/j.str.2014.10.022

Sibille, N., and Bernadó, P. (2012). Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. *Biochem. Soc. Trans.* 40, 955–962. doi: 10.1042/BST20120149

Song, J., Guo, L.-W., Muradov, H., Artemyev, N. O., Ruoho, A. E., and Markley, J. L. (2008). Intrinsically disordered γ-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1505–1510. doi: 10.1073/pnas.0709558105

Spera, S., and Bax, A. (1991). Empirical correlation between protein backbone conformation and Cα and Cβ ^{13}C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.* 113, 5490–5492. doi: 10.1021/ja00014a071

Strauss, D., and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *J. Am. Stat. Assoc.* 85, 204–212. doi: 10.1080/01621459.1990.10475327

Teilum, K., Kragelund, B., and Poulsen, F. (2002). Transient structure formation in unfolded acyl-coenzyme A-binding protein observed by site-directed spin labelling. *J. Mol. Biol.* 324, 349–357. doi: 10.1016/S0022-2836(02)01039-2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B (Methodol.).* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Unhelkar, M. H., Duong, V. T., Enendu, K. N., Kelly, J. E., Tahir, S., Butts, C. T., et al. (2017). Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*. *Biochim. Biophys. Acta* 1861, 636–643. doi: 10.1016/j.bbagen.2016.12.007

Urbanc, B., Betnel, M., Cruz, L., Bitan, G., and Teplow, D. (2010). Elucidation of amyloid β-protein oligomerization mechanisms: discrete molecular dynamics study. *J. Amer. Chem. Soc.* 132, 4266–4280. doi: 10.1021/ja9096303

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. New York, NY: Springer Science & Business Media.

Vivekanandan, S., Brender, J., Lee, S., and Ramamoorthy, A. (2011). A partially folded structure of amyloid-beta (1-40) in an aqueous environment. *Biochem. Biophys. Res. Commun.* 411, 312–316. doi: 10.1016/j.bbrc.2011.06.133

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Vol. 8. Cambridge, UK: Cambridge University Press.

Williamson, J. A., and Miranker, A. D. (2007). Direct detection of transient αhelical states in islet amyloid polypeptide. *Protein Sci.* 16, 110–117. doi: 10.1110/ps.062486907

Xiao, Y., Ma, B., McElheny, D., Parthasarathy, S., Long, F., Hoshi, M., et al. (2015). Aβ(1-42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* 22, 499–505. doi: 10.1038/nsmb.2991

Yaveroğlu, Ö., Fitzhugh, S., Kurant, M., Markopoulou, A., Butts, C., and Pržulj, N. (2015). ergm.graphlets: a package for erg modeling based on graphlet statistics. *J. Stat. Softw. Articles* 65, 1–29. doi: 10.18637/jss.v065.i12

Young, T., Abel, R., Kim, B., Berne, B. J., and Friesner, R. A. (2007). Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* 104, 808–813. doi: 10.1073/pnas.0610202104

Zanette, C., Bannan, C. C., Bayly, C. I., Fass, J., Gilson, M. K., Shirts, M. R., et al. (2018). Toward learned chemical perception of force field typing rules. *J. Chem. Theory Comput.* 15, 402–423. doi: 10.1021/acs.jctc.8b00821

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 13 August 2019
doi: 10.3389/fmolb.2019.00064

Using Small-Angle Scattering Data and Parametric Machine Learning to Optimize Force Field Parameters for Intrinsically Disordered Proteins

Omar Demerdash[1,2], *Utsab R. Shrestha*[1,2], *Loukas Petridis*[1,2,3], *Jeremy C. Smith*[2,3], *Julie C. Mitchell*[1,2] *and Arvind Ramanathan*[4,5]*

[1] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [2] University of Tennessee/Oak Ridge National Laboratory Center for Molecular Biophysics, Oak Ridge, TN, United States, [3] Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, United States, [4] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [5] Data Science and Learning Division, Argonne National Laboratory, Lemont, IL, United States

OPEN ACCESS

Edited by:
Gennady Verkhivker,
Chapman University, United States

Reviewed by:
Peng Tao,
Southern Methodist University, United States
Carter Tribley Butts,
University of California, Irvine, United States
Vladimir N. Uversky,
University of South Florida, United States

***Correspondence:**
Arvind Ramanathan
ramanathana@anl.gov

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 29 March 2019
Accepted: 16 July 2019
Published: 13 August 2019

Citation:
Demerdash O, Shrestha UR, Petridis L, Smith JC, Mitchell JC and Ramanathan A (2019) Using Small-Angle Scattering Data and Parametric Machine Learning to Optimize Force Field Parameters for Intrinsically Disordered Proteins. Front. Mol. Biosci. 6:64. doi: 10.3389/fmolb.2019.00064

Intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs) play important roles in many aspects of normal cell physiology, such as signal transduction and transcription, as well as pathological states, including Alzheimer's, Parkinson's, and Huntington's disease. Unlike their globular counterparts that are defined by a few structures and free energy minima, IDP/IDR comprise a large ensemble of rapidly interconverting structures and a corresponding free energy landscape characterized by multiple minima. This aspect has precluded the use of structural biological techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR) for resolving their structures. Instead, low-resolution techniques, such as small-angle X-ray or neutron scattering (SAXS/SANS), have become a mainstay in characterizing coarse features of the ensemble of structures. These are typically complemented with NMR data if possible or computational techniques, such as atomistic molecular dynamics, to further resolve the underlying ensemble of structures. However, over the past 10–15 years, it has become evident that the classical, pairwise-additive force fields that have enjoyed a high degree of success for globular proteins have been somewhat limited in modeling IDP/IDR structures that agree with experiment. There has thus been a significant effort to rehabilitate these models to obtain better agreement with experiment, typically done by optimizing parameters in a piecewise fashion. In this work, we take a different approach by optimizing a set of force field parameters simultaneously, using machine learning to adapt force field parameters to experimental SAXS scattering profiles. We demonstrate our approach in modeling three biologically IDP ensembles based on experimental SAXS profiles and show that our optimization approach significantly improve force field parameters that generate ensembles in better agreement with experiment.

Keywords: intrinsically disordered proteins, machine learning, optimization, force-field parameters, molecular dynamics

1. INTRODUCTION

Our understanding of classical structure-function paradigm of proteins was first established by recognizing a unique three-dimensional (3D) structure of specific amino acid sequence (Anfinsen, 1973). However, in the late '90s, it was reported that many proteins remain natively unfolded while biologically active (Wright and Dyson, 1999). Such intrinsically disordered proteins or regions (IDPs/IDRs) do not fold autonomously into stable 3D structures; however, they may possess short-transient secondary structure (Uversky, 2011; Das and Pappu, 2013; Latysheva et al., 2015). IDPs typically have an abundance of charged and polar residues while lacking hydrophobic groups. In addition, a recent study suggests IDPs, even with a low net charge, and high hydrophobicity, possess extended conformations in water (Riback et al., 2017). The 3D structure of IDPs is specifically influenced by their sequence, e.g., a linear sequence patterning of oppositely charged residues was found to govern the conformational dimension in polyampholytic IDPs (Das and Pappu, 2013).

Despite the interconverting ensemble of conformations and absence of structured region, IDPs play a vital role in many cell physiology, such as signal transduction and transcription (Habchi et al., 2014; Latysheva et al., 2015; Wright and Dyson, 2015; Mollica et al., 2016). Interest in IDPs also stems from their association with multiple diseases, such as cancers [p53 (Wells et al., 2008) and HPV (Uversky et al., 2006)], diabetes, cardiovascular, and neurodegenerative disorders (e.g., Alzheimer's and Parkinson's diseases) (Uversky et al., 2008; Knowles et al., 2014). Therefore, IDPs not only exemplify a new paradigm for understanding disorder-function relationships but also provide insights on pathological mutations that can lead to serious human diseases (Latysheva et al., 2015).

Nuclear magnetic resonance (NMR) spectroscopy (Wells et al., 2008; Pérez et al., 2009, 2013; Robustelli et al., 2012; Jensen et al., 2014; Arai et al., 2015; Lee et al., 2016; Arbesü et al., 2017), single-molecule Förster resonance energy transfer (smFRET) (Hofmann et al., 2012; Fuertes et al., 2017), cryo-electron microscopy (cEM) (Busch et al., 2015; Levine et al., 2015) and small-angle X-ray scattering (SAXS) (Wells et al., 2008; Receveur-Bréchot and Durand, 2012; Arbesü et al., 2017; Fuertes et al., 2017; Riback et al., 2017; Drulyte et al., 2018) are widely being used to study the disordered structures of IDPs. However, they lack a complete atomic or molecular description of disorder due to instrumental resolution and the ensemble-averaged nature of the measurements, which present a steep challenge to the unambiguous interpretation of the measurements (Fuertes et al., 2017; Kosciolek et al., 2017; Best et al., 2018; Drulyte et al., 2018; Riback et al., 2018). Therefore, molecular dynamics (MD) simulations are often combined with experiments for determining the ensemble of 3D structures of IDPs (Huang et al., 2017).

At the heart of running atomistic molecular dynamics (MD) simulations is a set of empirical potential energy functions from which forces are derived for characterizing the time evolution of a system (typically a protein, or a set of proteins, or other bio-molecules) (Karplus and McCammon, 2002). These potential energy functions are typically referred to as a force field (FF). The last four decades of FF development have been critical in enabling studies of bio-molecular systems in the context of ligand binding, enzyme reactions, protein folding/misfolding and other complex biological phenomena, such as self-assembly (Karplus, 2002).

Current FFs for proteins and other bio-molecules are mature in the sense that they have been rigorously validated for benchmark systems, have an underlying methodology for parameterization, and are being continuously improved upon as discrepancies between simulation results and experimental physical observables arise (Lopes et al., 2015). These deficiencies become particularly noticeable with current advances in sampling ability of MD on modern computer hardware and algorithmic improvements in the software, enabling limitations in sampling to be ruled out as the deficiency (Tiwary et al., 2015). One notable deficiency of standard, pairwise additive force fields is in their ability to correctly capture the experimentally observed properties of intrinsically disordered proteins (IDP) and partial disorder. While empirical force fields have demonstrated a high degree of success in reproducing experimentally derived physical properties of globular proteins, which are characterized by a few relevant, compact conformations, they are deficient in capturing the many transient conformational states and corresponding free energy minima characteristic of IDPs (Huang and MacKerell, 2018). This is best demonstrated in the tendency of empirical force fields to predict a small set of overly compact conformations, in contrast to experimental prediction of a large ensemble of more extended, less compact conformations where the protein interacts much more with solvent (Nettels et al., 2009; Best et al., 2014; Piana et al., 2014, 2015; Skinner et al., 2014). Indeed, this observation, as well as hydration free energy calculations on small molecules being observed to be too unfavorable (Shirts et al., 2003; Shirts and Pande, 2005) compared with experiment, have pointed to standard force fields being excessively solvophobic.

These observations have led researchers to tune the non-covalent energetic parameters in an effort to create a more balanced picture of protein-water interactions. While it could be argued that more complicated functional forms may be necessary, it is highly desirable to be able to preserve the current simple functional forms if possible, given their history of success in capturing an array of biophysical phenomena of interest, and their easy implementation on GPU and other high-performance platforms.

Efforts at rehabilitating FFs for use with IDP/IDR have focused on adjustment of short-ranged non-covalent contributions to protein-water interactions through tuning of van der Waals energetics, modeled in all cases by a Lennard-Jones potential with a 6–12 functional form (Best et al., 2014; Piana et al., 2015; Robustelli et al., 2018). In addition to reparameterization of protein-water interactions, closer attention has been paid to the underlying water model, recognizing the advantages of recently parameterized four-site water models, such as TIP4P-Ew (Horn et al., 2004) and TIP4P/2005 Vega and Abascal (2005), over simpler three-site models, such as TIP3P (Best and Mittal, 2010). Given the overly compact nature of simulated IDP, it was also considered natural to reparameterize the side-chain and backbone torsional parameters, and a number of groups

have pursued this line of research (Nerenberg and Head-Gordon, 2011; Rauscher et al., 2015; Huang et al., 2017; Song et al., 2017; Robustelli et al., 2018). Reparameterization of torsional potentials is likely necessary for a different reason, namely, the fact that torsional potentials implicitly have a degree of short-ranged non-bonded character. Despite the continuous progress in improving FF accuracy, our ability to recapitulate gross experimental observables, such as neutron reflectivity/scattering profiles from MD simulations has therefore remained extremely challenging.

For IDPs, small-angle X-ray and neutron scattering (SAXS and SANS, respectively) are ideal experimental methods for investigating the ensemble of IDP structures, as traditional imaging methods, such as X-ray crystallography or nuclear magnetic resonance (NMR), by themselves are not able to resolve the large number of rapidly interconverting structures of which the IDP ensemble is composed Bernado and Svergun (2012), Kikhney and Svergun (2015). Indeed, low-resolution methods, such as SANS/SAXS are ideal for conformationally polydisperse systems, such as IDP whose conformational ensemble is very large and consists of structures that are rapidly interconverting among themselves. SAXS and SANS are able to provide coarse structural information about the structural ensemble, such as compactness and overall size and shape. Due to the fact that the SAXS/SANS scattering intensities constitute an average over many different structures, these methods must be complemented by additional higher-resolution experimental data, such as NMR observables (Grishaev et al., 2005; Marsh et al., 2007; Marsh and Forman-Kay, 2009; Wang et al., 2009; Schwieters et al., 2010), or simulation-based methods (Bernado et al., 2007; Pelikan et al., 2009; Yang et al., 2010; Rozycki et al., 2011) to elucidate the structures of which the ensemble is composed. Therefore, given the important role of MD simulations as a complement to the interpretation of SAXS/SANS data, it is imperative that the underlying force field be accurate.

Here, we studied three IDPs with varying molecular weight and different charge-hydrophobicity characteristics (see **Figure 1A**): RS-peptide (24 residues), PaaA2 (63 residues), and SH4UD (95 residues). RS-peptide is highly charged IDR without any structured region in Serine/arginine-rich proteins, such as serine/arginine-rich splicing factor 1 (SRSF1) and plays a significant role in RNA metabolism, including transcription, RNA splicing and RNA export (Xiang et al., 2013). The phosphorylation of serine residues in RS repeats regulates peptide's interaction and subcellular localization, whereas it undergoes several cycles of phosphorylation and dephosphorylation during splicing (Xiang et al., 2013). PaaA2 is the antitoxin domain of toxin-antitoxin (TA) module in the human pathogen *E. coli* O157, which neutralizes the toxin domain such that TA module copes with different sources of stress (Sterckx et al., 2014, 2016). The TA module is also associated with the establishment of persister phenotype and virulence mechanisms (Sterckx et al., 2016). It has two preformed helices connected by a flexible linker in the absence of a binding partner, however is, classified as IDP due to a high degree of conformational flexibility from SAXS and NMR studies (Sterckx et al., 2014). Proto-oncogene non-receptor human tyrosine kinase c-Src is a multi-domain protein (Tatosyan and Mizenina, 2000; Pérez et al., 2009) that encompasses an N-terminal IDR containing the Src homology 4 (SH4) and unique (U) domains hereafter refer as SH4UD. Several studies suggest the high activity of the c-Src kinase in a wide variety of human cancers, such as colon, breast, pancreas, and brain (Wheeler et al., 2009). The phosphorylation in SH4UD induces a global electrostatic perturbation forcing c-Src kinase to untie from the membrane (Pérez et al., 2009).

In this work, we have implemented a method to optimize FF parameters against experimental SAXS and SANS intensities in ForceBalance (Wang et al., 2014)—these observables can be understood as ensemble-averaged properties with derivable gradients and Hessians with respect to force field parameters. Starting with the most recent and comprehensive reparameterization of an IDP force field (Robustelli et al., 2018) from the D. E. Shaw research group, we optimized the water and protein backbone Lennard-Jones σ and ϵ, as well as the barrier heights of protein backbone torsions, as was done in their study. We sought to determine whether we could systematically improve on the parameters they had derived, as our initial set of parameters was their optimized IDP force field named a99SB-disp. We found that through our systematic reparameterization using ForceBalance that we could achieve improved agreement with experimental SAXS profiles for 3 systems: RS-peptide, PaaA2, and SH4UD. We will henceforth refer to our version of the algorithm as *ForceBalance-SAS* (small-angle scattering). A key advantage of our approach is that nearly any experimental observable can be encoded as an ensemble-averaged property, for which analytic gradients and approximate Hessians with respect to force field parameters that are being optimized can be obtained.

2. METHODS

2.1. Parameter Optimization With ForceBalance-SAS

ForceBalance-SAS parameterization proceeds through an iterative non-linear least-squares minimization of the squared residual between experimental and calculated properties using analytical gradients and approximate Hessians (Gauss-Newton approximation whose term consists of a product of first derivatives) with respect to a set of FF parameters. A flowchart illustrating our approach is shown in **Figure 1**. Each iteration consists of a MD simulation with the current set of FF parameters, followed by a calculation of the objective function, gradient, and approximate Hessian (at the current set of FF parameter values), and an optimization step using Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) followed by a regularization to avoid overfitting.

The Levenberg-Marquardt algorithm is used, because it is both gradient- and Hessian-based. Moreover, if the initial parameters are far from the local minimum, it is able to converge faster than the Gauss-Newton algorithm. Lastly, the Levenberg-Marquardt algorithm is ideal due to its intrinsic ability to incorporate an adaptive trust radius (Dennis et al., 1981; More and Sorensen, 1983), effectively enabling the algorithm to change

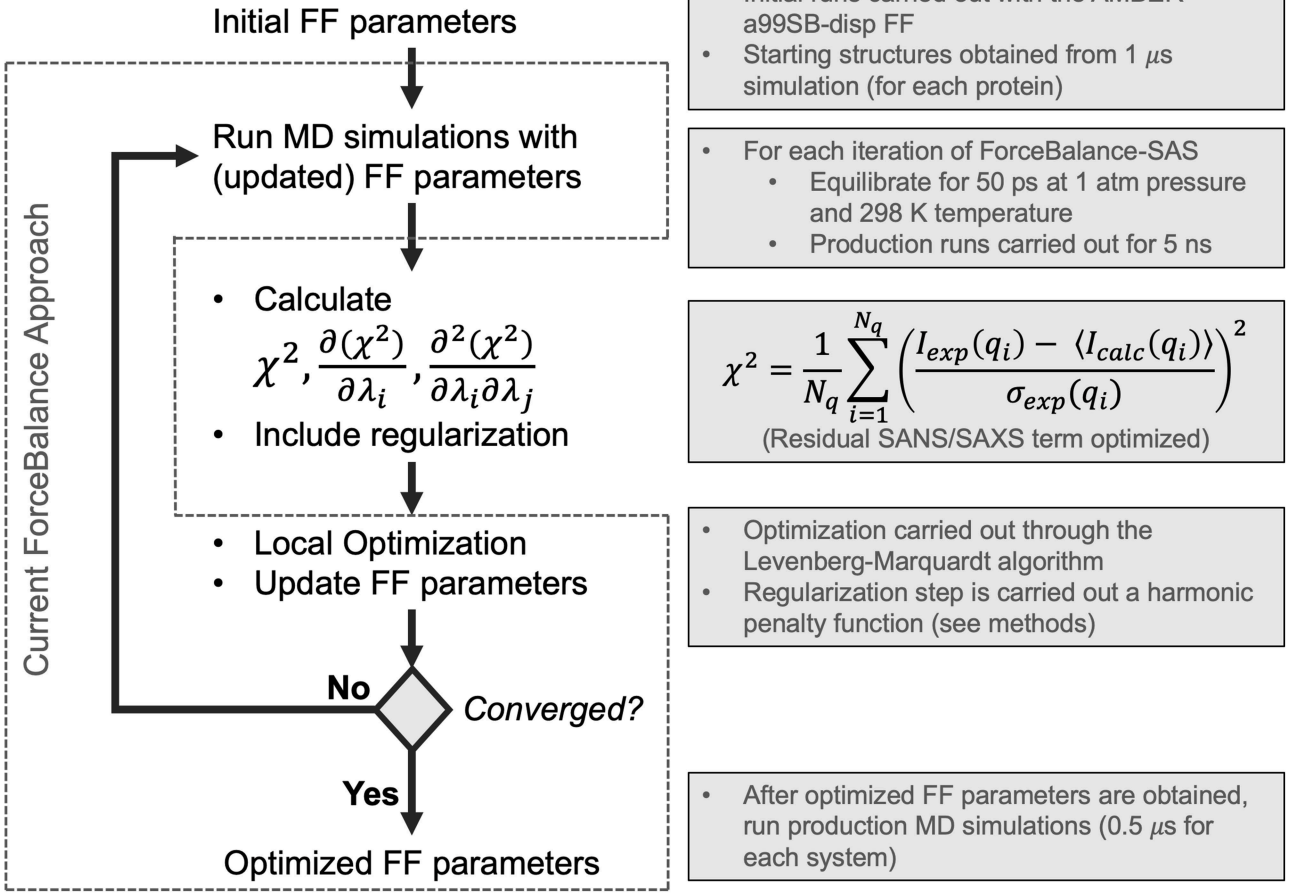


FIGURE 1 | Flowchart depicting the ForceBalance-SAS algorithm. An initial set of parameters is input, followed by MD simulation and calculation of ensemble-averaged small-angle scattering intensities. After the simulation stops, the residual between the simulated and experimental scattering intensities is calculated, along with the gradients and Hessians of the residual. If the desired convergence criteria are met, the algorithm stops, and the new force field parameters are output; if not, optimization is performed, a new set of parameters are obtained, and a simulation with the updated parameters are performed, completing the cycle. The current implementation of the ForceBalance (Wang et al., 2014) approach is demarcated from our approach using dotted lines.

the size of the step according to how well the objective function was improved in the previous step, as shown in the following equation framed in the context of the fitting task presented in this work:

$$(\mathbf{J}^{\mathbf{T}}\mathbf{J} + \gamma\mathbf{I})\delta = \mathbf{J}^{\mathbf{T}}(\mathbf{A}^{\mathbf{exp}} - \langle\mathbf{A}^{\mathbf{calc}}(\lambda)\rangle), \quad (1)$$

and

$$J_{ij} = \frac{\partial A_i^{calc}}{\partial\lambda_j}. \quad (2)$$

In the above equation, $\mathbf{A}^{\mathbf{exp}}$ is the set of experimentally measured observables, $\langle\mathbf{A}^{\mathbf{calc}}\rangle$ is corresponding calculated set of ensemble-averaged observables, λ are the parameters (here, FF parameters) whose values we are optimizing, δ is the step taken at the current step of the optimization, and γ is the parameter controlling the adaptive trust radius. In this work, the initial trust radius was set to 1.0, which is larger than the default of 0.1 in the standard ForceBalance approach. A minimum trust radius of 0.05 was allowed (the default in standard ForceBalance is 0.0). An adaptive damping factor controlling how much the trust region can vary from the initial value was set to the default value used in ForceBalance of 0.5. Regularization is achieved by means of a harmonic penalty function that constrains FF parameters to a physically reasonable range of values as follows:

$$R(\lambda) = \frac{\lambda^2}{\alpha^2}, \quad (3)$$

where $R(\lambda)$ is the harmonic penalty function, λ is the FF parameter, and α corresponds to the radius within which the parameter value can vary. In this work, α is determined by ForceBalance automatically according to the magnitudes of λ, and were 0.0529177, 2.4784, and 96.4853 for van der Waals σ, van der Waals ϵ, and torsional barrier heights, respectively.

If convergence criteria are met, the algorithm stops and the optimized FF parameters are output. If not, the cycle continues with a simulation at the new set of parameters.

Our method rests on the ability of ForceBalance-SAS to directly optimize a set of FF parameters with respect to the experimental SAXS and SANS scattering intensities. Any condensed phase observable can be calculated from rigorous statistical mechanical principles. In the isobaric-isothermal ensemble, the ensemble-averaged observable $\langle A \rangle$ (in our specific case, $\langle I(q) \rangle$, the small-angle scattering intensity—described in Equation 6), for all experimentally observed scattering vectors, $I(q)$ for a given set of FF parameters λ is:

$$\langle A \rangle_\lambda = \frac{1}{Q(\lambda)} \int A(r, V, \lambda) \exp(-\beta(E(r, V, \lambda) + PV)) dR dV, \quad (4)$$

where $Q(\lambda) = \int \exp(-\beta(E(r, V, \lambda) + PV)$ is the isothermal-isobaric partition function. Here, E is the potential energy, β is $\frac{1}{k_B T}$, T represents the temperature, P is the pressure, and V is the volume. In practice, $\langle A \rangle$ is not evaluated through a direct integration of Equation (4), but rather is sampled numerically by MD assuming ergodicity. Analytic gradients of properties A with respect to FF parameters λ can be obtained by analytically differentiating Equation (4):

$$\frac{\partial \langle A \rangle_\lambda}{\partial \lambda} = \langle \frac{\partial A}{\partial \lambda} \rangle_\lambda - \beta(\langle A \frac{\partial E}{\partial \lambda} \rangle_\lambda - \langle A \rangle_\lambda \langle \frac{\partial E}{\partial \lambda} \rangle_\lambda). \quad (5)$$

The above terms are calculated for each value of $I(q)$ in the experimental (and simulated) scattering profile. Thus, the primary objective of ForceBalance-SAS is to improve the agreement between experimental and calculated SAXS intensities by minimizing the following residual term:

$$\chi^2 = \frac{1}{N_q} \sum_{i=1}^{N_q} \left(\frac{I_{exp}(q_i) - \langle I_{calc}(q_i) \rangle}{\sigma_{exp}(q_i)} \right)^2, \quad (6)$$

where $I_{exp}(q_i)$ and $I_{calc}(q_i)$ are the experimental and calculated intensities, respectively, at a given wavenumber q_i, $\sigma_{exp}(q_i)$ is the experimental error in the measurement of $I_{exp}(q_i)$, and N_q is the number of observations of q_i obtained.

While the expression for the gradient of a property with respect to the FF parameters is analytic, gradients of the potential energy with respect to FF parameters are themselves calculated with three-point finite difference using a step size of 10^{-9}. In this work the FF parameters λ were the σ and ϵ of protein backbone Lennard-Jones, and the barrier heights of protein backbone torsions. The final simulation parameters were achieved for RS-peptide and PaaA2 after 18 and 4 cycles of ForceBalance-SAS (**Figure S1**), respectively, which amounted to the desired reduction in χ^2 of at least 50%.

2.2. SAXS/SANS Calculations

The experimental SAXS data for RS-peptide and PaaA2 were taken from (Rauscher et al., 2015) and (Sterckx et al., 2014), respectively. SH4UD SAXS data was provided by Hugh M. O'Neill, which was measured at X-Ray Laboratory, Spallation Neutron Source, Oak Ridge National Laboratory. SAXS/SANS scattering intensities $I(q)$ were calculated from MD snapshots using the crysol/cryson algorithms in the ATSAS package (Svergun et al., 1995; Franke et al., 2017). Since crysol/cryson are based on use of implicit solvent, it is essential that its parameter modeling the difference in solvation between the protein surface and bulk be optimized. To achieve this, we averaged the coordinates of all snapshots saved for the simulation of each iteration, and then fit the averaged coordinates to the experimental SAXS/SANS to optimize the solvation parameter; this optimization was done internally within crysol/cryson and details of how this is done can be found in (Svergun et al., 1995). This optimized value was used for the calculated SAXS/SANS of each of the snapshots. Since the calculated and experimental SAXS can have different number of q points, a spline-based interpolation of the calculated and experimental SAXS/SANS curves was used to match the number of q points between the two. Finally, the calculated SAXS/SANS intensities will necessarily have different amplitudes owing to aspects of the experiment not accounted for in the calculation. To match the amplitudes between calculation and experiment, a linear fit was performed between the SAXS/SANS $I(q)$ profile averaged over all snapshots and the corresponding experimental $I(q)$. These fitting parameters were then used for the calculated intensities $I(q)$ of the individual snapshots.

2.3. MD Simulations

The initial MD simulations (step 1 of **Figure 1**) of three systems (RS-peptide, PaaA2, and SH4UD) were conducted using GROMACS 5.1.2 (Van der Spoel et al., 2005; Hess, 2008; Abraham et al., 2015) using newly developed a99SB-disp FF parameter set (Robustelli et al., 2018). The energy of the system was minimized using 1,000 steepest decent steps, which was followed by 1 ns of equilibration using NVT and NPT ensembles. Finally, 1 μs of production runs were performed using the NPT ensemble. The snapshots saved at the end of the 1 μs simulations were further utilized for ForceBalance-SAS optimization.

For each cycle of ForceBalance-SAS, as part of our optimization procedure (step 2 in **Figure 1**), each protein was then simulated for 5 ns of production at each iteration in the isothermal-isobaric (NPT) ensemble at 1 atm and 298 K, preceded by 50 ps of equilibration. Achieving statistical convergence of the target scattering property is critical. Our choice of 5 ns of production for each iteration of ForceBalance-SAS was determined heuristically by running a single iteration at a range of production lengths from 0.5 ns to 50 ns. Scattering intensity and Kratky curves were calculated for each simulation length. We used the χ^2 metric (Equation 6 above) to quantitatively evaluate whether the global features of the scattering profiles at various time-windows from the simulation trajectory (50, 25, 10, 5, 2.5, 1, 0.5 ns) were sufficiently captured (see **Figure S2**). We found that a choice of 5 ns to have better χ^2 fit to the experimental data and our choice of 5 ns was an expedient compromise between computational cost and accuracy for each cycle of the optimization. Note that the choice of 5

ns production runs was made based prior to the start of the optimization step. We do note that this length of the simulations may affect the overall quality of fits obtained (see Discussion).

Thermostating (in steps 1 and 2 of **Figure 1**) was performed using GROMACS (Van der Spoel et al., 2005; Hess, 2008; Abraham et al., 2015) modified Berendsen thermostat (Berendsen

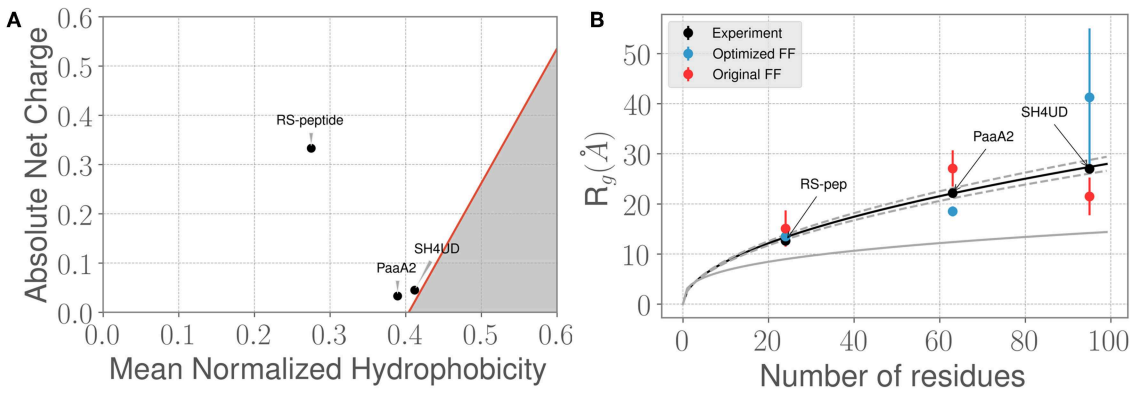


FIGURE 2 | Three prototypical IDP systems chosen for the ForceBalance-SAS approach indicate diverse structural characteristics. **(A)** The mean normalized hydrophobicity vs. the absolute net charge (Uversky) plots indicate that the RS-peptide system is more disordered than the other two systems. The red line is used to mark the boundary between disordered proteins vs. more folded/globular proteins; the gray highlighted area is indicative of the region that is enriched for folded/globular proteins (Uversky, 2011). **(B)** Comparison of the SAXS determined experimental radius of gyration (R_g) values vs. the R_g values predicted using simulations from the original FF (red dots) and the optimized FF (blue dots). The theoretical R_g values predicted from the Flory equation for IDPs (see Results section) is shown in black, along with expected standard deviations (gray dotted lines). The corresponding R_g values for a globular protein with the same number of amino acid residues is shown for reference (gray solid line). Additional details of the sequence/structural properties of the IDP ensembles considered here are provided in the supporting information.

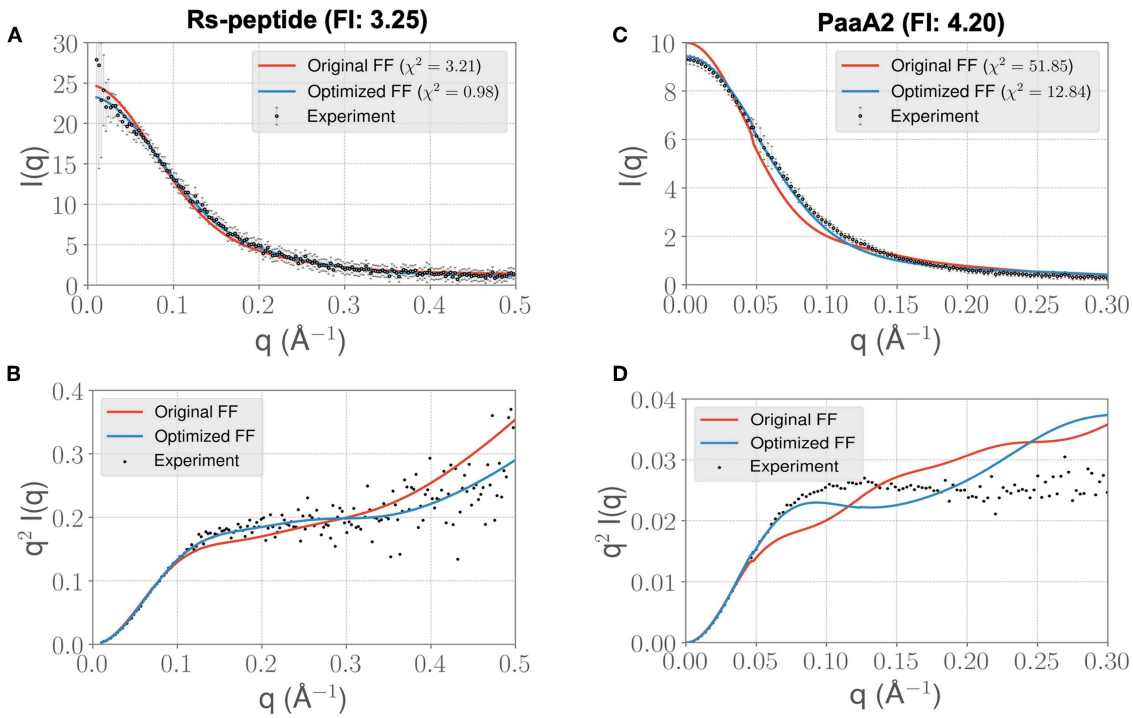


FIGURE 3 | ForceBalance-SAS based simulations generate IDP ensembles that are better fit to the experimental SAXS observables at shorter timescales. **(A)** The scattering profiles for RS-peptide showing the experimental data (black dots with error bars) along with the predicted SAXS scattering profiles from the original FF simulations (red lines) vs. the optimized FF simulations (blue lines). For clarity, the χ^2 values between the experiment and the respective simulations are shown in the legend. **(B)** The Kratky plot from experiments (black dots), and predicted profiles from simulations (red line corresponding to the original FF, blue line—optimized FF). For clarity, the error bars from the experiments are excluded. **(C,D)** Highlight the same comparison for the PaaA2 system. The factor improvement (FI) in the χ^2 values between the optimized and original FFs are listed above each protein system.

et al., 1984) with separate coupling of the protein and solvent to a heat bath at 298 K. Initial velocities assigned according to the Maxwell-Boltzmann distribution at 298 K. Barostating was performed with the Parrinello-Rahman method (Parrinello and Rahman, 1981). A 2-fs timestep was used, and covalent bonds between hydrogen and heavy atoms were constrained using the LINCS algorithm (Hess et al., 1997; Hess, 2008). A 12-Å distance cutoff was used for van der Waals and the real-space component of electrostatics. Long-range electrostatics were calculated using Particle Mesh Ewald (Darden et al., 1993) with a grid spacing of 1.6Å. Coordinate snapshots were saved every 100 ps. Simulations were run on a GPU-enabled version of Gromacs (v. 2019) on a single node equipped with two Tesla K80s.

2.4. Sequence-Structure Property Predictions

Per-residue disorder prediction was performed with the PONDR (Prediction of Natural Disordered Regions; Obradovic et al., 2003) algorithm using the VLXT model whose predictions are based on the integration of predictions made by three different neural networks. We used the web server CIDER (Holehouse et al., 2017) to ascertain relationships between the charged residue content of a sequence and its structural ensemble propensities.

3. RESULTS

SAXS and SANS scattering intensities were implemented as force field parameter fitting targets in ForceBalance-SAS. As the intensities are condensed-phase observables, much of the optimization machinery in ForceBalance-SAS was ideal for this purpose and modification to incorporate SAXS/SANS was straightforward. As our initial set of force field parameters, we used the most state-of-the-art IDP-specific force field, a99SB-*disp*, which has been developed and validated using a comprehensive IDP benchmark consisting of a range of protein systems and experimental observables. To have continuity with their work and previous efforts, we optimized the σ and ϵ of the water and protein backbone atoms' Lennard-Jones, as well as the protein backbone torsion barrier heights. Unlike previous efforts, we are able to optimize all of these simultaneously and, importantly, are able to directly target the agreement of calculated and experimental SAXS scattering profiles. This is an ideal experimental target, as it directly reports on how contracted or

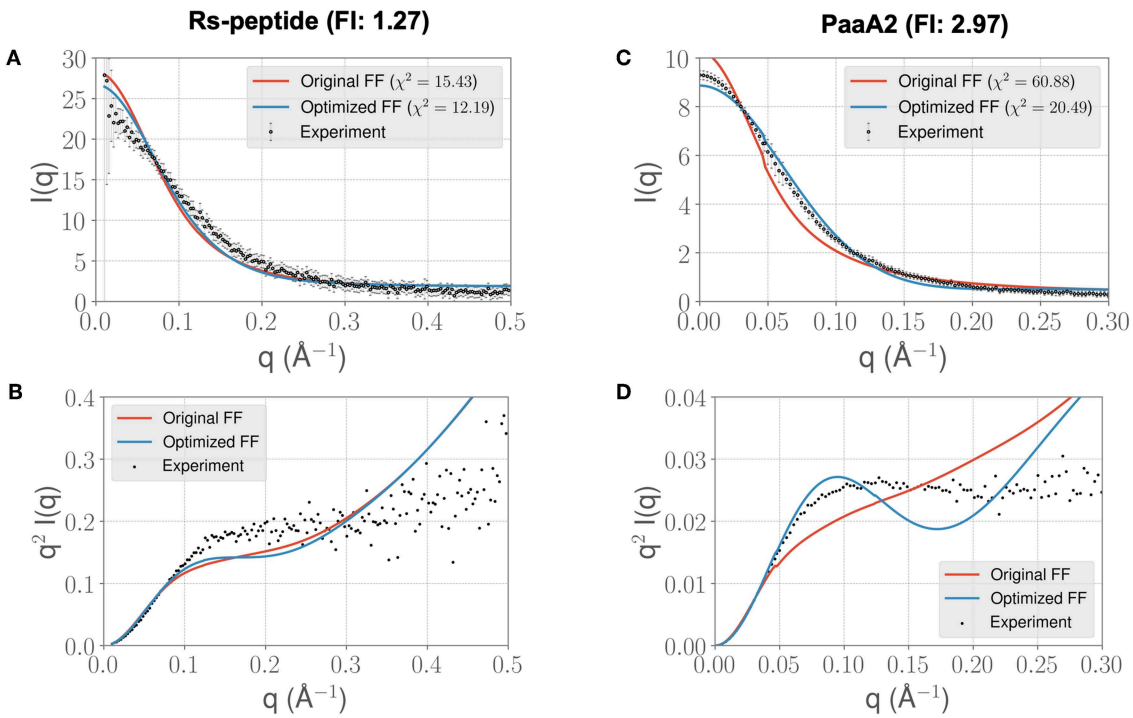


FIGURE 4 | Longer timescale simulations using the ForceBalance-SAS optimized FF parameters preserve improvement with experimentally observed scattering profiles. **(A)** The scattering profiles for RS-peptide showing the experimental data (black dots with error bars) along with the predicted SAXS scattering profiles from the original FF simulations (red lines) vs. the optimized FF simulations (blue lines). For clarity, the χ^2 values between the experiment and the respective simulations are shown in the legend. **(B)** The Kratky plot from experiments (black dots), and predicted profiles from simulations (red line corresponding to the original FF, blue line—optimized FF). For clarity, the error bars from the experiments are excluded. **(C,D)** Highlight the same comparison for the PaaA2 system. The factor improvement (FI) in the χ^2 values between the optimized and original FFs are listed above each protein system. Note that the FI in each simulation has decreased compared to the shorter timescales (**Figure 3**)—however, still preserves the overall trends. It is also notable that the fits of the MD simulations to the longer q values are poorer in both cases.

expanded protein conformations in the ensemble are, a protein property that force fields have notable difficulty in capturing.

