HHS Public Access

Author manuscript

*Curr Opin Struct Biol*. Author manuscript; available in PMC 2023 March 25.

Published in final edited form as:

*Curr Opin Struct Biol*. 2023 February ; 78: 102517. doi:10.1016/j.sbi.2022.102517.

# Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods

**Anupam Banerjee**[1], **Satyaki Saha**[1], **Nathan C. Tvedt**[1,2], **Lee-Wei Yang**[3,4], **Ivet Bahar**[1]

[1]Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh PA 15261, USA

[2]Computational and Applied Mathematics and Statistics, The College of William and Mary, Williamsburg, VA 23185, USA

[3]Institute of Bioinformatics and Structural Biology, and PhD Program in Biomedical Artificial Intelligence, National Tsing Hua University, Hsinchu 300044, Taiwan

[4]Physics Division, National Center for Theoretical Sciences, Taipei 106319, Taiwan

## Abstract

Proteins sample an ensemble of conformers under physiological conditions, having access to a spectrum of modes of motions, also called intrinsic dynamics. These motions ensure the adaptation to various interactions in the cell, and largely assist in, if not determine, viable mechanisms of biological function. In recent years, machine learning frameworks have proven uniquely useful in structural biology, and recent studies further provide evidence to the utility and/or necessity of considering intrinsic dynamics for increasing their predictive ability. Efficient quantification of dynamics-based attributes by recently developed physics-based theories and models such as elastic network models provides a unique opportunity to generate data on dynamics for training ML models towards inferring mechanisms of protein function, assessing pathogenicity, or estimating binding affinities.

## Introduction

Recent years have seen a growing number of machine learning (ML)-based approaches that assist in advancing structural biology research, especially in structure prediction [1]. A prime example is the development of a neural network (NN)-based tool, AlphaFold, by DeepMind [2,3]. This artificial intelligence (AI) system has now predicted the structures of 200 million proteins listed in UniProt, a breakthrough compared to the deposition of <200,000 structures in the Protein Data Bank since its inception in 1976 [4].



Corresponding author: Bahar, Ivet (bahar@pitt.edu).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ML applications in biology were originally made in the realm of sequence comparisons; this is how the field of bioinformatics emerged [5]. The extension to 3D structures naturally arose with ML's powerful pattern recognition and computer vision algorithms that apply equally well to 3D shapes, not only a string of letters. Further development of sequence-based deep learning (DL) methods, especially those based on sequence co-evolution [6,7] or genome-scale genetic data demonstrated remarkable success in structure prediction [8]. The purpose of this opinion is to discuss how the progress made in physics-based computational evaluation of *structural dynamics,* not only structure, can be leveraged if used in conjunction with ML or DL methods.

It is now established that knowledge of a single structure provides useful but incomplete insights into the mechanisms of function. A 3D structure only provides a single snapshot from amongst a multitude of conformations that the protein can sample during its function. The bridge between structure and function is structural dynamics [9,10], in accord with the sequence → structure → dynamics → function paradigm. Structural dynamics refers to fluctuations between (microscopic) conformations accessible near the folded (macro)state. These motions range from cooperative domain movements typical of allosteric machines to highly localized fluctuations at densely packed regions. They collectively form the *intrinsically accessible* 'spectrum of modes of motions' that enable adaptation to different environments and interactions and are often recruited to accomplish function while retaining the 3D fold [10,11].

Analyses based on elastic network models (ENMs) provide an effective description of the anisotropic fluctuations of proteins in agreement with experimental data [12], and robust evidence on the functional significance of structural dynamics [10,11,13,14]. Several examples of functional motions, e.g., opening and closure for binding and stabilizing a substrate, transition between outward-facing and inward-facing states of the transporters, allosteric structural changes in response to ligand binding, or simply breathing motions enabling the channeling of ions or small molecules through pores or cavities, are predictable by ENMs as highly probable modes of motions uniquely defined by the inter-residue contact topology. Such structure-encoded *intrinsic dynamics* pre-exists, independent of substrate/ligand binding. Ligand binding simply exploits these pre-existing mechanisms of reconfiguration to enable function.

Given the importance of protein dynamics, it is not surprising that ML studies that leverage structural dynamics data increasingly gain traction. The merger of ML and structural dynamics theory and methods is expected to be mutually beneficial (Fig. 1): dynamics data incorporated into ML algorithms can increase the accuracy of functional inferences, as exemplified by a recent pathogenicity predictor [15]; conversely, DL methods put to use for analyzing full-atomic simulations can help extract information on proteins' kinetics [16]. In both cases, the confluence of ML/DL and structural dynamics data is expected to generate knowledge that will help close the gap between experimental and computed quantities. We point below to such recent developments and their prospective utilities.

## Structural dynamics is a determinant of mutation pathogenicity

Single amino acid variants (SAVs) or point mutations are associated with more than half of human inherited diseases. They may directly compromise the active sites or have allosteric effects, and act as 'latent drivers' associated with cancer development and drug resistance [17]. The growing SAV data has led to the development of databases and ML frameworks for predicting pathogenicity [18-23]. We recently explored whether considering the *intrinsic dynamics* of the protein might improve the accuracy of SAV pathogenicity prediction [15]. A simple random forest (RF)-based ML tool benchmarked against >20,000 SAVs indeed confirmed that this is the case [15] (Fig. 2b). The tool, implemented in Rhapsody [20], shows that while sequence-based features make the largest contributions to predictions, dynamic features also make important contributions (e.g., larger than solvent-accessible surface area, SASA) (Fig. 2a). Among dynamics features, allosteric signaling effectiveness emerged as an important factor. More recently, the Rhapsody feature set has been expanded to incorporate sixteen additional features from Rosetta (excluding SASA) [24]; and dynamics-based attributes in Rhapsody have been used along with sequence- and structure-based attributes to construct an XGBoost classifier within the LYRUS tool to predict pathogenicity [25].

Fig. 2c-d illustrates the application of Rhapsody to p53. The heatmap in **c** displays the in silico saturation mutagenesis results for a 100-residue portion (see the complete map at Rhapsody interface; and detailed description in the figure caption). Several SAVs are 'newly' evaluated here, in the sense that they were not included in the training set. Panel **d** displays some of them (residues in green spheres) that are confirmed to be pathogenic according to the data reported in the literature; residues shown in green sticks are variants of unknown significance that are predicted to be pathogenic, waiting for validation.

ENMs such as the anisotropic network model (ANM) [12,29] and the Gaussian network model (GNM) [30] that are most broadly used in the literature (and in Rhapsody) provide residue-level information on structural dynamics and lack amino acid specificity. Atomic-level residue-specific information on the role of conformational dynamics in disease-causing missense mutations, on the other hand, can be inferred from MD trajectories, as illustrated in a recent study [31]. A recent ML- and MD-based study showed how the pathogenicity of mutations at sites that are neither evolutionarily conserved nor directly involved in biochemical activity could be explained by dynamic couplings [32]. Therein, the data from MD were used to train a feedforward NN for pathogenicity prediction. In another study, attributes extracted from MD were used in a rule-based classifier to predict the pathogenicity of unclassified variants of BRCA1 BRCT repeats [33]. Similarly, a K-nearest neighbor predictor of disease specificity was built for calmodulin variants, as well as Aβ peptide variants, using the distributions of ($\varphi$, $\psi$) angles and the root-mean-square deviations (RMSDs) and fluctuations (RMSFs) in atomic positions observed in MD simulations conducted for the two wild type proteins and the corresponding sets of variants [34]. The DL tool DiffNets, on the other hand, predicts the biochemical differences between variants using self-supervised autoencoders that learn the associated latent space from MD-sampled structural ensembles [35].

Note that residue-specificity can also be incorporated within the ENMs. DynaSig-ML is a package that utilizes elastic network contact model (ENCoM)-predicted modes and ML to predict biomolecular function and proved useful in a recent prediction of the evolutionary fitness of a bacterial enzyme [36]. ENCoM differs from the elastic network models ANM and GNM by the inclusion of residue-specificity through contact surface evaluation for interacting pairs [37]. It is worth noting that there are also other ENMs that are residue-specific, such as the ANM with inter-residue spring constants proportional to the number of atom—atom contacts, introduced in modeling the Markovian diffusion of allosteric signals [38], and even the original ENM of Tirion where uniform harmonic potentials were used between all atom pairs [39].

Overall, recent studies highlight the utility of using structural dynamics data, at either low or high resolution (e.g., ENMs or MD), in ML platforms, to accurately predict the effect of SAVs on (dys)function.

## Neural networks help learn molecular properties from MD trajectories

Recent studies point to the utility of NNs for learning from MD trajectories [16]. The NNs take as input the coordinates of all atoms (or α-carbons in coarse-grained MD), and output properties such as potential energy surface or operating force fields. ML models trained on MD trajectories can learn the “latent space” accessible to the protein, and predict new trajectories or conformations not observed in the original MD runs [40,41], allowing for rapid and more complete sampling of conformational space.

Methods of unsupervised learning applied to MD data further assist in transition kinetics modeling [42]. The last decade has seen a broad use of Markov state models (MSMs) for analyzing MD trajectories [43]. While the underlying stochastic theory and Master equation formalism go back to early 1900s, its use for characterizing ensembles of protein conformations to define states (or substates) and their transition kinetics took center stage in the last two decades. We demonstrated the utility of this formalism for mapping conformational space and kinetics in the early 2000s using toy models for proteins [44,45]. MSMs are now broadly used for extracting functional information from multiple or long MD runs. While their use originally required significant human input, DL algorithms can now automatically determine the significant features for defining representative states, e.g., VAMPNets [46] provides a fully automated framework to evaluate collective variables and MSM transition matrices [47,48]. Recently, Zhu and coworkers used GNNs to predict allosteric communication pathways from MD simulations [49].

## Confluence of structural dynamics and DL methods in drug discovery

Binding affinity at constant temperature and pressure, quantified as the Gibbs free energy change ΔG associated with binding, depends on the accompanying changes in enthalpy (H) and entropy (S) as $\Delta G = DH - T\Delta S$. An important source of ΔS is the conformational entropy change upon binding. ΔS scales with $\ln [\det(\boldsymbol{\sigma}_b)/\det(\boldsymbol{\sigma}_u)]$ [50] where $\det(\boldsymbol{\sigma}_b)$ is the determinant of the positional covariance matrix of the protein in ligand-bound form, and $\det(\boldsymbol{\sigma}_u)$ is that in the unbound form. Thus, the equilibrium dynamics of the protein, manifested by covariance in the fluctuations of residues/atoms, is an inseparable part of

any binding event. In fact, entropic effects are so salient that many proteins evolved to co-localize their ligand-binding sites with global hinge centers [14], associated rotation axes [51] or domain interfaces [52] where entropy loss upon ligand binding is minimal.