3.1. ForceBalance-SAS Enables Better Agreement Between Experimental and Simulated Ensembles

We chose three prototypical IDP systems that are of biological interest: (1) RS-peptide (Xiang et al., 2013), (2) prokaryotic type II antitoxin module PaaA2 from the human pathogen *E. coli* O157, and (3) the N-terminal regulatory region consisting of the SH4 unique domain (SH4UD) of the C-Src family of non-receptor tyrosine kinases. An examination of the mean hydrophobicity vs. net charge of these three IDP systems, also referred to as the Uversky plots (Uversky, 2011), shows that the RS-peptide system is more disordered than the other two systems (**Figure 2A**). Not surprisingly, the secondary structural content for the RS-peptide is significantly lower, given that its absolute charge is much higher compared to the other two IDP systems. Indeed from experimental data, such as circular dichroism (CD) and nuclear magnetic resonance (NMR), PaaA2 consists of at least two partially formed α-helices (Sterckx et al., 2014) and SH4UD consists of several transient helices (Pérez et al., 2009; Arbesü et al., 2017). We performed calculations with the CIDER (Classification of Intrinsically Disordered Ensemble Relationships) web server to further parse the sequence-structure relationships based on the fraction of positively and negatively charged residues in the sequence. The diagram of states generated by CIDER shows the propensity of some structure for both PaaA2 and SH4UD (**Figure S3**), in accord with CD and NMR predictions. RS-peptide presents an interesting case in that it is predicted to be collapsed or expanded, depending on context, but lies very close to the region corresponding to an expanded polyelectrolyte, which is supported by NMR and CD. The experimental observations from NMR and CD are further supported by predictions using the sequence-based prediction method PONDR (Prediction of Natural Disordered Regions), which predicts order for residues 16–35 and 52–75 for PaaA2 and SH4UD, respectively (**Figures S4A,B**); RS-peptide was too short in length for PONDR to make any prediction.

We next examined how the experimentally determined radius of gyration (R_g) varies with the amino-acid chain length. The experimental R_g values are obtained through Guinier fits to the scattering profiles. Notably, the experimentally determined R_g values for the three IDPs aligns closely with the theoretical predictions of R_g^{Flory} from the Flory equation: $R_g^{Flory} = (2.54 \pm 0.01) \times N^{(0.522 \pm 0.01)}$, where N represents the number of amino-acid residues in the IDP of interest. As shown in **Figure 2B**, the agreement between experimental R_g and R_g^{Flory} is quite remarkable. However, we note that when considering the simulated ensembles, the original a99SB-*disp* FF overestimates the R_g values for the PaaA2 protein where as the optimized FF underestimates the Rg for the SH4UD ensemble. On the other hand, the ForceBalance-SAS optimized FF overestimates the R_g values for the SH4UD ensemble, while being close to the experimentally observed R_g values for the RS-peptide and PaaA2 system. Note that for the SH4UD system, we did not explicitly optimize the FF parameters—we just took the optimized parameters from the PaaA2 simulation and used it to simulate the SH4UD system (see section 3.3).

The Guinier fits to the SAXS profiles for the three IDP systems provide a gross summary of their conformational ensembles; however, the R_g value by itself does not sufficiently capture all of the information contained in the scattering profiles. We therefore posited that even though the ForceBalance-SAS may underestimate the overall R_g values, its ability to fit the simulated ensembles to experimentally observed SAXS profiles may be better. To test this hypothesis, we used the χ^2 metric (Equation 6) to assess the quality of the fit. By optimizing the aforementioned set of force field parameters, we were able to reduce the discrepancy with experiment by a factor of 3.3 and 4.2 for RS-peptide and PaaA2, respectively, where the factor of improvement is simply the ratio of the χ^2 value obtained with the original parameters to that obtained with the optimized parameters.

Visual inspection of the $I(q)$ vs. q profile for RS-peptide (**Figure 3A**), as well as the Kratky plot (**Figure 3B**) of $q^2I(q)$ vs. q (**Figure 2**), reveal more information about the specific aspects of protein structure that have been improved. In general, the lower q values report on low-resolution protein behavior, such as how contracted or expanded the structures in the ensemble are, while larger q values can report more on finer scale detail. The Kratky plot is useful for quantifying disorder in a polymer chain. For the RS-peptide example, it is clear that the original FF predicts a more disordered ensemble, while both the experiment and the optimized FF based simulations predict some local structure

TABLE 1 | Original and optimized torsion angle parameters for RS-peptide.

Atom types comprising torsion	Original FF	Optimized FF	% Change
C–N–CT–C	0.142260	0.145503	2.280
C–N–CT–C	1.40164	1.40177	0.001
C–N–CT–C	2.27610	2.27026	−0.256
C–N–CT–C	0.334720	0.334548	−0.051
H1–CT–C–O	3.34720	3.34905	0.055
H1–CT–C–O	0.334720	0.331802	−0.872
H1–CT–C–OB	3.34720	3.34574	−0.044
H1–CT–C–OB	0.334720	0.334634	−0.026
HB–N–C–OB	8.36800	8.36773	−0.003
HB–N–C–OB	10.4600	10.4603	0.003
N–CT–C–N	0.824250	0.826095	0.224
N–CT–C–N	6.04588	6.05070	0.080
N–CT–C–N	2.00414	2.00474	0.030
N–CT–C–N	0.0799100	0.0797917	−0.148
N–CT–C–N	0.0167400	0.0197590	18.035

The left-hand label of each row indicates the four atom types of which each torsion is composed. C, backbone carbonyl carbon; N, backbone amide nitrogen; CT, aliphatic carbon (Cα in this context); O, backbone carbonyl oxygen; H1, hydrogen bound to Cα; HB, hydrogen bound to backbone amide nitrogen.

TABLE 2 | Original and optimized Lennard-Jones parameters for RS-peptide.

	Original FF		Optimized FF			
Atom type	σ	ϵ	σ	ϵ	**% Change σ**	**% Change ϵ**
C	0.339967	0.359824	0.339966	0.359787	–0.000235359	–0.0104181
H	0.106908	0.0656888	0.106908	0.0656513	–0.000374220	–0.0570937
HB	0.106908	0.0656888	0.106908	0.0657721	–0.000374220	0.126688
N	0.325000	0.711280	0.325000	0.711355	0.000123099	0.0105384
N3	0.325000	0.711280	0.324998	0.711156	–0.000492395	–0.0173983
OB	0.295992	0.878640	0.295992	0.878593	–0.000135163	–0.00539543
O2	0.295992	0.8786401	0.295992	0.878633	0.000135163	–0.000784472
OW-tip4pd	0.316500	0.998989	0.316502	0.998914	0.000505619	–0.00750471

C, backbone carbonyl carbon; H, hydrogen bound to N-terminal nitrogen; HB, hydrogen bound to backbone amide nitrogen; N, backbone amide nitrogen; N3, N-terminal amine nitrogen; OB, backbone carbonyl oxygen; O2, C-terminal carboxyl oxygen; OW-tip4pd, water oxygen of TIP4P-d model.

in the ensemble. It is interesting to note that the χ^2 value has also significantly improved (3.21 with the original FF vs. 0.98 with the optimized FF), indicating that the ensemble from the optimization process has indeed improved the similarity to the experimental data. For the RS-peptide there is evidence of improvement at high q values as well, indicating that fine-scale protein-solvent structural details have been improved.

The $I(q)$ vs. q plot for PaaA2 shows marked improvement for the optimized set of parameters in all parts of the profile (**Figures 3C,D**), and while an improvement is seen for RS-peptide the effect is not as strong (**Figure 3A**). As can be seen in **Figure 3C**, improvement is seen at lower q values for both RS-peptide and PaaA2, suggesting that the problem with predicting an overly compact ensemble has been remedied.

In light of the well-appreciated importance of sampling the rugged conformational landscape of IDPs, we extended our simulations of RS-peptide and PaaA2 using the parameters obtained from the shorter 5-ns simulation lengths to 0.459 and 0.512 μs, respectively. We found that the optimized parameters yield an improvement in χ^2, albeit more modest than that of the shorter simulation (**Figure 4**). We note too that the discrepancies between the experimental and simulated ensembles are more apparent at higher q ranges, indicating that fine scale interactions are not as well-modeled as global interactions. Nonetheless, this demonstrates that major features of the ensemble that inform the optimization, namely those reflecting large scale interactions, are captured at shorter timescales and are transferrable to longer timescales.

Given the improvements in agreement with experimental observables, it is instructive to ascertain which optimized parameters differed the most from their original values. For both RS-peptide (**Tables 1**, **2**) and PaaA2 (**Tables 3**, **4**), it was the torsional barrier heights that changed the most from their original values. Interestingly, the van der Waals parameters changed little from their original values. This is perhaps expected, given the relatively longer history of attention to balancing solute-solvent, and protein-water, interactions through these terms. This notion is supported by a separate set of calculations where we optimized only the van der Waals parameters for RS-peptide in PaaA2. When only the van der Waals parameters were optimized, the factors of improvement of the χ^2 values were only 1.98 and 1.3 for RS-peptide and PaaA2, respectively.

TABLE 3 | Original and optimized torsion angle parameters for PaaA2.

Atom types comprising torsion	**Original FF**	**Optimized FF**	**% Change**
C–N–CT–C	0.142260	0.144172	1.344
C–N–CT–C	1.401640	1.380281	−1.524
C–N–CT–C	2.276100	2.233383	−1.877
C–N–CT–C	0.334720	0.355767	6.288
H1–CT–C–O	3.347200	3.287138	−1.794
H1–CT–C–O	0.334720	0.356079	6.381
H1–CT–C–OB	3.347200	3.326153	−0.629
H1–CT–C–OB	0.334720	0.355767	6.288
HB–N–C–OB	8.368000	8.378679	0.128
HB–N–C–OB	10.460000	10.438641	−0.204
N–CT–C–N	0.824250	0.845297	2.553
N–CT–C–N	6.045880	6.088597	0.707
N–CT–C–N	2.004140	2.015231	0.553
N–CT–C–N	0.079910	0.068819	−13.880
N–CT–C–N	0.016740	0.023640	41.219

*Refer to the **Table 1** legend for an explanation of the atom types.*

3.2. ForceBalance-SAS Improves Agreement With NMR Chemical Shift Observables for PaaA2

These observations also led us to the next question: *do the optimized FF parameters allow us to improve agreement with other (independent) experimental observables, such as NMR?* We posited that the improvement in agreement with respect to the gross structural details of the IDPs from SAXS data should also translate to agreement between NMR and MD simulations using the optimized FF. To test this hypothesis, we examined the PaaA2 system in greater detail. While previous work (Sterckx et al., 2014) used both NMR and SAXS data to refine conformational ensembles of PaaA2 using the Flexible-Meccano (Charavay et al.,

TABLE 4 | Original and optimized Lennard-Jones parameters for PaaA2.

	Original FF		Optimized FF			
Atom type	σ	ϵ	σ	ϵ	**% Change σ**	**% Change ϵ**
C	0.339967	0.359824	0.339979	0.360922	0.0034457	0.30501
H	0.106908	0.0656888	0.106920	0.0654144	0.010798	–0.41769
HB	0.106908	0.0656888	0.106920	0.0656251	0.010957	–0.097042
N	0.325000	0.711280	0.324998	0.710731	–0.00064537	–0.077150
N3	0.325000	0.711280	0.325006	0.711554	0.0018022	0.038575
OB	0.295992	0.878640	0.295998	0.879181	0.0019788	0.061544
O2	0.295992	0.878640	0.295969	0.879181	–0.0079152	0.061544
OW-tip4pd	0.316500	0.998989	0.316512	0.998715	0.0036472	–0.027465

*Refer to the **Table 2** legend for an explanation of the atom types.*

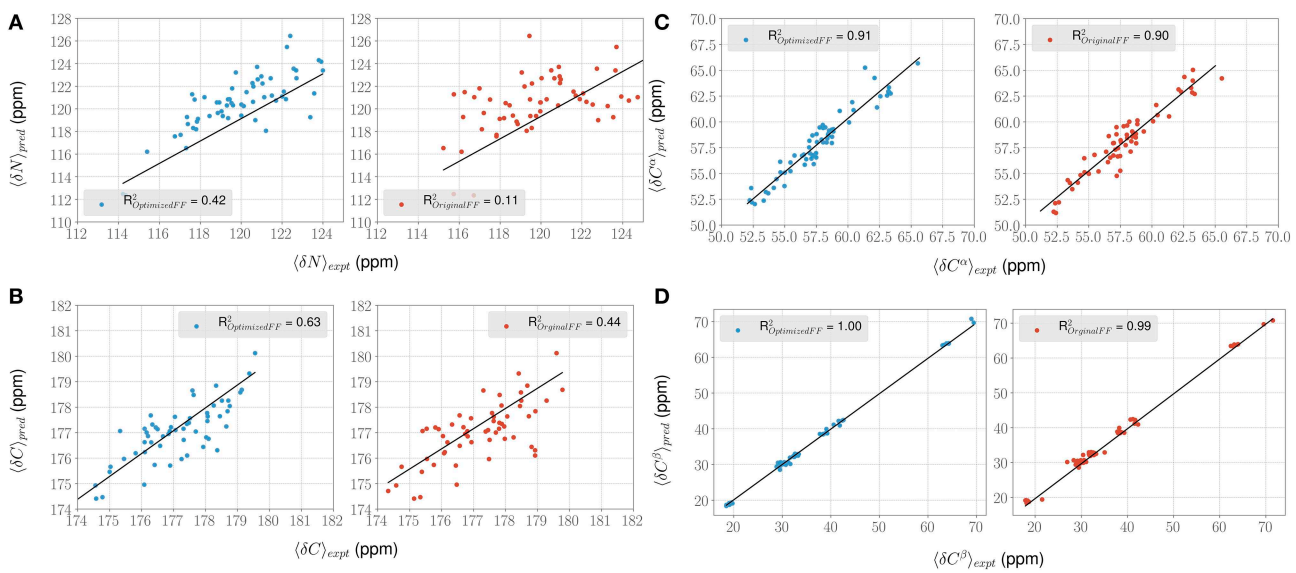


FIGURE 5 | ForceBalance-SAS improves agreement with NMR chemical shift observables for the PaaA2 ensemble. **(A–D)** Panels highlight the comparison between the average experimental (x-axis) chemical shift vs. predicted (y-axis) chemical shift values for N, C^{α}, C, and C^{β} atoms, respectively. Predictions from the optimized FF (blue dots) are compared with the original FF (red dots) simulations along with the $\mathbf{R^2}$ value for the fits (shown as black lines). Error bars are not highlighted for clarity.

2012) approach, here we used the optimized FF parameters to recapitulate the NMR chemical shift observables.

For each conformer in the MD trajectories from the original FF and the optimized FF, we used the program ShiftX2 (Han et al., 2011) to determine the chemical shifts of the backbone atoms: N, C^{α}, and C, along with the side-chain: C^{β}. We then plotted the agreement between the average experimental chemical shifts with the predicted chemical shifts. As shown in **Figures 5A–D**, the ForceBalance-SAS optimized FF parameters result in ensembles that are in better agreement with the experimental data, notably for C^{α} and C^{β} atoms. The agreement for the backbone Nitrogen atoms is also significantly improved compared to the original FF, indicating that our approach results in ensembles that better agree with NMR data. Further, for each of the atom types, a non-parametric bootstrap test (*p*-values) for significance also indicated that these correlations are significant (**Table 5**).

TABLE 5 | Summary of the statistical significance in comparing NMR observed chemical shifts with the FF parameters (original and optimized) for PaaA2 system.

	Original FF			Optimized FF		
Atom type	$\mathbf{R^2}$	**Standard error**	***p*-value**	$\mathbf{R^2}$	**Standard error**	***p*-value**
N	0.11	0.123	1.31E-05	0.42	0.072	1.23E-14
C^{α}	0.84	0.056	5.67E-27	0.91	0.039	5.68E-35
C^{β}	0.99	0.009	4.52E-72	1.00	0.005	5.53E-85
C	0.44	0.108	9.64E-10	0.63	0.090	5.42E-14

These were calculated using the `sckit.learn` *package (Pedregosa et al., 2011; Buitinck et al., 2013).*

This led us to further examine the generated ensembles. Each ensemble in **Figure 6** is colored using the R_g value corresponding to that conformation. The experimentally determined ensemble

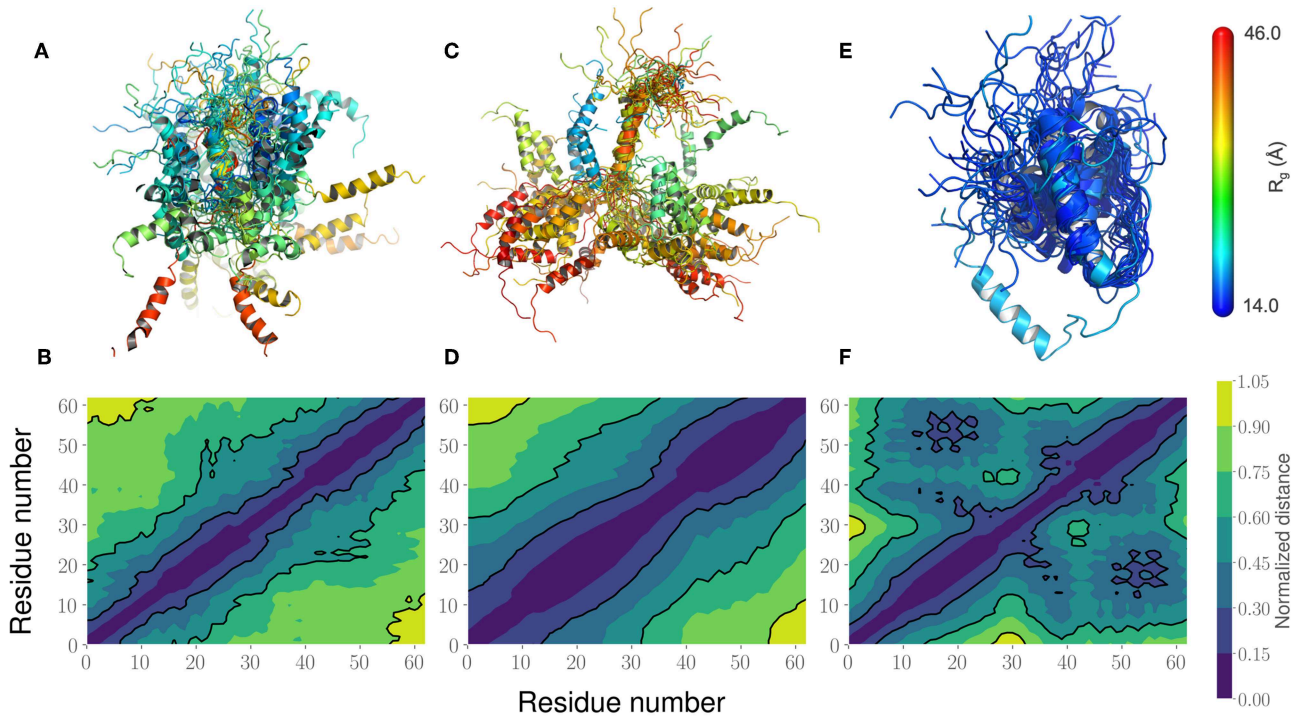


FIGURE 6 | Comparison of the PaaA2 Ensembles determined from experiments and simulations highlight regions of long-range interactions between the two α-helices. **(A)** A cartoon depiction of the PaaA2 ensemble determined using SAXS and NMR techniques using Flexible-Meccano. **(B)** Shows the normalized mean distance matrix showing the various interactions between residues; shades of blue indicate proximity in the chain—implying the increased likelihood of interactions. **(C)** Cartoon representation of the PaaA2 ensemble from the original FF along with the **(D)** normalized mean distance matrix. Note that many of the conformations are in the extended state—indicating less likelihood of interactions between the α-helices. **(E)** Cartoon representation from the PaaA2 ensemble from the optimized FF simulation along with the **(F)** normalized mean distance matrix. The conformations generated by the optimized FF are more compact than the other two datasets mainly because the sampling from the optimization runs are limited.

(Flexible-Meccano, **Figure 6A**) shows the presence of large-scale fluctuations in the orientation between the two α-helices. Each conformer in the ensemble is colored using its R_g value to highlight the nature of compactness (darker shades of red indicate larger R_g, implying less compact states). To better characterize the nature of these fluctuations, we chose to examine the average (normalized) distance matrix for the experimental ensemble (**Figure 6B**). This provides us a qualitative measure of the long-range interactions between specific regions of the PaaA2 ensemble. The MD simulations from the original FF capture some of the large-scale fluctuations, however is not fully representative of the experimental data (**Figure 6C**). Notably, within the experimental ensemble, there are some interactions between the two α-helices, which are not represented in the original FF simulations (**Figure 6D**). Although visually the average distance matrices look similar, the ensemble generated from the MD simulations using the original FF is dominated by mostly extended states (thus de-emphasizing the interactions between the two α-helices). The simulations from the optimized FF, on the other hand highlight mostly compact conformations (**Figure 6E**). An examination of the distance matrix (**Figure 6F**) also shows that there are significantly larger number of interactions between the two α-helices and only localized fluctuations in their relative orientations. We posit that this observation may be a consequence of limited sampling of the conformational landscape (∼5 ns every iteration of the optimization).

3.3. ForceBalance-SAS Optimized FF Parameters Are Partially Transferable at Shorter Timescales

We lastly sought to determine whether our optimized parameters would improve the experimental SAXS agreement for an independent test case. We hypothesized that an appropriate test case would be a protein with a similar charge/hydrophobicity (Uversky) profile, as this has been shown to predict relative disorder/order. For the training system PaaA2, a protein close on the Uversky plot is SH4UD. For this system, we were able to observe a reduction in χ^2 from 9.7 to 7.2 (**Figure 7A**), with improvements in agreement seen in the mid-range to high q regions of the Kratky plot (**Figure 7B**). Note that this simulation (with the PaaA2 FF parameters) was carried out only for 5 ns—corresponding to the same timescales of the optimization cycle. Although the improvement in the χ^2 value is somewhat limited in the high q values, we still observe that the ensembles have a better agreement with the SAXS observables.

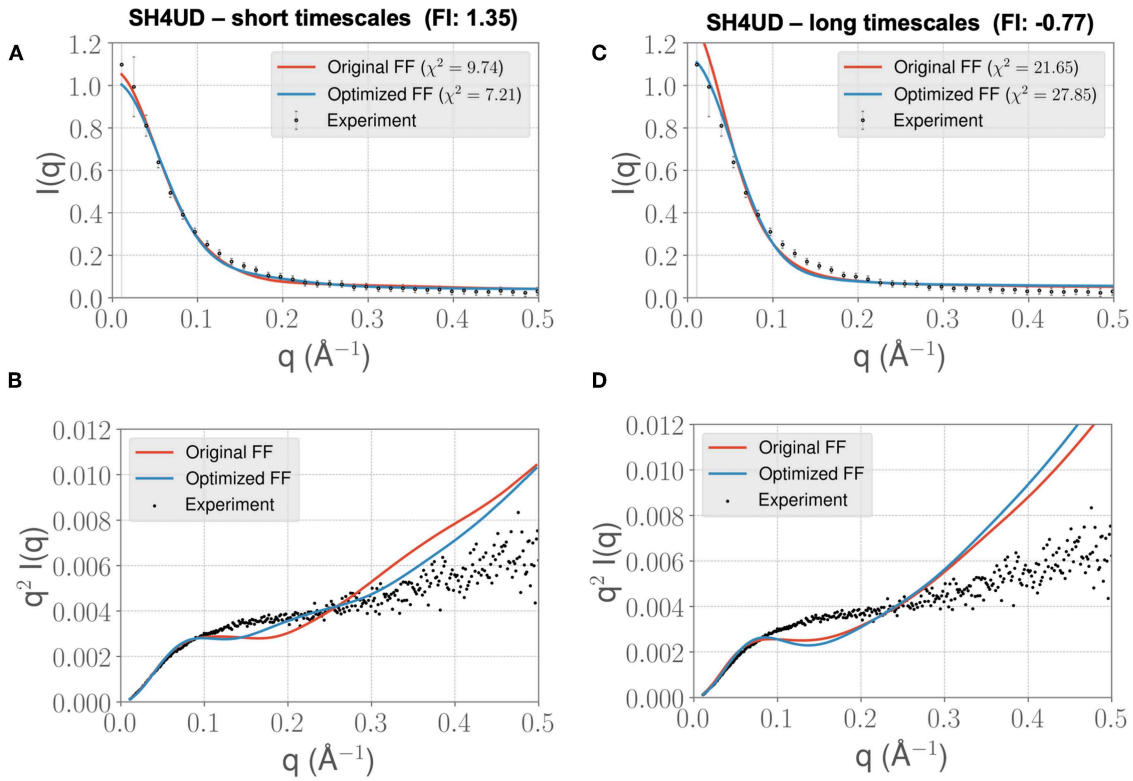


FIGURE 7 | FF parameters learned from the PaaA2 simulations used to simulate the SH4UD IDR improves the fit to experimental SAXS data. Although the factor of improvement (FI) is lower than the other two systems **(A)**, the fit to the experimental data as seen from the Kratky plot **(B)** shows better agreement in the mid-*q* range. This allows us to determine that the parameters learned from one simulation can be used reasonably on other proteins as well. Further fine-tuning may be essential to obtain better fits (especially with solvent-protein interactions). **(C)** and **(D)** highlight the same information as in **(A)** and **(B)** but for longer timescales. Note that the factor of improvement has reversed.

However, when we extend the simulations to about 0.3 μs, we find that the agreement between experimental SAXS and the MD ensemble deteriorates (see **Figures 7C,D**). This observation is significant, given the fact that the PaaA2 ensemble consists of two well-defined α-helices (a feature is mostly well-described by existing FFs) and the SH4UD consists of only transient helices, which are not fully captured at the timescales of our current simulations. Further studies would be necessary to validate these simulations (and the transferability of the FF parameters at longer timescales) against available experimental data.

4. DISCUSSION

We have presented a proof-of-concept demonstration to optimize a set of FF parameters using small-angle scattering data on a protein-by-protein basis. We started with a few assumptions, including that (1) simulations would be initiated from a single starting structure (for e.g., from an experimental crystal structure), (2) MD simulations would be performed under some equilibrium conditions without necessitating enhanced sampling techniques, such as replica exchange, and (3) longer time-scale simulations ($O(\mu s)$ would not be accessible for all systems of interest. Such assumptions, especially in the context of IDP systems may seem limiting, given that both enhanced sampling and ensemble MD simulation techniques are known to improve the overall ability of MD simulations to "match" experimental observations (Lee and Chen, 2016; Holehouse et al., 2017; Bhattacharya and Lin, 2019). We believe that the optimization scheme outlined here can be extended in a straightforward way for ensemble MD strategies, and it would need some modifications for enhanced sampling strategies. This is a direction that we will pursue in the near future.

The fact that our method seemed to change the torsional parameters much more than the van der Waals is noteworthy. As mentioned previously, the torsional components are covalent energetic degrees of freedom, but also implicitly contain a degree of non-covalent character, given the larger 1-4 separation of the atoms (as opposed to the 1–2 and 1–3 separations for bond stretching and angle bending, which can more definitively be considered purely covalent). It is therefore likely that short-ranged non-covalent energetics that are not explicitly accounted for in typical force field functional forms are being folded into the torsional term.

We note that the fitting procedure used in ForceBalance-SAS improves the agreement with independent observations, such as NMR chemical shifts. NMR chemical shifts represent effective local measurements for conformational changes in an ensemble and provide a powerful technique to characterize

IDP/IDR ensembles in the context of their biological function (Pérez et al., 2009; Sterckx et al., 2014; Arbesü et al., 2017). Our optimization procedure takes into account only the SAS measurements. However, by fitting our MD ensembles to SAS curves, we also found that it consequently improved the agreement of local measurements. In the context of modeling IDP/IDR ensembles, our approach therefore represents a complementary approach to using multiple experimental methods to capture atomistic details of these systems. Whereas approaches such as Flexible-Meccano (and other tools) utilize all of the available experimental data to model IDP/IDR ensembles, our iterative approach can be modified to take into account gross structural features first, and then followed by further tuning FF parameters to recapitulate fine-grained features.

We also showed that the optimized FF parameters developed for an IDP could be transferred (in a limited manner) to other IDPs. Although the improvement in agreement between experiments and simulations was only marginal, we were still able to recapitulate some of the finer grained details of the SH4UD ensemble better than the original FF at short simulation length. The parameters that get optimized most likely depend on the amount of sampling carried out at each iteration. While preliminary testing indicated that calculated SAXS profiles appeared to converge at about 5 ns for each iteration, it is likely that this may not hold for all IDP systems of interest, especially those that are larger than the systems studied here. Indeed, the rugged free energy/conformational landscapes of IDP are very different from those of systems such as neat water to which the parent ForceBalance method had been previously applied (Wang et al., 2013, 2014; Laury et al., 2015). Nonetheless, the fact that longer simulations at 100s of nanoseconds performed with parameters obtained from a 5-ns simulation length still show improved agreement of the MD ensemble with the experimental SAXS supports the view that major signatures of the full ensemble are captured and can be optimized against to yield the observed improvement at longer timescales. Further work on the reproducibility of our approach is also needed, especially in the context of benchmark IDP/IDR ensembles that have been recently made available (Varadi et al., 2013). To this end, the effect of the simulation length in ForceBalance-SAS on the resulting parameters will be investigated in the future.

We are endeavoring to enhance this method on a number of fronts. We plan on addressing the sampling issue by deploying this method on more powerful supercomputers so that longer simulations in each cycle of the algorithm are less onerous. We also note that in all cases, the ability to optimize in the higher q range was poorer than in the low q range, as is best depicted in the Kratky plots. This indicates that in the current regime, we are optimizing global scale interactions better than fine scale interactions. Therefore, it is only natural that a worthwhile objective is to differentially weight the contributions of different q regions to the objective function during the optimization. Furthermore, current work is focused on optimizing FF parameters using the experimental data of multiple protein targets simultaneously, which should enhance the transferability of the optimized parameters. Nonetheless, for those who are interested in detailed simulation studies of specific systems, the current system-by-system approach is useful.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

OD, AR, JM, JS, and LP conceived the project. OD developed methodology, implemented and tested the techniques and ran simulations. US and LP ran simulations. OD and AR contributed analysis tools and analyzed the data. All authors wrote, edited, and approved the manuscript.

ACKNOWLEDGMENTS

We would like to thank Hugh O'Neill, Puneet Juneja, and Sai Venkatesh Pingali of the X-ray Laboratory at the Spallation Neutron Source, Oak Ridge National Laboratory for kindly providing the experimental SAXS for SH4UD. We thank Heng Ma for his assistance in preparing **Figure 5**. We would also like to thank Lee-Ping Wang for providing advice and comments in our algorithmic development. The authors acknowledge the support of the Genomic Science Program, Office of Biological and Environmental Research (OBER), U. S. Department of Energy, under Contract FWP ERKP300.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00064/full#supplementary-material

Supporting information consists of additional figures for the generated IDP ensembles. The MD simulation datasets, as well as the analysis codes are available upon request.

REFERENCES

Abraham, M. J., Murtola, T., Schulz, R., Pall, S., Smith, J. C., Hess, B., et al. (2015). Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25. doi: 10.1016/j.softx.2015.06.001

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181, 223–230. doi: 10.1126/science.181.4096.223

Arai, M., Sugase, K., Dyson, H. J., and Wright, P. E. (2015). Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci. U.S.A.* 112, 9614–9619. doi: 10.1073/pnas.1512799112

Arbesü, M., Maffei, M., Cordeiro, T. N., Teixeira, J. M., Pérez, Y., Bernadó, P., et al. (2017). The unique domain forms a fuzzy intramolecular complex in src family kinases. *Structure* 25, 630–640.e4. doi: 10.1016/j.str.2017.02.011

Berendsen, H. J. C., Postma, J. P. M., Vangunsteren, W. F., Dinola, A., and Haak, J. R. (1984). Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684–3690. doi: 10.1063/1.448118

Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle x-ray scattering. *J. Am. Chem. Soc.* 129, 5656–5664. doi: 10.1021/ja069124n

Bernadó, P., and Svergun, D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Mol. Biosyst.* 8, 151–167. doi: 10.1039/c1mb05275f

Best, R. B., and Mittal, J. (2010). Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B* 114, 14916–14923. doi: 10.1021/jp108618d

Best, R. B., Zheng, W., Borgia, A., Buholzer, K., Borgia, M. B., Hofmann, H., et al. (2018). Comment on "innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water." *Science* 361:eaar7101. doi: 10.1126/science.aar7101

Best, R. B., Zheng, W. W., and Mittal, J. (2014). Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theor. Comput.* 10, 5113–5124. doi: 10.1021/ct500569b

Bhattacharya, S., and Lin, X. (2019). Recent advances in computational protocols addressing intrinsically disordered proteins. *Biomolecules* 9:E146. doi: 10.3390/biom9040146

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (Prague), 108–122.

Busch, D. J., Houser, J. R., Hayden, C. C., Sherman, M. B., Lafer, E. M., and Stachowiak, J. C. (2015). Intrinsically disordered proteins drive membrane curvature. *Nat. Commun.* 6:7875. doi: 10.1038/ncomms8875

Charavay, C., Bauer, F., Huang, J.-R., Salmon, L., Jensen, M. R., Blackledge, M., et al. (2012). Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28, 1463–1470. doi: 10.1093/bioinformatics/bts172

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald–an n.log(n) method for ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

Das, R. K., and Pappu, R. V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13392–13397. doi: 10.1073/pnas.1304749110

Dennis, John E., J., Gay, D. M., and Welsh, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Softw.* 7, 369–383.

Drulyte, I., Johnson, R. M., Hesketh, E. L., Hurdiss, D. L., Scarff, C. A., Porav, S. A., et al. (2018). Approaches to altering particle distributions in cryo-electron microscopy sample preparation. *Acta Crystallogr. D Struct. Biol.* 74(Pt 6), 560–571. doi: 10.1107/S2059798318006496

Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., et al. (2017). Atsas 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* 50, 1212–1225. doi: 10.1107/S1600576717007786

Fuertes, G., Banterle, N., Ruff, K. M., Chowdhury, A., Mercadante, D., Koehler, C., et al. (2017). Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in saxs vs. fret measurements. *Proc. Natl. Acad. Sci. U.S.A.* 114, E6342–E6351. doi: 10.1073/pnas.1704692114

Grishaev, A., Wu, J., Trewhella, J., and Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle x-ray scattering and nmr data. *J. Am. Chem. Soc.* 127, 16621–16628. doi: 10.1021/ja054342m

Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014). Introducing protein intrinsic disorder. *Chem. Rev.* 114, 6561–6588. doi: 10.1021/cr400514h

Han, B., Liu, Y., Ginzinger, S. W., and Wishart, D. S. (2011). Shiftx2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* 50:43. doi: 10.1007/s10858-011-9478-4

Hess, B. (2008). P-lincs: a parallel linear constraint solver for molecular simulation. *J. Chem. Theor. Comput.* 4, 116–122. doi: 10.1021/ct700200b

Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997). Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18, 1463–1472.

Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., and Schuler, B. (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16155–16160. doi: 10.1073/pnas.1207719109

Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G., and Pappu, R. V. (2017). Cider: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. *Biophys. J.* 112, 16–21. doi: 10.1016/j.bpj.2016.11.3200

Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., et al. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *J. Chem. Phys.* 120, 9665–9678. doi: 10.1063/1.1683075

Huang, J., and MacKerell, A. D. (2018). Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 48, 40–48. doi: 10.1016/j.sbi.2017.10.008

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* 14, 71–73. doi: 10.1038/Nmeth.4067

Jensen, M. R., Zweckstetter, M., Huang, J. R., and Blackledge, M. (2014). Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using nmr spectroscopy. *Chem. Rev.* 114, 6632–6660. doi: 10.1021/cr400688u

Karplus, M. (2002). Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.* 35, 321–323. doi: 10.1021/ar020082r

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652. doi: 10.1038/nsb0902-646

Kikhney, A. G., and Svergun, D. I. (2015). A practical guide to small angle x-ray scattering (saxs) of flexible and intrinsically disordered proteins. *FEBS Lett.* 589, 2570–2577. doi: 10.1016/j.febslet.2015.08.027

Knowles, T. P., Vendruscolo, M., and Dobson, C. M. (2014). The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* 15:384. doi: 10.1038/nrm3810

Kosciolek, T., Buchan, D. W. A., and Jones, D. T. (2017). Predictions of backbone dynamics in intrinsically disordered proteins using *de novo* fragment-based protein structure predictions. *Sci. Rep.* 7:6999. doi: 10.1038/s41598-017-07156-1

Latysheva, N. S., Flock, T., Weatheritt, R. J., Chavali, S., and Babu, M. M. (2015). How do disordered regions achieve comparable functions to structured domains? *Protein Sci.* 24, 909–922. doi: 10.1002/pro.2674

Laury, M. L., Wang, L. P., Pande, V. S., Head-Gordon, T., and Ponder, J. W. (2015). Revised parameters for the amoeba polarizable atomic multipole water model. *J. Phys. Chem. B* 119, 9423–9437. doi: 10.1021/jp510896n

Lee, C., Kim, D. H., Lee, S. H., Su, J., and Han, K. H. (2016). Structural investigation on the intrinsically disordered n-terminal region of hpv16 e7 protein. *BMB Rep.* 49, 431–436. doi: 10.5483/BMBRep.2016.49.8.021

Lee, K. H., and Chen, J. (2016). Multiscale enhanced sampling of intrinsically disordered protein conformations. *J. Comput. Chem.* 37, 550–557. doi: 10.1002/jcc.23957

Levenberg, K. Q. (1944). A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.* 2, 164–168.

Levine, Z. A., Larini, L., LaPointe, N. E., Feinstein, S. C., and Shea, J. E. (2015). Regulation and aggregation of intrinsically disordered peptides. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2758–2763. doi: 10.1073/pnas.1418155112

Lopes, P. E., Guvench, O., MacKerell, and Alexander D, J. (2015). Current status of protein force fields for molecular dynamics simulations. *Methods Mol. Biol.* 1215, 47–71. doi: 10.1007/978-1-4939-1465-4_3

Marquardt, D. W. (1963). An algortithm for least-squares estimation of nonlinear parameters. *J. Soc. Indus. Appl. Math.* 11, 431–441.

Marsh, J. A., and Forman-Kay, J. D. (2009). Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.* 391, 359–374. doi: 10.1016/j.jmb.2009.06.001

Marsh, J. A., Neale, C., Jack, F. E., Choy, W. Y., Lee, A. Y., Crowhurst, K. A., et al. (2007). Improved structural characterizations of the drkn sh3 domain unfolded

state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.* 367, 1494–1510. doi: 10.1016/j.jmb.2007.01.038

Mollica, L., Bessa, L. M., Hanoulle, X., Jensen, M. R., Blackledge, M., and Schneider, R. (2016). Binding mechanisms of intrinsically disordered proteins: theory, simulation, and experiment. *Front. Mol. Biosci.* 3:52. doi: 10.3389/fmolb.2016.00052

More, J. J., and Sorensen, D. C. (1983). Computing a trust region step. *SIAM J. Sci. Stat. Comput.* 4, 553–572.