ML-aided drug screening approaches have shown success in recent years. For example, a generative tensorial reinforcement learning model [53] outperformed adversarial models and identified novel potent inhibitors of DDR1 kinase, which were experimentally validated. In a later version, generating 3D structures of drug-like compounds were conditioned on a receptor binding site and validated on receptors with mutations [54]. However, these and most of ML approaches for drug screening do not adequately consider the protein intrinsic dynamics (while the translational and rotational motions of the small molecule are thoroughly sampled). One of the exceptions is AtomNet [55], a structure-based deep-convolutional NN, trained on multiple poses of active compounds bound to a single target site as well as experimentally verified inactive compounds, which proved to outperform established scoring functions. In another study, inclusion of MD descriptors was shown to improve the discrimination of good caspase-8 inhibitors from poor ones [56].

In contrast, recent ML algorithms for predicting drug-resistant mutations do utilize structural dynamics data from MD [57,58] to generate dynamics-based features that are trained, often by RF and feedforward multilayer perceptron, using labels of ~1000 drug-sensitive or -resistant mutants extracted from the Platinum database [59]. Similarly, SUSPECT-RIF uses a residue-specific version of ENM (ENCoM) [60] to take account of dynamics in predicting drug resistance [61]. ENMs provide an efficient platform to uniformly generate data on equilibrium dynamics by virtue of their applicability at omics scale without bias [62], as validated by comparisons with NMR-sampled equilibrium dynamics [63] and X-ray crystallographically solved ensembles of structures for the same protein in different states [13].

ML-based methods are not necessarily more accurate than MM/GBSA evaluation of binding free energy [64]. The value of ML-based methods in drug screening lies in their efficiency. ML methods may also be subject to challenges including applicability to new cases as well as uncertainties in experimental data (e.g., affinity data reported for the same protein-ligand complex may differ by orders of magnitude). Yet, ML-based docking [65] or re-ranking scheme [55] could be efficiently adopted together with a conformational sampling scheme for a first screening (also called ensemble docking) before performing (for selected cases) MD-based free energy evaluations. The latter could be even used for screening purposes [66], empowered by the accelerations enabled by GPUs. Finally, data-driven ML efforts could be advantageously redirected to areas where first principles of physical sciences fall short, such as predictions of cell toxicity and PK/PD in animals, which are important elements of drug development.

## Ensemble analyses based on equilibrium dynamics are yet to be routinely used in ML-based stability predictions

Despite advances in computational methods, in silico predictions for changes in Gibbs free energy (ΔΔG) of folding associated with SAVs still suffer from limitations and challenges [67]. As mentioned above, the entropic contribution to ΔΔG, which directly relates to the

distribution of residue fluctuations, or the curvature of the energy minimum near equilibrium is often overlooked. Among ML studies for predicting the effect of mutations on stability, DynaMut [68] is distinguished, as it utilizes the changes in vibrational entropy predicted by ENCoM [60].

For illustrative purposes, we examined how well a simple gradient boosting regressor exclusively trained on equilibrium dynamics data predicts the changes in stability (ΔΔG values) associated with point mutations (Fig. 3a). The regressor uses ESSA scores as a measure of the impact of mutation on the global modes' frequency dispersion or on the energy minimum curvature in the subspace of essential motions [69], in addition to residue MSFs, and perturbation response scanning [70] and mechanical stiffness data trained on 2,298 mutations in the S2648 database [71]. The predictions tested on 350 mutations yield a PCC of 0.61 with experimental data together with an RMSE of 1.24 kcal/mol. Around 77% of tested cases are correctly identified to be destabilizing (ΔΔG <0). The fact that this much can be achieved by equilibrium dynamics-based attributes exclusively points to the unexploited potential of equilibrium dynamics in improving the predictions of ΔΔG (or ΔΔS).

We further posit ML models may be improved upon considering properties based on ensemble of conformers accessible under equilibrium conditions, as opposed to a single structure. For example, in Fig. 3b, show that the SASA of hen egg-white lysozyme Asn46 varies from 0.04 to 0.86 across two intrinsically accessible states (the full range is $0 \leq$ SASA $\leq 1$, from completely buried and completely exposed). A regressor trained on a fixed SASA score in this case would thus be misleading. Fig. 3c displays the distribution of the difference in SASAs for all residues in all proteins in the S2648 database, obtained by generating 40 conformers for each case using ClustENMD. Similarly, Fig. 3 panels **d** and **e** show that the hydrophobic packing density and the number of hydrogen bonds near mutating residues may vary considerably across intrinsically accessible states. The availability of hybrid simulation methods [72] enables a high-throughput generation of such ensembles of conformers. Using such ensembles, the variance of different physicochemical attributes, not only their means, may be used in ML algorithms to help design better ΔΔG predictors.

The prediction of the change in stability due to insertions/deletions (InDels) of amino acids in proteins is another important but neglected area. The ML tool PROFOUND [73] was recently introduced to predict the change in stability associated with multiple (contiguous) amino acids deletions. As the dependence on residue specificity is lower for deletions than substitutions, we explored the change in stability predicted by PROFOUND using ENM-based attributes. We found that the dynamics-based attributes could adequately predict the change in stability (recall = 78.0% on 10-fold cross-validation on positive-unlabeled-learning). Upon combining dynamics-based features with the PROFOUND features, we could achieve a recall of 84.3%. Fig. 3f shows that equilibrium dynamics/ensemble-based attributes contribute 72.3% to classification. This further highlights the importance of considering equilibrium dynamics for in silico prediction of change in stability.