Nerenberg, P. S., and Head-Gordon, T. (2011). Optimizing protein-solvent force fields to reproduce intrinsic conformational preferences of model peptides. *J. Chem. Theor. Comput.* 7, 1220–1230. doi: 10.1021/ct2000183

Nettels, D., Muller-Spath, S., Kuster, F., Hofmann, H., Haenni, D., Ruegger, S., et al. (2009). Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20740–20745. doi: 10.1073/pnas.0900622106

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins* 53, 566–572. doi: 10.1002/prot.10532

Parrinello, M., and Rahman, A. (1981). Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* 52, 7182–7190. doi: 10.1063/1.328693

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pelikan, M., Hura, G. L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle x-ray scattering. *Gen. Physiol. Biophys.* 28, 174–189. doi: 10.4149/gpb_2009_02_174

Pérez, Y., Gairí, M., Pons, M., and Bernadó, P. (2009). Structural characterization of the natively unfolded n-terminal domain of human c-src kinase: insights into the role of phosphorylation of the unique domain. *J. Mol. Biol.* 391, 136–148. doi: 10.1016/j.jmb.2009.06.018

Pérez, Y., Maffei, M., Igea, A., Amata, I., Gairi, M., Nebreda, A. R., et al. (2013). Lipid binding by the unique and sh3 domains of c-src suggests a new regulatory mechanism. *Sci. Rep.* 3:1295. doi: 10.1038/srep01295

Piana, S., Donchev, A. G., Robustelli, P., and Shaw, D. E. (2015). Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* 119, 5113–5123. doi: 10.1021/jp508971m

Piana, S., Klepeis, J. L., and Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 24, 98–105. doi: 10.1016/j.sbi.2013.12.006

Rauscher, S., Gapsys, V., Gajda, M. J., Zweckstetter, M., de Groot, B. L., and Grubmuller, H. (2015). Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theor. Comput.* 11, 5513–5524. doi: 10.1021/acs.jctc.5b00736

Receveur-Bréchot, V., and Durand, D. (2012). How random are intrinsically disordered proteins? a small angle scattering perspective. *Curr. Protein Pept. Sci.* 13, 55–75. doi: 10.2174/138920312799277901

Riback, J. A., Bowman, M. A., Zmyslowski, A., Knoverek, C. R., Jumper, J., Kaye, E. B., et al. (2018). Response to comment on "innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water"? *Science* 361:eaar7949. doi: 10.1126/science.aar7949

Riback, J. A., Bowman, M. A., Zmyslowski, A. M., Knoverek, C. R., Jumper, J. M., Hinshaw, J. R., et al. (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* 358, 238–241. doi: 10.1126/science.aan5774

Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* 115, E4758–E4766. doi: 10.1073/pnas.1800690115

Robustelli, P., Stafford, K. A., and Palmer, A. G. (2012). Interpreting protein structural dynamics from nmr chemical shifts. *J. Am. Chem. Soc.* 134, 6365–6374. doi: 10.1021/ja300265w

Rozycki, B., Kim, Y. C., and Hummer, G. (2011). Saxs ensemble refinement of escrt-iii chmp3 conformational transitions. *Structure* 19, 109–116. doi: 10.1016/j.str.2010.10.006

Schwieters, C. D., Suh, J. Y., Grishaev, A., Ghirlando, R., Takayama, Y., and Clore, G. M. (2010). Solution structure of the 128 kda enzyme i dimer from escherichia coli and its 146 kda complex with hpr using residual dipolar couplings and small- and wide-angle x-ray scattering. *J. Am. Chem. Soc.* 132, 13026–13045. doi: 10.1021/ja105485b

Shirts, M. R., and Pande, V. S. (2005). Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* 122:134508. doi: 10.1063/1.1877132

Shirts, M. R., Pitera, J. W., Swope, W. C., and Pande, V. S. (2003). Extremely precise free energy calculations of amino acid side chain analogs: comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* 119, 5740–5761. doi: 10.1063/1.1587119

Skinner, J. J., Yu, W., Gichana, E. K., Baxa, M. C., Hinshaw, J. R., Freed, K. F., et al. (2014). Benchmarking all-atom simulations using hydrogen exchange. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15975–15980. doi: 10.1073/pnas.14042 13111

Song, D., Luo, R., and Chen, H. F. (2017). The idp-specific force field ff14idpsff improves the conformer sampling of intrinsically disordered proteins. *J. Chem. Inform. Model.* 57, 1166–1178. doi: 10.1021/acs.jcim.7b00135

Sterckx, Y. G., Jové, T., Shkumatov, A. V., Garcia-Pino, A., Geerts, L., De Kerpel, M., et al. (2016). A unique hetero-hexadecameric architecture displayed by the escherichia coli o157 paaa2-pare2 antitoxin-toxin complex. *J. Mol. Biol.* 428, 1589–1603. doi: 10.1016/j.jmb.2016.03.007

Sterckx, Y. G. J., Volkov, A. N., Vranken, W. F., Kragelj, J., Jensen, M. R., Buts, L., et al. (2014). Small-angle x-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin paaa2. *Structure* 22, 854–865. doi: 10.1016/j.str.2014.03.012

Svergun, D., Barberato, C., and Koch, M. H. J. (1995). Crysol–a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28, 768–773. doi: 10.1107/S00218898950 07047

Tatosyan, A. G., and Mizenina, O. A. (2000). Kinases of the src family: structure and functions. *Biochemistry* 65, 49–58. Available online at: http://protein.bio.msu.ru/biokhimiya/contents/v65/pdf/bcm_0049.pdf

Tiwary, P., Limongelli, V., Salvalaglio, M., and Parrinello, M. (2015). Kinetics of protein–ligand unbinding: predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U.S.A.* 112, E386–E391. doi: 10.1073/pnas.1424461112

Uversky, V. N. (2011). Intrinsically disordered proteins from a to z. *Int. J. Biochem. Cell Biol.* 43, 1090–1103. doi: 10.1016/j.biocel.2011.04.001

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

Uversky, V. N., Roman, A., Oldfield, C. J., and Dunker, A. K. (2006). Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in e6 and e7 oncoproteins from high risk hpvs. *J. Proteome Res.* 5, 1829–1842. doi: 10.1021/pr0602388

Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). Gromacs: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718. doi: 10.1002/jcc.20291

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A. K., et al. (2013). pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* 42, D326–D335. doi: 10.1093/nar/gkt960

Vega, C., and Abascal, J. L. F. (2005). Relation between the melting temperature and the temperature of maximum density for the most common models of water. *J. Chem. Phys.* 123:144504. doi: 10.1063/1.2056539

Wang, J. B., Zuo, X. B., Yu, P., Byeon, I. J. L., Jung, J. W., Wang, X. X., et al. (2009). Determination of multicomponent protein structures in solution using global orientation and shape restraints. *J. Am. Chem. Soc.* 131, 10507–10515. doi: 10.1021/ja902528f

Wang, L.-P., Head-Gordon, T., Ponder, J. W., Ren, P., Chodera, J. D., Eastman, P. K., et al. (2013). Systematic improvement of a classical molecular model of water. *J. Phys. Chem. B* 117, 9956–9972. doi: 10.1021/jp403802c

Wang, L. P., Martinez, T. J., and Pande, V. S. (2014). Building force fields: an automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* 5, 1885–1891. doi: 10.1021/jz500737m

Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., et al. (2008). Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proc. Natl. Acad. Sci. U.S.A.* 105, 5762–5767. doi: 10.1073/pnas.0801353105

Wheeler, D. L., Iida, M., and Dunn, E. F. (2009). The role of src in solid tumors. *Oncologist* 14, 667–678. doi: 10.1634/theoncologist.2009-0009

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293, 321–331.

Wright, P. E., and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signaling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29. doi: 10.1038/nrm3920

Xiang, S., Gapsys, V., Kim, H. Y., Bessonov, S., Hsiao, H. H., Möhlmann, S., et al. (2013). Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. *Structure* 21, 2162–2174. doi: 10.1016/j.str.2013.09.014

Yang, S. C., Blachowicz, L., Makowski, L., and Roux, B. (2010). Multidomain assembled states of hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. U.S.A.* 107, 15757–15762. doi: 10.1073/pnas.1004569107

Disclaimer: This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 June 2019
doi: 10.3389/fmolb.2019.00044

Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations

Steve Agajanian[1], *Odeyemi Oluyemi*[1] *and Gennady M. Verkhivker*[1,2]*

[1] Graduate Program in Computational and Data Sciences, Schmid College of Science and Technology, Chapman University, Orange, CA, United States, [2] Department of Biomedical and Pharmaceutical Sciences, Chapman University School of Pharmacy, Irvine, CA, United States

OPEN ACCESS

Edited by:
Shozeb Haider,
University College London,
United Kingdom

Reviewed by:
Arvind Ramanathan,
Argonne National Laboratory (DOE),
United States
Debsindhu Bhowmik,
Oak Ridge National Laboratory (DOE),
United States

***Correspondence:**
Gennady M. Verkhivker
verkhivk@chapman.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *28 February 2019*
Accepted: *23 May 2019*
Published: *11 June 2019*

Citation:
Agajanian S, Oluyemi O and Verkhivker GM (2019) Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations. Front. Mol. Biosci. 6:44. doi: 10.3389/fmolb.2019.00044

Development of machine learning solutions for prediction of functional and clinical significance of cancer driver genes and mutations are paramount in modern biomedical research and have gained a significant momentum in a recent decade. In this work, we integrate different machine learning approaches, including tree based methods, random forest and gradient boosted tree (GBT) classifiers along with deep convolutional neural networks (CNN) for prediction of cancer driver mutations in the genomic datasets. The feasibility of CNN in using raw nucleotide sequences for classification of cancer driver mutations was initially explored by employing label encoding, one hot encoding, and embedding to preprocess the DNA information. These classifiers were benchmarked against their tree-based alternatives in order to evaluate the performance on a relative scale. We then integrated DNA-based scores generated by CNN with various categories of conservational, evolutionary and functional features into a generalized random forest classifier. The results of this study have demonstrated that CNN can learn high level features from genomic information that are complementary to the ensemble-based predictors often employed for classification of cancer mutations. By combining deep learning-generated score with only two main ensemble-based functional features, we can achieve a superior performance of various machine learning classifiers. Our findings have also suggested that synergy of nucleotide-based deep learning scores and integrated metrics derived from protein sequence conservation scores can allow for robust classification of cancer driver mutations with a limited number of highly informative features. Machine learning predictions are leveraged in molecular simulations, protein stability, and network-based analysis of cancer mutations in the protein kinase genes to obtain insights about molecular signatures of driver mutations and enhance the interpretability of cancer-specific classification models.

Keywords: cancer driver mutations, machine learning classifiers, ensemble-based machine learning features, random forest, deep learning, convolutional neural networks, drug discovery

INTRODUCTION

Deep sequencing studies have enabled a detailed characterization of cancer genomes and unveiled important gene-specific signatures of somatic mutations (Davies et al., 2002; Bardelli et al., 2003; Futreal et al., 2004; Samuels et al., 2004; Stephens et al., 2004, 2005; Wang et al., 2004; Sjoblom et al., 2006; Greenman et al., 2007; Wood et al., 2007; Vogelstein et al., 2013; Watson et al., 2013). The steadily growing amount of data generated in cancer genomic studies and next-generation sequencing (NGS) have been the impetus behind formation of international cancer genomic projects and development of large bioinformatics data resources such as Cancer Genome Atlas (TCGA), Genomics Data Commons Portal (https://portal.gdc.cancer.gov/) (Weinstein et al., 2013; Jensen et al., 2017), COSMIC database (http://cancer.sanger.ac.uk) (Forbes et al., 2015), and the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010; Zhang et al., 2011; Klonowska et al., 2016; Hinkson et al., 2017). The Cancer Gene Census of the Catalog of Somatic Mutations in Cancer (COSMIC) database has grown from 291 well-characterized cancer genes (Futreal et al., 2004) to more than 500 entries (Forbes et al., 2015) where some cancer genes can be commonly mutated across cancer types, while other genes are predominantly cancer-specific. The cBio Cancer Genomics Portal (https://www.cbioportal.org/) is an open-access resource for exploration of large cancer genomics data sets (Cerami et al., 2012; Gao et al., 2013). These datasets have allowed for comprehensive genome-wide analyses of genetic alterations in multiple tumor types (Poulos and Wong, 2018). A relatively small fraction of somatic variants known as driver mutations have considerable functional effects and can be acquired over time as a result of a range of mutational processes, rather than inherited (Haber and Settleman, 2007; Lawrence et al., 2013; Vogelstein et al., 2013). A comprehensive analysis of cancer driver genes and mutations has provided classification of 751,876 unique missense mutations, producing a dataset of 3,442 functionally validated driver mutations (Bailey et al., 2018). Another significant dataset of 1,049 experimentally tested and functionally validated driver mutations (Ng et al., 2018) has expanded our knowledge of cancer-causing variants in oncogenes and tumor suppressor genes. TCGA organized the Multi-Center Mutation Calling in Multiple Cancers (MC3) network project which generated a comprehensive and consistent collection of somatic mutation calls for the 10,437 tumor samples dataset (Ellrott et al., 2018). Computational approaches that assess the impact of somatic mutations are often characterized by different basic assumptions, types of input information, models, and prediction targets such as driver gene or driver mutation (Gonzalez-Perez et al., 2013; Cheng et al., 2016).

A number of somatic variant callers based on various statistical and machine learning approaches are now available for somatic mutation detection, including MuTect2 (Cibulskis et al., 2013), MuSE (Fan et al., 2016), VarDict (Lai et al., 2016), VarScan2 (Koboldt et al., 2012), Strelka2 (Kim et al., 2018), SomaticSniper (Larson et al., 2012), and SNooPer (Spinella et al., 2016). A deep convolutional neural network (CNN) approach termed DeepVariant can identify genetic variation in NGS data by discerning statistical relationships around putative variant sites (Poplin et al., 2018). To facilitate systematic and standardized somatic variant refinement from cancer sequencing data, random forest (RF) models and deep learning (DL) approach were utilized, showing that these machine learning techniques could achieve high and similar classification performance across all variant refinement classes (Ainscough et al., 2018). A machine learning approach called Cerebro increased the accuracy of calling validated somatic mutations in tumor samples and outperformed several other somatic mutation detection methods (Wood et al., 2018).

Many computational methods have been proposed for prediction of cancer driver genes. Some of these approaches use cohort-based analysis to detect driver genes, including ActiveDriver (Reimand and Bader, 2013), MutSigCV (Lawrence et al., 2013), MuSiC (Dees et al., 2012), OncodriveCLUST (Tamborero et al., 2013), OncodriveFM (Gonzalez-Perez and Lopez-Bigas, 2012), and OncodriveFML (Mularoni et al., 2016). The success of hybrid methods for scoring coding variants has indicated that integration of different tools may enhance predictive accuracy for both coding and non-coding variants (Li et al., 2015). A deep learning-based method (deepDriver) predicts driver genes by CNN trained with mutation-based feature matrix constructed using similarity networks (Luo et al., 2019). Since many methods are often found to predict distinct or partially overlapping subsets of cancer driver genes, a consensus-based strategy was recently proposed, showing considerable promise and outperforming the individual approaches (Bertrand et al., 2018). A unified machine learning-based evaluation framework for analysis of driver gene predictions compared the performance of these methods, showing that the driver genes predicted by individual tools can vary widely (Tokheim C. et al., 2016; Tokheim C. J. et al., 2016).

Computational methods designed to identify driver mutations have become increasingly important to facilitate an automated assessment of functional and clinical impacts (Gnad et al., 2013; Ding et al., 2014; Martelotto et al., 2014; Raphael et al., 2014; Cheng et al., 2016). Functional computational prediction methods include Sorted Intolerant From Tolerant (SIFT) (Sim et al., 2012), PolyPhen-2 (Adzhubei et al., 2010), Mutation Assessor (Reva et al., 2011), MutationTaster (Schwarz et al., 2010), CONsensus DELeteriousness score of missense mutations (Condel) (Gonzalez-Perez and Lopez-Bigas, 2011), Protein Variation Effect Analyzer (PROVEAN) (Choi et al., 2012), and Functional Analysis Through Hidden Markov Models (FATHMM) (Shihab et al., 2013). Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM) (Carter et al., 2009; Douville et al., 2013; Masica et al., 2017), Cancer Driver Annotation (CanDrA) (Mao et al., 2013), and FATHMM (Shihab et al., 2013). Many new approaches have recently addressed a problem of locating driver mutations within the non-coding genome regions (Piraino and Furney, 2016). The identification of cancer mutation hotspots in protein structures has been a fruitful approach for identifying driver mutations (Dixit et al., 2009; Dixit and Verkhivker, 2011; Gao et al., 2013; Gauthier et al., 2016; Niu et al., 2016; Tokheim C. et al., 2016; Tokheim C. J. et al., 2016). To consolidate

functional annotation for SNVs discovered in exome sequencing studies, a database of human non-synonymous SNVs (dbNSFP) was developed (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016). This resource allows for computation of a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches and 15 conservation features (Wu et al., 2016). In our recent investigation, two cancer-specific machine learning classifiers were proposed that utilized 48 functional scores from dbWGFP server in classification of cancer driver mutations (Agajanian et al., 2018).

In this work, we explore and integrate RF and DL/CNN machine learning approaches for prediction and classification of cancer driver mutations. We first explore the ability of CNN models to identify and classify cancer driver mutations directly from raw nucleotide sequence information without relying on specific functional scores. The performance of these classifiers was compared to RF and gradient boosted tree (GBT) methods to provide a comparative analysis of various classification models. These raw sequence-derived scores are advantageous because they can be obtained for any mutation with a known chromosome and position, whereas the functional scoring features can be limited to subsets of genomic mutations. By developing a successful classification scheme that could leverage information from raw DNA sequences, the universe of classifiable mutations can be greatly expanded leading to more general and robust machine learning tools. The results of this study reveal that CNN models can learn high importance features from genomic information that are complementary to the ensemble-based predictor scores traditionally employed in machine learning classification of cancer mutations**.** We show that integration of the DL-derived predictor score with only several ensemble-based features can recapitulate the results obtained with a large number of functional features and improve performance in capturing driver mutations across a spectrum of machine learning classifiers. Machine learning predictions are leveraged in biophysical simulations and network analysis of protein kinase oncogenes to obtain more detailed functional information about molecular signatures of activating driver mutations, aiding in the interpretability of cancer mutation classifiers.

MATERIALS AND METHODS

Mutational Datasets and Feature Selection

In our earlier study (Agajanian et al., 2018) we used RF classifier to predict cancer driver mutations using a combination of two golden datasets (Mao et al., 2013; Martelotto et al., 2014). Here, we expanded this dataset by adding the predicted cancer driver mutations and passengers from the analysis of missense mutations in Cbioportal database (Agajanian et al., 2018). By leveraging the earlier analysis, we created a dataset consisting of functionally validated 6,389 cancer driver mutations and 12,941 passenger mutations. The driver/passenger classifications for 2,570 of these mutations were present in the two aforementioned golden datasets, and our RF classifier made predictions on the remaining 16,760 missense mutations from the Cbioportal database. Given the performance level of our model (Agajanian et al., 2018), we conjectured that a combination of the two golden datasets and the missense mutations in the Cbioportal database would yield an informative dataset for the current study. The initially selected features for RF predictions were obtained from dbWGFP web server (Wu et al., 2016) of functional predictions for human whole-genome single nucleotide variants (**Supplementary Table S1**). A total of 32 sequence-based, evolutionary and functional features identified in our previous study (Agajanian et al., 2018) were initially used for machine learning experiments with the new dataset of cancer mutations. In cancer driver mutation predictions, traditional input data contain distinct features that cannot be directly applied to CNN models due to their lack of spatial meaning. Using the chromosome and the position on that chromosome that corresponded to the mutated nucleotide, we could retrieve the surrounding nucleotides of the mutation of interest to perform classification with only this raw string of nucleotides. To represent the original nucleotide and its mutated version, we placed two nucleotide sequences on top of each other, one containing the original string, and the other contained the mutated version. This would only result in a one nucleotide difference between the two, allowing to effectively utilizing the sliding window format of the CNN models. The schematic workflow diagram of the CNN approach employed in this study is presented in **Figure 1**.

To create this dataset, we parsed information from University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) (Tyner et al., 2017) which takes a chromosome (CHR) and a position (POS) on that chromosome as arguments and returns back all nucleotides within the sequence. Using the dataset consisting of 6,389 driver mutations and 12,941 passengers, we created 5 different datasets of various window sizes around each given CHR/POS pair. The explored window sizes (10, 50, 100, 500, and 5,000) produced nucleotide strings of length 21, 101, 201, 1,001, and 10,001, respectively. To represent the type of mutation (A->C, A->G, etc.) we stacked two of the same nucleotide sequences on top of each other, having one contain the original nucleotide at the position passed in initially, and the other containing the mutated version (**Figure 2A**). This operation resulted in a total input matrix size of (2, 21), (2, 101), (2, 201), (2, 1001), and (2, 10001), respectively. Three different preprocessing techniques were then applied to the dataset to allow it to be passed into the CNN model in the numerical form: label encoding (**Figure 2B**), one-hot encoding (**Figure 2C**; Goh et al., 2017), and embedding (**Figure 2D**). Label encoding involves assigning each nucleotide its own unique ID (A->0, C->1, etc.) This imposes an ordering on the nucleotide sequences that may have implications for the neural network learning (**Figure 2B**). This technique was implemented using the Scikit-learn LabelEncoder package for the Python programming language. We also tried one-hot encoding the dataset by assigning each nucleotide its own bit encoded string (A -> [0,0,0,0,1], C-> [0,0,0,1,0]) (**Figure 2C**). This tends to be a favorable preprocessing function for weight-based classifiers because no artificial ordering is imposed on the samples. This technique tends to be the default representation choice for categorical variables due to how it is interpreted. Because each nucleotide gets its own index in a 5 bit string, a 1 in any particular index means that nucleotide is present in that location.

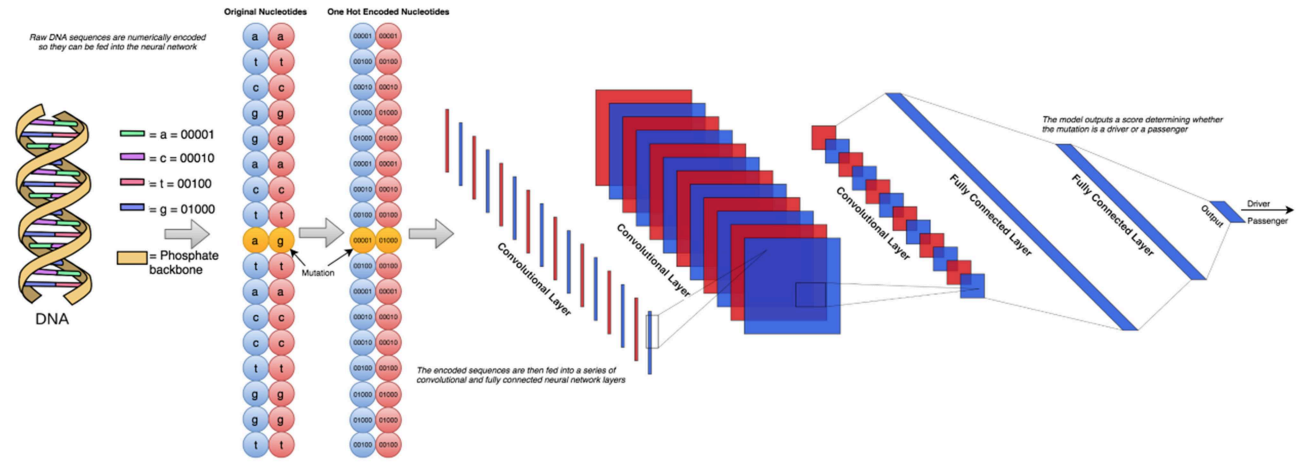


FIGURE 1 | The schematic workflow diagram of the CNN approach employed in this study. To determine the optimal architecture, we performed a grid search over a total of 72 different neural network architectures. These 72 architectures consisted of between 1 and 3 convolutional layers and 1–3 fully connected layers following. The number of nodes in each of these layers was also varied between 2 and 256 in powers of 2. The simplest architecture covered in this search contains 1 convolutional layer with 2 filters feeding into 1 fully connected layer with 2 nodes, and the most complex would have 3 convolutional layers feeding into 3 fully connected layers, all containing 256 nodes.

For example, since A->[0,0,0,0,1], this can essentially be read as "There are 0 'n,' 0 'g,' 0 't,' 0 'c,' and 1 'a' nucleotides present at this location." Since the one-hot encoding preprocessing technique lengthens the string, the resulting dimensionalities were (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005), respectively. The final preprocessing technique employed for the DNA sequences involved learned embeddings created with the word2vec algorithm (Mikolov et al., 2013). This technique analyzes the sequential context of the nucleotides assigning them a numeric representation in vector space. Using this representation, the nucleotide segments with similar meaning in the word2vec model would yield similar vectors in an N-dimensional representation. This technique was implemented using the Word2Vec model from the genism library for the Python programming language. Since the vocabulary in this application is fairly small, consisting of only 5 bit components, we chose to convert the nucleotide to 2 dimensional vectors which is sufficient to effectively encode this set. This resulted in the input sizes (2, 42), (2, 202), (2, 402), (2, 2002), and (2, 20002), respectively (**Figures 1**, **2**). The implementation and execution of these three preprocessing techniques provides adequate and efficient nucleotide representations for the CNN classifier.

Machine Learning Models

We used and compared performance of tree based classifiers and DL/CNN machine learning models. For the tree based methods, we used previously established protocol for obtaining hyper-parameters (Agajanian et al., 2018). The model training and tuning was done using Scikit-learn free software machine learning library for the Python programming language (Pedregosa et al., 2011; Biau, 2012). The Keras framework was used for training, validation and testing of CNN models (Erickson et al., 2017). We initially held out 20% of the data in a stratified manner as a testing set so that it had the same distribution of passengers/drivers as the total dataset. We then used the remaining 80% of the dataset as the training set to learn and tune its hyper-parameters. To choose between the hyper-parameters attempted, we test our model out on unseen data so that we have an unbiased estimate of its performance. To do this, we performed 3-fold cross validation, splitting the training set up into three equal sized portions. The model trains on two of them, and makes predictions on the third. This is repeated three times so that each of the three portions has been predicted on. A workflow diagram of the CNN approach (**Figure 1**) was carefully engineered to determine the optimal architecture. For this, we performed a grid search over a total of 72 different neural network architectures. These 72 architectures consisted of between 1 and 3 convolutional layers and 1–3 fully connected layers following. The number of nodes in each of these layers was also varied between 2 and 256 in powers of 2. The simplest architecture covered in this search contains 1 convolutional layer with 2 filters feeding into 1 fully connected layer with 2 nodes, and the most complex would have 3 convolutional layers feeding into 3 fully connected layers, all containing 256 nodes. The ReLU activation function was used, which returns max (0, X). All 72 different architectures (**Table 1**) were tested using this cross-validation algorithm and the architecture that had the highest F1 score across all 3-folds was chosen. Our neural networks were trained for 100 epochs, which means that they will pass through the entire dataset 100 times to complete their training. In between each epoch, the model recorded its predictions on the validation fold, and the epoch with the best performance on the validation set was recorded. Dropout was applied in between layers, so that inputs into a layer are randomly set to 0 with a certain probability. This prevents the neural network from overfitting, forcing it to learn without random features present. The best architecture was used for predictions on the test set.

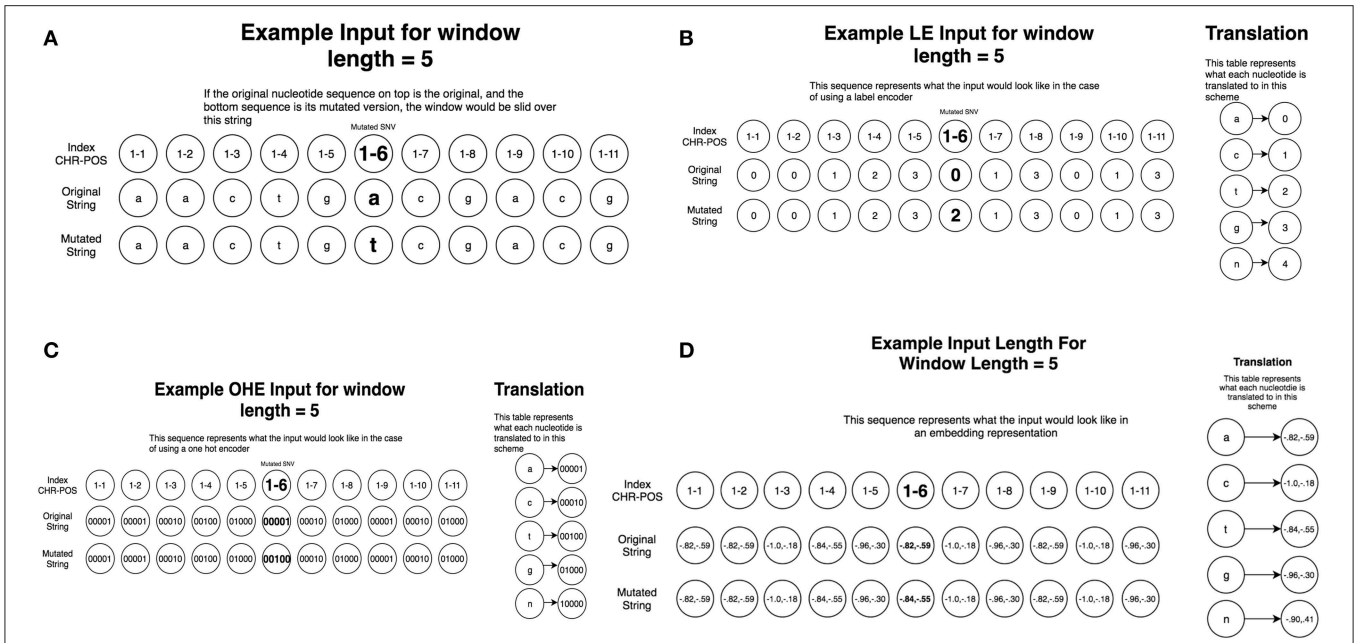


FIGURE 2 | Preprocessing of the nucleotide information for CNN machine learning of cancer driver mutations. Two different preprocessing techniques were then applied to the dataset to allow it to be passed into the CNN model in the numerical form: label encoding and one-hot encoding. **(A)** A schematic diagram of window sliding protocol. To represent the original nucleotide and its mutated version, two nucleotide sequences are placed on top of each other, one containing the original string, and the other contained the mutated version. This representation allows to utilize the sliding window format of the CNN models. **(B)** A schematic diagram of label encoding preprocessing protocol. Label encoding assigns each nucleotide its own unique ID (A->0, C->1 etc.) This imposes an ordering on the nucleotide sequences. **(C)** A schematic diagram of one hot encoding preprocessing protocol. One-hot encoding assigns each nucleotide its own bit encoded string (A -> [0,0,0,0,1], C-> [0,0,0,1,0]). This tends to be a favorable preprocessing function for weight-based classifiers because no artificial ordering is imposed on the samples. **(D)** A schematic diagram of embedding preprocessing scheme created with the word2vec algorithm.

TABLE 1 | The parameters of displayed CNN architectures in classification of cancer driver mutations.

Architecture	# Layers	# Nodes per layer
0	2	32,2
1	3	16,8,2
2	3	16,16,2
3	3	32,16,2
4	3	32,8,2
5	3	64,32,2
6	3	64,16,2
7	4	64,64,16,2
8	4	128,64,16,2
9	4	128,64,32,2
10	5	128,64,32,16,2

To assess the performance of each model, Accuracy, Recall, Precision, and F1 score were calculated to measure the performance of classification models. These parameters are defined as follows:

$$Accuracy = \frac{TP + TN}{all};\ Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN};\ F_1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{2}$$

True Positive (TP) and True Negative (TN) are defined as the number of mutations that are classified correctly as driver and passenger mutations, respectively. False Positive (FP) and False Negative (FN) are defined as the number of mutations that are misclassified into the other mutational classes. Precision is defined as the amount of positive samples the model predicts correctly (true positives) divided by the true positives plus the false positives. Recall is defined as true positives divided by true positives plus false negatives. The model performance was evaluated using receiver operating characteristic area under the curve. The receiver operating curve (ROC) is a graph where sensitivity is plotted as a function of 1-specificity. The area under the ROC is denoted AUC. The sensitivity or true positive rate (TPR) is defined as the percentage of non-neutral mutations that are correctly identified as driver mutations:

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{3}$$

The specificity or true negative rate (TNR) is defined as the percentage of mutations that are correctly identified as passengers:

$$Specificity = TNR = \frac{TN}{TN + FP} \tag{4}$$

In combination, these scores allow us to differentiate models by providing evaluation options to properly asses a model's performance. We relied on the F1 score, precision and recall as the primary discriminatory measures that can assess the quality of classification more reliably than accuracy. Under this data distribution, a model that only predicted passenger would yield an accuracy of 66.95%, but an F1 score of 0. In the case that two model's exhibited the same F1 score, we used the AUC measure to break the tie. The AUC measure is derived from the fact that the output of these classification models is a likelihood value between 0 and 1. A powerful classifier learns a likelihood function that consistently maps instances of the negative class to likelihoods lower than the positive class. A model that is reliable able to do this would receive an AUC of 1, whereas a model that only predicted the negative class would also receive an AUC of 0.

Bimolecular Simulations of Cancer Mutation Effects: Rigidity Decomposition and Protein Stability Analysis

We used FIRST (Floppy Inclusion and Rigid Substructure Topography) approach (Jacobs et al., 2001; Rader et al., 2002; Chubynsky and Thorpe, 2007) and the Python-based Constraint Network Analysis (CNA) interface (Hespenheide et al., 2002; Kruger et al., 2013; Pfleger et al., 2013a,b) to analyze partition of rigid and flexible regions in a set of protein kinases with the predicted cancer driver mutations. The employed parameters are consistent with our previous studies of protein kinases (Stetz et al., 2017). Protein stability computations that evaluated the effect of cancer driver mutations on the functional forms of the ErbB kinases were performed using CUPSAT (Cologne University Protein Stability Analysis Tool) (Parthiban et al., 2006, 2007). This approach was successfully adopted for the energetic analysis of cancer mutation hotspots (Dixit et al., 2009; Dixit and Verkhivker, 2011). We also employed the Foldx method (Guerois et al., 2002; Schymkowitz et al., 2005; Tokuriki et al., 2007; Van Durme et al., 2011) that allows for robust assessment of mutational effects on protein stability. These calculations were done with the user interface for the FoldX force field calculations (Schymkowitz et al., 2005) implemented as a plugin for the YASARA molecular graphics suite (Van Durme et al., 2011).

Protein Structure Network Analysis

For network-based analysis, a graph-based representation of protein structures is employed in which residues are treated as network nodes and inter-residue edges represent residue interactions (Sethi et al., 2009; Vijayabaskar and Vishveshwara, 2010; Stetz and Verkhivker, 2017). NAPS approach (Chakrabarty and Parekh, 2016) was used for construction of the residue interaction networks and subsequent residue-based network centrality analysis. For our analysis, an interaction strength-based graph representation of protein structures was used in which a residue is considered as node in the network and an edge is constructed if the interaction strength between two residues is more than the threshold of 4%. The pair of residues with the interaction I_{ij} greater than a user-defined cut-off ($I_{\min}$) are connected by edges and produce a protein structure network graph for a given interaction cutoff $I_{\min}$. The interaction strength I_{ij} is considered as edge weight. The edges in the residue interaction networks were weighted based on the defined interaction strength and dynamic residue correlations couplings (Sethi et al., 2009; Stetz and Verkhivker, 2017). Using the constructed protein structure networks, the residue-based betweenness parameters were also computed with the NAPS server (Chakrabarty and Parekh, 2016). The betweenness of residue i is defined to be the sum of the fraction of shortest paths between all pairs of residues that pass through residue i:

$$C_b(n_i) = \sum_{j<k}^{N} \frac{g_{jk}(i)}{g_{jk}} \tag{5}$$

g_{jk} denotes the number of shortest geodesics paths connecting j and k, and $g_{jk}(i)$ is the number of shortest paths between residues j and k passing through the node n_i. Residues with high occurrence in the shortest paths connecting all residue pairs have a higher betweenness values. For each node n, the betweenness value is normalized by the number of node pairs excluding n given as $(N-1)(N-2)/2$, where N is the total number of nodes in the connected component that node n belongs to.

RESULTS

Deep Learning Classification of Cancer Driver Mutations From Nucleotide Information

We began with an attempt to recapitulate our predictions by using various DL/CNN architectures informed by raw nucleotide sequence data evaluated the ability to make predictions based solely on raw genomic information. The inclusion of the three different preprocessing techniques allowed us to select the most informative representation of the nucleotides. The one hot encoded sequences yielded the model with the best performance, and for clarity of presentation we report only the dimensions and performance of the one hot encoded model. This preprocessing model resulted in input matrices of size (2, 105), (2, 505), (2, 1005), (2, 5005), and (2, 50005) corresponding to the different window sizes (10, 50, 100, 500, 1,000) surrounding the original nucleotide. It is worth noting that the embedding algorithm also learned meaningful representations of the nucleotides. The missing place indicator, "n," was predictably separated from the original nucleotides, which were arranged in 2 neat clusters (**Figure 2D**). Cluster 1 consisted of the adenine and tyrosine nucleotides, and cluster 2 consisted of the guanine and cytosine nucleotides. These two clusters are easily identified due to the fact that their constituent components are very close to each other while simultaneously being far away from the other cluster.

We employed 72 different DL architectures (**Table 1**) and the results for the window size of 10 are presented since they revealed more variance (**Figure 3**). The figures below display the 10 best performing models out of the 72 attempted. The training accuracy continued to increase for the duration of training (**Figure 3A**), while on the validation testing set of

cancer mutations, the best DL/CNN architecture achieved an average validation accuracy of 86.68% with an F1 score of 0.61 (**Figure 3B**). Interestingly, we found that the DL model seemed to learn early on, overfitting with each successive epoch (**Figure 3B**). In fact, the model achieved its highest validation accuracy on the first epoch, and proceeds to decline as learning proceeds in subsequent epochs. Furthermore, the AUC score of the model as well as the F1 score consistently stayed the same throughout all of the process. This is further contextualized by the tree based method's performance on the same dataset. The GBT classifier exhibited an F1 score of 0.57 with an average validation accuracy of 66.59%, and the RF classifier exhibited an F1 score of 0.58 and an average validation accuracy of 69.86%. We analyzed predictions by the DL/CNN model by assigning the predicted values for the entire dataset as a separate new feature termed DL score. Although we probed a variety of different architectures and several nucleotide-encoding protocols, a direct brute-force application of DL/CNN models to predict driver mutations only as a function of surrounding nucleotides appeared to be challenging. As a result, we suggested that a diverse set of more informative features may be required to recapitulate the level of robust performance achieved in our earlier work with sequence-based conservation and functional features (Agajanian et al., 2018).

We first used the RF classifier on the cancer mutation dataset with functional and conservation features obtained from dbWGFP server and adopted in our previous study (Agajanian et al., 2018). A database of human non-synonymous SNVs (dbNSFP) was developed as a one-stop resource for analysis of disease-causing mutations (Liu et al., 2011, 2013, 2016; Dong et al., 2015; Wu et al., 2016) storing 8.58 billion possible human whole-genome SNVs, with capabilities to compute a total of 48 functional prediction scores for each SNV, including 32 functional prediction scores by 13 approaches, 15 conservation features from 4 different tools including ensemble-based predictors RadialSVM, LR, and MSRV scores. The initially selected features were obtained from dbWGFP web server of functional predictions for human whole-genome single nucleotide variants that provided 32 functional prediction scores and 15 evolutionary features (Agajanian et al., 2018). Functional prediction scores refer to scores that predict the likelihood of a given SNV to cause a deleterious functional change in the protein, and evolutionary scores refer to scores providing different conservation measures of a given nucleotide site across multiple species (**Supplementary Table S1**). Some of the score features (SIFT, PolyPhen, LRT, Mutation Assessor, MutationTaster, FATHMM, RadialSVM, LR, MSRV, and SinBaD) can be applied only to SNVs in the protein coding regions, while other scores (Gerp++, SiPhy, PhyloP, Grantham, CADD, and GWAVA) can evaluate SNVs spreading over the whole genome (**Supplementary Table S1**). The ensemble-based scores RadialSVM and LR are integrated features that used machine learning approaches to combine information from 10 individual component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, Gerp++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) (Agajanian et al., 2018).

In this baseline experiment we evaluated feature performance of 32 input features on the expanded dataset (**Figure 4A**). Similar to our previous investigation (Agajanian et al., 2018), we found that the ensemble-based scores LR and RadialSVM considerably overshadowed the contributions of other features (**Figure 4**). By adding DL score to the original 32 features, we applied the RF model for predicting cancer driver mutations with this expanded set of features. The first question was to analyze feature importance of the RF model with the DL score included and determine whether the nucleotide-based scoring feature can contribute to the prediction performance in a meaningful and appreciable way (**Figure 4**). In the second round of RF classification experiments, we added DL score to the original list of 32 features (**Figure 4B**). Strikingly, the DL score ranked third following the ensemble-based LR and RadialSVM scores (**Figure 4B**). Moreover, it was evident that these three feature scores completely dominated feature importance distribution, with the DL score contributing almost as much

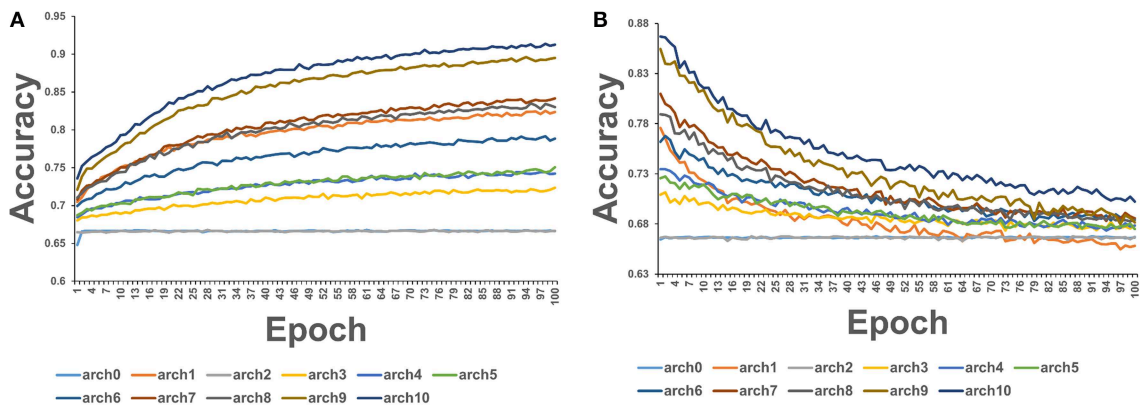


FIGURE 3 | The average accuracy of CNN model using exclusively nucleotide information. **(A)** Average accuracy across all 3-folds on an epoch by epoch basis on the training set with the sliding window size = 10. **(B)** Average accuracy across all 3-folds on an epoch by epoch basis on the validation set with the sliding window size = 10.

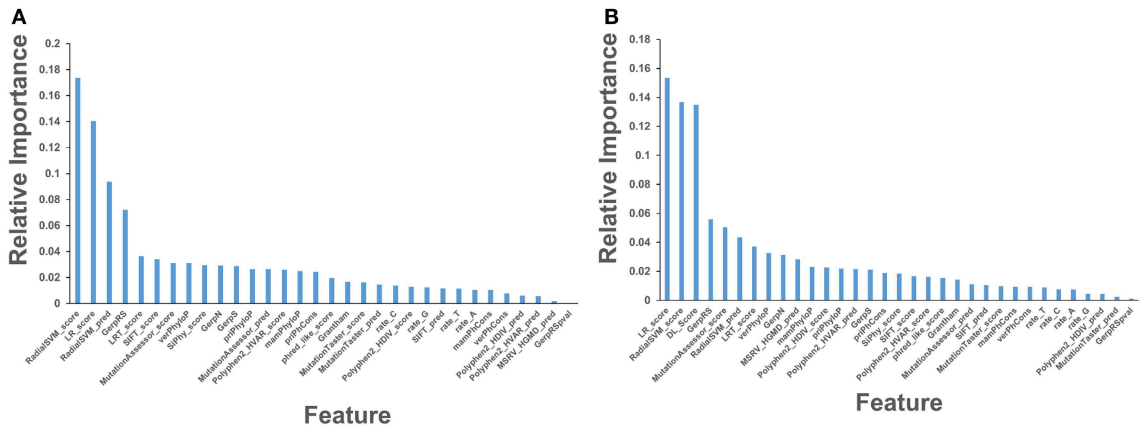


FIGURE 4 | Feature importance of the RF machine learning model on the cancer mutation dataset. The dataset consists of functionally validated 6,389 cancer driver mutations and 12,941 passenger mutations. The initially selected features for RF predictions were obtained from dbWGFP web server (Wu et al., 2016) of functional predictions for human whole-genome single nucleotide variants (**Supplementary Table S1**). The test set contained 20% of the samples from the original dataset, ensuring that the distribution of drivers and passengers was equivalent to that of the original dataset. The training set was subjected to recursive feature elimination process, resulting in a final dataset of 32 features. **(A)** Feature importance of 32 functional and sequence conservation features with DL score feature produced by CNN model excluded. **(B)** Feature importance of 33 features with the DL score included in the RF classification. The feature importance values are shown in blue filled bars and annotated. Feature importance is measured using the information value and weight of evidence criteria.

as the ensemble-based RadialSVM feature (**Figure 4B**). Quite remarkably, the DL-based score derived by CNN exclusively from primary nucleotide information can deliver significant information content and enrich predictions.

Using Spearman's rank correlation coefficient, we computed the pairwise correlations between different prediction scores (**Figure 5**). In this analysis, we found that the two dominant feature scores RadialSVM and LR are only moderately correlated with DL score, with the correlation coefficient of 0.486 and 0.423, respectively. Interestingly, RadialSVM and LR scores are more significantly correlated, suggesting that these ensemble-based features could be complementary with the nucleotide-based DL score. Accordingly, we argued that a combination of these dominant and yet complementary scores may allow for feature reduction and more robust performance of the RF classification models.

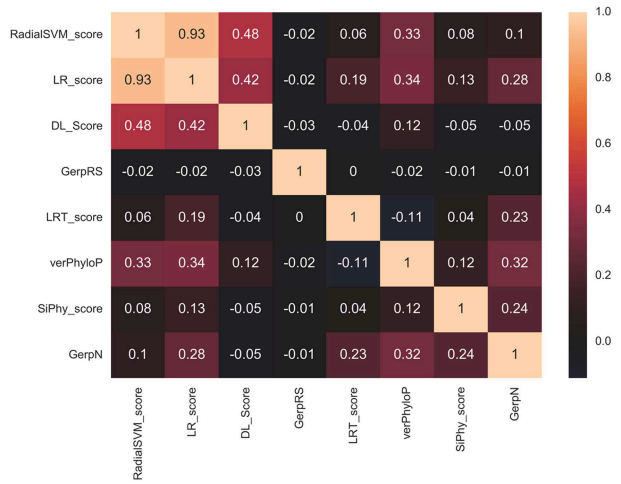


FIGURE 5 | The pairwise Spearman's rank correlation heat map between different prediction scores. The heat map of pairwise Spearman's rank correlation coefficients is shown for top 8 ranking features in the RF classification of cancer mutations with a total of 33 features with DL score included. The high ranking features include ensemble-based RadialSVM, LR scores along with DL score produced by CNN model solely from the raw nucleotide information.

Integration of CNN Predictions With Ensemble-Based Features in Classification Models of Cancer Driver Mutations

Based on these findings, we evaluated feature selection again aiming to recreate the same accuracy with only 8 features: RadialSVM score, LR score, DL score, GerpRS, LRT score, verPhyloP, SiPhy score, GerpN (**Figure 6A**). The RF model with only 8 features produced a similar ranking in which the ensemble-based scores and DL score contributed the most (**Figure 6A**). Other contributing features included evolutionary conservation scores derived from multiple sequence alignments and reflecting functional specificity, such as GerpRS (Davydov et al., 2010), SiPhy (Garber et al., 2009), and PhyloP (Garber et al., 2009) also showed appreciable information score values (**Figure 6A**). We then tested the performance of the RF model and feature importance by performing machine learning of cancer driver mutations using only 3 top features (**Figure 6B**).

The predictive performance of the RF models with different set of features was examined using area under the curve (AUC) plots (**Figure 7**). First, we examined difference in the AUC curves for RF-based classification with 32 functional features and with additional DL score (**Figure 7A**). The results showed a very similar high-level prediction performance with AUC =

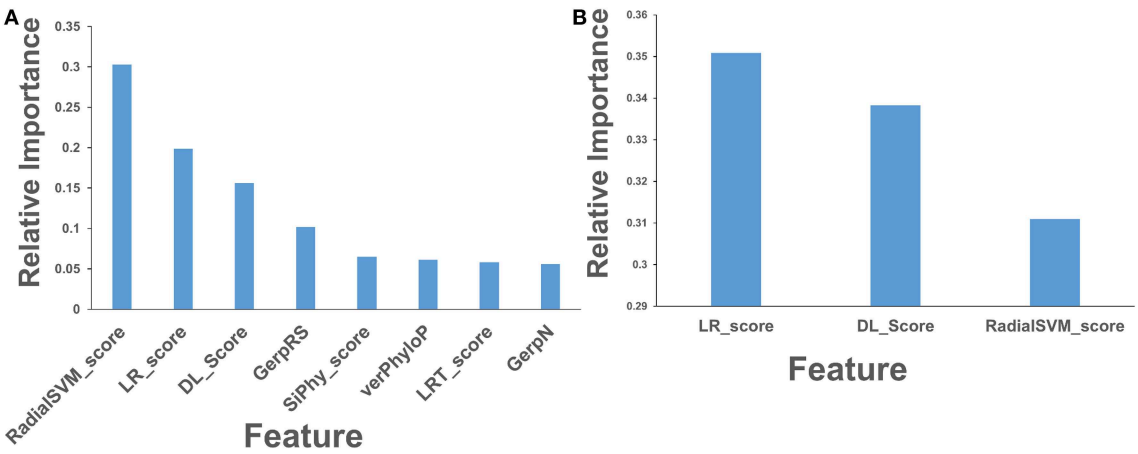


FIGURE 6 | Feature importance of the RF model on the cancer mutation dataset with the reduced number of features. **(A)** Feature importance ranking based on RF classification with only 8 most informative features. **(B)** Feature importance ranking based on RF classification with only 3 top features that included ensemble-based RadialSVM, LR scores, and DL score produced by CNN model. The feature importance values are shown in blue filled bars and annotated. Feature importance is measured using the information value and weight of evidence criteria.

0.95–0.96. It is worth noting that due to high AUC value for RF classification with 32 informative functional features, the addition of DL could not significantly enhance it. However, we showed that this nucleotide-derived predictor score provides an additional information content and is complementary to the ensemble-based RadialSVM score and LR score. In this context, it was instructive to observe that addition of DL score may marginally improve separation between TPR and FPR at higher values of these parameters (**Figure 7A**).

Strikingly, RF learning model that relied on only 3 top features (RadialSVM score, LR score, and DL score) yielded AUC = 0.94, thereby showing that these features may be sufficient to achieve robust classification of cancer driver mutations on a fairly large dataset of somatic mutations employed in this study. Combined with the findings that DL score only weakly correlated with the ensemble-based scores, we concluded that unexpectedly few highly informative parameters can achieve high level of performance (**Figure 7**). We then tested several machine learning models including RF, GBTs and support vector machine (SVM) on the dataset with the top 8 features to benchmark performance against the original RF model with 32 features (Agajanian et al., 2018). The performance of classification models was carefully assessed (**Table 2**). All methods achieved a high classification accuracy of ∼90%. The sensitivity values were higher for the SVM and RF models, but all methods yielded similar high performance classification on the dataset with only limited number of major features that included DL score (**Table 2**).

To summarize, our results supported the notion that machine learning-derived ensemble functional predictors may play a central role in classification of cancer driver mutations. The central finding of these machine learning experiments was that combination of ensemble-based features and DL score derived by CNN model from nucleotide information are complementary and when combined can yield classification accuracy comparable and often exceeding the one obtained with a full set of features. The important lesson from this analysis is that integrated high-level features derived by machine learning approaches from primary nucleotide and protein sequence information may be sufficient to predict an important functional phenotype. Although structure-derived features and other functional scores contribute to feature importance ranking and tightly linked with the mutational phenotype, the success of machine learning tools in deciphering predictive features from primary sequence information is encouraging and should be further explored in other applications.

Leveraging Machine Learning Predictions in Structure-Functional Analysis of Molecular Signatures of Driver Mutations in Oncogenic Protein Kinases

Machine learning driver/passenger classifications typically consider activating, inactivating and inhibitory (or resistant) mutations as drivers, often leaving aside a more detailed characterization and assignment of driver positions. Direct predictions of these specific classes may not be adequately suited for machine learning tools due to smaller datasets. To expand our predictions and aim at extracting a more granular functional information about driver mutations, we conducted rigidity decomposition simulations and analyzed conformational flexibility of the predicted driver positions in protein kinase genes. The objective of this analysis was to facilitate functional validation and interpretation of machine learning results through coarse-grained biophysical simulations as an effective post-processing tool of machine learning classification. In fact, the proposed simulation analysis of mobility at the driver positions allows to expand classification of driver mutations further and characterize activating drivers. Previous studies have suggested that conformational mobility of many oncogenic kinases may be linked with preferential localization of activating

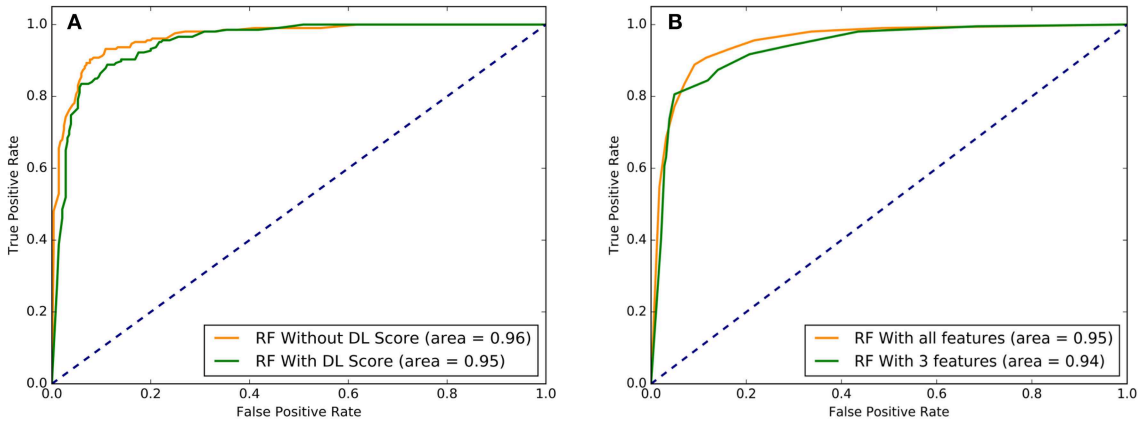


FIGURE 7 | The ROC plots of sensitivity (TPR) as a function of 1-specificity, where specificity (TNR). **(A)** The ROC curves for overall performance of the RF model with 32 functional features excluding DL score (in green) and 33 features that included DL score (in red). **(B)** The ROC curves for the RF model with all 33 features (in green) and with the top 3 performing features that included LR score, Radial_SVM score, and DL score (in red). Higher AUC score indicates better performance. These plots illustrated a comparative performance of machine learning models for top prediction scores.

TABLE 2 | The relative performance metrics and statistics of various machine learning models in classification of cancer driver mutations with the top 8 features.

	Boosted trees	SVM	Random forest
Accuracy	0.896	0.890	0.896
F1 score	0.900	0.890	0.900
Precision	0.900	0.890	0.900
Recall	0.900	0.890	0.900
True positive rate	0.850	0.949	0.857
False positive rate	0.112	0.797	0.123
True negative rate	0.115	0.016	0.107
False negative rate	0.913	0.748	0.907

cancer mutations in flexible functional regions (Paladino et al., 2015; Kiel et al., 2016; Stetz et al., 2017).

We examined flexibility of specific functional regions targeted by driver mutations in oncogenic protein kinases and probed functional propensity of these drivers to promote transitions to constitutively active states. The primary focus of this analysis is on the family of the ErbB protein tyrosine kinases (Lemmon and Schlessinger, 2010; Roskoski, 2014). A number of human cancers are associated with mutations causing the increased expression of the ErbB kinases. A large number of activating and drug resistance EGFR mutations have been extensively studied at the molecular and functional levels (Paez et al., 2004; Kobayashi et al., 2005; Zhou et al., 2009; Eck and Yun, 2010). Oncogenic kinase mutants are known to act by destabilizing the inactive dormant kinase form while promoting conformational transitions and stabilization of a constitutively active kinase state—a salient functional characteristic linked with the initiation or progression of cancer (Carey et al., 2006; Wang et al., 2011). We used the crystal structures of the EGFR, ErbB2, ErbB3, and ErbB4 kinases that constitute this family to perform rigidity decomposition and then align the positions of the predicted cancer driver mutations with the structural mobility maps (**Figure 8**). We examined how the predicted driver mutations for ErbB protein kinases are distributed on the rigidity/flexibility map of the catalytic core and whether the dynamic preferences of mutational sites can be linked with their primary function as activating drivers. To explore these questions, we examined the predicted cancer driver mutations for the ErbB kinase family. Structural mapping of these cancer mutations onto the crystallographic ErbB conformations showed that activating driver mutations are preferentially localized in the flexible regions and target positions where they can readily promote conformational changes to the active form without severely compromising thermodynamic stability (**Figure 8**).

To quantify these arguments further, we also characterized the free energy differences between wild-type and cancer-driver mutations for the ErbB proteins in both inactive and active kinase forms (**Figure 9**). Since both CUPSAT and FoldX approaches yielded similar results, we illustrated our findings by presenting FoldX-derived protein stability changes (**Figure 9**). The results of this simulation-driven functional classification of predicted driver mutations were compared with the biochemical and mutagenesis data. The analysis of driver mutations in EGFR confirmed that L858 and L861 positions target flexible regions as can be manifested by classical activating driver mutations L858R and L861Q (Littlefield and Jura, 2013; Red Brewer et al., 2013). The energetics of these activating drivers is consistent with a common mechanism of the constitutive activation of kinases by driver mutations (**Figure 9A**). This mechanism reflects a combined effect of activating mutations producing a more significant destabilization of the inactive state as compared to the active state, triggering shift of the thermodynamic equilibrium toward the active conformation. We found that some EGFR mutations such as T854A are mapped onto more stable regions of the kinase (**Figure 8A**) and showed similar destabilization in the inactive and active forms. Accordingly, this predicted cancer driver mutation is likely not

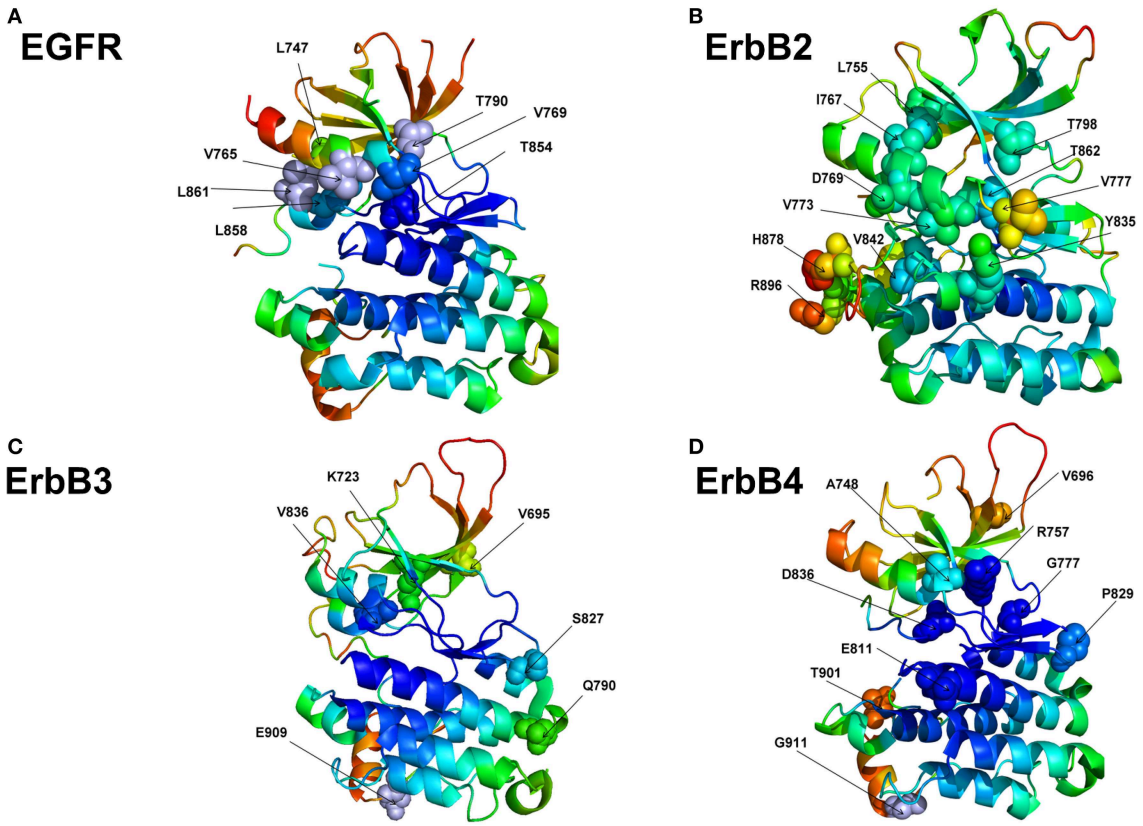


FIGURE 8 | Structural maps of rigidity decomposition and mobility signatures of cancer mutation drivers in the ErbB protein kinases. Structural mapping of rigidity and flexibility regions in the crystal structure of EGFR (pdb id 1XKK) **(A)**, crystal structure of ErbB2 kinase (pdb id 3PP0) **(B)**, crystal structure of ErbB3 kinase (pdb id 3KEX) **(C)**, and crystal structure of ErbB4 kinase (pdb id 3BBT) **(D)**. Crystallographic conformations are colored using a color range from red (highest flexibility) to blue (highest rigidity). The positions of predicted in machine learning cancer driver mutations are shown in spheres (colored according to their mobility level) and annotated.

activating but rather may be attributed to inhibitory or resistant mutations. Indeed, the recent experimental studies showed that T854A mutation is the acquired mutation causing resistance to known drugs (Bean et al., 2008). Another EGFR mutation V769M/L showed an intermediate level of mobility (**Figure 8A**) and greater stabilization of the active state. These results are in line with recent functional experiments showing that EGFR-V769M mutation is indeed activating that may explain the role of this driver mutation in the development of multiple lung cancers in a pool of lung cancer patients (Deng et al., 2018).

The positions of almost all predicted driver mutations in ErbB2 kinase target highly flexible regions and can be assigned in our model to activating driver mutations (**Figures 8B**, **9B**). Our previous biophysical simulations and network analysis of activation mechanisms in the ErbB proteins similarly indicated that almost all oncogenic ErbB2 variants are localized in the mobile αC-β4 loop and highly dynamic in their inactive states promoting transition to the active form and causing an uncontrollable activity (James and Verkhivker, 2014). These findings are consistent with the experimental studies (Fan et al., 2008; Aertgeerts et al., 2011). While the majority of somatic mutations in the EGFR and ErbB2 kinases increase the kinase activity, a number of the classified ErbB4 cancer mutants have been shown to inhibit or reduce the kinase activity (Tvorogov et al., 2009). In particular, some cancer-associated mutations of ErbB4 can promote loss of ErbB4 kinase activity as these alterations weaken the important functional interactions in the catalytic core and may interfere with the protein stability. According to experimental data, some cancer mutations have only minor or no effect on kinase activity (V696I, E785K, A748S, P757Q, P829Q, and T901M), while K726R abolishes kinase activity and D818N and D836Q are known as kinase-dead mutations (Tvorogov et al., 2009). We found that predicted cancer driver mutations are mapped onto more stable regions in ErbB4, owing to the greater rigidity of this catalytic domain (**Figures 8D**, **9D**). Accordingly, the respective driver mutations cannot function as activating but rather may cause significant distortions of the kinase structure, causing abolishment of kinase activity which is the functional signature of most cancer drivers in ErbB4 kinase. The performed simulation-driven post-processing of machine learning predictions facilitated *in silico* functional characterization of cancer mutations and allowed to properly assign activating or inhibiting phenotypic effects to a pool of pathogenic kinase variants.

To provide more quantitative insights, we used the predicted cancer mutations in the ErbB kinases and conducted protein

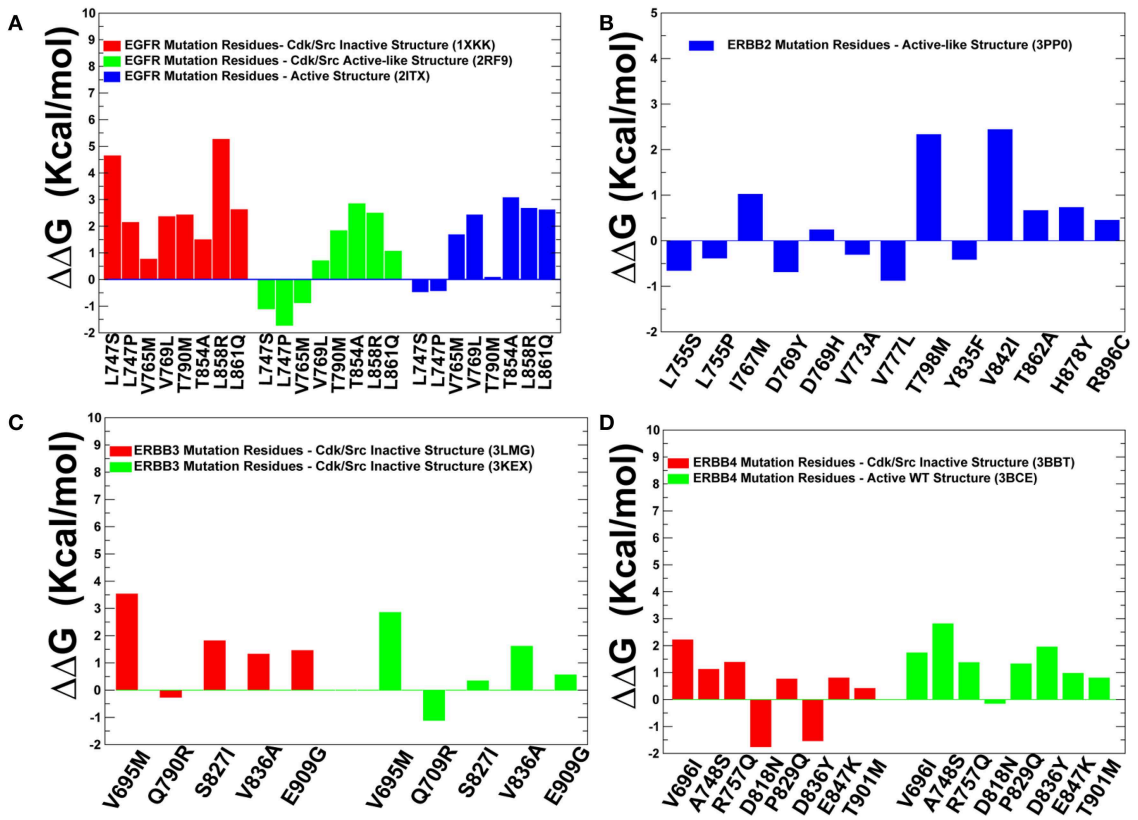


FIGURE 9 | Protein stability analysis of the predicted cancer driver mutations. Protein stability differences calculated between the wild-type and mutants for predicted cancer driver mutations in the ErbB kinases using FOLDx approach. Protein stability changes induced by cancer driver mutations in the inactive and active states of EGFR kinase **(A)**, ErbB2 kinase **(B)**, ErbB3 kinase **(C)**, and ErbB4 kinase **(D)**. Positive values of protein stability changes correspond to destabilizing mutations.

structure network analysis to identify whether positions of deleterious mutations would overlap with the global mediating nodes in the interaction networks. The betweenness of a residue node is defined as the number of shortest paths that can go through that node, thus estimating the contribution of the node to the global communication flow in the system. High betweenness nodes can influence the spread of information through the network by facilitating, hindering, or altering the communication between others. According to our hypothesis, cancer mutations may preferentially target the essential mediating residues with a high centrality that play an important role in activity and signaling of protein kinase genes.

The centrality analysis revealed important differences in the distribution of mediating centers in the ErbB kinase structures (**Figure 10**). We particularly observed that the betweenness of the active form of EGFR (**Figure 10A**) and ErbB4 (**Figure 10D**) was on average higher than for the inactive states. Importantly, the location of the properly classified EGFR mutations with the highest oncogenic potential (L858R, T790M, L838V, V742A, V851A, I853T) corresponds to some of the high centrality peaks of the profile (**Figure 10A**). In addition, these residues showed appreciable differences in the betweenness values between the inactive to the active states, as the residue centrality in these positions typically increased in the functional active form (**Figures 10A,D**). These findings suggested that a number of key activating mutations in the ErbB kinases target mediating sites of global allosteric communication in the protein structures. We believe that by adding this significant additional component to our study, we have been able to further quantify and explain the protein rigidity/flexibility analysis of predicted cancer mutations in the kinase genes. In our view, by complementing machine learning predictions with the structural and network-based analyses we can obtain useful insights into mechanisms underlying effects of cancer mutations and also identify limitations of classification models and ways to improve interpretability and trustability of machine learning model approaches.

DISCUSSION

As large-scale biological data are available from high-throughput assays, and methods for learning the thousands of network parameters have matured, we can now assess feasibility and practicality of using specialized neural network architectures as classification tools for recognizing cancer-causing variants and associated cancer types. Given rapid proliferation and increasing popularity of deep learning tools to address various biological problems, there are several fundamental questions arising in the

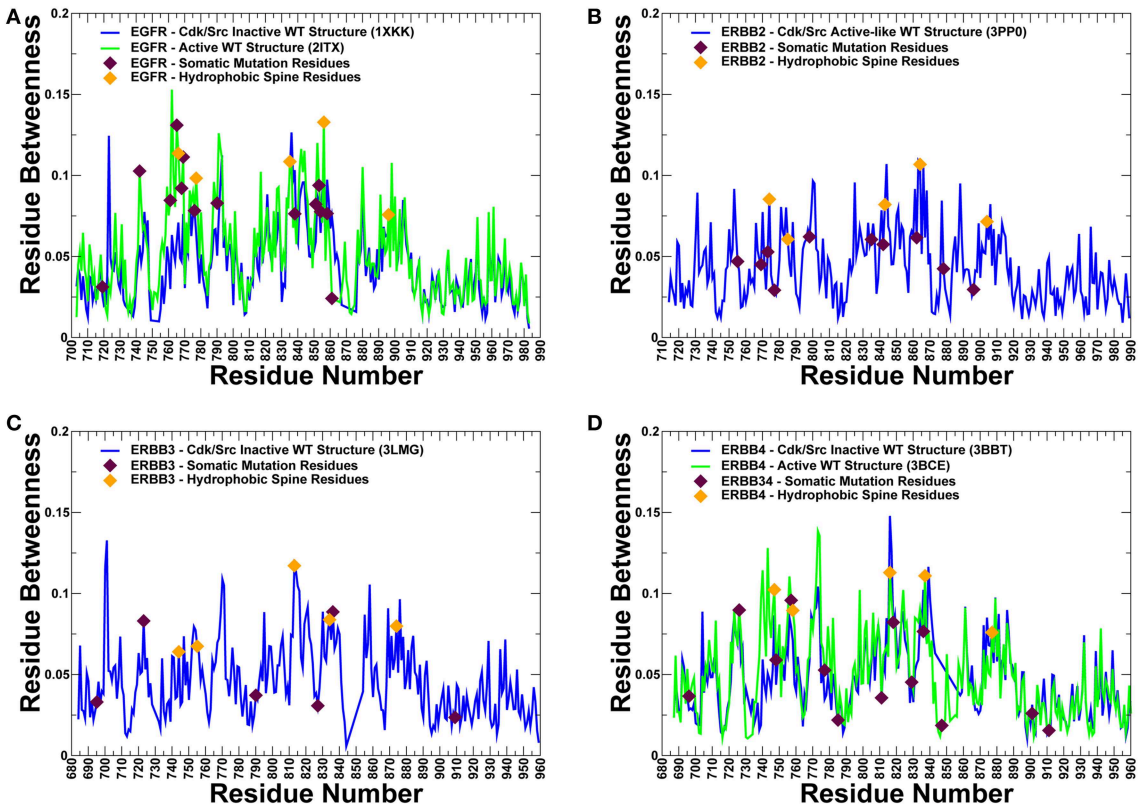


FIGURE 10 | The residue-based betweenness profiles of the ErbB kinase structures. The residue betweenness (residue centrality) profiles for the inactive and active crystal structure states of EGFR **(A)**, ErbB2 **(B)**, ErbB3 **(C)**, and ErbB4 kinases **(D)**. For ErbB2 and ErbB3 only crystal structures of the inactive-type states were available for the analysis. The positions of somatic mutations predicted by machine learning experiments are shown in maroon-colored filled triangles, and residue positions corresponding to the hydrophobic spine residues are shown in orange-colored filled triangles. Protein kinase activation is controlled by two networks of mostly hydrophobic residues that form a regulatory spine (R-spine) and a catalytic spine (C-spine). The EGFR R-spine residues include L777 from the β4-strand, M766 from the C-terminal end of the αC-helix, F856 of the DFG motif in the activation segment, H835 of the HRD motif of the catalytic loop, and D896 of the αF-helix.). The R-spine residues in ErbB2 are M774, L785, F864, H843, and D904. The R-spine residues in ErbB3 are I744, L755, F843, H813, and D874. The R-spine residues in ErbB4 are M747, L758, H816, F837, and D877.

context of classification of cancer driver mutations. Will deep learning make all other models obsolete? Can deep learning models achieve robust classification and recognition of cancer driver mutations based solely on nucleotide information? What is the role of many functional and structural predictors derived from biophysical perspective in this context? In this work, we have explored and integrated different machine learning approaches for prediction and classification of cancer driver mutations. We first explored the ability of CNN models to identify and classify cancer driver mutations directly from raw nucleotide sequence information without relying on specific functional scores.

The results of this study have demonstrated that while CNN models can learn high level features from genomic information that has sufficiently high importance, accurate classification of cancer mutation driver phenotype using exclusively nucleotide data continues to be challenging. This problem is admittedly more complex than the experimental design suggests, due to the complex nature of protein interactions in the human body. This experimental setup considered only the primary sequence form of the nucleotides, which could only ever partially explain the onset of cancer. The secondary, tertiary, and quaternary form of these same strings would certainly contain more information, due to the folding processes that occur in these steps. Additionally, this technique ignores all of the possible interactions that can be had with other structures in the body, which further dilutes the informational value present in the dataset. As such it's unreasonable to assume that our solely primary sequence based dataset would be able to explain all of the variance present in a complex problem like determining a single mutation's level of effect on the onset of cancer. The experimental inclusion of the different window sizes was also an attempt to allow increasing numbers of surrounding nucleotides to have an influence on our chosen mutation's effect. An obvious assumption here is that more nucleotides would in fact bring in more information. This, however, proved not to hold up as the only dataset that provided any significant variance in performance was the window size = 10 dataset. This suggests that more nucleotides only confuse the model and disallow it from learning informative patterns. This problem could possibly be combatted in future research by testing out larger architectures.

The benefits of integrating CNN-derived predictors obtained from nucleotide information with protein sequence features, evolutionary and functional scores were then carefully examined. By exploring various encoding techniques and an array of different CNN architectures, we have found that neural networks can quickly learn an important functional signal, but can rarely steadily improve the initial performance spike with the number of additional epochs. The juxtaposition of monotonically increasing training accuracy with monotonically decreasing validation accuracy is a telltale sign of overfitting. This suggests that there is only a small amount of useful information that can be learned very early on, and subsequent epochs only cause the model to learn noisy patterns that are only exhibited in the training set. It is difficult to determine exactly what was learned by the model due to the black box nature of neural networks, however due to the short path to optimality it is safe to say that any learned concepts cannot be overly complex. We have pursued a synergistic strategy in which the prediction score generated by CNN models was integrated with physics-based functional, structural and evolutionary conservation features. The important lesson of this analysis was the revelation that CNN-derived features may be complementary to the ensemble-based predictors often employed for classification of cancer mutations. These other scores are not calculated from raw sequence based techniques, which supports this DL score as a novel inclusion into a portfolio of scores due to its unique derivation.

By combining deep learning-generated score with only two main ensemble-based functional features, we were able to achieve a high performance level for cancer driver mutations. The robustness of this approach was verified by several traditional machine learning classifiers, including RF, SVM, and GBTs. We have found that integration of CNN-derived predictor score with only several ensemble-based features can recapitulate the results obtained with a large number of functional features and improve performance in capturing driver mutations across a spectrum of machine learning classifiers. Our findings have also demonstrated that synergy of nucleotide-based deep learning scores and integrated metrics derived from protein sequence conservation scores can allow for robust classification of cancer driver mutations with a reduced number of highly informative features. This is an interesting and highly informative result, as the law of parsimony holds for machine learning models so simpler models with comparable performance are typically preferred over their more complex counterparts. Part of this model complexity includes the number of features that a model relies on. As such a reduction in features is a universally positive outcome. In addition to the improved quality of the model, it also expands the universe of predictable nucleotides that are available to us since we depend only on the presence of two ensemble-based scores. The DL score can be derived for any mutation with known coordinates so this is not a limiting factor. In this respect our initial goal of expanding the nucleotides we can make predictions for was partially achieved. This increase in the generalization of these models facilitates the logical conclusion of driver classification efforts, accurately classifying all known nucleotides.

While machine learning approaches can often produce robust and accurate predictors, the ultimate goal of research is fundamental understanding of the underlying phenomena which requires a mechanistic model of the world. In this context, machine learning predictions are leveraged in biomolecular simulations to enable analysis of cancer mutation mechanisms and obtain a more specific information about an important subset of cancer mutations, activating drivers. The results of our investigation suggested that through integration of machine learning classification and biomolecular simulations of cancer mutations we can often validate the predictions and facilitate a more detailed functional analysis of activating driver mutations. These findings can provide insight and new angle to the problem of interpretability of "black box" machine learning results. By carefully inspecting predictions of machine learning models in the context of dynamic and energetic signatures of mutational sites for oncogenic protein kinases, this study offered instructive strategy for simulation-based post-processing of machine learning predictions and detailed functional specification of cancer driver mutations. The proposed synergistic integration of machine learning and biomolecular simulations into a single computational platform allows to rapidly process large datasets and make robust predictions on functionally significant cancer drivers. The results of this study may also inform and guide design of targeted and personalized therapeutic agents combating a spectrum of mutational changes occurring in cancer.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://www.cbioportal.org/.

AUTHOR CONTRIBUTIONS

GV and SA conceived and designed the research. SA and OO performed the research. SA, OO, and GV analyzed the results and wrote the manuscript. GV wrote the final version of the manuscript and supervised the project.