## Conclusion

ML methods have been taking advantage of the rapid growth in structural data in the PDB, and soon they may take advantage of the accumulation of ensembles of structures for a given protein. With the increased generation of alternative conformers for a given protein, using either full atomic simulations such as MD, coarse-grained approaches such as those based on elastic network models, or hybrid models, it is conceivable that NNs will be trained for each protein or homologous proteins on their ensemble of conformers. Such models would predict the changes in distances, potentially at multiple scales, not only between amino acid pairs, but also entire domains or subunits, and enable a more comprehensive mapping of the space of conformational dynamics, thus providing new tools for bridging structure and function. Integration of ML techniques with data on structural dynamics are likely to uncover disease mechanisms that are otherwise intractable by experiments alone, as demonstrated in a recent study [74].

Notably, ENM-predicted structural dynamics depend on inter-residue contact topology. The latter has been pointed out to be a major a descriptor that discriminates pathogenic human variants [22]. The significance of contact topology is also borne out by the predictive performance of Rhapsody that incorporates ENM-derived descriptors (Fig. 2b). ENMs further reveal the pre-existing paths of collective reconfiguration of proteins.

It is important to note, however, that ENMs such as the ANM and the GNM are agnostic to the chemical nature of amino acids and cannot predict the effect of specific substitutions/mutations in amino acids. Despite its lack of amino acid specificity, the GNM proved useful when used within pathogenicity predictors such as Rhapsody. One way to explain this dichotomy is that ENMs identify critical sites or specific positions in the 3D structure, which could play an important mechanical role, or a critical site for allosteric communication, and thus would not tolerate mutations irrespective of the specific substitution. We previously demonstrated for example that the location of enzyme active sites can be inferred from GNM-predicted mechanical hinges even in the absence of the coordinates of amino acid side chains [14,62]; and such hinge sites tend to be evolutionarily conserved [62]. Mutations at those sites, either enhancing or reducing local flexibility, may impair the enzyme activity [75]. Dynamics-based features thus provide information on the adaptability of specific positions to substitutions purely based on 3D topology (a collective structural property). Sequence conservation, substitution, or co-evolution properties further shed light on the likelihood of (or tolerance to) specific substitutions at those positions along the sequence (or on the 3D structure). In other words, structural dynamics provides an overall estimate for the specific position (as can be seen by the vertical red or blue slabs that show little dependence on specific amino acid type in Fig. 2C), and sequence information further discriminates between different types of substitutions.

It is conceivable that if a structure is available for the mutant, and if this structure is sufficiently different from the native fold, ENMs could be applied to both structures to deduce the effect of mutation. However, some amino acid substitutions may have minimal effect on the protein backbone while altering the protein specificity or functionality. For example, at a solvent-exposed substrate-recognition site, the change in amino acid

may not affect the overall fold/topology but may impair the substrate recognition and cause a loss of function. While sequence-specific ENMs (such as ENCoM) may help in discriminating between amino acid types, an assessment of gain or loss of function may further require knowledge of substrate-binding sites, or protein-substrate interaction interfaces. Existing databases permit us to learn about protein—protein interaction interfaces using ML methods. Thus, inclusion of such knowledge of interaction interfaces between pairs of proteins, learned by ML methods, may be a future direction for further improving our evaluation of missense variants.

In summary, ENMs may indirectly contain the effects of sequences and structure following the sequence-encodes-structure-encodes-dynamics paradigm, and this may partially explain their success in predicting the pathogenicity of SAVs or the effect of indels on stability. However, a direct consideration of sequence properties, including both conservation and co-evolution behavior, which is best achieved by ML methods, is essential to the success of existing predictors. Comprehensive mapping of the conformational dynamics space, potentially at multiple scales, will help build increasingly powerful ML tools for bridging structure and function.

## Acknowledgements

Support by NIH (R01 GM139297) is gratefully acknowledged by IB.

## Data availability

Data will be made available on request.

## References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

** of outstanding interest

1. Pearce R, Zhang Y: Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr Opin Struct Biol 2021, 68:194–207. [PubMed: 33639355]
2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D: Highly accurate protein structure prediction with AlphaFold. Nature 2021, 596:583–589. [PubMed: 34265844]
3. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zidek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S: AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022, 50:D439–D444. [PubMed: 34791371]
4. AlQuraishi M: Machine learning in protein structure prediction. Curr Opin Chem Biol 2021, 65:1–8. [PubMed: 34015749]

5. Gauthier J, Vincent AT, Charette SJ, Derome N: A brief history of bioinformatics. Briefings Bioinf 2019, 20:1981–1996.

6. Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CPI, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, Kang C, Sheridan R, Draizen EJ, Dallago C, Sander C, Marks DS: The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 2019, 35:1582–1584. [PubMed: 30304492]

7. Ju F, Zhu J, Shao B, Kong L, Liu TY, Zheng WM, Bu D: CopulaNet: learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. Nat Commun 2021, 12:2535. [PubMed: 33953201]

8. Braberg H, Echeverria I, Kaake RM, Sali A, Krogan NJ: From systems to structure - using genetic data to model protein structures. Nat Rev Genet 2022, 23:342–354. [PubMed: 35013567]

9**. Alderson TR, Kay LE: NMR spectroscopy captures the essential role of dynamics in regulating biomolecular function. Cell 2021, 184:577–595. [PubMed: 33545034] This NMR-based study successfully shows the importance of dynamics between the gap–junction structure and function. Many examples are provided in this article to uncover the importance of structural dynamics for the function and misfunction of proteins.