FUNDING

This work was partly supported by institutional funding from Chapman University.

ACKNOWLEDGMENTS

The authors acknowledge the technical assistance of Schmid College Grand Challenge Initiative Postdoctoral Fellow Dr. Anne Sonnenschein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00044/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Aertgeerts, K., Skene, R., Yano, J., Sang, B. C., Zou, H., Snell, G., et al. (2011). Structural analysis of the mechanism of inhibition and allosteric activation of the kinase domain of HER2 protein. *J. Biol. Chem.* 286, 18756–18765. doi: 10.1074/jbc.M110.206193

Agajanian, S., Odeyemi, O., Bischoff, N., Ratra, S., and Verkhivker, G. M. (2018). Machine learning classification and structure-functional analysis of cancer mutations reveal unique dynamic and network signatures of driver sites in oncogenes and tumor suppressor genes. *J. Chem. Inf. Model.* 58, 2131–2150. doi: 10.1021/acs.jcim.8b00414

Ainscough, B. J., Barnell, E. K., Ronning, P., Campbell, K. M., Wagner, A. H., Fehniger, T. A., et al. (2018). A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet.* 50, 1735–1743. doi: 10.1038/s41588-018-0257-y

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e318. doi: 10.1016/j.cell.2018.02.060

Bardelli, A., Parsons, D. W., Silliman, N., Ptak, J., Szabo, S., Saha, S., et al. (2003). Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* 300:949. doi: 10.1126/science.1082596

Bean, J., Riely, G. J., Balak, M., Marks, J. L., Ladanyi, M., Miller, V. A., et al. (2008). Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. *Clin. Cancer Res.* 14, 7519–7525. doi: 10.1158/1078-0432.CCR-08-0151

Bertrand, D., Drissler, S., Chia, B. K., Koh, J. Y., Li, C., Suphavilai, C., et al. (2018). Consensus driver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Res.* 78, 290–301. doi: 10.1158/0008-5472.CAN-17-1345

Biau, G. (2012). Analysis of a random forest model. *J. Mach. Learn. Res.* 13, 1063–1095. Available online at: http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Carey, K. D., Garton, A. J., Romero, M. S., Kahler, J., Thomson, S., Ross, S., et al. (2006). Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Res.* 66, 8163–8171. doi: 10.1158/0008-5472.CAN-06-0453

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., et al. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* 69, 6660–6667. doi: 10.1158/0008-5472.CAN-09-1133

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chakrabarty, B., and Parekh, N. (2016). NAPS: network analysis of protein structures. *Nucleic Acids Res.* 44, W375–W382. doi: 10.1093/nar/gkw383

Cheng, F., Zhao, J., and Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinformatics* 17, 642–656. doi: 10.1093/bib/bbv068

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7:e46688. doi: 10.1371/journal.pone.0046688

Chubynsky, M. V., and Thorpe, M. F. (2007). Algorithms for three-dimensional rigidity analysis and a first-order percolation transition. *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.* 76:041135. doi: 10.1103/PhysRevE.76.041135

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. doi: 10.1038/nbt.2514

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature* 417, 949–954. doi: 10.1038/nature00766

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6:e1001025. doi: 10.1371/journal.pcbi.1001025

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111

Deng, Q., Xie, B., Wu, L., Ji, X., Li, C., Feng, L., et al. (2018). Competitive evolution of NSCLC tumor clones and the drug resistance mechanism of first-generation EGFR-TKIs in Chinese NSCLC patients. *Heliyon* 4:e01031. doi: 10.1016/j.heliyon.2018.e01031

Ding, L., Wendl, M. C., McMichael, J. F., and Raphael, B. J. (2014). Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* 15, 556–570. doi: 10.1038/nrg3767

Dixit, A., and Verkhivker, G. M. (2011). The energy landscape analysis of cancer mutations in protein kinases. *PLoS ONE* 6:13. doi: 10.1371/journal.pone.0026071

Dixit, A., Yi, L., Gowthaman, R., Torkamani, A., Schork, N. J., and Verkhivker, G. M. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS ONE* 4:14. doi: 10.1371/journal.pone.0007485

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi: 10.1093/hmg/ddu733

Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P. D., et al. (2013). CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* 29, 647–648. doi: 10.1093/bioinformatics/btt017

Eck, M. J., and Yun, C. H. (2010). Structural and mechanistic underpinnings of the differential drug sensitivity of EGFR mutations in non-small cell lung cancer. *Biochim. Biophys. Acta* 1804, 559–566. doi: 10.1016/j.bbapap.2009.12.010

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 6, 271–281.e277. doi: 10.1016/j.cels.2018.03.002

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., and Philbrick, K. (2017). Toolkits and libraries for deep learning. *J. Digit Imag.* 30, 400–405. doi: 10.1007/s10278-017-9965-6

Fan, Y., Xi, L., Hughes, D. S., Zhang, J., Futreal, P. A., Wheeler, D. A., et al. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* 17:178. doi: 10.1186/s13059-016-1029-6

Fan, Y. X., Wong, L., Ding, J., Spiridonov, N. A., Johnson, R. C., and Johnson, G. R. (2008). Mutational activation of ErbB2 reveals a new protein kinase autoinhibition mechanism. *J. Biol. Chem.* 283, 1588–1596. doi: 10.1074/jbc.M708116200

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 43, D805–811. doi: 10.1093/nar/gku1075

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:pl1. doi: 10.1126/scisignal.2004088

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25, i54–62. doi: 10.1093/bioinformatics/btp190

Gauthier, N. P., Reznik, E., Gao, J., Sumer, S. O., Schultz, N., Sander, C., et al. (2016). MutationAligner: a resource of recurrent mutation hotspots in protein domains in cancer. *Nucleic Acids Res.*. 44, D986–991. doi: 10.1093/nar/gkv1132

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 14 (Suppl 3):S7. doi: 10.1186/1471-2164-14-S8-S7

Goh, G. B., Hodas, N. O., and Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.* 38, 1291–1307. doi: 10.1002/jcc.24764

Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* 88, 440–449. doi: 10.1016/j.ajhg.2011.03.004

Gonzalez-Perez, A., and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* 40:e169. doi: 10.1093/nar/gks743

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., et al. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* 10, 723–729. doi: 10.1038/nmeth.2562

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387. doi: 10.1016/S0022-2836(02)00442-4

Haber, D. A., and Settleman, J. (2007). Cancer: drivers and passengers *Nature* 446, 145–146. doi: 10.1038/446145a

Hespenheide, B. M., Rader, A. J., Thorpe, M. F., and Kuhn, L. A. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.* 21, 195–207. doi: 10.1016/S1093-3263(02)00146-8

Hinkson, I. V., Davidsen, T. M., Klemm, J. D., Kerlavage, A. R., and Kibbe, W. A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front. Cell. Dev. Biol.* 5:83. doi: 10.3389/fcell.2017.00083

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins* 44, 150–165. doi: 10.1002/prot.1081

James, K. A., and Verkhivker, G. M. (2014). Structure-based network analysis of activation mechanisms in the ErbB family of receptor tyrosine kinases: the regulatory spine residues are global mediators of structural stability and allosteric interactions. *PLoS ONE* 9:e113488. doi: 10.1371/journal.pone.0113488

Jensen, M. A., Ferretti, V., Grossman, R. L., and Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 130, 453–459. doi: 10.1182/blood-2017-03-735654

Kiel, C., Benisty, H., Llorens-Rico, V., and Serrano, L. (2016). The yin-yang of kinase activation and unfolding explains the peculiarity of Val600 in the activation segment of BRAF. *Elife* 5:e12814. doi: 10.7554/eLife.12814

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Kallberg, M., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* 15, 591–594. doi: 10.1038/s41592-018-0051-x

Klonowska, K., Czubak, K., Wojciechowska, M., Handschuh, L., Zmienko, A., Figlerowicz, M., et al. (2016). Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget* 7, 176–192. doi: 10.18632/oncotarget.6128

Kobayashi, S., Boggon, T. J., Dayaram, T., Janne, P. A., Kocher, O., Meyerson, M., et al. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* 352, 786–792. doi: 10.1056/NEJMoa044238

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kruger, D. M., Rathi, P. C., Pfleger, C., and Gohlke, H. (2013). CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. *Nucleic Acids Res.* 41, W340–W348. doi: 10.1093/nar/gkt292

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., et al. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44:e108. doi: 10.1093/nar/gkw227

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., et al. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28, 311–317. doi: 10.1093/bioinformatics/btr665

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213

Lemmon, M. A., and Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell* 141, 1117–1134. doi: 10.1016/j.cell.2010.06.011

Li, J., Drubay, D., Michiels, S., and Gautheret, D. (2015). Mining the coding and non-coding genome for cancer drivers. *Cancer Lett.* 369, 307–315. doi: 10.1016/j.canlet.2015.09.015

Littlefield, P., and Jura, N. (2013). EGFR lung cancer mutants get specialized. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15169–15170. doi: 10.1073/pnas.1314719110

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32, 894–899. doi: 10.1002/humu.21517

Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* 34, E2393–2402. doi: 10.1002/humu.22376

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10:13. doi: 10.3389/fgene.2019.00013

Mao, Y., Chen, H., Liang, H., Meric-Bernstam, F., Mills, G. B., and Chen, K. (2013). CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLoS ONE* 8:e77945. doi: 10.1371/journal.pone.0077945

Martelotto, L. G., Ng, C. K., De Filippo, M. R., Zhang, Y., Piscuoglio, S., Lim, R. S., et al. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol.* 15:484. doi: 10.1186/s13059-014-0484-1

Masica, D. L., Douville, C., Tokheim, C., Bhattacharya, R., Kim, R., Moad, K., et al. (2017). CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Res.* 77, e35–e38. doi: 10.1158/0008-5472.CAN-17-0338

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient Estimations of Word Representations in Vector Space*. arXiv:1301.3781 [cs.CL]. Available online at: https://arxiv.org/abs/1301.3781

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* 17:128. doi: 10.1186/s13059-016-0994-0

Ng, P. K., Li, J., Jeong, K. J., Shao, S., Chen, H., Tsang, Y. H., et al. (2018). Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* 33, 450–462.e410. doi: 10.1016/j.ccell.2018.01.021

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* 48, 827–837. doi: 10.1038/ng.3586

Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500. doi: 10.1126/science.1099314

Paladino, A., Morra, G., and Colombo, G. (2015). Structural stability and flexibility direct the selection of activating mutations in epidermal growth factor receptor kinase. *J. Chem. Inf. Model.* 55, 1377–1387. doi: 10.1021/acs.jcim.5b00270

Parthiban, V., Gromiha, M. M., Abhinandan, M., and Schomburg, D. (2007). Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. *BMC Struct. Biol.* 7:54. doi: 10.1186/1472-6807-7-54

Parthiban, V., Gromiha, M. M., and Schomburg, D. (2006). CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 34, W239–242. doi: 10.1093/nar/gkl190

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pfleger, C., Radestock, S., Schmidt, E., and Gohlke, H. (2013a). Global and local indices for characterizing biomolecular flexibility and rigidity. *J. Comput. Chem.* 34, 220–233. doi: 10.1002/jcc.23122

Pfleger, C., Rathi, P. C., Klein, D. L., Radestock, S., and Gohlke, H. (2013b). Constraint Network Analysis (CNA): a python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. *J. Chem. Inf. Model.* 53, 1007–1015. doi: 10.1021/ci400044m

Piraino, S. W., and Furney, S. J. (2016). Beyond the exome: the role of non-coding somatic mutations in cancer. *Ann. Oncol.* 27, 240–248. doi: 10.1093/annonc/mdv561

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* 36, 983–987. doi: 10.1038/nbt.4235

Poulos, R. C., and Wong, J. W. H. (2018). Finding cancer driver mutations in the era of big data research. *Biophys. Rev.* 11, 21–29. doi: 10.1007/s12551-018-0415-6

Rader, A. J., Hespenheide, B. M., Kuhn, L. A., and Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3540–3545. doi: 10.1073/pnas.062492699

Raphael, B. J., Dobson, J. R., Oesper, L., and Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* 6:5. doi: 10.1186/gm524

Red Brewer, M., Yun, C. H., Lai, D., Lemmon, M. A., Eck, M. J., and Pao, W. (2013). Mechanism for activation of mutated epidermal growth factor receptors in lung cancer. *Proc. Natl. Acad. Sci. U.S.A.* 110, E3595–3604. doi: 10.1073/pnas.1220050110

Reimand, J., and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* 9:637. doi: 10.1038/msb.2012.68

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39:e118. doi: 10.1093/nar/gkr407

Roskoski, R. Jr. (2014). The ErbB/HER family of protein-tyrosine kinases and cancer. *Pharmacol. Res.* 79, 34–74. doi: 10.1016/j.phrs.2013.11.002

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., et al. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science* 304:554. doi: 10.1126/science.1096502

Schwarz, J. M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576. doi: 10.1038/nmeth0810-575

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–388. doi: 10.1093/nar/gki387

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA: protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6620–6625. doi: 10.1073/pnas.0810961106

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi: 10.1002/humu.22225

Sim, N. L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–457. doi: 10.1093/nar/gks539

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268–274. doi: 10.1126/science.1133427

Spinella, J. F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., et al. (2016). SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics* 17:912. doi: 10.1186/s12864-016-3281-2

Stephens, P., Edkins, S., Davies, H., Greenman, C., Cox, C., Hunter, C., et al. (2005). A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* 37, 590–592. doi: 10.1038/ng1571

Stephens, P., Hunter, C., Bignell, G., Edkins, S., Davies, H., Teague, J., et al. (2004). Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* 431, 525–526. doi: 10.1038/431525b

Stetz, G., Tse, A., and Verkhivker, G. M. (2017). Ensemble-based modeling and rigidity decomposition of allosteric interaction networks and communication pathways in cyclin-dependent kinases: differentiating kinase clients of the Hsp90-Cdc37 chaperone. *PLoS ONE* 12:e0186089. doi: 10.1371/journal.pone.0186089

Stetz, G., and Verkhivker, G. M. (2017). Computational analysis of residue interaction networks and coevolutionary relationships in the Hsp70 chaperones: a community-hopping model of allosteric regulation and communication. *PLoS Comput. Biol.* 13:e1005299. doi: 10.1371/journal.pcbi.1005299

Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013). OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* 29, 2238–2244. doi: 10.1093/bioinformatics/btt395

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., et al. (2016). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* 76, 3719–3731. doi: 10.1158/0008-5472.CAN-15-3190

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14330–14335. doi: 10.1073/pnas.1616440113

Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D. S. (2007). The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol.* 369, 1318–1332. doi: 10.1016/j.jmb.2007.03.069

Tvorogov, D., Sundvall, M., Kurppa, K., Hollmen, M., Repo, S., Johnson, M. S., et al. (2009). Somatic mutations of ErbB4: selective loss-of-function phenotype affecting signal transduction pathways in cancer. *J. Biol. Chem.* 284, 5582–5591. doi: 10.1074/jbc.M805438200

Tyner, C., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., et al. (2017). The UCSC genome browser database: 2017 update. *Nucleic Acids Res.* 45, D626–D634. doi: 10.1093/nar/gkw1134

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., and Rousseau, F. (2011). A graphical interface for the FoldX forcefield. *Bioinformatics* 27, 1711–1712. doi: 10.1093/bioinformatics/btr254

Vijayabaskar, M. S., and Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophys. J.* 99, 3704–3715. doi: 10.1016/j.bpj.2010.08.079

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Wang, Z., Longo, P. A., Tarrant, M. K., Kim, K., Head, S., Leahy, D. J., et al. (2011). Mechanistic insights into the activation of oncogenic forms of EGF receptor. *Nat. Struct. Mol. Biol.* 18, 1388–1393. doi: 10.1038/nsmb.2168

Wang, Z., Shen, D., Parsons, D. W., Bardelli, A., Sager, J., Szabo, S., et al. (2004). Mutational analysis of the tyrosine phosphatome in colorectal cancers. *Science* 304, 1164–1166. doi: 10.1126/science.1096096

Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* 14, 703–718. doi: 10.1038/nrg3539

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Wood, D. E., White, J. R., Georgiadis, A., Van Emburgh, B., Parpart-Li, S., Mitchell, J., et al. (2018). A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.* 10:eaar7939. doi: 10.1126/scitranslmed.aar7939

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108–1113. doi: 10.1126/science.1145720

Wu, J., Wu, M., Li, L., Liu, Z., Zeng, W., and Jiang, R. (2016). dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. *Database* 2016:baw024. doi: 10.1093/database/baw024

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database* 2011:bar026. doi: 10.1093/database/bar026

Zhou, W., Ercan, D., Chen, L., Yun, C. H., Li, D., Capelletti, M., et al. (2009). Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature* 462, 1070–1074. doi: 10.1038/nature08622

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 19 June 2019
doi: 10.3389/fmolb.2019.00046

Using Dimensionality Reduction to Analyze Protein Trajectories

Gareth A. Tribello[1]* *and Piero Gasparotto*[2]

[1] Atomistic Simulation Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, United Kingdom,
[2] Department of Physics and Astronomy, Thomas Young Centre, University College London, London, United Kingdom

In recent years the analysis of molecular dynamics trajectories using dimensionality reduction algorithms has become commonplace. These algorithms seek to find a low-dimensional representation of a trajectory that is, according to a well-defined criterion, optimal. A number of different strategies for generating projections of trajectories have been proposed but little has been done to systematically compare how these various approaches fare when it comes to analysing trajectories for biomolecules in explicit solvent. In the following paper, we have thus analyzed a molecular dynamics trajectory of the C-terminal fragment of the immunoglobulin binding domain B1 of protein G of Streptococcus modeled in explicit solvent using a range of different dimensionality reduction algorithms. We have then tried to systematically compare the projections generated using each of these algorithms by using a clustering algorithm to find the positions and extents of the basins in the high-dimensional energy landscape. We find that no algorithm outshines all the other in terms of the quality of the projection it generates. Instead, all the algorithms do a reasonable job when it comes to building a projection that separates some of the configurations that lie in different basins. Having said that, however, all the algorithms struggle to project the basins because they all have a large intrinsic dimensionality.

Keywords: molecular dynamics, dimensionality reduction, machine learning, trajectory analysis, computer simulation, clustering

OPEN ACCESS

Edited by:
Vojtech Spiwok,
University of Chemistry and Technology in Prague, Czechia

Reviewed by:
Peng Tao,
Southern Methodist University, United States
Shuanghong Huo,
Clark University, United States
Evangelos Coutsias,
Stony Brook University, United States

***Correspondence:**
Gareth A. Tribello
g.tribello@qub.ac.uk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *15 March 2019*
Accepted: *31 May 2019*
Published: *19 June 2019*

Citation:
Tribello GA and Gasparotto P (2019) Using Dimensionality Reduction to Analyze Protein Trajectories. Front. Mol. Biosci. 6:46. doi: 10.3389/fmolb.2019.00046

1. INTRODUCTION

For many years researchers have sought to determine whether it is possible to predict the tertiary structure of a protein from the amino acid sequence alone. Numerous structure prediction algorithms have been developed to solve this problem and these algorithms have then been tested in the biennial, community-wide blind tests to predict the unknown structures of proteins. The tertiary structure of the protein is only a part of the story, however. To truly understand how these molecules function in the cell we must also understand their dynamical behavior (Dunker et al., 2008; Constanzi, 2010; Goldfeld et al., 2011; Kmiecik et al., 2015). In fact, a whole new class of intrinsically disordered proteins (IDP) that do not have the same familiar and relatively permanent tertiary structures has been discovered (Dyson and Wright, 2005).

Molecular dynamics (MD) simulations with force fields that model the interactions between the atoms in the biomolecule have emerged as a useful tool for investigating the dynamical structure of proteins. This technique is, in fact, particularly important for IDPs as the experiments alone often

do not provide sufficient information on the conformers adopted by the biomolecules. Detailed structural information is thus obtained by formulating constraints based on the experimental data and by then performing constrained MD simulations (Bonomi et al., 2017). There is a problem, however, when it comes to visualizing the results from these MD simulations. Biomolecules, unlike simpler chemical systems such as solids, do not undergo transitions that involve a change of symmetry. Instead, they undergo transitions between various low symmetry structures, which makes it difficult to know how to analyze the trajectories that emerge from MD simulations.

During the last few decades, many researchers have sought to solve the problems outlined in the previous paragraph by analyzing their MD trajectories using dimensionality reduction algorithms. The theory behind such approaches is that the computer can determine what features of the data are important and what features are simply noise. Many different algorithms have been used to analyze MD trajectories and some have even been developed with this particular purpose in mind (Garcia, 1992; Amadei et al., 1993; Balsera et al., 1996; Yuguang et al., 2005; Das et al., 2006; Konrad, 2006; Plaku et al., 2007; Spiwok et al., 2007; Zhuravlev et al., 2009; Stamati et al., 2010; Sutto et al., 2010; Ceriotti et al., 2011; Spiwok and Kralova, 2011; Tribello et al., 2012; Noé and Clementi, 2015, 2017; Tiwary and Berne, 2016; Sultan and Pande, 2017; Chen and Ferguson, 2018; Sultan et al., 2018). Much less work has been done, however, to compare the performance of the various dimensionality reduction algorithms although there are a few notable examples of work on systems in implicit solvent in the literature (Duan et al., 2013).

One reason why few systematic comparisons between the projections of trajectories generated using different dimensionality reduction have been performed is that it is difficult to formulate an appropriate method to test the quality of a projection. After all, if we knew what the appropriate method for analyzing our trajectory was we most likely wouldn't be reliant on dimensionality reduction algorithms. Duan et al. (2013) argue that one feature of a good projection is that the distances between the projections of the points are similar to the true distances between the trajectory frames. This criterion is undoubtedly sensible but it is also the criterion that is used when optimizing the projection. What we thus find out when it is measured is the extent to which the algorithm was able to satisfy the constraints of the optimization problem. As Duan et al. (2013) point out it is much more difficult to unequivocally say that the assumptions of method X are appropriate. Particularly so when it comes to the non-linear methods. Nevertheless, in what follows we use a number of different algorithms to analyze the trajectory of a biomolecule in explicit solvent. We show two-dimensional projections of the trajectory that are obtained using each of these algorithms and perform various analyses to compare how well these projections have encoded the information in the trajectory in section 3. Before getting onto this, however, we provide some background information on the various algorithms that we have used in section 2 and the trajectory we have analyzed in section 3.

2. BACKGROUND

A molecular dynamics trajectory for a set of N atoms is essentially an ordered set of $3N$-dimensional vectors. Furthermore, if we assume that the simulated system is equilibrated and if we are only interested in static properties then the order the vectors are in is not particularly important. The problem of analysing the trajectory thus reduces down to one of simply visualizing the position of each frame in the trajectory relative to all the others. Obviously, however, we cannot draw a diagram illustrating the position of each trajectory frame in the $3N$-dimensional vector space of atomic positions and are thus forced to discard (ideally) all but two of these dimensions.

Oftentimes decisions as to how to plot the relationships between the trajectory frames are made using chemical or physical intuition about the problem under study. In these cases, some function/s of the atomic positions - usually referred to as collective variables or CVs - is computed for each of the trajectory frames. The positions of each of the trajectory frames in the low-dimensional CV space can then be plotted so that conclusions can be drawn about the parts of space that were sampled in the trajectory.

Dimensionality reduction algorithms adopt a similar approach. Instead of using chemical/physical intuition to decide on the appropriate low dimensional space in which to visualize the data, however, dimensionality reduction algorithms introduce a loss function. Optimization algorithms are then used to ensure that a low-dimensional representation of the data that minimizes the value of this loss function is found.

To understand how these algorithms work in practice consider how the multidimensional scaling (MDS) algorithm (Borg and Groenen, 2005) would produce a low dimensional representation, $\{\mathbf{x}^{(i)}\}$ of the N M-dimensional vectors in the set, $\{\mathbf{X}^{(i)}\}$. The first step is to compute the dissimilarity between each pair of trajectory frames using Pythagoras theorem:

$$||\mathbf{X}^{(i)} - \mathbf{X}^{(j)}|| = \sqrt{\sum_{k=1}^{M} (X_k^{(i)} - X_k^{(j)})^2} \tag{1}$$

The low dimensional representation is then found by optimizing the loss function:

$$\chi(\{x^{(i)}\}) = \sum_{i \neq j} \left\{ ||\mathbf{X}^{(i)} - \mathbf{X}^{(j)}|| - ||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|| \right\}^2 \tag{2}$$

where $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are the low dimensional representations of points i and j respectively. In other words, the MDS algorithm works by endeavoring to arrange the points in the low dimensional space so that the distances between the projections are the same as the dissimilarities between the trajectory frames.

All of the dimensionality reduction algorithms that have been used to analyze molecular dynamics trajectories work using a strategy that is similar to the one described above. In short, some features that describe how the data is arranged across the high dimensional space are computed. Points are then arranged in the low dimensional space in a way that reproduces the high-dimensional features as closely as possible. A large number of

algorithms exist, however, because one can perform numerous variations on this theme. The best way to understand these variants is to consider the choices that must be made in order to analyze a trajectory using one of these algorithms:

- **How to represent the trajectory frames** - The simplest representation to use for the trajectory frames is a list of atomic coordinates. Using the atomic coordinates is not particularly sensible, however, as one essentially discards all chemical and physical intuition about the problem at hand. It is thus often better to embed the known physical details about the problem when one is calculating the high-dimensional vectors that represent each of the trajectory frames. As a case in point, it may well be better to input vectors of backbone dihedral angles into the dimensionality reduction algorithm when one is examining a trajectory of a biomolecule. Alternatively, a number of general purpose representations of atomic structures have been developed in the context of fitting potentials based on the results of density functional theory calculations (Behler, 2011; Bartók et al., 2013; Willatt et al., 2018). These representations have the advantage of providing a systematically convergent description of the chemical environment and were used in De et al. (2016); Bartók et al. (2017); Musil et al. (2018).
- **Whether to use landmark points and if so how to choose these landmarks** - Dimensionality reduction algorithms scale quadratically with the number of input vectors. It is thus often not feasible to perform a dimensionality reduction with a whole trajectory as input. Researchers therefore often adopt a more computationally-efficient strategy whereby they run the algorithm on a small number of so-called landmark frames. Projections for all the frames in the trajectory are then constructed using an out-of-sample procedure. Obviously, using this procedure entails choosing an appropriate number of landmark points and devising a strategy for selecting these landmarks. Typically, however, one of two strategies is employed either (i) landmarks frames are selected randomly so the distribution of landmarks resembles the distribution of trajectory frames or (ii) landmarks frames are selected using farthest point sampling so as to have frames from all the regions of configuration space that were explored. A third option that combines the strengths of these two approaches is discussed in Ceriotti et al. (2013)
- **How to construct the loss function** - By changing the way the loss function is defined one can change what features from the high-dimensional space the algorithm is endeavoring to reproduce as it arranges the projections in the low-dimensional space. Minimizing Equation 2 for instance is akin to attempting to reproduce the Euclidean distances between the high-dimensional vectors. It is possible to use a different method for calculating the dissimilarity between the trajectory frames, however. For example, in isomap (Tenenbaum et al., 2000) dissimilarities between the high-dimensional frames are computed by using Dijkstra's shortest path algorithm to compute the approximate geodesic distances between frames. In an isomap projection, it is thus the geodesic distances between frames that are reproduced. Other algorithms, sketch-map for example (Ceriotti et al., 2011), have a loss function that is designed so that only a particular subset of the dissimilarities between the trajectory frames are reproduced. Yet another option is to design the loss function so that a matrix of non-linear kernels is reproduced rather than a matrix of dissimilarities (Schölkopf et al., 1998; Schölkopf et al., 1999). One final option is to design a loss function that reproduces the distribution of points in the neighborhood of each of the high-dimensional points (van der Maaten and Hinton, 2008).
- **How to optimize the loss function** - As with any optimization problem there is a concern when finding the low dimensional projection that the minimum found is a local rather than the global optimum. In many of the commonly used algorithms, this problem is sidestepped by insisting that the distances between the projections should constitute the best linear approximation of the dissimilarities. This approximation simplifies matters considerably as finding the projections simply becomes a matter of diagonalizing a matrix. The fact remains, however, that the algorithm used to optimize the loss function may have an effect on the final projection produced.

The rationale that should be born in mind when making these decisions is not always clear. In other fields, decisions are often made based on an understanding of what the high-dimensional data looks like and on an understanding of what features in the high-dimensional data set the users of these algorithms would like to reproduce (Rosman et al., 2010). One might therefore, suspect that by trying a range of algorithms and by determining how well each one performs one might be able to get some insight into the structure of the data in the high dimensional space.

3. METHODS

To test how effective various dimensionality reduction algorithms are at projecting data from biomolecular trajectories we took the data from the parallel tempering metadynamics trajectories of the 16-residue C-terminal fragment of the immunoglobulin binding domain B1 of protein G of Streptococcus that was generated in Ardevol et al. (2015) and projected it using a range of different algorithms. Within that work, the protein was simulated using Gromacs-4.5.5 (Hess et al., 2008), the AMBER99SB-ILDN* force field (Lindorff-Larsen et al., 2010) and an explicit solvation model. The protein and surrounding water molecules were then simulated for 300 ns/replica with metadynamics biases that acted on the radius of gyration and the number of hydrogen bonds between backbone atoms. The same protein was studied in the work in the work on comparing different dimensionality reduction algorithms by Duan et al. (2013) but an implicit solvent model was used in that work rather than the explicit model that we have used.

The wild-type trajectory in the work of Ardevol et al. (2015) that we have analyzed in this work contains 150,000 trajectory frames. Running each of the dimensionality reduction algorithms on this large number of trajectory frames would be computationally prohibitive so we selected a subset of 25,311 to analyse with each of the algorithms by sampling configurations

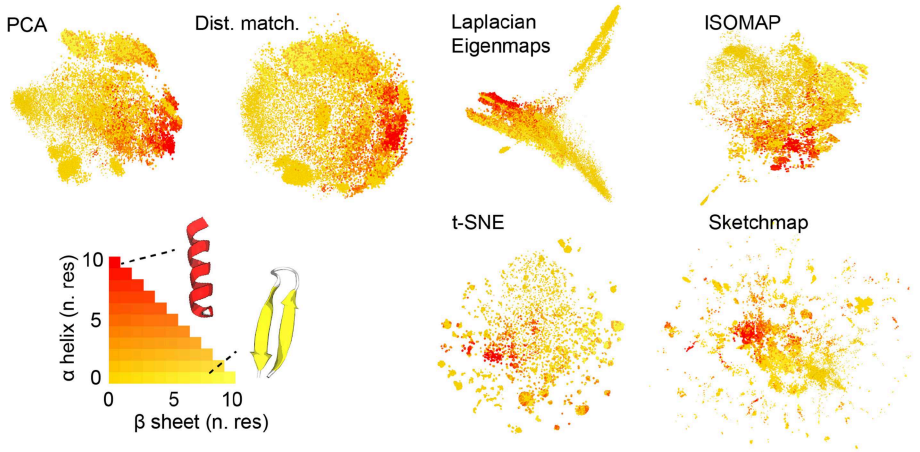


FIGURE 1 | The projections of the trajectory that were generated using each of the various dimensionality reduction algorithms. As discussed in the text the projections in these figures are colored in accordance with the secondary structure content of the corresponding trajectory frame. As you can see the majority of the algorithms do a reasonable job of separating those frames that have a configuration that resembles an alpha helix from those that have a configuration that resembles a beta sheet.

from the trajectory of the lowest-temperature replica at random. For each of these configurations, we computed the full set of 32 torsional backbone dihedral angles. Two-dimensional projections for each of these 32-dimensional vectors were then generated using the implementations of the various algorithms that are available in SciKit Learn (Pedregosa et al., 2011) and the sketch-map code (Ceriotti et al., 2011). Detailed step-by-step instructions showing how these projections were generated using these tools can be found in the supporting information (**Data Sheet 1**).

The algorithms that we used to project the data were principal component analysis (PCA) (Jolliffe, 2002), Laplacian Eigenmaps (Belkin and Niyogi, 2003), isomap (Tenenbaum et al., 2000), t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) and sketch-map (Ceriotti et al., 2011). In addition, we also generated a projection by simply minimizing Equation 2 using conjugate gradients (dist. match).

We performed PCA on the dihedral angles in a way that is sympathetic to the periodicity of these variables by using the method described in Yuguang et al. (2005); Konrad (2006); Altis et al. (2007). Meanwhile, for all the remaining algorithms we simply incorporated periodic boundaries when calculating the vectors connecting the positions of the trajectory frames in the space of dihedral angles. For isomap and Laplacian Eigenmaps we constructed graphs in the high dimensional space by connecting each point to its 15 nearest neighbors. For Laplacian Eigenmaps we then used a Gaussian kernel with a gamma parameter of 1. When using t-SNE we employed the Barnes-Hut implementation with a perplexity of 110 and a theta value of 0.5. Lastly, for sketch-map, we selected 1000 landmark point using the well-tempered farthest point sampling algorithm that is described in the appendix of Ceriotti et al. (2013) and a gamma parameter of 0.1. Weights for each of these landmarks were generated using a Voronoi procedure and the sketch-map stress function with parameters $\sigma = 6$, $A = 8$, $B = 8$, $a = 2$ and $b = 8$ was then optimized to find projections. Once projections for these landmarks had been found the rest of the trajectory was projected using the out of sample procedure described in Tribello et al. (2012).

4. RESULTS

Figure 1 shows the projections of the trajectories for immunoglobulin binding domain B1 of protein G that we obtained. Before projecting the trajectory we used the STRIDE algorithm discussed in Frishman and Argos (1995) to determine the secondary structure content in each of the frames that were analyzed. In particular, we counted the number of residues that had a structure that was similar to an alpha helix and the number of residues that had a structure that was similar to a beta sheet. When constructing the projections in **Figure 1** we thus colored the projections according to the number of residues in the corresponding trajectory frames that appeared to be in an alpha helix configuration and the number of residues that appeared to be in a configuration that resembled a beta sheet. Coloring the projections in this way gives us a qualitative way to compare how well each of the algorithms does when it comes to projecting the trajectory data. What we see is that all the algorithms do a reasonable job of separating the configurations that are predominantly alpha-helix-like from those that have a structure that is predominantly composed of beta sheets. In this sense at least then the algorithms all give a reasonable projection of the high-dimensional data.

There are additional observations to be made based on the results in **Figure 1**. For PCA the distances between the projections are systematically shorter than the dissimilarities between the corresponding trajectory frames. This fact is illustrated in **Figure 2**, which shows pair distribution functions for the dissimilarities, R_{ij}, between the frames and the distances, r_{ij}, between their corresponding projections for each of the

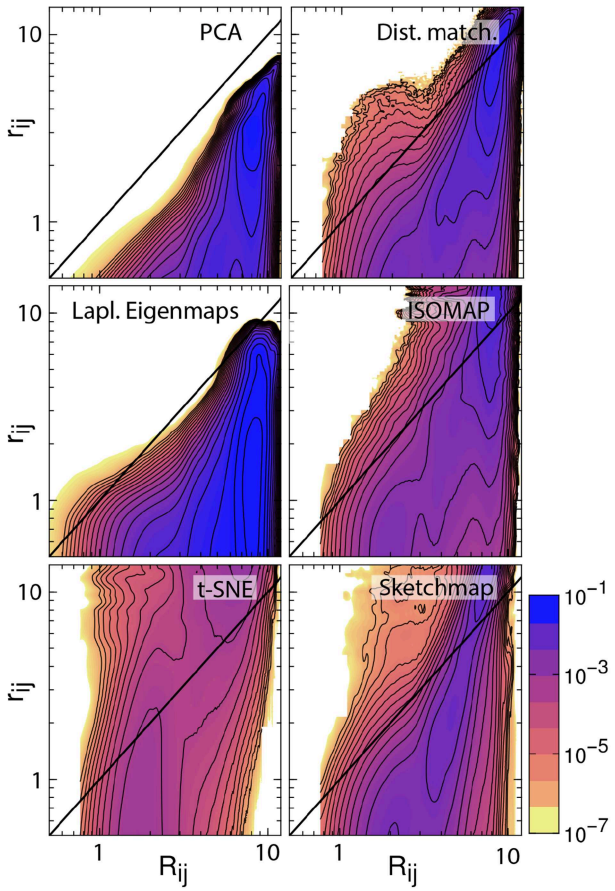


FIGURE 2 | Histograms illustrating the joint probability density function for the dissimilarities between the configurations in the trajectory and the distances between the corresponding projections of these trajectory frames. The particular projections that have been analyzed here are those that are shown in **Figure 1**. The black line in each of these figures is the line $R_{ij} = r_{ij}$. For an ideal projection all the density in these histograms would lie on this line.

representations of the data shown in **Figure 1**. To compute the dissimilarities, R_{ij}, we inserted the vector of backbone dihedral angles for each pair of trajectory frames into Equation 1 and made suitable dispensations for the periodicity of these variables. The distances, r_{ij}, were, meanwhile, computed by using Equation 1 on the projections of these points. With these two quantities computed we then determined these probability density functions by using:

$$P(D,d) = \frac{1}{0.5N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i} \delta(R_{ij} - D)\delta(r_{ij} - d)$$

where N is the number of frames that were projected and where δ is a Dirac delta function.

If a projection is perfect the distances between the projections and the dissimilarities between the trajectory frames are identical and all the density in the joint distribution function is concentrated on the line $y = x$. For all the probability density functions shown in **Figure 2**, however, we see that the density away from $y = x$ is substantial and we, therefore, know that the projections are thus imperfect. Furthermore, for PCA we see that all the density lies underneath the line $y = x$, which is why we are able to state that the distances between the projections are systematically shorter than the dissimilarities between the corresponding trajectory frames. The fact that these distances are shorter when we use this algorithm is unsurprising, however. This algorithm projects the high-dimensional data into a linear subspace. Any differences between configurations that are along directions that are orthogonal to this subspace are thus discarded when projections are constructed in this way.

It is possible to formulate PCA as a linear optimization of Equation 2. In other words, this algorithm projects the data in the linear subspace that is best able to reproduce the dissimilarities between frames, which, incidentally, is why the distances between the projections are systematically shorter than the dissimilarities between the trajectory frames. We can avoid producing a projection in which the distances between points are systematically shorter than the dissimilarities between the corresponding trajectory frames by minimizing Equation 2 using an iterative algorithm such as conjugate gradient. The top right panel of **Figure 2** illustrates that when we do so the joint probability distribution for the distances and dissimilarities is then peaked around the line $y = x$ so some distances are projected further apart than they should be while others are projected closer together. The effect this has on the appearance of the projection is illustrated in **Figure 1**. Essentially the projections of the points are spread more uniformly over the low dimensional space than they would be if the projection had been constructed using PCA. Given that one of our aims is to identify the various basins in the free energy landscapes that were sampled during the trajectory this spreading out of the projections is clearly disadvantageous as it may cause different basins to overlap with each other.

The other algorithms that have been tested in **Figures 1**, **2** all use different criteria when constructing the projections. In other words, these algorithms do not seek to generate projections in which the Euclidean distances between the trajectory frames are reproduced in the projection. Instead, Laplacian Eigenmaps (Belkin and Niyogi, 2003) and isomap (Tenenbaum et al., 2000) seek to reproduce the diffusion distances and geodesic distances respectively and calculate these distances by using ideas from graph theory. t-SNE (van der Maaten and Hinton, 2008) and sketch-map (Ceriotti et al., 2011), meanwhile, do not try to generate projections in which the distances are reproduced at all. Instead, sketch-map seeks to ensure that points that are far apart in the high-dimensional space are projected far apart, while simultaneously ensuring points that are close together in the high-dimensional space are projected near to each other. t-SNE, meanwhile, endeavors to generate a low-dimensional projection that reproduces the distribution of neighbors around each point in the high-dimensional space. The projections generated using each of these algorithms that are shown in **Figure 1** differ starkly from those that are generated using the two forms of distance matching that were described previously. Reassuringly, however, all four algorithms do a reasonable job when it comes to

separating the configurations that resemble an alpha helix from those that resemble a beta sheet.

The non-Euclidean-distance-matching algorithms: isomap, Laplacian Eigenmaps, t-SNE, and sketch-map make assumptions about the structure of the manifold from which the trajectory frames are sampled, which may or may not be valid. These assumptions affect the dissimilarities that these algorithms attempt to reproduce and can thus affect the distances between the projections. In fact, Duan et al. (2013) showed that projections generated using isomap do not preserve the neighborhood structure in the high dimensional space even when 50-dimensional projections are constructed precisely because of the way in which the geodesic distances are constructed. Furthermore, Brown et al. (2008) showed that these algorithms can give incorrect estimates for the dimensionality of energy landscapes precisely because they assume that a manifold-like structure exists that may not be there. To investigate the effect these assumptions are having on the appearance of the projections the bottom four joint probability distributions in **Figure 2** illustrate how each of the non-distance-matching algorithms performs when it comes to generating a projection in which the Euclidean distances between the various trajectory frames are reproduced. As you can see only Laplacian Eigenmaps produces a projection in which the distances between projections are systematically shorter than the dissimilarities between the corresponding trajectory frames. All the remaining algorithms generate projections in which only some distances are underestimated. The distances between the projections of the points in these representations are, in contrast to the other algorithms, predominantly larger than the dissimilarities between the corresponding trajectory frames. This "stretching out" of the distances in the projections is arguably a good thing as it ensures that the representations of the various basins in the low dimensional space do not overlap. At the same time, however, it may be that this stretching out of space causes basins to appear split into smaller pieces in the projection, which may give one the impression that there are more features in the energy landscape than there are in actuality.

To better understand the various projections in **Figure 1** we performed an analysis of the trajectory that was similar to that performed in Gasparotto et al. (2018). To generate the images shown in **Figure 3** we analyzed the high-dimensional data using the probabilistic analysis of molecular motifs (PAMM) method that is discussed in Gasparotto and Ceriotti (2014) and Gasparotto et al. (2018). This clustering method works by first selecting a sparse grid of points in the high dimensional space. The probability density at each of these grid points is then computed using kernel density estimation (KDE). Once the density at each of these points has been estimated the Quick-Shift algorithm is used to connect points on the grid to nearby points that have higher probability densities unless a stopping criterion is satisfied. The points at which the stopping criteria are satisfied are then assumed to correspond to the various local maxima in the probability density. In Gasparotto et al. (2018) dimensionality reduction was performed in order to better understand the clusters output by PAMM. In this work, however, we performed PAMM in order to assign each of the structures in our trajectory to the nearest local maximum in the high-dimensional probability distribution so that the way in which each of these features is represented in each of the projections could be visualized. **Figure 3**, therefore, shows representative configurations from each of the eleven modes that were identified using PAMM. The projections in **Figure 3** are then colored according to the particular mode from which the corresponding trajectory frame was sampled. Once again we find that all the algorithms do a reasonable job of projecting the data. In most of the projections, the different modes are projected in different parts of the low dimensional space and there is little overlap between the projections of the modes.

There are a number of specific things that are worth noting about the projections shown in **Figure 3**. The first is that all the algorithms clearly struggle to project the PAMM motif that is shown in light blue in the figures. In all the projections the blue points are split into multiple distinct clusters. If, as PAMM is telling us, these points are all from the same mode you would hope that they would all be clustered in together in a single feature of the projection. In other words, you would hope that the points that are clustered together in the high-dimensional space would appear clustered together in the projection. Given that this appears to not be the case it would be unwise to run a clustering algorithm on the low-dimensional projection.

The fact that the blue PAMM motif appears to be so diffuse in the projections is surprising as if you look at the representative structure at the top right of the figure the structures in this basin would appear to resemble alpha helices. One would expect such a structure to be in a deep energetic minimum in the free energy landscape and one might further expect that the entropy of the configuration - and hence the size of the feature - to be small. A comparison of **Figures 1** and **3** clears these matters up, however. In **Figure 1**, remember, the points were colored red if they had a configuration resembling an alpha helix. It is clear that the red regions in **Figure 1** are considerably smaller than the blue regions in **Figure 3**. It is thus clear that the mode that is colored light blue in **Figure 3** must also contain structures that do not have a configuration that resembles an alpha helix. What thus seems likely is that the alpha helix configuration is at the bottom of a broad funnel in the free energy landscape, which is why these features appear to be so diffuse in these projections.

The fact that the alpha helix lies at the base of a broad funnel in the free energy landscape is further confirmed by the numbers that are given underneath the representative configurations in **Figure 3**. These numbers were computed from the covariance matrices for each of the various clusters that were identified using PAMM. The first row of numbers in **Figure 3** gives the determinant of the covariance matrices scaled by the determinant of the covariance of the largest cluster. As you can see the alpha-helical cluster that is shown in light-blue has the largest determinant and this cluster is thus the mode that takes up the largest volume of the high-dimensional space by some considerable margin. Furthermore, the second largest mode is the other folded configuration; namely, the purple-colored basin that includes the beta-hairpin configuration.

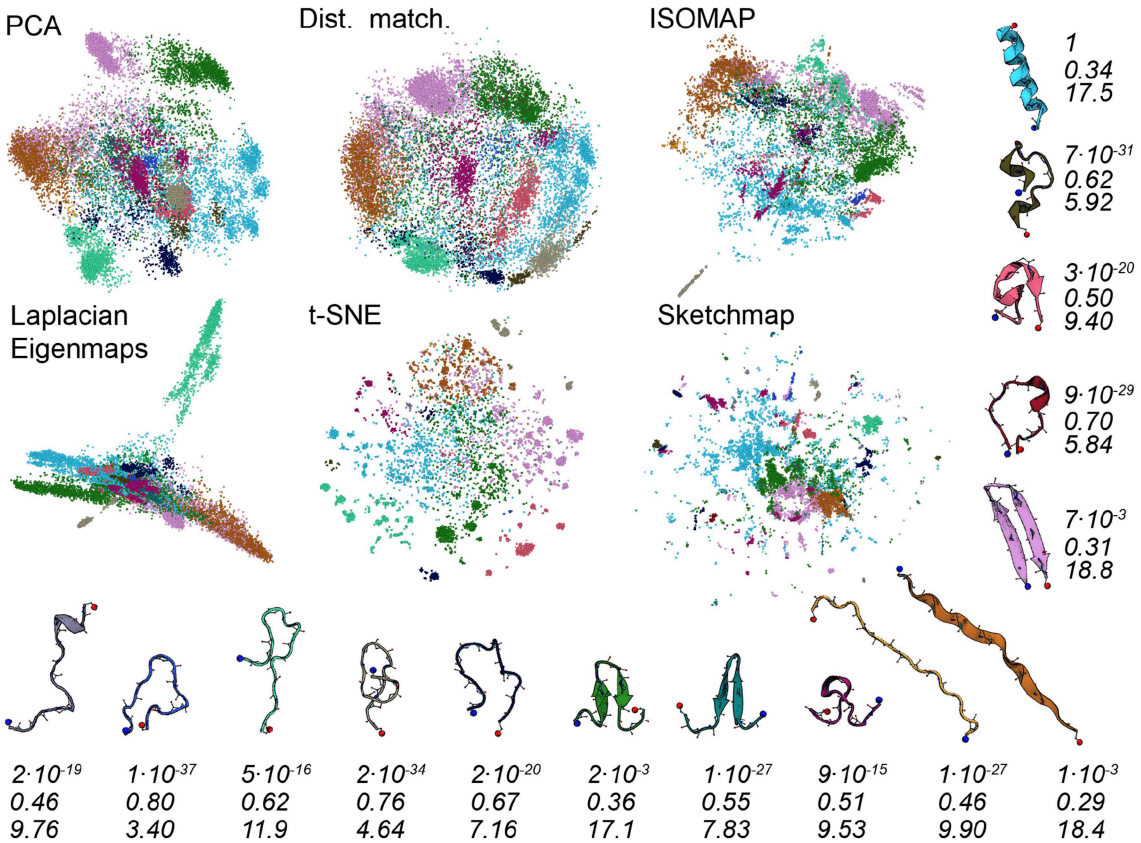


FIGURE 3 | As discussed in the text, we used the PAMM clustering algorithm to identify clusters that correspond to free energy minima in the high dimensional space. In this figure, we have thus shown the projections again but this time with the points colored according to the particular cluster the PAMM algorithm identifies each of the high dimensional configurations to lie within. Representative and similarly-colored configurations from each of the various clusters are shown below and to the right of the projections. The three rows of numbers beneath the representative configurations provide information on the covariance matrix for each of the clusters that PAMM identified. The top row of numbers contains the determinants of this matrix but these numbers have been scaled so that the cluster with the largest determinant has a determinant of 1. To compute the second row of numbers, meanwhile, we divided the sum of the first two eigenvalues of these covariance matrices by the sum of all the eigenvalues of these matrices. The final row of numbers contains the estimated dimensionality of each of the clusters, which was computed using Equation 3.

To compute the second row of numbers in **Figure 3** we diagonalized the covariance matrix for each of the PAMM clusters and calculated the sum of the largest two eigenvalues of this matrix divided by the sum of all the eigenvalues. These numbers thus give a measure of how much of the variance of each particular feature can be represented in a two-dimensional linear subspace of the high dimensional space. It is apparent from this analysis that each of the PAMM features that we have identified is not well represented in a two-dimensional space as much of the variation within each of the basins is in directions that are orthogonal to these two principal eigenvectors. We should thus perhaps not be surprised to find that the algorithms struggle to project these clearly-high-dimensional features correctly.

Further information on the features that are difficult to reproduce in a two-dimensional space is given in the third row of numbers in **Figure 3**. This row of numbers contains the dimension of each basin which was estimated using:

$$D_i = \exp\left(-\sum_{k=1}^{M} \eta_k \log \eta_k\right) \qquad \text{where} \qquad \eta_k = \frac{\lambda_k}{\sum_{j=1}^{M} \lambda_j} \quad (3)$$

and $\{\lambda_k\}$ is the eigenvalue spectrum of the covariance matrix for the ith PAMM feature. As you can clearly see all the algorithms do a good job of projecting clusters that have an estimated dimensionality that is less than around seven. These features appear as a single cluster in the low dimensional space. It is those features that have an estimated dimension that is higher than around seven that represent a problem. The projections of points from these clusters are often spread across multiple separated clusters, which makes it difficult to realize that these configurations are all part of a single basin.

It is interesting to note from **Figure 3** how strongly the projection generated using Laplacian Eigenmaps differs from the others. The projection generated using this algorithm has the light green motif separated strongly from all the other motifs, which appear squeezed together. This same squeezing together of some of the motifs and pulling apart of others is not observed in the other representations of the trajectory. The representative structure for the light green motif offers a tantalizing explanation as to why this particular behavior might be observed for this particular algorithm. The light green motif is the only structure containing no secondary structure content. One might therefore

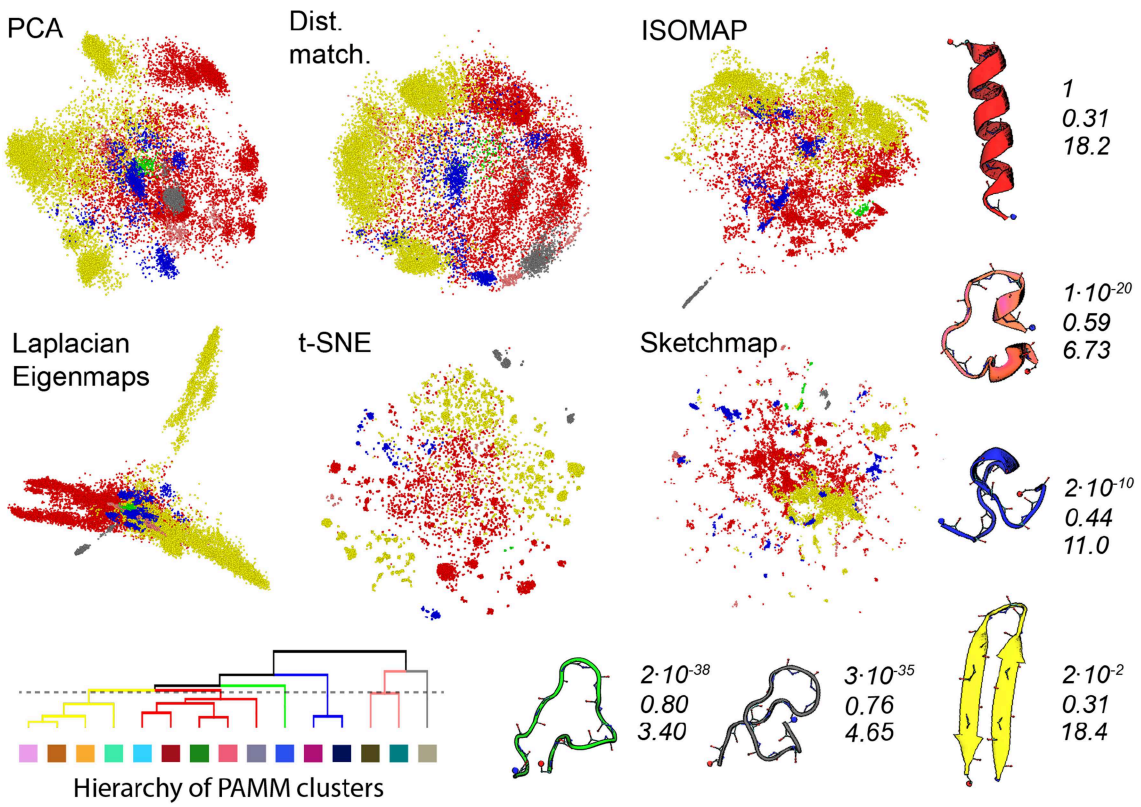


FIGURE 4 | As discussed in the text, we can use an agglomerate procedure based on the error in the KDE procedure that is at the heart of PAMM to construct the hierarchy shown in the bottom left corner of this figure for the PAMM clusters that were identified and shown in **Figure 3**. This hierarchy would seem to indicate that we can reduce the 11 PAMM clusters shown in **Figure 3** down to 6 macro clusters. In the projections above the points are therefore colored according to the particular macro cluster the PAMM algorithm identities each of the high dimensional configurations to lie within. As in **Figure 3** representative and similarly-colored configurations from each of the various clusters are shown below and to the right of the projections together with scaled values for determinant of the covariance matrices for each of the clusters (**Top**), ratios of the sum of the largest two eigenvalues of these matrices to the sum of all the eigenvalues of these matrices (**Middle**) and the estimated dimensionality of each of the features (**Bottom**).

suppose that this motif corresponds to a random coil, that the diffusion distance, which Laplacian Eigenmaps endeavors to reproduce when it constructs a projection, between these states and the other folded configurations might well be quite large, and that the transitions between this random coil state and the other folded states might thus be quite infrequent. As we will show in what follows it is not clear that this interpretation is correct, however.

Gasparotto et al. (2018) discuss how bootstrapping can be used to judge the statistical significance of the clusters identified by PAMM. Furthermore, in analysing the errors in this way a distance between clusters that determines whether or not clusters get merged in some of the 41 bootstrap samples that we took from the trajectory can be defined. We have used this distance measure in **Figure 4** to generate a tree-like plot that illustrates the results of a hierarchical clustering procedure performed on the eleven clusters that were identified in **Figure 3**. The clusters that are connected in this plot are those that are likely to be merged in the bootstrap samples. We thus also re-show the projections generated using each of the algorithms in **Figure 4** but this time we have reduced the number of PAMM clusters from eleven to six by using the connectivity that is identified in the tree diagram.

The PAMM analysis shown in **Figure 4** makes clear that the free energy landscape for this protein contains two broad funnels and three additional, much-smaller funnels. The beta-hairpin and alpha-helical configurations lie at the bases of the two broad funnels while the three narrower funnels have three unfolded structures at their base that have much higher energies. We can speak of the size of the funnels because we have, once again, performed an analysis of the covariance matrices and because the determinant of the clusters that contain the alpha-helix and beta-sheet are both considerably larger than those of the other identified features. The other aspects of this analysis demonstrate that if we take the ratio of the sum of the two largest eigenvalues of the covariance matrix to the sum of all the eigenvalues of this matrix we find that many of the fluctuations that take place within these two funnels take place along directions that are orthogonal to the direction of the eigenvectors that correspond to these two largest eigenvalues of the funnel. Furthermore, the estimated dimensionalities of these two features are substantial.

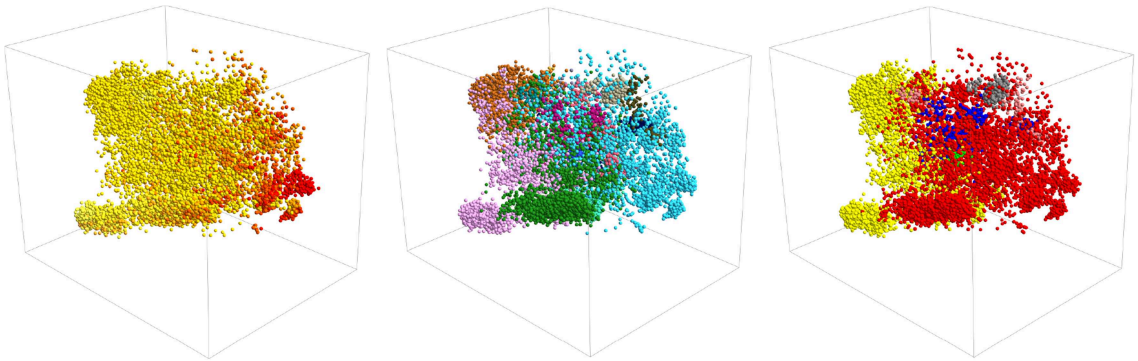


FIGURE 5 | A three-dimensional projection of the trajectory that was generated using PCA. The positions of projections in the three panels above are all identical. In the left-most panel, however, points are colored in accordance with their secondary structure as they are in **Figure 1**. In the middle panel, they are colored as they are in **Figure 3** and in the right panel they are colored as in **Figure 4**.

When we look at the various projections in **Figure 4** we see that all the algorithms do a good job of separating the red configurations that come from the funnel that has the alpha helix at is base from the yellow configurations that come from the funnel that has the beta sheet at its base. Furthermore, the fact that these two features have a larger spatial extent than the other clusters identified by PAMM is clear from the way these features are projected using all the different algorithms. Having said all that though there is no way that one would be able to identify these two features in the landscape if one were just given the projections generated using one of these algorithms. In all these various representations of the trajectory, the red and yellow points appear divided up into multiple separate clusters. The most dramatic example of this is in the projection generated using Laplacian Eigenmaps which divides the yellow points between two very distinct clusters. It is clear from a comparison of the projections in **Figures 3** and **4** that the configurations in these two clusters are separated in all the other projections of the trajectory. Furthermore, the hierarchy of clusters shown in **Figure 4** also indicates that these two clusters are likely to be separate features. It may well be, therefore, that the rate of transition between these two parts of configuration space is slow because there is perhaps a kinetic trap on the folding funnel for the beta hairpin. This observation does, however, raise an interesting question when it comes to selecting which dimensionality reduction algorithm to use when constructing a projection. Using the slow degrees of freedom to construct a low-dimensional representation of the data makes physical sense but it may well be that the projections generated using algorithms that work by constructing a low dimensional representation of a trajectory in which the dissimilarities between trajectory frames are reproduced may give one a clearer sense of the various different structural possibilities in the ensemble.

It is interesting to ask if we can construct a clearer visualization of the structure in the data by producing a three-dimensional projection. **Figure 5** shows three representations of a three dimensional PCA projection of the trajectory with the points colored as in **Figures 1**, **3**, **4**. It is clear from this figure that the points in this three dimensional PCA projection are spread out over all three coordinates and certainly not split into distinct clusters. Furthermore, when it comes to distinguishing configurations with different secondary structures the projection is OK but there is still a substantial overlap between the regions of space where the structure has a lot of alpha-helical content and the regions of space where the structure more closely resembles a beta-hairpin as was the case for the two dimensional projection in **Figure 1**. In addition, each of the various PAMM features identified in **Figures 3**, **4** does not appear as a single cluster that is well separated from each of the other features. Instead, the points belonging to each of these features appear split between multiple apparently distinct clusters much like they appeared in the two-dimensional projections shown in **Figures 3**, **4**. In short, a three-dimensional projection of this trajectory does not provide much greater insight than the two-dimensional projections that we have shown thus far and is considerably harder to visualize and interpret.

5. CONCLUSIONS

In the preceding sections, we have analyzed a molecular dynamics trajectory for a short protein molecule using a number of different dimensionality reduction algorithms. The results we have are in some senses reassuring as all the algorithms do a reasonable job when it comes to giving a representation of the trajectory that gives a sense of the structural diversity that one observes in the trajectory. In all the projections if two configurations have markedly different structures they are projected in different parts of the low dimensional space. Furthermore, configurations that are structurally similar are for the most part projected close together. In other words, even projections constructed using the easier to apply dimensionality reduction algorithms such as PCA and MDS, which have no parameters that need to be tuned, can provide one with a useful visualization of the high-dimensional data.

When one of these dimensionality reduction algorithms clearly outperforms the others it is often because the data has some structure that only one of the algorithms can recognize. For instance, isomap will outperform PCA when it comes to projecting data that lies on a curved manifold because PCA assumes the data lies on a linear manifold in the high dimensional space. The fact that all the algorithms perform similarly well and that no algorithm outshines the other thus perhaps simply reflects the fact that we do not fully understand how the trajectory data is distributed across the high-dimensional space. In other words, none of the data distribution models underlying these various algorithms provides a complete description of the structure of the data in the high-dimensional space. It seems that the data does not all lie on a low-dimensional linear or non-linear manifold and similarly there perhaps isn't a single length scale that separates configurations that lie in different basins in the free energy landscape. Perhaps then, given that all these algorithms are imperfect, the appropriate strategy for analysing an MD trajectory is to try something similar to the approach that has been taken in this paper. In short, analyze the trajectory using a range of different dimensionality reduction and clustering algorithms and consider what the result from each analysis is telling you by comparing the results obtained.

DATA AVAILABILITY

The trajectories for the immunoglobulin binding domain B1 of protein G that we have analyzed within this article can be found online at https://github.com/cosmo-epfl/sketchmap/tree/master/examples/protein.

AUTHOR CONTRIBUTIONS

GT designed the work and wrote the text of the paper. PG analyzed the trajectories and produced the figures.

FUNDING

GT acknowledges funding from EP/L025124/1.

ACKNOWLEDGMENTS

GT and PG thank Michele Ceriotti for useful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00046/full#supplementary-material

REFERENCES

Altis, A., Nguyen, P. H., Hegger, R., and Stock, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* 126:244111. doi: 10.1063/1.2746330

Amadei, A., Linssen, A. B. M., and Berendsen, H. J. (1993). Essential dynamics of proteins. *PROTEINS Struct. Funct. Gen.* 17:412.

Ardevol, A., Tribello, G. A., Ceriotti, M., and Parrinello, M. (2015). Probing the unfolded configurations of a β-hairpin using sketch-map. *J. Chem. Theory. Comput.* 11, 1086–1093. doi: 10.1021/ct500950z PMID: 26579758.

Balsera, M. A., Wriggers, W., Oono, Y., and Schulten, K. (1996). Principal component analysis and long time protein dynamics. *J. Phys. Chem.* 100, 2567–2572.

Bartók, A. P., De, S., Poelking, C., Bernstein, N., Kermode, J. R., Csányi, G., et al. (2017). Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* 3:e1701816. doi: 10.1126/sciadv.1701816

Bartók, A. P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. *Phys. Rev. B* 87:184115. doi: 10.1103/PhysRevB.87.219902

Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* 134:074106. doi: 10.1063/1.3553717

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Bonomi, M., Helloer, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi: 10.1016/j.sbi.2016.12.004

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. New York, NY: Springer-Verlag.

Brown, W. M., Martin, S., Pollock, S. N., Coutsias, E. A., and Watson, J. P. (2008). Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.* 129:064118. doi: 10.1063/1.2968610

Ceriotti, M., Tribello, G. A., and Parrinello, M. (2011). Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13023–13029. doi: 10.1073/pnas.1108486108

Ceriotti, M., Tribello, G. A., and Parrinello, M. (2013). Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.* 9, 1521–1532. doi: 10.1021/ct3010563

Chen, W., and Ferguson, A. L. (2018). Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* 40, 2079–2102. doi: 10.1002/jcc.25520

Constanzi, S. (2010). Modeling g protein-coupled receptors: a concrete possibility. *Chim. Oggi.* 28, 26–31.

Das, P., Moll, M., Stamati, H., Kavraki, L. E., and Clementi, C. (2006). Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. U.S.A.* 103, 9885–9890. doi: 10.1073/pnas.0603553103

De, S., Bartok, A. P., Csanyi, G., and Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 18, 13754–13769. doi: 10.1039/C6CP00415F

Duan, M., Fan, J., Li, M., Han, L., and Huo, S. (2013). Evaluation of dimensionality-reduction methods from peptide folding-unfolding simulations. *J. Chem. Theory Comput.* 9, 2490–2497. doi: 10.1021/ct400052y

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756–764. doi: 10.1016/j.sbi.2008.10.002

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208. doi: 10.1038/nrm1589

Frishman, D., and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Prot. Struct. Funct. Bioinform.* 23, 566–579.

Garcia, A. E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68, 2696–2699.

Gasparotto, P., and Ceriotti, M. (2014). Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* 141:174110. doi: 10.1063/1.4900655

Gasparotto, P., MeiÅŸner, R. H., and Ceriotti, M. (2018). Recognizing local and global structural motifs at the atomic scale. *J. Chem. Theory Comput.* 14, 486–498. doi: 10.1021/acs.jctc.7b00993

Goldfeld, D. A., Zhu, K., Beuming, T., and Friesner, R. A. (2011). Successful prediction of the intra- and extracellular loops of four g-protein-coupled receptors. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8275–8280. doi: 10.1073/pnas.1016951108

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). Gromacs 4: algorithms for highly efficient, load-balanced and scalable molecular simulation. *J. Chem. Theory Comput.* 4, 435–447. doi: 10.1021/ct700301q

Jolliffe, I. (2002). *Principal Component Analysis*. New York, NY: Springer-Verlag.

Kmiecik, S., Jamroz, M., and Kolinski, M. (2015). Structure prediction of the second extracellular loop in g-protein-coupled receptors. *Biophys. J.* 106, 2408–2416. doi: 10.1016/j.bpj.2014.04.022

Konrad, H. (2006). Comment on: "energy landscape of a small peptide revealed by dihedral angle principal component analysis." *Prot. Struct. Funct. Bioinform.* 64, 795–797. doi: 10.1002/prot.20900

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., et al. (2010). Improved side-chain torsion potentials for the amber ff99sb protein force field. *Prot. Struct. Funct. Bioinformat.* 78, 1950–1958. doi: 10.1002/prot.22711

Musil, F., De, S., Yang, J., Campbell, J. E., Day, G. M., and Ceriotti, M. (2018). Machine learning for the structure-energy-property landscapes of molecular crystals. *Chem. Sci.* 9, 1289–1300. doi: 10.1039/C7SC04665K

Noé, F., and Clementi, C. (2015). Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* 11, 5002–5011. doi: 10.1021/acs.jctc.5b00553

Noé, F., and Clementi, C. (2017). Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* 43, 141–147. doi: 10.1016/j.sbi.2017.02.006

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* 12, 2825–2830. Available online at: http://www.jmlr.org/papers/v12/pedregosa11a.html

Plaku, E., Stamati, H., Clementi, C., and Kavraki, L. E. (2007). Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Prot. Struct. Funct. Bioinform.* 67, 897–907. doi: 10.1002/prot.21337

Rosman, G., Bronstein, M. M., Bronstein, A. M., and Kimmel, R. (2010). Nonlinear dimensionality reduction by topologically constrained isometric embedding. *Int. J. Comput. Vision* 89, 56–58. doi: 10.1007/s11263-010-0322-1

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computat.* 10, 1299–1319.

Schölkopf, B., Smola, A., and Muller, K.-R. (1999). "Kernel principal component analysis," in *Advances in Kernel Methods-Support Vector Learning* eds B. Schölkopf, C. J. C. Burges, and A. J. Smola (Cambridge, MA: MIT Press), 327–352.

Spiwok, V., and Kralova, B. (2011). Metadynamics in the conformational space nonlinearly dimensionally reduced by isomap. *J. Chem. Phys.* 135:224504. doi: 10.1063/1.3660208

Spiwok, V., Lipovová, P., and Králová, B. (2007). Metadynamics in essential coordinates: free energy simulation of conformational changes. *J. Phys. Chem. B* 111, 3073–3076. doi: 10.1021/jp068587c

Stamati, H., Clementi, C., and Kavraki, L. E. (2010). Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Prot. Struct. Funct. Bioinform.* 78, 223–235. doi: 10.1002/prot.22526

Sultan, M. M., and Pande, V. S. (2017). tica-metadynamics: accelerating metadynamics by using kinetically selected collective variables. *J. Chem. Theory Comput.* 13, 2440–2447. doi: 10.1021/acs.jctc.7b00182

Sultan, M. M., Wayment-Steele, H. K., and Pande, V. S. (2018). Transferable neural networks for enhanced sampling of protein dynamics. *J. Chem. Theory Comput.* 4, 1887–1894. doi: 10.1021/acs.jctc.8b00025

Sutto, L., Dâbramo, M., and Gervasio, F. L. (2010). Comparing the efficiency of biased and unbiased molecular dynamics in reconstructing the free energy landscape of met-enkephalin. *J. Chem.. Theory Comput.* 6, 3640–3646. doi: 10.1021/ct100413b

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323. doi: 10.1126/science.290.5500.2319

Tiwary, P., and Berne, B. J. (2016). Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2839–2844. doi: 10.1073/pnas.1600917113

Tribello, G. A., Ceriotti, M., and Parrinello, M. (2012). Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5196–5201. doi: 10.1073/pnas.1201152109

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: http://www.jmlr.org/papers/v9/vandermaaten08a.html

Willatt, M. J., Musil, F., and Ceriotti, M. (2018). Atom-density representations for machine learning. *J. Chem. Phys.* 150:154110. doi: 10.1063/1.5090481

Yuguang, M., H., Nguyen, P. H., and Gerhard, S. (2005). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Prot. Struct. Funct. Bioinform.* 58, 45–52. doi: 10.1002/prot.20310

Zhuravlev, P. I., Materese, C. K., and Papoian, G. A. (2009). Deconstructing the native state: energy landscapes, function and dynamics of globular proteins. *J. Phys. Chem. B* 113, 8800–8812. doi: 10.1021/jp810659u

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 24 May 2019
doi: 10.3389/fmolb.2019.00036

Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times

Daria B. Kokh[1]*, *Tom Kaufmann*[1,2], *Bastian Kister*[1,2] and *Rebecca C. Wade*[1,3,4,5]*

[1] Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany, [2] Department of Biosciences, Heidelberg University, Heidelberg, Germany, [3] Zentrum für Molekulare Biologie der Universität Heidelberg, DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg, Germany, [4] Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany, [5] Department of Physics, Heidelberg University, Heidelberg, Germany

Drug-target residence times can impact drug efficacy and safety, and are therefore increasingly being considered during lead optimization. For this purpose, computational methods to predict residence times, τ, for drug-like compounds and to derive structure-kinetic relationships are desirable. A challenge for approaches based on molecular dynamics (MD) simulation is the fact that drug residence times are typically orders of magnitude longer than computationally feasible simulation times. Therefore, enhanced sampling methods are required. We recently reported one such approach: the τRAMD procedure for estimating relative residence times by performing a large number of random acceleration MD (RAMD) simulations in which ligand dissociation occurs in times of about a nanosecond due to the application of an additional randomly oriented force to the ligand. The length of the RAMD simulations is used to deduce τ. The RAMD simulations also provide information on ligand egress pathways and dissociation mechanisms. Here, we describe a machine learning approach to systematically analyze protein-ligand binding contacts in the RAMD trajectories in order to derive regression models for estimating τ and to decipher the molecular features leading to longer τ values. We demonstrate that the regression models built on the protein-ligand interaction fingerprints of the dissociation trajectories result in robust estimates of τ for a set of 94 drug-like inhibitors of heat shock protein 90 (HSP90), even for the compounds for which the length of the RAMD trajectories does not provide a good estimation of τ. Thus, we find that machine learning helps to overcome inaccuracies in the modeling of protein-ligand complexes due to incomplete sampling or force field deficiencies. Moreover, the approach facilitates the identification of features important for residence time. In particular, we observed that interactions of the ligand with the sidechain of F138, which is located on the border between the ATP binding pocket and a hydrophobic transient sub-pocket, play a key role in slowing compound dissociation. We expect that the combination of the τRAMD simulation procedure with machine learning analysis will be generally applicable as an aid to target-based lead optimization.

Keywords: drug-protein residence time, machine learning, drug-target binding kinetics, structure-kinetic relationships (SKRs), heat shock protein 90 (HSP90), molecular dynamics simulation, tauRAMD

OPEN ACCESS

Edited by:
Vojtech Spiwok, University of Chemistry and Technology in Prague, Czechia

Reviewed by:
Natallia Kulik, Institute of Microbiology (ASCR), Czechia
Gareth Aneurin Tribello, Queen's University Belfast, United Kingdom

***Correspondence:**
Daria B. Kokh
daria.kokh@h-its.org
Rebecca C. Wade
rebecca.wade@h-its.org

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *27 March 2019*
Accepted: *02 May 2019*
Published: *24 May 2019*

Citation:
Kokh DB, Kaufmann T, Kister B and Wade RC (2019) Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times. Front. Mol. Biosci. 6:36. doi: 10.3389/fmolb.2019.00036

INTRODUCTION

The binding affinity of small compounds to their target is commonly used as a selection criterion in drug design pipelines, both for the early screening of chemical libraries and for the subsequent lead optimization. Recent studies have, however, shown that drug efficacy often correlates better with the residence time than with the binding affinity of drugs (Copeland et al., 2006; Schuetz et al., 2017). These observations suggest that the optimization of the kinetic properties of drug candidates at an early stage of the drug design process would be advantageous.

The computation of drug-target binding kinetics by using MD simulations is more challenging than the computation of binding affinity (Romanowska et al., 2015). A major problem in using conventional MD simulations for computing binding kinetic parameters is the need to sample the intermediate transition states between the bound and unbound states, which is not required for the calculation of binding affinity. This poses tremendous challenges for brute-force conventional MD sampling, whose application is so far limited to computation of the binding kinetics of small molecules to small proteins, e.g., benzamidine to trypsin, which still requires extensive millisecond simulations (Dror et al., 2011; Wu et al., 2016). Reconstruction of a single dissociation event for a pharmacologically relevant compound, which typically occurs on the time-scale of minutes or hours, is currently not feasible from conventional MD simulations. To overcome this limitation, a range of enhanced sampling techniques has been explored recently (Bruce et al., 2018). Some of them are aimed at the reduction of the configurational space to be sampled for the computation of binding kinetic rates, e.g., metadynamics (Tiwary et al., 2015, 2017), weighted ensemble methods (Dickson and Lotz, 2016; Dixon et al., 2018), or milestoning (Tang and Chang, 2017) [a detailed review can be found elsewhere (Mollica et al., 2016; Dickson et al., 2017)]. Although these methods are designed for the prediction of the absolute values of binding and unbinding rates within a reasonable computation time, they are still very computationally demanding and require high user expertise, which impedes the implementation of these methods in drug design pipelines. Furthermore, in addition to the limitations arising from the selection of the sub-space to be sampled, intrinsic limitations of the underlying physical model of molecular interactions, such as the force field and the water model, may affect the accuracy of the computed rates.

While absolute values are difficult to attain, it has been demonstrated recently that the relative values of unbinding rates for a series of ligands of a particular target are more robust to these limitations (Marques et al., 2019). In line with this finding, computationally efficient approaches that provide estimates of the relative residence times for a set of compounds have been reported. Instead of deriving the residence time from the energetic profile of dissociation paths, these techniques allow estimation of relative τ values from the times required for ligand egress during enhanced sampling simulations. The residence times obtained can then be scaled for direct comparison with experimental data. One example of this approach is scaled MD (Mollica et al., 2015; Schuetz et al., 2018a) in which the potential of the system is rescaled during simulations. Another approach, recently developed in our group, is the τRAMD method (Kokh et al., 2018), which employs multiple short random acceleration MD, RAMD, simulations to generate ligand dissociation trajectories. Relative drug-protein residence times are estimated from the times required for the ligand to leave the binding pocket in simulations started from the structures of protein-ligand complexes. In RAMD (Lüdemann et al., 2000), an additional randomly oriented force is applied to the ligand's center of mass and its direction is altered during the simulations, depending on the motion of the ligand. RAMD was originally developed to explore ligand egress routes from protein binding sites [see e.g., (Winn et al., 2002; Schleinkofer et al., 2005)], where simulated trajectories were employed to explore ligand unbinding pathways and mechanisms. In the τRAMD procedure, many trajectories are generated (usually more than 40 for each compound) and each trajectory contains hundreds of thousands of snapshots that may contain important information for the ligand unbinding rate. The value of extracting molecular features from MD simulations as fingerprints for building machine learning (ML) models to predict molecular properties has been demonstrated in Re. (Riniker, 2017). Here, we explore whether fingerprint-based ML techniques can aid the detection of features important for drug-target residence time in RAMD trajectories and, furthermore, improve the robustness of the estimated residence times.