10*. Zhang Y, Doruker P, Kaynak B, Zhang S, Krieger J, Li H, Bahar I: Intrinsic dynamics is evolutionarily optimized to enable allosteric behavior. Curr Opin Struct Biol 2020, 62:14–21. [PubMed: 31785465] This review describes that allosteric responses would be induced only if the protein were already predisposed to certain conformation changes. This intrinsic dynamics actually acts as evolutionary adaptation mechanism to facilitate allosteric communication.

11. Wingert B, Krieger J, Li H, Bahar I: Adaptability and specificity: how do proteins balance opposing needs to achieve function? Curr Opin Struct Biol 2021, 67:25–32. [PubMed: 33053463]

12. Eyal E, Chennubhotla C, Yang LW, Bahar I: Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. Bioinformatics 2007, 23:i175–i184. [PubMed: 17646294]

13. Bakan A, Bahar I: The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proc Natl Acad Sci U S A 2009, 106:14349–14354. [PubMed: 19706521]

14. Yang LW, Bahar I: Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 2005, 13:893–904. [PubMed: 15939021]

15. Ponzoni L, Bahar I: Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci U S A 2018, 115:4164–4169. [PubMed: 29610305]

16*. Noe F, Tkatchenko A, Muller KR, Clementi C: Machine learning for molecular simulation. Annu Rev Phys Chem 2020, 71:361–390. [PubMed: 32092281] This review discusses recent advances and future goals in researching the application of ML to MD simulations. Topics discussed include different types of energies to be predicted from trajectories, generation of new conformations, and different types of neural networks used.

17. Nussinov R, Tsai CJ: Latent drivers' expand the cancer mutational landscape. Curr Opin Struct Biol 2015, 32:25–32. [PubMed: 25661093]

18. Adzhubei I, Jordan DM, Sunyaev SR: Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013, 76(1):7–20.

19. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, Marks DS: Mutation effects predicted from sequence co-variation. Nat Biotechnol 2017, 35:128–135. [PubMed: 28092658]

20*. Ponzoni L, Penaherrera DA, Oltvai ZN, Bahar I: Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics 2020, 36:3084–3092. [PubMed: 32101277] The article introduces the Rhapsody tool that uses dynamics-based attributes along with sequence and structure-based properties to predict the pathogenicity of human single amino acid mutations.

21. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y: MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun 2021, 12:510. [PubMed: 33479230]

22*. Woodard J, Iqbal S, Mashaghi A: Circuit topology predicts pathogenicity of missense mutations. Proteins 2022 Sep;90(9):1634–1644. [PubMed: 35394672] In this study the inverse parallel

and cross relation circuit topologies of the mutated residue are shown to discriminate between pathogenic and benign human missense variants. On comparing with several structural attributes, the authors establish that circuit topology provide non-redundant information on protein structures and pathogenicity of mutations.

23. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y, Marks DS: Disease variant prediction with deep generative models of evolutionary data. Nature 2021, 599:91–95. [PubMed: 34707284]
24. Wu TH, Lin PC, Chou HH, Shen MR, Hsieh SY: Pathogenicity prediction of single amino acid variants with machine learning model based on protein structural energies. IEEE ACM Trans Comput Biol Bioinf 2021.
25. Lai J, Yang J, Gamsiz Uzun ED, Rubenstein BM, Sarkar IN: LYRUS: a machine learning model for predicting the pathogenicity of missense variants. Bioinform Adv 2022, 2:vbab045. [PubMed: 35036922]
26. General IJ, Liu Y, Blackburn ME, Mao W, Gierasch LM, Bahar I: ATPase subdomain IA is a mediator of interdomain allostery in Hsp70 molecular chaperones. PLoS Comput Biol 2014, 10, e1003624. [PubMed: 24831085]
27. Chen Y, Dey R, Chen L: Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer. Structure 2010, 18:246–256. [PubMed: 20159469]
28. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, …Sunyaev SR: A method and server for predicting damaging missense mutations. Nature methods 2010, 7(4):248–249. [PubMed: 20354512]
29. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I: Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001, 80:505–515. [PubMed: 11159421]
30. Bahar I, Atilgan AR, Erman B: Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding Des 1997, 2:173–181.
31. Kots E, Mlynarczyk C, Melnick A, Khelashvili G: Conformational transitions in BTG1 antiproliferative protein and their modulation by disease mutants. Biophys J 2022 Oct, 121(19):3753–3764. [PubMed: 35459639]
32**. Ose NJ, Butler BM, Kumar A, Kazan IC, Sanderford M, Kumar S, Ozkan SB: Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants. PLoS Comput Biol 2022, 18, e1010006. [PubMed: 35389981] This article discusses missense variants that are neither well conserved nor fall in any known functional domains but are known to be pathogenic. The authors establish that the allosteric dynamic coupling between these residues and known functional sites are plausibly responsible for the pathogenicity of such mutations.
33. Sinha S, Wang SM: Classification of VUS and unclassified variants in BRCA1 BRCT repeats by molecular dynamics simulation. Comput Struct Biotechnol J 2020, 18:723–736. [PubMed: 32257056]
34. McCoy MD, Hamre J 3rd, Klimov DK, Jafri MS: Predicting genetic variation severity using machine learning to interpret molecular simulations. Biophys J 2021, 120:189–204. [PubMed: 33333034]
35*. Ward MD, Zimmerman MI, Meller A, Chung M, Swamidass SJ, Bowman GR: Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with DiffNets. Nat Commun 2021, 12:3023. [PubMed: 34021153] This paper shows how a new method- DiffNets automatically identifies the deterministic/specific structural features from the entire ensemble of structures that a protein adopts, to predict biological differences between protein families using MD simulation data. In this dimensionality reduction algorithm, self-supervised autoencoder architecture is used to learn features of a protein's structural ensemble.
36. Mailhot Om F, Najmanovich R: The DynaSig-ML Python package: automated learning of biomolecular dynamics-function relationships. bioRxiv 2022.
37. Frappier V, Najmanovich RJ: A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol 2014, 10, e1003569. [PubMed: 24762569]

38. Chennubhotla C, Bahar I: Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Comput Biol 2007, 3:1716–1726. [PubMed: 17892319]

39. Tirion MM: Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996, 77:1905–1908. [PubMed: 10063201]

40. Hoseini P, Zhao L, Shehu A: Generative deep learning for macromolecular structure and dynamics. Curr Opin Struct Biol 2021, 67:170–177. [PubMed: 33338762]

41. Degiacomi MT: Coupling molecular dynamics and deep learning to mine protein conformational space. Structure 2019, 27:1034–1040 e1033. [PubMed: 31031199]

42. Glielmo A, Husic BE, Rodriguez A, Clementi C, Noe F, Laio A: Unsupervised learning methods for molecular simulation data. Chem Rev 2021, 121:9722–9758. [PubMed: 33945269]

43. Husic BE, Pande VS: Markov state models: from an art to a science. J Am Chem Soc 2018, 140:2386–2396. [PubMed: 29323881]

44. Ozkan SB, Dill KA, Bahar I: Computing the transition state populations in simple protein models. Biopolymers 2003, 68:35–46. [PubMed: 12579578]

45. Ozkan SB, Dill KA, Bahar I: Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. Protein Sci 2002, 11:1958–1970. [PubMed: 12142450]

46. Mardt A, Pasquali L, Wu H, Noe F: VAMPnets for deep learning of molecular kinetics. Nat Commun 2018, 9:5. [PubMed: 29295994]

47. Konovalov KA, Unarta IC, Cao S, Goonetilleke EC, Huang X: Markov state models to study the functional dynamics of proteins in the wake of machine learning. JACS Au 2021, 1:1330–1341. [PubMed: 34604842]

48. Noe F, De Fabritiis G, Clementi C: Machine learning for protein folding and dynamics. Curr Opin Struct Biol 2020, 60:77–84. [PubMed: 31881449]

49. Zhu J, Wang J, Han W, Xu D: Neural relational inference to learn long-range allosteric interactions in proteins from molecular dynamics simulations. Nat Commun 2022, 13:1661. [PubMed: 35351887]

50. Karplus M, Kushick JN: Method for estimating the configurational entropy of macromolecules. Macromolecules 1981, 14:325–332.

51. Yang LW: Models with energy penalty on interresidue rotation address insufficiencies of conventional elastic network models. Biophys J 2011, 100:1784–1793. [PubMed: 21463592]

52. Li H, Sakuraba S, Chandrasekaran A, Yang LW: Molecular binding sites are located near the interface of intrinsic dynamics domains (IDDs). J Chem Inf Model 2014, 54:2275–2285. [PubMed: 25089914]

53. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, Volkov Y, Zholus A, Shayakhmetov RR, Zhebrak A, Minaeva LI, Zagribelnyy BA, Lee LH, Soll R, Madge D, Xing L, Guo T, Aspuru-Guzik A: Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol 2019, 37:1038–1040. [PubMed: 31477924]

54. Ragoza M, Masuda T, Koes DR: Generating 3D molecules conditional on receptor binding sites with deep generative models. Chem Sci 2022, 13:2701–2713. [PubMed: 35356675]

55*. Wallach I, Dzamba M, Heifets A: AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv; 2015, 151002855. A leading algorithm AtomNet considers structural features of the drug–protein complex during training, while ligand dynamics is implicitly presented as multiple binding modes (poses) of ligands localized in the binding site. However, protein dynamics is not (explicitly) considered in the training.

56. Jamal S, Grover A, Grover S: Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against alzheimer's disease. Front Pharmacol 2019, 10:780. [PubMed: 31354494]

57. Wang DD, Ou-Yang L, Xie H, Zhu M, Yan H: Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. Comput Struct Biotechnol J 2020, 18:439–454. [PubMed: 32153730]