ML has been applied for drug-target τ prediction in several studies. Qu et al. (2016) derived quantitative structure-kinetics relationships (QSKRs) for a set of HIV-1 protease inhibitors by using Volsurf descriptors. Chiu and Xie (2016) went beyond a static model by accounting for flexibility with a coarse-grained normal mode analysis to classify HIV-1 protease inhibitors in binding kinetics classes using a multi-target ML approach. Comparative Binding Energy (COMBINE) analysis (Ortiz et al., 1995; Perez et al., 1998), in which PLS (Partial Linear Regression Projection to Latent Structures) is used to reweight components of the bound protein-ligand interaction energies to predict binding properties, has recently been applied to datasets of HSP90 and HIV-1 protease inhibitors (Ganotra and Wade, 2018) and was found to give models with good predictive ability for residence time. It should be noted that the COMBINE analysis method was originally developed for the prediction of binding affinity for congeneric series of compounds. While compounds with a common scaffold are required for good prediction of the equilibrium dissociation constant, K_D, a good prediction of the off-rate could be obtained for a dataset of diverse compounds from analysis of the bound protein-ligand complexes (Ganotra and Wade, 2018) suggesting that differences in the unbound state are less important for off-rate than for binding affinity. Huang et al. (2019) applied PLS analysis to interaction-energy fingerprints extracted from snapshots of steered MD ligand dissociation trajectories to obtain a predictive model for residence time for a set of HIV-1 protease inhibitors and found that important interactions for determining τ were in the first half of the dissociation processes. This is consistent with a previous steered MD study of HIV-1 protease inhibitor dissociation in which the strength of the ligand-protein hydrogen

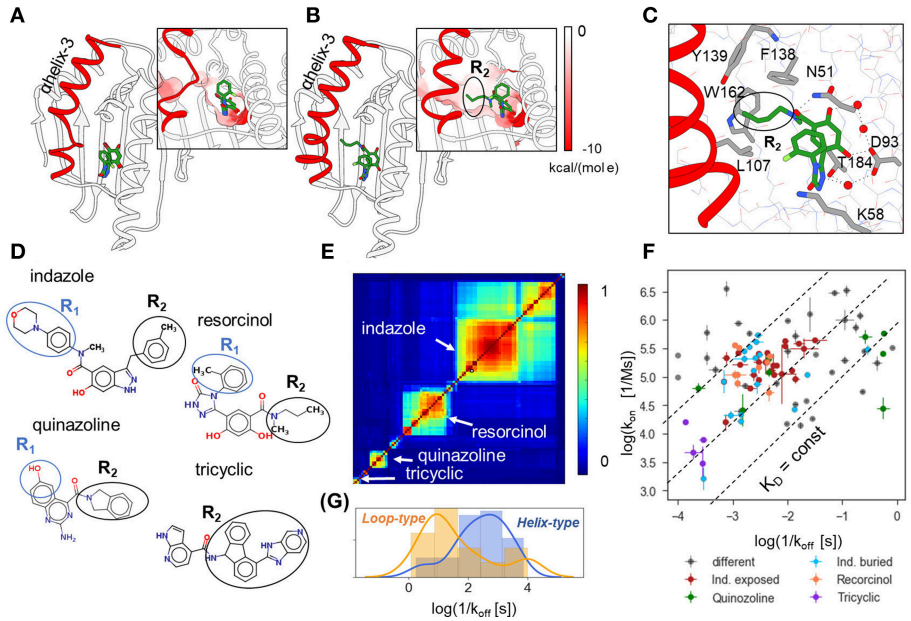


FIGURE 1 | Structural and binding kinetic properties of the dataset of 94 N-HSP90 inhibitors. **(A,B)** Two conformations of the ATP binding site in N-HSP90 with a bound ligand shown in stick representation with coloring by atom type; α-helix3 (highlighted in red) can be distorted in the middle (loop-type conformation **(A)**, compound **5** PDB ID 5J2X) or complete (helix-type conformation **(B)**, compound **13**, PDB ID 5J9X) (Amaral et al., 2017); the molecular surface of the binding pocket colored by the Coulomb potential is shown in insets for both conformations: the ATP binding site has predominantly negative charge (red), whereas the transient sub-pocket under α-helix3 is mostly hydrophobic. **(C)** Protein-ligand contacts for helix-binding compounds are illustrated for compound **13,** (PDB ID 5J9X): the ligand-protein binding network consisting of D93, T184, and three water molecules (red spheres) is common to all compounds; compounds bound to the helix-conformation of the binding site also interact with F138 and may interact with residues in the hydrophobic pocket, such as W162 and Y139. **(D)** 2D representation showing the four main groups of compounds discussed in the text. **(E)** Similarity matrix of the 90 N-HSP90 inhibitors generated using Maestro [(Schrödinger, 2019); see text]. **(F)** Distribution of the experimental binding rate constants of the entire set of compounds. The three largest groups of compounds are colored as denoted in the legend: "Ind. exposed"—indazole-based compounds with different R_1 fragments, "Ind. buried"—indazole compounds with different R_2 fragments, compounds with resorcinol and quinazoline scaffolds, as well as bulky compounds with a tricyclic fragment and different ATP-pocket binding core. **(G)** Distribution of residence times of the helix-binding and loop-binding compounds.

bond network of the bound state was found to be crucial for the dissociation process (Li et al., 2011), as well as with the above-mentioned models based solely on analysis of the bound state.

In the present study, we use our previously published τRAMD simulation results for a data set of 70 inhibitors of the cancer target HSP90 for which off-rates were measured by surface plasmon resonance (SPR) (Amaral et al., 2017; Kokh et al., 2018). These compounds bind in the ATP binding site of the N-terminal domain of human HSP90 (N-HSP90α, residues 9-236; NP_005339). The τRAMD procedure gave predictions of relative residence times with an accuracy of about 2.3τ for 78% of the compounds and <2.0τ within congeneric series. It was found that the computed residence times were sensitive to the quality of the underlying MD simulations of the protein-ligand complexes. For some compounds, deficiencies in the force field or inaccuracies in the docking pose led to notable underestimation of the residence time, although within a series of compounds with the same binding scaffold and small fragment substitutions, the ranking of the residence time was well-reproduced. The latter result suggests that the inaccuracy of the simulations of the bound state may be overcome in τRAMD simulations if the transition state is the main determinant of the variation in residence time within a congeneric series of compounds.

Here, we have performed τRAMD simulations for an additional 25 HSP90 inhibitors, whose binding kinetics were recently reported (Schuetz et al., 2018b). We have then combined these simulations with our previous simulations (Kokh et al., 2018), and applied ML approaches to the combined dataset of simulated trajectories for 94 HSP90 inhibitors.

N-HSP90 is a challenging target for the prediction of binding kinetics, as it has a flexible ATP binding site lined by the unstable α-helix3 that can adopt either "helical" or "loop" conformations (see **Figures 1A,B**), depending on the ligand bound. The "helical" conformation contains an additional hydrophobic sub-pocket adjacent to the ATP binding site, which provides space for substitutions on 'helix-binders' (fragment R_2, see **Figures 1C,D**), while this fragment is absent in the compounds bound to the "loop" conformation ('loop-binders'). It has been recently demonstrated that the binding kinetics of resorcinol inhibitors of HSP90 is related to the protein binding site conformation in the bound complex, and that the R_2 substitution can effectively stabilize α-helix3 and result in lower binding and unbinding rates for ligands with such fragments (Amaral et al., 2017). In particular, ligands with large R_2 substitutions, such as tricyclic

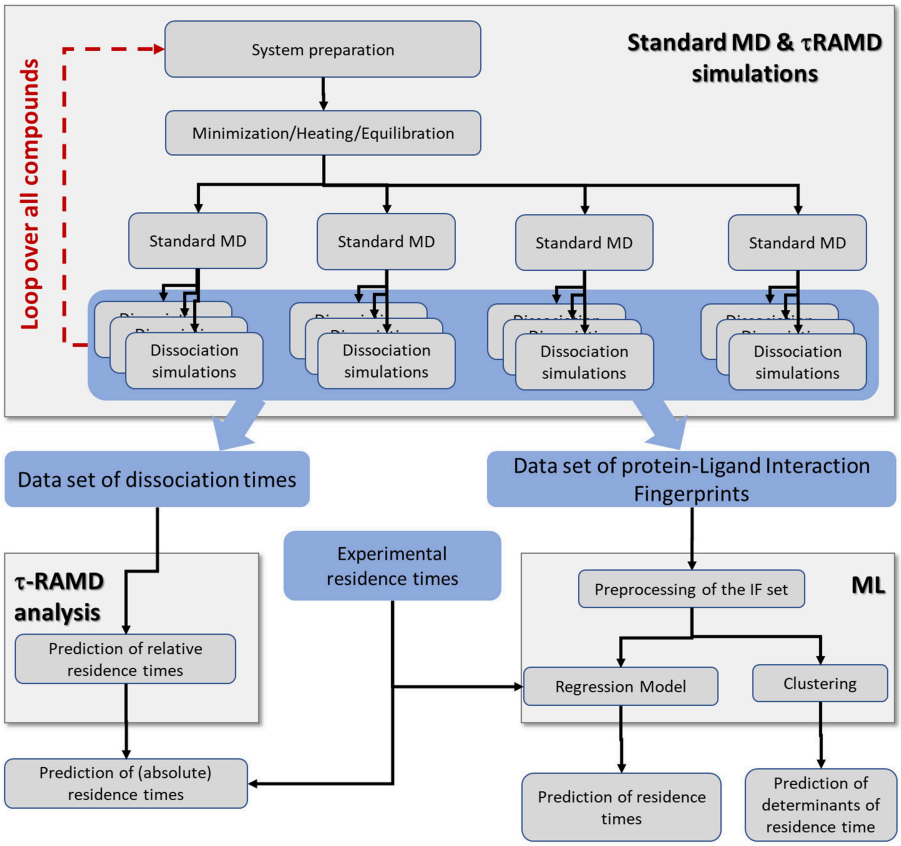


FIGURE 2 | Workflow incorporating the simulation protocol for τRAMD simulations and the ML analysis. The τRAMD simulations provide (i) computed relative residence times, and (ii) trajectories that are used for analysis of protein-ligand contacts and building a ML regression model for prediction of residence times and determining the factors governing residence time (see section Methods and Materials); data sets generated and elements of simulation workflow are highlighted by blue and gray background, respectively.

compounds (**Figure 1D**), generally have the slowest binding and unbinding kinetics (**Figure 1F**).

The set of 94 compounds considered in the present study contains molecules with 11 different scaffolds: resorcinol (28), hydroxyindazole (47), benzamide (3), aminoquinazoline (8), aminopyrrolopyrimidine (2), 7-azaindole (2), aminothienopyridine (1), imidazopyridine (1), 6-hydroxyindole (1), and adenine (1) (with the number of compounds given in brackets; see **Supplementary Tables 1** and **2**; SMILES of all studied compounds are given in **Supplementary Table 4**). The scaffold occupies the ATP binding pocket and binds to D93 as illustrated in **Figure 1C** for an indazole-based compound. The three most populated scaffolds are shown in **Figure 1D**, along with an example of compounds with different binding scaffolds but a common tricyclic group, which will be discussed below. Further, the resorcinol compounds with triazole and 2-methylbenzyl solvent-exposed groups and different buried fragments, illustrated in **Figure 1D**, build a sub-group of 8 compounds. Following Schuetz et al. (2018b), one can also distinguish two sub-groups of indazole compounds: (i) indazole-exposed: 24 compounds with a 3-methylbenzyl R_2 moiety in the hydrophobic sub-pocket and different exposed R_1 fragments, and (ii) indazole-buried: 17 compounds with an exposed 4-(4-morpholinyl) phenyl R_1 fragment and different buried R_2 fragments (see **Figure 1D**). The rest of the compounds is quite diverse, as can be seen from the 2D similarity plot generated using Maestro software (Schrödinger, 2019) by hierarchical clustering of compounds based on their 2D fingerprint similarity in **Figure 1E**. There are both loop- and helix-binders of different scaffolds, though the sub-set of loop-binders is much smaller (only 13) than the helix-binders.

The experimental binding kinetics data for the full compound set (Amaral et al., 2017; Kokh et al., 2018; Schuetz et al., 2018b) are plotted in **Figure 1F**. Both off-rates ($k_{off} = 1/\tau$) and on rates (k_{on}) vary by several orders of magnitude and there is no clear correlation between them, indicating that both the height of the transition barrier and the free energy of the bound state vary across the compound set. Notably, the helix-binders generally have longer residence times than the loop-binding compounds (**Figure 1G**).

Here, we built ML models based on the τRAMD dissociation trajectories for this data set aimed at: (i) investigating whether residence time can be deduced from the protein-ligand contact

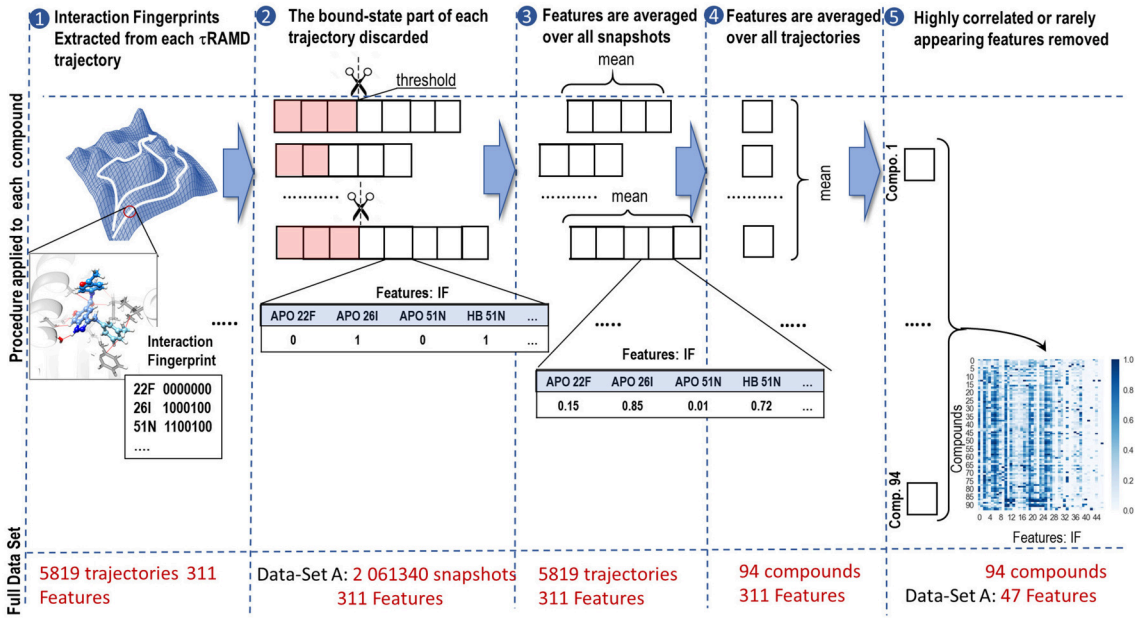


FIGURE 3 | Workflow illustrating the generation of features from simulated τRAMD trajectories: (1) Extraction of interaction fingerprints as features for ML; (2) Discarding of the bound state part of the trajectory (highlighted in pink). The discarded part of the trajectory depends on the threshold used, resulting in data-sets A, B, and C (see text for details); (3) Averaging of the features over all snapshots in each trajectory; (4) Averaging of the features over all trajectories for each compound; (5) Removal of features that appear rarely or are strongly correlated with other features. The size of each set of data at each stage of the workflow is indicated in red.

occurrence in τRAMD ligand dissociation trajectories, in particular for the cases where the relative residence times derived from the lengths of τRAMD trajectories are consistently underestimated; and (ii) identifying molecular properties that affect ligand residence time and that can be used to guide the design of ligands with altered binding kinetics.

METHODS AND MATERIALS

An overview of the simulation workflow is given in **Figure 2**. For each compound, the τRAMD procedure was performed, which consists of the preparation of the solvated protein-ligand complex, the equilibration of the system using multiple replicas of standard MD simulation, and then the simulation of multiple RAMD ligand dissociation trajectories. The τRAMD relative residence times are obtained using the protocol reported by Kokh et al. (2018). In the second part of the workflow, the protein-ligand contacts (referred to hereafter as interaction fingerprints, IFs) are extracted from τRAMD dissociation trajectories. Then, for all compounds, the IFs are transformed into a set of features for the ML analysis, which includes the clustering of the ligand dissociation properties and the building of regression models for residence time based on available experimental binding kinetics data (see the next section). The workflow is described in detail in the following sections.

Kinetic and Structural Data for the Dataset of HSP90 Inhibitors

We employed 69 of the 70 compounds with structural and kinetic data in Kokh et al. (2018). One compound [**70** in Kokh et al. (2018)] was eliminated from the dataset because its complex with N-HSP90 was structurally unstable during MD equilibration. For two compounds with affinities and long residence times beyond the measurement range (PDB ID 2VCI and 5NYI, compounds **1** and **4**, see **Supplementary Tables 1**, **2**), we used the lower limit values of $k_{off} = 10^{-4}$ s^{-1} and $K_D = 10^{-9}$ M^{-1}. Additionally, we studied 25 compounds from Schuetz et al. (2018b). Since there are no crystal structures of protein-ligand complexes available for these 25 compounds yet, the ligands were modeled in the N-HSP90 binding site using (MOE., 2017) on the basis of similarity to available crystal structures for similar compounds: PDB ID 5OCI and 6EFU for the indazole compounds, and PDB ID 5J86 for the resorcinol compounds.

MD and RAMD Simulations

The τRAMD protocol as described by Kokh et al. (2018) was followed. Here, we outline this protocol briefly for completeness. First, the starting structure of each protein-ligand complex was protonated at pH 7. The ligand was protonated using MOE (MOE., 2017) and the protein was protonated using PDB2PQR (Unni et al., 2011). The atomic partial charges of the ligands were assigned using the RESP approach (Bayly et al., 1993) with the molecular electrostatic potential computed using *ab initio* quantum mechanical calculations performed at the HF level with a 6-31G*(1d) basis set using the Gamess software (Gordon and Schmidt, 2005). The protonated protein-ligand complex was solvated in a periodic box of TIP3P water molecules and Na^+ and Cl^- ions at an ionic strength of about 150 mM. Crystallographic water molecules were retained. The system was energy minimized, gradually heated and shortly equilibrated

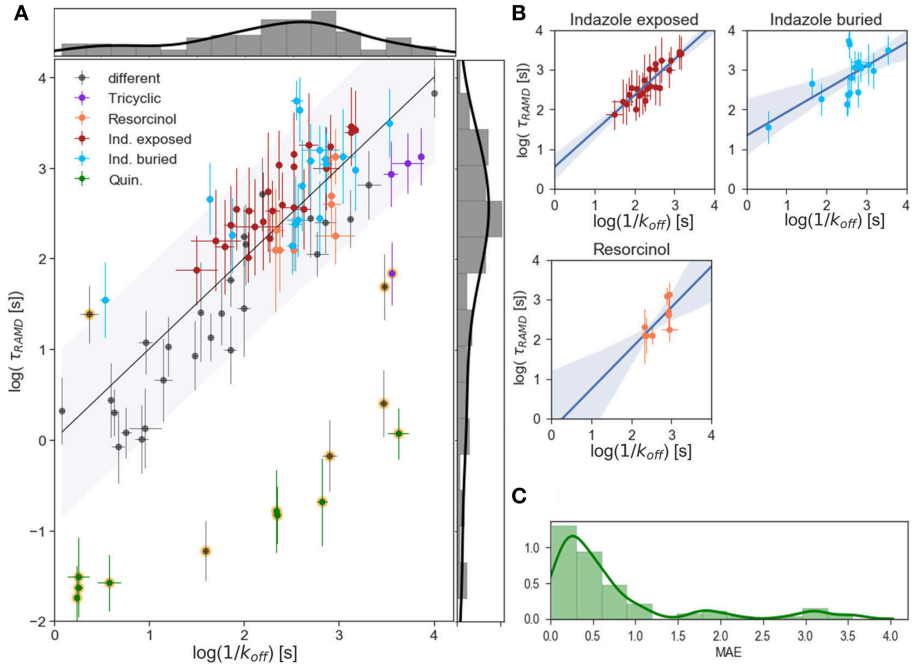


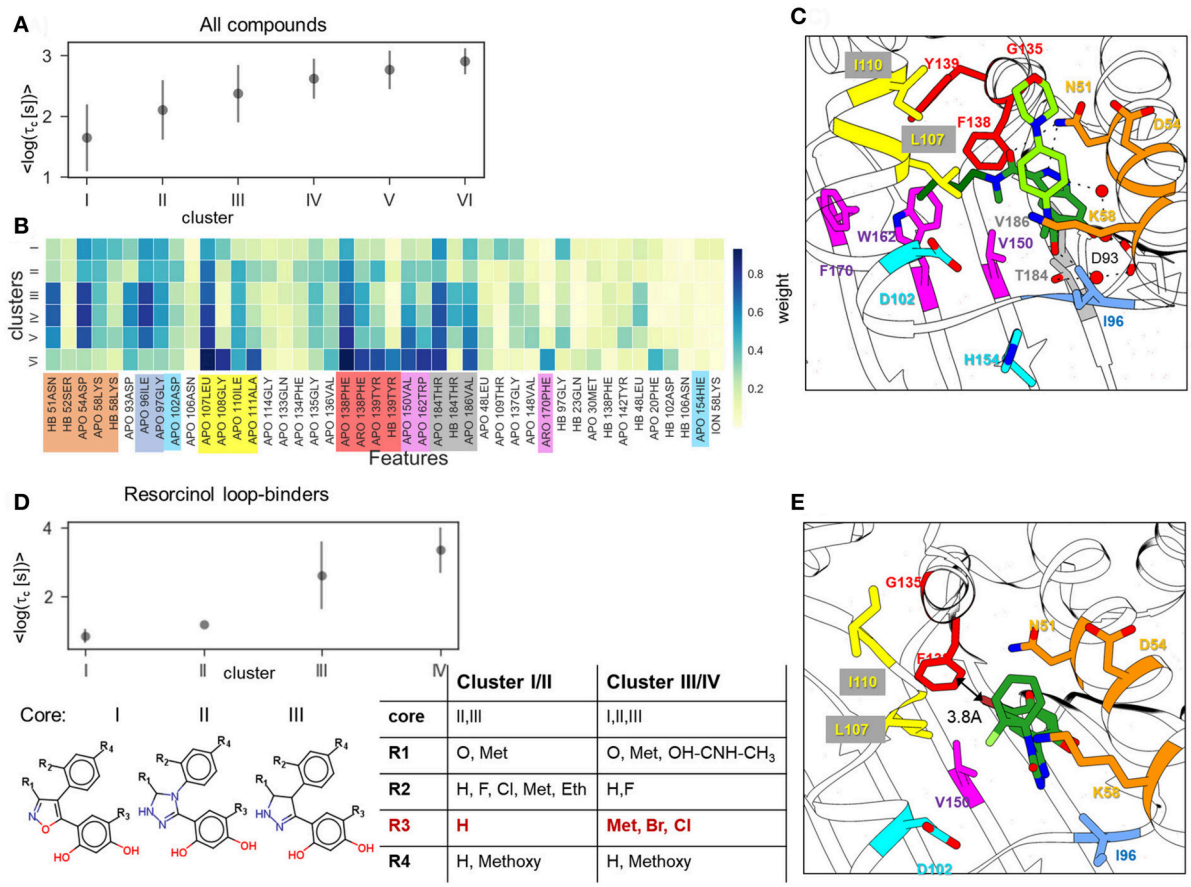
FIGURE 4 | Results of τRAMD simulations. **(A)** Scaled τRAMD residence times plotted vs. measured log ($1/k_{off}$) values on a logarithmic scale for the complete set of compounds. The τRAMD residence times are scaled according to the linear fitting ($\log(\tau_{RAMD}) = 0.39*\log(1/k_{off}) - 0.52$) of all compounds except for the 13 outliers identified in Kokh et al. (2018) (shown with background yellow circles). Two groups of indazoles (with different exposed R_1 and buried R_2 fragments, see **Figure 1D**), resorcinol and quinazoline compounds, as well as compounds with a tricyclic fragment, are colored as denoted in the legend; other compounds are shown in gray. The black line shows the one-to-one correspondence of the computed and experimental data and the interval within 1.5-fold of the mean of the residuals (0.9 log([s])) is shown by the gray area. The distributions of measured and τRAMD computed residence times are shown along the X and Y axes, respectively. **(B)** Linear fitting with 95% confidence interval for several sub-groups of compounds highlighted in **(A)**: indazole exposed, indazole buried, and resorcinol colored as in **(A)**. **(C)** Histogram showing the distribution of the mean absolute error, MAE, of τRAMD residence times relative to measured values; the long tail arises from the outliers.

with gradually decreasing restraints on all non-hydrogen atoms of the protein, ligand, and crystallographic water molecules using the AMBER molecular dynamics simulation software (Case et al., 2016). Simulations were run under NPT conditions (Langevin thermostat and barostat). Then the coordinates of the preliminary equilibrated binding complex were transferred to the NAMD program (Phillips et al., 2005) and used as the input for heating and equilibrating the system. The coordinates and velocities obtained after 30–40 ns of equilibration were used to initiate simulations of ligand dissociation using the RAMD method with a randomly oriented force on the ligand with a constant magnitude of 14 $kcalmol^{-1}Å^{-1}$. Every 100 fs, the orientation of the force was randomly re-initialized if the center of mass of the ligand had moved <0.025Å. The simulations were stopped when the center of mass of the ligand had moved 30 Å from the original bound position.

At least four MD equilibration replicas were prepared and from each replica 10–20 RAMD dissociation trajectories were generated. The relative residence time was defined as the time when a dissociation event was observed in 50% of the trajectories. It was computed for each starting replica and then averaged over all replicas simulated. Sufficient sampling to compute residence time was ensured by increasing the number of equilibration replicas and/or the number of dissociation trajectories if necessary as discussed in Kokh (2018).

Feature Generation

The feature generation procedure is illustrated in **Figure 3**. First, a set of interaction fingerprints (IF) was obtained from the τRAMD dissociation trajectories (40–100 trajectories for each compound) using the following protocol: (i) the position of the center of mass of the ligand and the coordinates of the protein and the ligand atoms were extracted from each trajectory frame and stored using a tcl script for the VMD program (Humphrey et al., 1996) (snapshots illustrating egress routes and residues contacting the ligand during dissociation are visualized in **Supplementary Figure 1**); (ii) the coordinates extracted in (i) were used to generate interaction fingerprints for each frame using an OpenEye's OEChem Toolkit (OpenEye., 2018) as 7-bit strings encoding hydrophobic, aromatic face-to-face and edge-to-face, H-bond donor/acceptor and cationic/anionic interaction types (Marcou and Rognan, 2007; Mysinger et al., 2012). Then the interaction fingerprints were grouped into four categories of protein-ligand contacts: hydrogen-bond (HB), aromatic (ARO), ionic (IP), and apolar (APO) interactions, and each category was assigned a value of 1 or 0 according to whether the contact type was, respectively, present or not; (iii) finally, the bound-state part of the trajectory was removed and only the part of the trajectory covering the transition of the ligand from the bound to the unbound state was used for further analysis (step 2 in **Figure 3**). Since the threshold for the separation between the bound- and



	Cluster I/II	Cluster III/IV
core	II,III	I,II,III
R1	O, Met	O, Met, OH-CNH-CH_3
R2	H, F, Cl, Met, Eth	H,F
R3	H	Met, Br, Cl
R4	H, Methoxy	H, Methoxy

FIGURE 5 | Result of clustering analysis based on the IFs of the ligand dissociation trajectories. **(A,B)** Clustering of the complete data set of 94 compounds: **(A)** mean and standard deviation of log residence times in each cluster obtained in 50 clustering runs; **(B)** weights of IFs for each cluster. HB, ION, ARO, and APO mean hydrogen bond (donor or acceptor), ionic, aromatic, and apolar interactions, respectively; **(C,E)** Position of indazole compound bound to the helix-type conformation of the binding pocket (PDB ID:5LNZ), and **(C)** of resorcinol compound bound to the loop-type conformation (PDB ID: 5J2X) **(E)**; residues that contribute to the protein-ligand contacts along the ligand dissociation trajectories are shown in stick representation and colored by protein region consistently with **(B)**. **(D)** Clustering of the resorcinol loop-binders (see compound list in **Supplementary Table 2**) showing mean and standard deviation of the log residence time in each cluster (above) and cluster composition (below).

transition parts can be defined arbitrarily, we explored three possible threshold definitions (these will be referred to as data sets, A, B and C, hereafter): (A) when two IF observed in the bound state (i.e., in the first frame of a trajectory) are lost, or (B) when 20%, or (C) when 60% of the bound-state contacts are lost (the size of each data set is given in **Supplementary Table 3**).

Although the sequence of interaction events may bear important information about the ligand dissociation mechanism, preliminary tests showed that the RAMD trajectories generated did not permit us to build a reliable time-dependent model, probably due to having insufficient number of snapshots along the ligand dissociation trajectories as the artificial random force accelerated dissociation. Therefore, we eliminated time dependence in our data by computing the occurrence of each type of contact in each trajectory and averaging them over all trajectories for a particular compound (steps 3 and 4 in **Figure 3**). This provided us with a matrix of 94 labels (compounds) x 311 features (fingerprints). This matrix was further reduced by partial elimination of the noise in the data set. In particular, since we did not expect that a very rare contact would affect dissociation rate, we excluded features that were found in fewer than 5% of the frames for any compound. This reduced the number of features to 68/69/75 for the complete A/B/C data-sets, respectively. Then, we performed preliminary correlation analysis and removed one of the features from each pair that had a correlation $R^2 > 0.9$, thus further reducing the number of features to 47/48/57 for the data-sets A/B/C, respectively (see **Supplementary Table 3**).

To explore the influence of molecular properties on the residence time, we additionally generated a set of molecular features, MFs, for all compounds using MOE (MOE., 2017). The MFs include the number of bonds of different types, the number of atoms with hydrogen-bond properties, the number of heavy atoms, and the solvation energy (the complete list is given in **Supplementary Table 2**). For testing the importance of these molecular features, they were either added to the IFs of data-set A or used as a separate feature set.

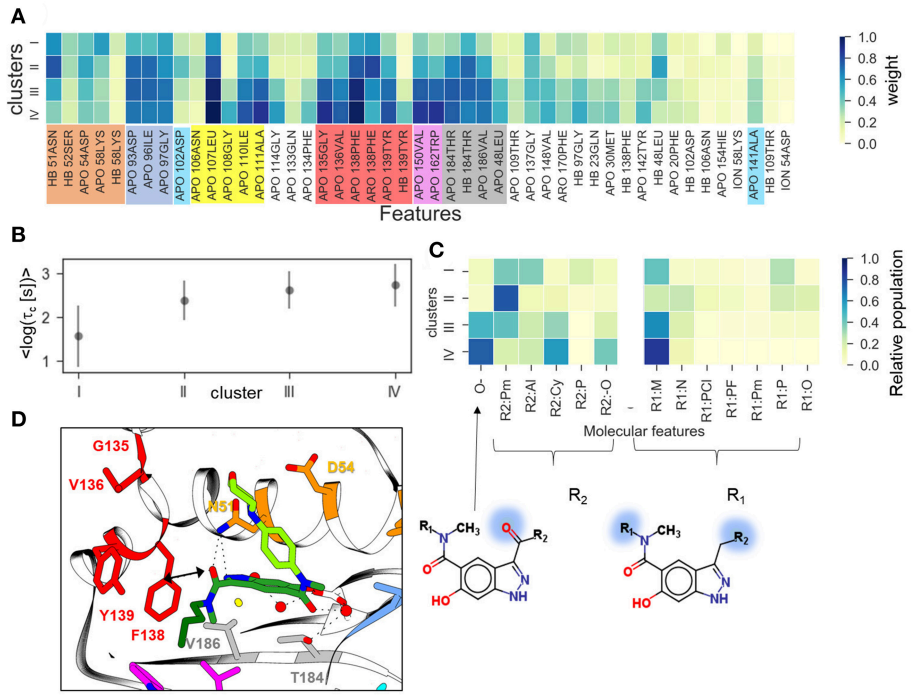


FIGURE 6 | Clustering of indazole compounds: **(A)** weights of IFs for each cluster (coloring scheme and labels as in **Figure 4**); **(B)** mean and standard deviation of log residence times of compounds in each cluster; **(C)** population of selected molecular fragments in each cluster (see **Supplementary Figure 6** for naming convention); the structures of two compounds discussed in the text are shown below (fragment substitutions are highlighted in blue); **(D)** Position of indazole compound **37** in the binding pocket, the main contact residues are shown in sticks and colored as in **(A)**.

Machine Learning Protocol

The scikit-learn Python library (Pedregosa et al., 2011) was used for all machine learning (ML) procedures.

Regression Analysis

The data sets were normalized by transforming each feature vector to the interval [0:1]. The ML models were trained and tested against measured $\log(1/k_{off})$ values. Two regression models (RM), one linear—Ridge Linear Regression with L^2 regularization terms (LR)—and one non-linear—Support Vector Regression (SVR)—were found to be more balanced and slightly more stable in cross-validation than the other methods tested (Partial Least Squares, Random Forest and Gaussian Boosting Regression). Additionally, a dummy regression model with the mean value of the training set as a null-hypothesis (referred to as Dummy Regressor hereafter) was used as a control.

The modeling workflow consisted of the following steps (as illustrated in **Supplementary Figure 2**):

(i) **Split the data set into a training (internal) set and an external test set**. For the test set, we selected 20% of compounds from the data set while ensuring that the test set contained 2 randomly selected compounds from the outlier subset of 8 quinazolines (compounds **58–65**) and six other compounds (**11**, **17**, **30**, **66**, **67**, **69**) as defined in Kokh et al. (2018); these compounds are highlighted in yellow in **Supplementary Table 2**), and 20% (i.e., at least 9 compounds) from the subset of indazole compounds (compound scaffolds are given in **Supplementary Table 2**). The rest of the test set was selected randomly from the remaining compounds. The purpose of this selection was two-fold: (1) to test the prediction accuracy for compounds that were considered as outliers in τRAMD simulations; and (2) to avoid over-representation of the indazole compounds in the training set, since they constitute almost 50% of all compounds in the data set.

(ii) **Selection of hyperparameters for the two regression models, LR and SVR** (this block is zoomed in in **Supplementary Figure 2**). The internal training set was used for the selection of hyperparameters. The following parameters were optimized: coefficient of the regularization term for the LR model; kernel coefficient (the RBF kernel was used), parameter of the loss function, and coefficient of the error term for the SVR model. We employed exhaustive grid-search with 10-fold cross-validation (using random permutation splitting with a validation test set size of 20%). The results of the optimization procedure are given in **Supplementary Data** and illustrated in **Supplementary Figures 3, 4**.

(iii) **Training and testing of the models.** After the hyperparameters were selected, 10 cross-validation runs were performed on the internal training set. In each round, two regression models, LR and SVR, were trained on a sub-set of the internal training set and then the mean absolute error, MAE, and the Q^2_{F3} metric, reported as the most reliable

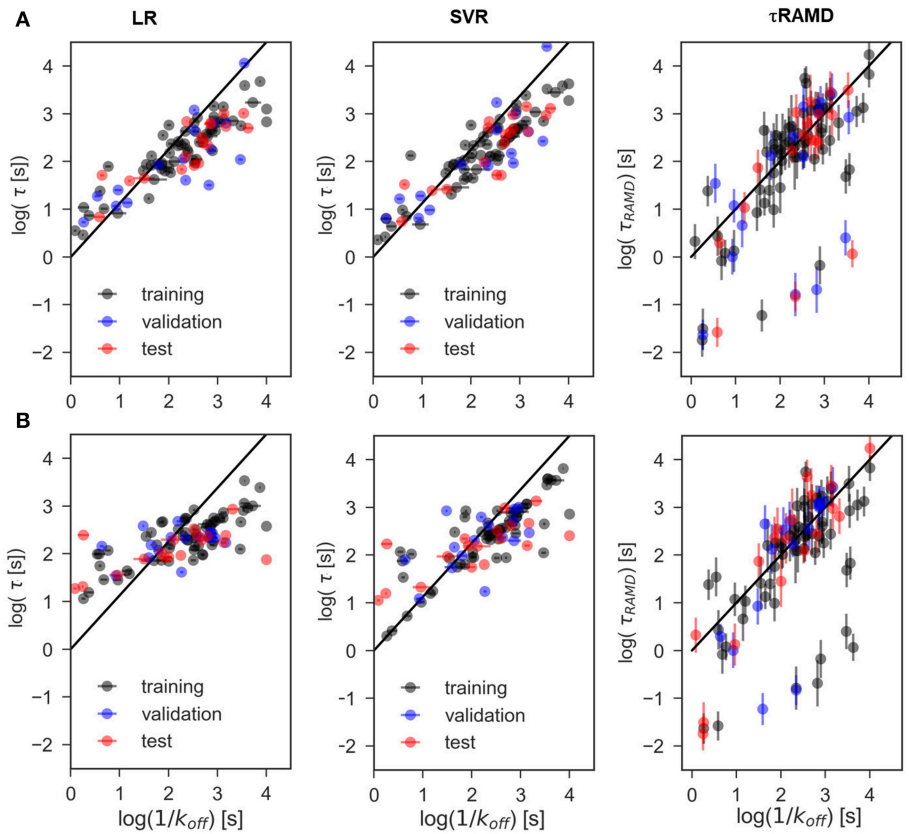


FIGURE 7 | Representative examples of computed vs. experimental residence times obtained for data-sets. **(A)** A and **(B)** C using linear (LR) and non-linear (SVR) ML models as well as from the τRAMD residence time estimation procedure. Black/blue and red points belong to the training/validation and external test sets, respectively.

metric for the evaluation of the regression models (Todeschini et al., 2016), were computed for the training and validation sub-sets (generated using random permutation splitting with a validation sub-set size of 20%), as well as for the external test set (all for the residence time on a $\log_{10}$ scale; for more details, see **Supplementary Information**). Additionally, the same data sub-sets were used to evaluate the Dummy model and the τRAMD simulations.

Then new internal training/external test set combinations were generated step (i) and the steps (ii–iii) were repeated. All MAE and Q^2_{F3} values obtained in these calculations were stored. Altogether, we performed 200 computation rounds, each with a different split of training and test sets, to gain proper statistics. The histograms of the MAE distributions obtained for each ML method were compared with those for the Dummy model for control; histograms of MAE and Q^2_{F3} were compared with the corresponding distributions obtained from the τRAMD protocol. The complete procedure for 100 rounds takes about 1.5 h on a laptop with an Intel Core i5-5200U, 2.2 GHz processor.

Clustering

We employed a Gaussian Mixture Model (GMM) for the classification of the compounds by their IFs in the data sets A for all compounds and for the sub-set of indazole-based compounds only. The feature set was normalized by transforming to the interval [0:1], as for the regression models. For the scikit-learn GMM function, we used an option where each component has its own multivariate covariance matrix. To estimate the optimal number of clusters, we used the Akaike information criterion (see **Supplementary Information** for details). Following a scan of cluster size, 6 clusters were chosen on the basis of minimum loss of information for the complete data set of 94 compounds (A) and 4 clusters for the indazole sub-set of the data set A (Ind) (see **Supplementary Figures 5A,B**). For each dataset, 50 independent repeats of clustering were performed. For each clustering round, the clusters were ordered by increasing average residence time of the inhibitors belonging to each cluster, and the weights of all features in each cluster were stored. Finally, for each dataset, the mean cluster residence time, τ_c, over the 50 clusterings was computed for each of the clusters (from their average residence times), with the first having the shortest τ_c.

Further, for the indazole subset (Ind), we explored how some selected structural properties of the compounds are distributed over the clusters. For this, we selected two sets of small fragments that might affect the dissociation rate constant (see **Supplementary Figure 6**): (i) seven types of solvent-exposed fragments (i.e., different classes of the R_1 substitution (**Figure 1D**) and six types of buried fragments (i.e., R_2, placed

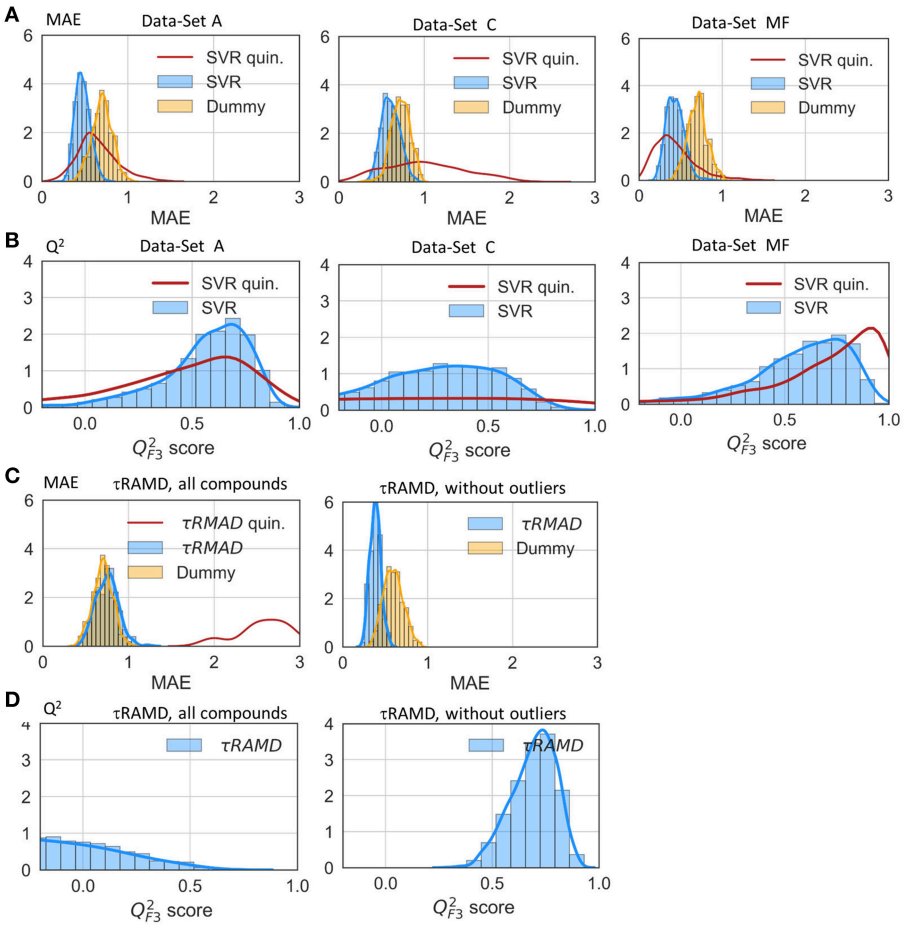


FIGURE 8 | Assessment of the RM quality. Histograms of mean absolute error, MAE **(A)** and Q^2_{F3} score **(B)** of the external test set obtained in 200 repeated test/training set splitting using RMs and the same values computed from τRAMD simulations **(C,D)** are shown in blue along with results for the Dummy model (orange); results for the sub-set of only quinazoline compounds (from the full data set A) are shown by red lines; in τRAMD simulations Q^2_{F3} values **(D)** are negative for quinazoline compounds; in the right-hand plot of panels **(C,D)** all quinazoline compounds were removed as outliers. The data-set used are denoted in each plot: A and C data-sets, MF—data-set from molecular descriptors only.

in the hydrophobic sub-pocket, see **Figure 1C**). The number of compounds in each cluster with the corresponding R_1 and R_2 fragments was computed and normalized by the cluster size.