58. Hamre J r R, Klimov DK, McCoy MD, Jafri MS: Machine learning-based prediction of drug and ligand binding in BCL-2 variants through molecular dynamics. Comput Biol Med 2021, 140, 105060. [PubMed: 34920365]
59. Pires DE, Blundell TL, Ascher DB: Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. Nucleic Acids Res 2015, 43:D387–D391. [PubMed: 25324307]
60. Frappier V, Chartier M, Najmanovich RJ: ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic Acids Res 2015, 43:W395–W400. [PubMed: 25883149]
61. Portelli S, Myung Y, Furnham N, Vedithi SC, Pires DEV, Ascher DB: Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches. Sci Rep 2020, 10, 18120. [PubMed: 33093532]
62. Li H, Chang YY, Lee JY, Bahar I, Yang LW: DynOmics: dynamics of structural proteome and beyond. Nucleic Acids Res 2017, 45:W374–W380. [PubMed: 28472330]
63. Yang LW, Eyal E, Bahar I, Kitao A: Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. Bioinformatics 2009, 25:606–614. [PubMed: 19147661]
64. Wang DD, Zhu M, Yan H: Computationally predicting binding affinity in protein-ligand complexes: free energy-based simulations and machine learning-based scoring functions. Briefings Bioinf 2021 May, 22(3):bbaa107.
65. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, Sunseri J, Koes DR: Gnina 1.0: molecular docking with deep learning. J Cheminf 2021, 13:43.
66. Takemura K, Sato C, Kitao A: ColDock: concentrated ligand docking with all-atom molecular dynamics simulation. J Phys Chem B 2018, 122:7191–7200. [PubMed: 29993242]
67. Iqbal S, Li F, Akutsu T, Ascher DB, Webb GI, Song J: Assessing the performance of computational predictors for estimating protein stability changes upon missense mutations. Briefings Bioinf 2021 Nov, 22(6):bbab184.
68. Rodrigues CHM, Pires DEV, Ascher DB: DynaMut2: assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci 2021, 30:60–69. [PubMed: 32881105]
69. Kaynak BT, Bahar I, Doruker P: Essential site scanning analysis: a new approach for detecting sites that modulate the dispersion of protein global motions. Comput Struct Biotechnol J 2020, 18:1577–1586. [PubMed: 32637054]
70. Atilgan C, Atilgan AR: Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS Comput Biol 2009, 5, e1000544. [PubMed: 19851447]
71. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M: Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009, 25:2537–2543. [PubMed: 19654118]
72. Krieger JM, Doruker P, Scott AL, Perahia D, Bahar I: Towards gaining sight of multiscale events: utilizing network models and normal modes in hybrid methods. Curr Opin Struct Biol 2020, 64:34–41. [PubMed: 32622329]
73. Banerjee A, Kumar A, Ghosh KK, Mitra P: Estimating change in foldability due to multipoint deletions in protein structures. J Chem Inf Model 2020, 60:6679–6690. [PubMed: 33225697]
74. Martin W, Sheynkman G, Lightstone FC, Nussinov R, Cheng F: Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: applications to Alzheimer's disease. Curr Opin Struct Biol 2022, 72:103–113. [PubMed: 34628220]
75. Gotz A, Mylonas N, Hogel P, Silber M, Heinel H, Menig S, Vogel A, Feyrer H, Huster D, Luy B, Langosch D, Scharnagl C, Muhle-Goll C, Kamp F, Steiner H: Modulating hinge flexibility in the APP transmembrane domain alters gamma-secretase cleavage. Biophys J 2019, 116:2103–2120. [PubMed: 31130234]

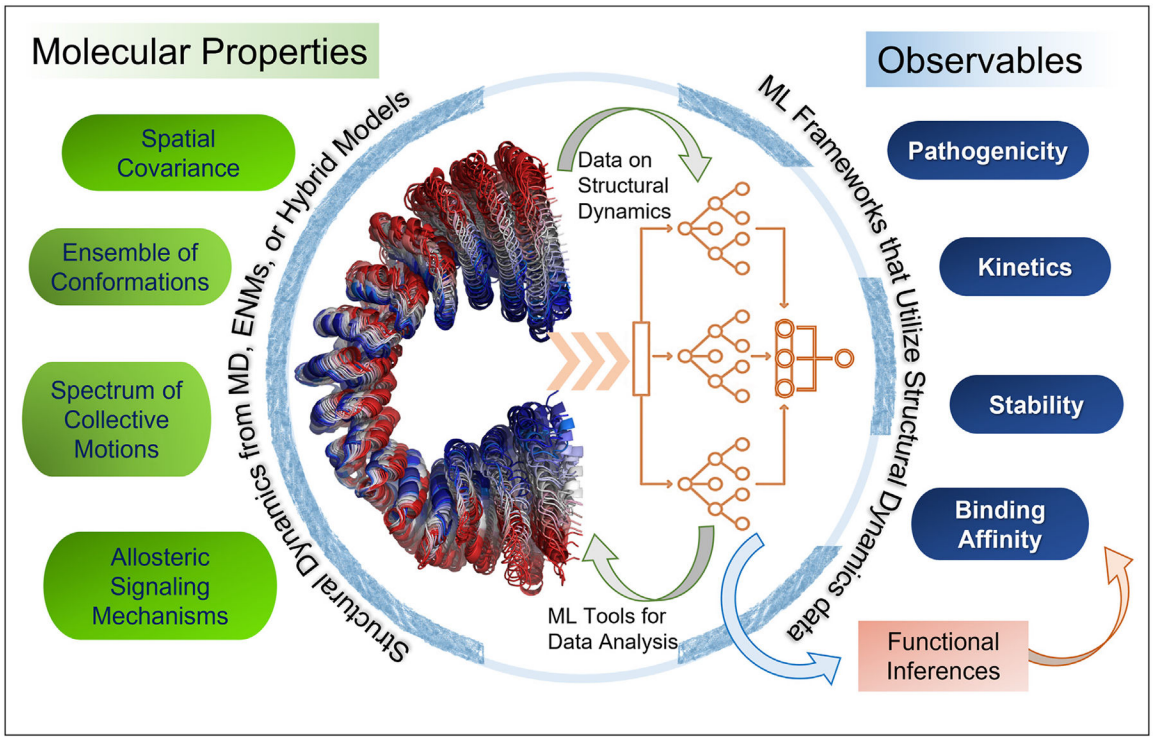


**Figure 1. Utility of combining microscopic time-dependent features predicted by structure-based models, methods, and simulations with ML frameworks for making functional inferences.**

The confluence of the structural dynamics (molecular) data (*left part* of the circle) with ML methods (*right part* of the circle) enables us to evaluate macroscopic properties (observables) from molecular properties.

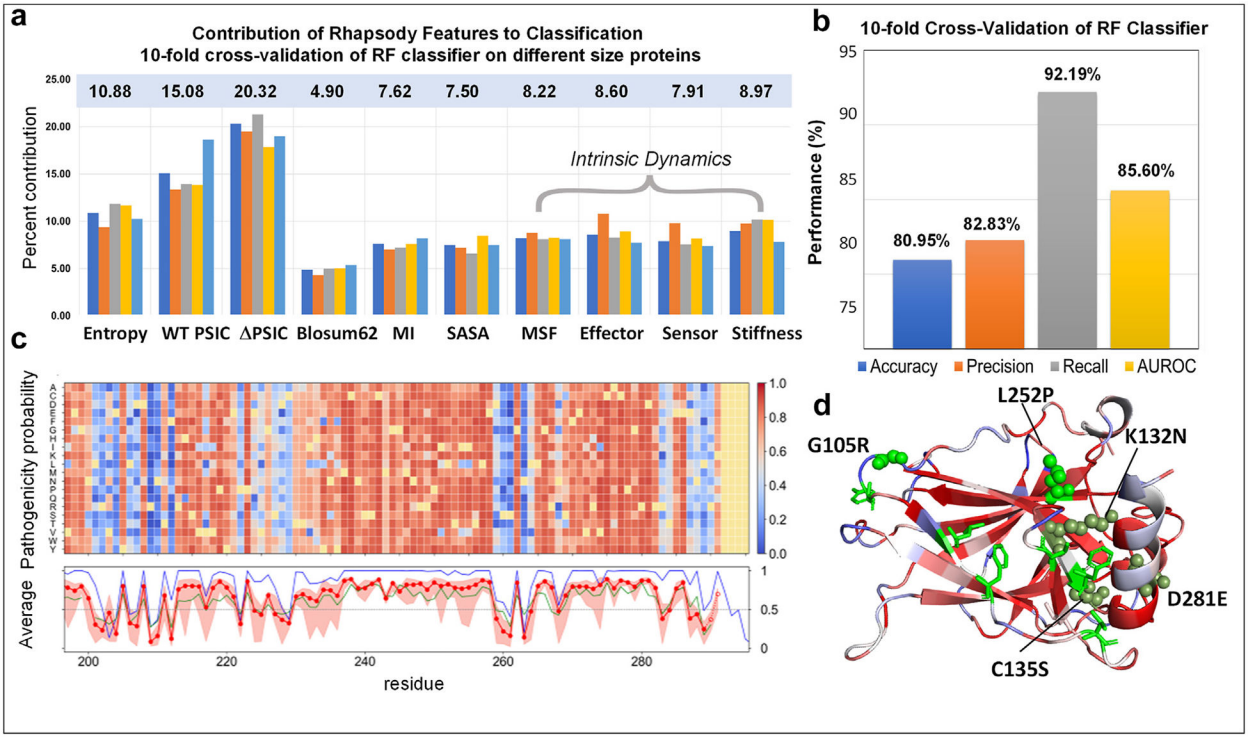


**Figure 2. ML-based pathogenicity predictor Rhapsody takes account of structural dynamics, yields highly accurate predictions, and can be used to generate *in silico saturation mutagenesis* heatmaps.**

**(a)** Descriptors used in Rhapsody: sequence (conservation (entropy), position specific independent counts (PSIC) and change in PSIC (ΔPSIC), amino acid substitution (Blosum62), mutual information (MI)), structure (SASA) and structural dynamics (mean-square fluctuations (MSF) of mutated residue, propensity to serve as effector or sensor of allosteric signals (see previous work [26]), and mechanical stiffness). The *bars* display the percent contribution of these descriptors to the trained classifier. A set of five bars is displayed for each descriptor, corresponding to subsets of proteins of different sizes, with numbers of residues lying in the ranges [150–249] (*orange*), [250–361] (*gray*), [362–520] (*yellow*), and [521 –3636] (*light blue*). The first bar (*dark blue)* in each group refers to the entire set. The corresponding percent contributions of different features are listed in the *light blue box.* (**b**) *Prediction* performance based on different metrics. (**c**) In silico saturation mutagenesis heatmap. These are pathogenicity probabilities (see the scale bar on the right) evaluated for all 19 substitutions (ordinate) at each residue position (abscissa), shown here for a 100-residue segment of p53. Structural and dynamic features are based on the tetrameric structure (PDB id: 3KMD) [27]. The curves underneath are the averages over all 19 substitutions for each residue, predicted by Rhapsody (*red dots*), PolyPhen-2 (*dark blue*) [28] and EVMutation (*green*) [19]. The Pearson correlation coefficient (*PCC*) between each pair of results is around 0.74; whereas that between PolyPhen-2 and EVMutation is 0.58. (**d**) Color-coded pathogenicity results for p53 monomer. Mutations at sites colored red are highly susceptible to be pathogenic. A few such residues are labeled. These are reported in ClinVar to be pathogenic (*green spheres*), likely pathogenic (*olive spheres*), or unknown (*green sticks).*