RESULTS AND DISCUSSION

τRAMD Simulations

Computed relative residence times obtained from the τRAMD simulations for the 94 compounds are shown vs. measured $1/k_{off}$ values on the logarithmic scale in **Figure 4A**. As discussed in our previous study (Kokh et al., 2018), 14 compounds from the dataset are outliers: compounds **11**, **17**, **30**, **66**, **67**, **69**, and **8** quinazoline compounds (highlighted in yellow in **Figure 4A**). Without the outliers, i.e. for 80 compounds (85% of the data set), the correlation coefficient $R^2 = 0.75$, MAE $= 0.39 \pm 0.06$, and the mean prediction uncertainty, MPU, is 3.1 τ on average, which is somewhat higher than in the set of 70 compounds studied previously (Kokh et al., 2018) ($R^2 = 0.86$ and MPU $= 2.3\tau$ for 78% of the compounds, i.e. 55 compounds after omission of outliers).

To understand the reason for this difference, one has to look at the simulation results for the indazole compounds since most of the added compounds are indazoles. 17 out of the 25 additional compounds have an indazole scaffold with a buried 3-methylbenzyl R_2 substituent and different exposed R_1 fragments (shown in dark red in **Figures 1F**, **4A**). This group has a computed τ that is systematically longer by approximately 0.5 log units than the value from the linear fit for the other compounds, despite showing a good correlation with the experimental τ values within the group ($R^2 = 0.86$, MAE $= 0.34$, **Figure 4B**). In contrast, variation of the buried R_2 fragment in the indazoles leads to a large and non-specific deviation of computed τ values from the fit. Specifically, a series with 4-(4-Morpholinyl) phenyl substitutions in indazole compounds (group colored in cyan in **Figures 1F, 4A,B**) has a correlation coefficient with experimental data of $R^2 = 0.67$,

TABLE 1 | Results of evaluation tests for different models: mean of MAE and Q^2_{F3} score obtained from 200 rounds of simulations (the standard deviation is given in parentheses) for the external test sets.

	RM	A	B	C	A*	MF	Ind
MAE	LR	0.47(0.08)	0.51(0.09)	0.60(0.11)	0.43(0.08)	0.51(0.10)	0.39(0.10)
	SVR	0.48(0.09)	0.53(0.10)	0.60(0.11)	0.43(0.08)	0.45(0.11)	0.39(0.11)
	τRAMD	0.76(0.12)	0.39(0.06)	–	0.38(0.08)		
	Dummy	0.71(0.11)	0.61(0.11)	0.71(0.11)	0.55(0.14)		
Q^2_{F3}	LR	0.57(0.21)	0.44(0.30)	0.29(0.30)	0.54(0.23)	0.36(0.52)	0.41(0.52)
	SVR	0.56(0.22)	0.44(0.30)	0.28(0.30)	0.51(0.25)	0.52(0.30)	0.38(0.58)
	τRAMD	−0.41(0.47)	0.69(0.10)	–	0.57(0.23)		

Calculations were done for data-sets A, B, and C (see main text) are based on the complete set of 94 compounds. The test sets in these three cases were required to contain some of the outliers found by applying the τRAMD procedure to estimate relative residence times, see Methods for details. A–data-set of 80 compounds with outliers discarded. MF—based on molecular property features only. Ind—only IFs of indazole compounds from data-set A are included. For data-set A, the quinazoline compounds (8 compounds) have a mean MAE = 0.60 ± 0.2/0.61 ± 0.2 and Q^2_{F3} = 0.44 ± 0.4/0.41 ± 0.4 for LR and SVR models, respectively; for the data set MF quinazoline compounds have a mean MAE = 0.59 ± 0.21/0.43 ± 0.25 and Q^2_{F3} = 0.45 ± 0.39/0.65 ± 0.42 for LR and SVR models, respectively; for the Dummy model Q^2_{F3} = 0.*

MAE = 0.43. Similarly, a subgroup of 6 resorcinol compounds shown with different R_2 (shown in **Figure 1D**, their residence times are colored in orange in **Figures 4A,B**) substituents has a low correlation, $R^2 = 0.72$, MAE = 0.32. The mean prediction uncertainties for the latter three groups are 2.3, 4.3, and 2.2 τ, respectively.

One possible explanation for the poorer correlations for subgroups of compounds with different R_2 fragments is uncertainty regarding the structure of the bound-state of the protein-ligand complex. All 21 indazole and 6 resorcinol compounds mentioned above were modeled using a template structure since crystal structures were not available for these complexes. Some of these compounds require a relatively large substituent to be modeled in, leading to uncertainty in the protein and ligand conformations and in the position of the compound, particularly when the fragment fits tightly in the hydrophobic binding sub-pocket and adaptation of the protein structure is necessary. The 40 ns MD equilibration carried out might not be sufficient for achieving an optimal ligand-protein configuration, which may affect the computed residence time.

Another possible reason can be deduced from the observation that sets of compounds with different buried fragments R_2 demonstrate inhomogeneous deviations from the general linear fitting of the complete set, while sets of compounds with the same buried fragment show very similar deviations. This implies the systematic omission of a specific contribution to the observed residence time. In RAMD, conformational changes of the protein induced by the ligand's motions on the nanosecond timescale of the simulations are captured rather well, but the longer time scale motions of the protein are not fully sampled and these can be expected to modulate the ligand dissociation times. For example, if backbone changes, such as the unfolding of a helix, are needed for ligand egress, then this is likely to be captured to a lesser extent than side chain rotations in RAMD simulations. Such long-time motions may facilitate ligand dissociation, and therefore poor sampling of these motions may result in the overestimation of residence times with the τRAMD procedure.

Elucidation of the Molecular Features Affecting Residence Time From Simulated Ligand Dissociation Trajectories

As discussed above, the relative τ value is obtained in the τRAMD procedure from the computed ligand dissociation times that are assumed to be longer for the slower dissociating compounds and shorter for the faster dissociating ones. By building a feature set of protein-ligand IFs from the ligand dissociation trajectories, we deliberately omitted information on the trajectory length (see section Methods and Materials). Instead, we assessed whether the pattern of protein-ligand contacts in the ligand dissociation trajectories contains information on the ligand dissociation mechanism and whether it can be used to deduce how ligand substituents affect residence time prolongation.

To explore this, we employed the largest data-set, A, for clustering of all 94 compounds by the similarity of their IF features. We found that the optimal number of clusters was 6 (see Methods and Materials for details). Although in some clusters, the distributions of residence times are quite wide, there is a clear difference in their mean residence times, so that the clusters can be ranked by their mean τ value, τ_c (see **Figure 5A**). The average cluster properties obtained from 50 repeated clusterings mainly reflect the general structural similarity of compounds. The composition of the clusters and their order is mostly preserved in all 50 clustering rounds: the cluster with the longest average residence time comprises compounds with a tricyclic fragment, whereas the two clusters with the shortest average residence times consist mainly of loop-binders and fast unbinding compounds, such as quinazolines; in the two intermediate clusters, one contains indazoles and one contains resorcinols. From the IF weights in each cluster (**Figure 5B**), one can see that most of the contacts associated with large τ_c values arise from residues lining the hydrophobic sub-pocket formed due to α-helix3 stabilization: specifically, residues that belong to α-helix3 (L107- A111, marked in yellow in **Figure 5B**), those located in the hydrophobic sub-pocket at its entrance (F138, Y139, V150, W162, F170, shown in red and magenta in **Figure 5B**), and two residues at the bottom

of the ATP binding pocket (V186 and T184, highlighted in gray). These residues are shown in **Figure 5C** in the same color as in **Figure 5B**. It is noteworthy, that the weights of several residues located at the entrance of the hydrophobic sub-pocket, specifically F138, V150, and L107, gradually increase with the residence time. This result agrees with the conclusion of our previous study that steric hinderance at the egress channel for compounds partially located in the hydrophobic sub-pocket is an important factor in increasing the transition state energy and thus prolonging the residence time (Kokh et al., 2018). The interaction with exposed residues lining the entrance to the ATP binding pocket (polar residues N51, D54) has a large contribution for the clusters III-V with intermediate residence times. However, they do not show a notable correlation with the residence time in this cluster splitting.

Overall, the splitting of the 94 compounds into just six clusters reveals several very general tendencies, showing that the interactions of the compound fragment located in the hydrophobic sub-pocket generally promote slower dissociation, while the interactions with exposed residues lining the entrance to the ATP binding pocket may affect the residence time, but without showing a systematic trend. Increasing the number of clusters leads to a general reduction of the residence time diversity in each cluster (see **Supplementary Figure 5C**), which suggests that the similarity of the IFs in dissociation trajectories does generally correlate with the residence time. However, to obtain a more detailed understanding of dissociation mechanisms, one has to consider clustering of specific compound sub-sets. For example, clustering of the 11 resorcinol-based loop-binders from cluster I effectively separates the faster dissociating compounds from the slower dissociating compounds (**Figure 5D**). Interestingly, although the cluster composition varies during repeated clustering, the main difference between the slower dissociating compounds (clusters III and IV) and the faster dissociating ones (cluster I and II) is retained: either a halogen (Cl or Br) or an aliphatic fragment (for example, a methyl group) on the resorcinol group (fragment R3 in **Figure 5D**) is always associated with longer residence time. All other fragments (R1, R2, and R4) appear in both groups with short and long residence times (clusters I/II and III/IV, respectively). We therefore surmise that the interaction with F138 (in particular from the Cl atom) is one of the important factors for prolongation of the residence time even though this interaction is not clearly established in the bound state (see structure shown in **Figure 5E**).

Furthermore, we have performed clustering on the largest subset of compounds available (indazole compounds bound to the helix-type conformation). The averaged weights of different types of IFs that distinguish the four clusters are shown in **Figure 6A**. The mean residence time variation over the clusters (**Figure 6B**) shows that there is a significant gap between the fastest dissociating compounds in cluster I and the slower dissociating ones in clusters II-IV. As we observed for the complete set of compounds, the slowest dissociating clusters are characterized by a large contribution of the IF from residues lining the hydrophobic sub-pocket located at α-helix3 (L107, G108, I110, A111) or at the entrance of or inside the hydrophobic pocket (F138, V150, T139, W184). Additionally, residues G135 and V136, located between the entrance to the hydrophobic sub-pocket and the ATP binding pocket, contribute (**Figure 6D**). These residues may interact with the solvent-exposed part, R_1, of the ligand, a 4-(40morpholinyl) phenyl fragment (see **Figure 1D**). To obtain a more detailed understanding of these protein-ligand interactions, we selected several molecular fragments that predominantly define structural variance in the indazole set (see **Supplementary Figure 6**) and computed the average occurrence of these fragments in each cluster (**Figure 6C**). It can be seen that all compounds with a carbonyl oxygen at the R_2 fragment (located between N51 and F138 in the bound complex, see **Figure 6D**), belong to the long-residence time clusters III and IV. On the other hand, although N51 can form an H-bond with the carbonyl oxygen, this interaction does not have a large contribution to the slowest unbinding clusters (see **Figure 6A**). The results suggest that the carbonyl oxygen plays a similar role to the halogen atom in the loop-binders discussed above, and forms transient interactions with F138. Also, all compounds with alicyclic (and methoxy) groups in the hydrophobic binding pocket (indicated in **Figure 6C** as R2:Cy and R2:O, respectively) appear in the clusters with the longest residence times. Consistently, the hydrogen bonding (HB) interaction with the buried Y139 appears only in the slowest dissociating cluster and can be associated with a polar (carboxyl) group at the R_2 fragment. Finally, the effect of the exposed R_1 fragment on the residence time is less well-defined than the buried R_2 fragment (apart from a large contribution of the 4-(40morpholinyl) phenyl fragment, R1:M, which is present in about half of the indazole compounds).

Regression Models for the Prediction of Residence Time

The results of two regression models, Linear Regression with a regularization term (LR) and Support Vector Regression (SVR), to different data-sets are shown in **Figures 7**, **8**, and the computed model quality metrics are given in **Table 1**. In particular, **Figure 7** shows representative plots of computed against experimental residence times for the data-sets A and C. The linear and non-linear regression methods provide very similar results. Moreover, the predictions of the two methods were strongly correlated (similar under- or over-estimation of the residence times), which indicates that the data set quality, not the complexity of the RM chosen, poses the main limitation on the accuracy. Consistently, the MAE distributions for both methods obtained from 200 different test sub-sets are similar, as shown in **Figure 8**. The mean MAE value for the test sets are about 0.47 ± 0.08 for both RMs, while the Dummy model yields 0.71 ± 0.11 (see **Table 1**; the MAE histogram for the training and validation sets are shown in **Supplementary Figure 7**). The predictions have a $Q^2_{F3} = 0.57/0.56 \pm 0.2$ for LR and SVR RMs, respectively, which indicates that the model quality is acceptable, albeit with a relatively large standard deviation. Note, that in this model we included all compounds, even those that were considered as outliers in τRAMD simulations in Kokh et al. (2018) and each test set was required to contain at least 2 quinazoline compounds,

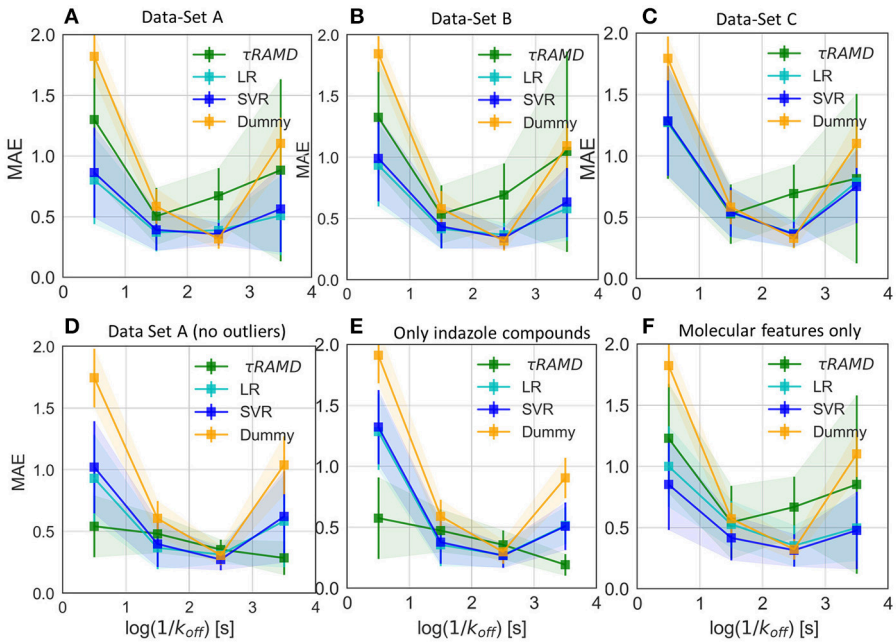


FIGURE 9 | Average value of MAE for the sub-set of compounds with experimental residence times in the ranges of (<1s), (1s-2s), (2s-3s), and (>3s) as obtained in 100 simulations for different test sets and compared with the Dummy-model (null-hypothesis) and τRAMD for the same set of compounds. **(A–C)** For the complete set of compounds in models **(A–C)**, respectively; **(D)** For the data-set model A* (model A without outliers); **(E)** Only a sub-set of indazole compounds from the data-set A was used; **(F)** Only molecular features were used.

whose τ is strongly underestimated in τRAMD simulations, as can be seen in **Figure 4A**. Therefore, the τ estimated directly from the τRAMD simulations has a large mean MAE of 0.76 ± 0.12 (the MAE distribution is shown in **Figure 4C**).

To gain deeper insight into the determinants of the quality of the RMs, we split the τ interval into four regions and plotted the mean of the MAE distributions for each region (**Figure 9**). Both RMs have almost identical results and they clearly outperform τRAMD for all four intervals used if all the compounds are considered (**Figure 9A**). However, the τRAMD method yields better prediction accuracy than the RMs for the shortest and longest residence time intervals if the 14 outliers (highlighted in **Figure 4A**) are not included in the compound set (**Figure 9D**, data-set A without outliers), with a mean of MAE $= 0.39 \pm 0.06$ and $Q^2_{F3} = 0.69 \pm 0.10$, see **Table 1**. On the other hand, the quality of the RMs is only slightly changed on removal of the outliers, see **Table 1**. This is likely due to the much larger number of ligands with intermediate τ values than those with short or long τ, as can be seen from the histogram in **Figure 4A**, which ensures better training of RMs in the middle of the interval but difficulties in the prediction of more extreme values.

To further assess the ability of the RMs to correctly predict the residence times of the compounds that appear as outliers in τRAMD simulations, we computed the MAE distribution for a test subset consisting of quinazoline compounds only, which yielded a mean value of MAE $= 0.60 \pm 0.2$ (MAE distribution from the model dataset A is shown by a red line in **Figure 8**) and a mean $Q^2_{F3} = 0.44 \pm 0.4$. This result is worse than for the whole set of compounds, probably because of the small number of quinazoline compounds in the training set: 6, and in the external test set, 2. Nonetheless, the estimation of τ from RMs is much better for these compounds than that obtained from τRAMD simulations of the residence time based on the trajectory length, which results in underestimation of τ by several orders of magnitude. This is an important result suggesting that the residence time can be reasonably well-predicted by RMs trained on diverse compounds whereas τRAMD simulations cannot always be used to rank τ computed for compounds with different scaffolds. In Kokh et al. (2018), it was hypothesized that the main reason for the underestimation of the residence time of the quinazoline compounds in τRAMD simulations was the deficiency of the bound state representation in MD simulations. Following this hypothesis, one may assume that the robustness of ML models for such compounds is a consequence of the data preprocessing, where the major part of the trajectory in which the bound-state is sampled is discarded (i.e., the main bound-state IFs are still considered but the exact length of the bound-state trajectory is not retained).

To explore the importance of the bound state IFs for RMs, we applied the same protocol using trajectories starting from snapshots where 20% and 60% of the bound-state contacts were lost (model data-sets B and C, respectively), which corresponds to loss of 2–3 and 5–16 contacts, depending on the compound size. Data-set B yielded only slightly worse prediction accuracy than data-set A, whereas the predictive ability for data-set C was notably worse and closer to the null hypothesis (see **Figures 7–9** and **Table 1**), especially for compounds with short residence times, **Figure 9C**. The Q^2_{F3} score of the RMs drops from 0.57 to

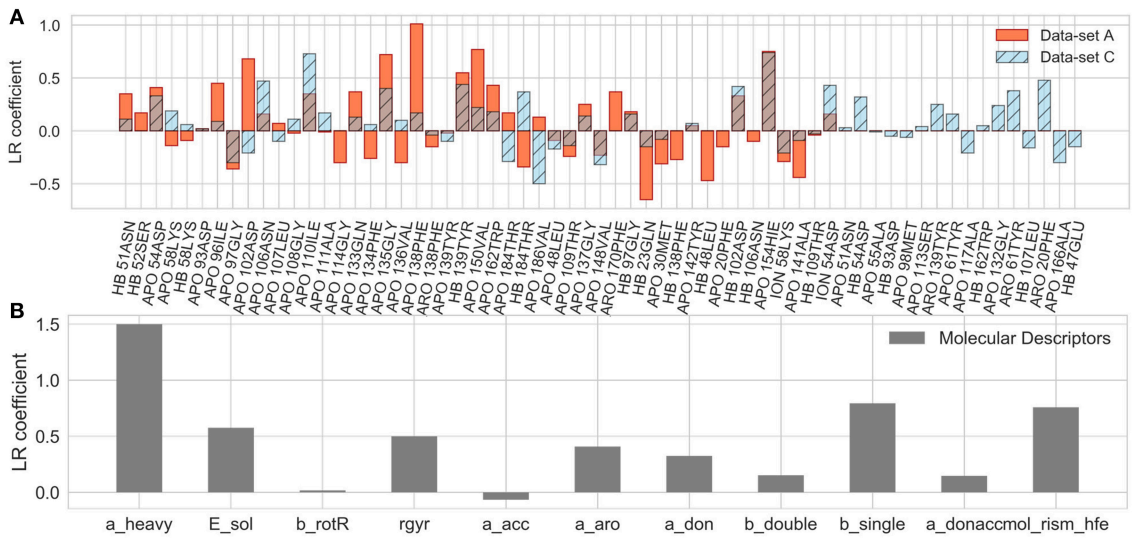


FIGURE 10 | Coefficients of the LR model in the test set averaged over 200 different splitting of the training and external test sets for the A and C data-sets **(A)** and for the LR built on molecular descriptors only **(B)**, as denoted in each plot.

0.44 and then to 0.29 for the data sets A, B, and C, respectively, with SD values increasing, indicating a strong dependence of the model performance on the test subset selected.

The coefficients of the IFs in the LR model on the data-set A and C are compared in **Figure 10A**. The features that have major contributions are quite similar for the data-sets A and B (data for the set B are not shown). The largest contribution comes from several residues lining the binding pocket and located at the entrance of the hydrophobic sub-pocket (F138, V150, G135), which is generally consistent with the clustering analysis given above. Additionally, several more distant residues, such as D102 and H154, appear to be important for the LR model. It is noteworthy that in both the clustering analysis and LR, the interaction with F138 plays a major role and correlates with longer residence times. For the data-set C, however, the hydrophobic sub-pocket residues do not contribute essentially. Instead, the role of polar residues around the pocket entrance (D54, N106, K58) and more distant residues, such as I110 and T61, or even F20 (located at the exit of the hydrophobic sub-pocket) increases. These results suggest that: (i) the presence of the bound state IFs in the feature set is crucial for the quality of RMs for prediction of residence times, although the RMs do not seem to be very sensitive to the exact duration of the bound state, (ii) dissociation pathways may be very diverse, which makes it difficult to build a consistent model from transition state information only.

Notably, the residues that make the main contributions to the LR and to the clustering models in the present study are quite similar to those reported for COMBINE analysis of HSP90 inhibitors (Ganotra and Wade, 2018). They include residues of the part of the α-helix3 fragment that lines the ATP binding pocket (L107-A111), as well as some polar residues surrounding the ATP binding site (N51, D54, D93, G97, D102), and several residues inside the hydrophobic sub-pocket (Y139 and T184). This agreement supports the main trend in the dissociation kinetics of the HSP90 inhibitors studied, namely that large compounds that bind in the hydrophobic sub-pocket formed by αhelix3 are generally slower dissociators. The importance of the interaction of the ligand with F138 was not highlighted by the COMBINE analysis, likely because this residue does not always directly interact with the ligand in the bound state. On the other hand, some polar residues, such as K58, N51, and D54, seem to have less importance when the complete dissociation trajectory is considered. For example, although a H-bond between some ligands and K58 is observed in the crystal structures, it is quite unstable in MD simulations and its contribution is negligible to both the LR and the clustering models.

RMs built for the congeneric series of 45 indazole compounds (data-set Ind) demonstrate similar performance for the mid- and long-range residence times to those for the complete data set (**Table 1**, **Figure 9E** and **Supplementary Figure 8**). For the region with $k_{off} > 0.01\ s^{-1}$, however, the model quality is poor because only 3 indazole compounds belong to this region.

Finally, we considered whether the model could be improved by the inclusion of parameters describing the molecular features of the ligands or even by training the model solely on ligand parameters. Thus, we added several molecular descriptors, such as solvation energy, number of heavy atoms, single, double and aromatic bonds, hydrogen donors and acceptors, and radius of gyration (see **Supplementary Table 2**) to the set of IF features. Although the RMs were not notably improved (data not shown), the number of heavy atoms appeared as a major term in the LR model. We therefore went further and trained RMs on molecular descriptors alone. Surprisingly, the SVR model based on just molecular descriptors demonstrated a good performance ($Q^2_{F3} = 0.52 \pm 0.30$), comparable to

that for data-set A, albeit with a larger SD, and better than the LR model ($Q^2_{F3} = 0.36 \pm 0.52$) on the same dataset (see also MAE and Q^2_{F3} histograms in **Figure 9F**). The latter is mostly driven by the number of the heavy atoms in the molecule (**Figure 10B**), which is an expected result since there is a clear correlation between the residence time and the number of heavy atoms ($R^2 = 0.74$, **Supplementary Figure 9A**). The number of single bonds and solvation energy are the next most important factors, where the dependence on the solvation energy is mostly driven by the compounds with different buried fragments, in particular, indazole compounds (**Supplementary Figure 9C**) while variation of the exposed fragment does not have much effect (the correlation of solvation energy with $\log(1/k_{off})$ for different sub-sets is shown in **Supplementary Figures 9B–F**).

CONCLUSIONS

In the present study, we propose a protocol for estimating drug-target residence times and for exploring which protein-ligand interactions affect the residence time. We performed a machine learning analysis of ligand dissociation trajectories obtained from τRAMD simulations. For the evaluation of the method, we analyzed the ligand dissociation trajectories of 94 inhibitors of HSP90 [previously published for 69 compounds (Kokh et al., 2018) and simulated for an additional 25 compounds from Schuetz et al. (2018b)]. We excluded from the analysis the first part of each simulated trajectory where the majority of protein-ligand interactions were retained as in the starting complex structure. We considered three different thresholds for defining the minimum number of protein-ligand contacts that must be lost to assign a snapshot to the transition part of the trajectory: (i) 2 contacts, (ii) 20%, and (iii) 60% of all bound-state contacts (data-sets A, B, and C, respectively). A collection of protein-ligand interaction fingerprints, IFs, extracted from the transition part of each dissociation trajectory as defined above, was employed to build a set of features for machine learning analysis.

We first explored the possibility to obtain insights into key protein-ligand contacts and to reveal ligand fragments that influence the ligand residence time using a clustering algorithm and the data-set A. Then, we built regression models, RMs, for the prediction of ligand dissociation rates using experimental data. We tested different data models, as well as a data sub-set containing indazole compounds only, and a set of molecular descriptors. We systematically compared the predictive performance of the RMs with the null-hypothesis, as well as with the results of the τRAMD method, where relative residence times were estimated based on the lengths of the dissociation trajectories for each compound. We found that RMs have good predictive ability for residence times, even for compounds where the τRAMD method fails because of deficiencies in the modeling of the ligand-protein bound state due to force field or sampling issues.

Comparison of the three data-sets, with different definitions of the transition part of the trajectory, shows that the residence time strongly depends on the interaction of the ligand with residues of the binding cavity, when most of the bound state protein-ligand contacts are still preserved. This is in accord with the recent calculations of relative residence times for HIV-1 protease inhibitors (Huang et al., 2019) and HIV-1 protease and HSP90 inhibitors (Ganotra and Wade, 2018), which demonstrated that protein-ligand contacts in the complex could be used to deduce ligand residence times. From the linear regression model, as well as from clustering analysis, we found out that the interaction of the ligand with F138 is very important. Although F138 is not always directly contacting the ligands in their bound states, it forms transient interactions with aromatic groups as well as with polar groups of the binding core (either halogen or carbonyl oxygen) present in most of the compounds, and thereby promotes prolongation of the ligand residence time.

As expected, the quality of the ML models strongly depends on the range and the homogeneity of the distribution of kinetic rate constants for the compounds studied, and the size of the set of compounds with similar scaffolds but different substitutions. In particular, the quality of the present models is strongly affected by the fact that about 50% of the compounds have intermediate residence times, while there are much fewer compounds with short or long values of τ.

Finally, we demonstrated that the LR model based only on the molecular features of the compounds reproduced the general trend in τ reasonably well. It showed an increase of τ with molecular size, but was less reliable for the prediction of the dissociation rates of compounds with short τ values, for which the determinants of the dissociation kinetics are more complex. On the other hand, the SVR model trained on the molecular features shows surprisingly good performance (similar to that obtained when the model was trained on the complete set of IFs), albeit with a larger variation in the performance for different sub-sets of compounds.

Overall, this study demonstrates that the proposed machine learning procedures can effectively extend the value of the τRAMD procedure by making corrections for outliers, improving the predictive ability for ligand residence time, and giving information on key determinants of the ligand dissociation mechanism and the ligand functional groups that are critical for residence time prolongation.

DATA AVAILABILITY

The IF data set and Python Jupyter Notebook scripts and a data set are accessible at https://zenodo.org/record/2652166#.XMMg6YWxU5k (doi: 10.5281/zenodo.2652166), the raw data are available from the authors upon request, without undue reservation, to any qualified researcher. Kinetic data can be found in Kokh et al. (2018), Schuetz et al. (2018b), and Amaral et al. (2017) along with the Protein Databank identifiers of crystal structures of protein-ligand complexes. The 2D structures of the compounds used in the study are given in SMILES format in the Microsoft Excel **Supplementary Table 4** as a separate file.

AUTHOR CONTRIBUTIONS

DK and RW conceived and designed the study. DK and BK carried out the MD simulations. DK, TK, and BK performed the machine learning analysis. DK wrote the first draft of the manuscript. All authors contributed to manuscript revision, and read and approved the submitted version.

FUNDING

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2). We also acknowledge financial support by the EU/EFPIA Innovative Medicines Initiative (IMI) Joint Undertaking, K4DD (grant no. 115366), the Klaus Tschira Foundation, and Deutsche Forschungsgemeinschaft within the funding programme Open Access Publishing, by the Baden-Württemberg Ministry of Science, Research and the Arts and by Ruprecht-Karls-Universität Heidelberg.

ACKNOWLEDGMENTS

We thank Albert J. Kooistra and Chris de Graaf for the help in developing of the fingerprint analysis procedure and Lorenzo Fabbri for testing scripts and revising documentation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00036/full#supplementary-material

REFERENCES

Amaral, M., Kokh, D. B., Bomke, J., Wegener, A., Buchstaller, H. P., Eggenweiler, H. M., et al. (2017). Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Commun.* 8:2276. doi: 10.1038/s41467-017-02258-w

Bayly, C. C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* 97, 10269–10280. doi: 10.1021/j100142a004

Bruce, N. J., Ganotra, G. K., Kokh, D. B., Sadiq, S. K., and Wade, R. C. (2018). New approaches for computing ligand–receptor binding kinetics. *Curr. Opin. Struct. Biol.* 49, 1–10. doi: 10.1016/j.sbi.2017.10.001

Case, D. A., Betz, R. M., Botello-Smith, W., Cerutti, D. S., Cheatham, T. E. III, Darden, T. A., et al. (2016). *AMBER 2016 Reference Manual*.

Chiu, S. H., and Xie, L. (2016). Toward high-throughput predictive modeling of protein binding/unbinding kinetics. *J. Chem. Inf. Model.* 56, 1164–1174. doi: 10.1021/acs.jcim.5b00632

Copeland, R. A., Pompliano, D. L., and Meek, T. D. (2006). Drug-target residence time and its implications for lead optimization. *Nat. Rev. Drug Discov.* 5, 730–739. doi: 10.1038/nrd2082

Dickson, A., and Lotz, S. D. (2016). Ligand Release Pathways Obtained with WExplore: Residence Times and Mechanisms. *J. Phys. Chem. B* 120, 5377–5385. doi: 10.1021/acs.jpcb.6b04012

Dickson, A., Tiwary, P., and Vashisth, H. (2017). Kinetics of ligand binding through advanced computational approaches: a review. *Curr. Top. Med. Chem.* 17, 2626–2641. doi: 10.2174/1568026617666170414142908

Dixon, T., Dickson, A., and Lotz, S. D. (2018). Predicting ligand binding affinity for the SAMPL6 challenge from on- and off-rates using weighted ensembles of trajectories. *J. Comput. Aided. Mol. Des.* 32, 1001–1012. doi: 10.1007/s10822-018-0149-3

Dror, R. O., Pan, A. C., Arlow, D. H., Borhani, D. W., Maragakis, P., Shan, Y., et al. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc Natl Acad Sci U.S.A.* 108, 13118–13123. doi: 10.1073/pnas.1104614108/r1104614108

Ganotra, G. K., and Wade, R. C. (2018). Prediction of drug-target binding kinetics by comparative binding energy analysis. *ACS Med. Chem. Lett.* 9, 1134–1139. doi: 10.1021/acsmedchemlett.8b00397

Gordon, M. S., and Schmidt, M. W. (2005). "Chapter 41: Advances in electronic structure theory: GAMESS a decade later," in *Theory and Applications of Computational Chemistry: The First Forty Years*, eds C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Amsterdam: Elsevier), 1167–1189. doi: 10.1016/B978-044451719-7/50084-6

Huang, S., Zhang, D., Mei, H., Kevin, M., Qu, S., Pan, X., et al. (2019). SMD-based interaction-energy fingerprints can predict accurately the dissociation rate constants of HIV-1 protease inhibitors. *J. Chem. Inf. Model.* 59, 159–169. doi: 10.1021/acs.jcim.8b00567

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *J. Mol. Graph.* 14, 33–38. doi: 10.1016/0263-7855(96)00018-5

Kokh, D., Amaral, M., Bomke, J., Graedler, U., Musil, D., Buchstaller, H., et al. (2018). Estimation of drug-target residence times by τ–random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.* 14, 3859–3869. doi: 10.1021/acs.jctc.8b00230

Kokh, D. B. (2018). TauRAMD. Available online at: https://www.h-its.org/downloads/ramd/

Li, D., Ji, B., Hwang, K.-C., and Huang, Y. (2011). Strength of hydrogen bond network takes crucial roles in the dissociation process of inhibitors from the HIV-1 protease binding pocket. *PLoS ONE* 6:e19268. doi: 10.1371/journal.pone.0019268

Lüdemann, S. K., Lounnas, V., and Wade, R. C. (2000). How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *J. Mol. Biol.* 303, 813–30. doi: 10.1006/jmbi.2000.4155

Marcou, G., and Rognan, D. (2007). Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* 47, 195–207. doi: 10.1021/ci600342e

Marques, S. M., Bednar, D., and Damborsky, J. (2019). Computational study of protein-ligand unbinding for enzyme engineering. *Front. Chem.* 6, 1–15. doi: 10.3389/fchem.2018.00650

MOE. (2017). *Molecular Operating Environment (MOE), 2013.08*. Chemical Computing Group Inc.

Mollica, L., Decherchi, S., Zia, S. R., Gaspari, R., Cavalli, A., and Rocchia, W. (2015). Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Sci. Rep.* 5, 11539. doi: 10.1038/srep11539

Mollica, L., Theret, I., Antoine, M., Perron-Sierra, F., Charton, Y., Fourquez, J. M., et al. (2016). Molecular dynamics simulations and kinetic measurements to estimate and predict protein-ligand residence times. *J. Med. Chem.* 59, 7167–7176. doi: 10.1021/acs.jmedchem.6b00632

Mysinger, M. M., Weiss, D. R., Ziarek, J. J., Gravel, S., Doak, A. K., and Karpiak, J. (2012). Structure-based ligand discovery for the protein – protein interface of chemokine receptor CXCR4. *PNAS* 109, 5517–5522. doi: 10.1073/pnas.1120431109

OpenEye. (2018). *OEChem Toolkit 2018.Oct.1 OpenEye Scientific Software*. Santa Fe, NM. Available online at: http://www.eyesopen.com

Ortiz, A. R., Pisabarro, M. T., Gago, F., and Wade, R. C. (1995). Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* 38, 2681–2691. doi: 10.1021/jm00014a020

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Perez, C., Pastor, M., Ortiz, A. R., and Gago, F. (1998). Comparative binding energy analysis of HIV-1 protease inhibitors : incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Mater. Chem.* 2623, 836–852. doi: 10.1021/jm970535b

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–802. doi: 10.1002/jcc.20289

Qu, S., Huang, S., Pan, X., Yang, L., and Mei, H. (2016). Constructing interconsistent, reasonable, and predictive models for both the kinetic and thermodynamic properties of HIV-1 protease inhibitors. *J. Chem. Inf. Model.* 56, 2061–2068. doi: 10.1021/acs.jcim.6b00326

Riniker, S. (2017). Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences. *J. Chem. Info. Model.* 57, 726–741. doi: 10.1021/acs.jcim.6b00778

Romanowska, J., Kokh, D. B., Fuller, J. C., and Wade, R. C. (2015). "Computational Approaches for Studying Drug Binding Kinetics," in *Thermodynamics and Kinetics of Drug Binding*, eds G. M. Keserü and D. C. Swinney (Weinheim: KGaA and Wiley-VCH Verlag GmbH & Co), 211–235. doi: 10.1002/9783527673025.ch11

Schleinkofer, K., Winn, P. J., Lüdemann, S. K., and Wade, R. C. (2005). Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? *EMBO Rep.* 6, 584–589. doi: 10.1038/sj.embor.7400420

Schrödinger, L. (2019). *Small-Molecule Drug Discovery Suite 2019–1*.

Schuetz, D. A., Bernetti, M., Bertazzo, M., Musil, D., Recanatini, M., Masetti, M., et al. (2018a). Predicting residence time and drug unbinding pathway through scaled molecular dynamics. *J. Chem.* 59, 535–549. doi: 10.1021/acs.jcim.8b00614

Schuetz, D. A., de Witte, W. E. A., Wong, Y. C., Knasmueller, B., Richter, L., Kokh, D. B., et al. (2017). Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discov. Today* 22, 896–911. doi: 10.1016/j.drudis.2017.02.002

Schuetz, D. A., Richter, L., Grandits, M., Graedler, U., Buchstaller, H., Eggenweiler, H., et al. (2018b). Ligand desolvation steers on-rate and impacts drug residence time of heat shock protein 90 (Hsp90) inhibitors. *J. Med. Chem.* 90, 4397–4411. doi: 10.1021/acs.jmedchem.8b00080

Tang, Z., and Chang, C. A. (2017). Energy barriers, molecular motions, and residence time in ligand dissociation: a computational study on type II inhibitors binding to CDK8/CycC. *BioRxiv* 7, 1–24. doi: 10.1101/169607

Tiwary, P., Limongelli, V., Salvalaglio, M., and Parrinello, M. (2015). Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proc. Natl. Acad. Sci. U.S.A.* 112, E386–E391. doi: 10.1073/pnas.1424461112

Tiwary, P., Mondal, J., and Berne, B. J. (2017). How and when does an anticancer drug leave its binding site ? *Sci. Adv.* 3:e170001431. doi: 10.1126/sciadv.1700014

Todeschini, R., Ballabio, D., and Grisoni, F. (2016). Beware of unreliable Q2! A comparative study of regression metrics for predictivity assessment of QSAR models. *J. Chem. Inf. Model.* 56, 1905–1913. doi: 10.1021/acs.jcim.6b00277

Unni, S., Huang, Y., Hanson, R. M., Tobias, M., Krishnan, S., Li, W. W., et al. (2011). Web servers and services for electrostatics calculations with APBS and PDB2PQR. *J. Comput. Chem.* 32, 1488–1491. doi: 10.1002/jcc.21720

Winn, P. J., Lüdemann, S. K., Gauges, R., Lounnas, V., and Wade, R. C. (2002). Comparison of the dynamics of substrate access channels in three cytochrome P450s reveals different opening mechanisms and a novel functional role for a buried arginine. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5361–6. doi: 10.1073/pnas.082522999

Wu, H., Paul, F., Wehmeyer, C., and Noé, F. (2016). Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 113, E3221–E3230. doi: 10.1073/pnas.1525092113

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership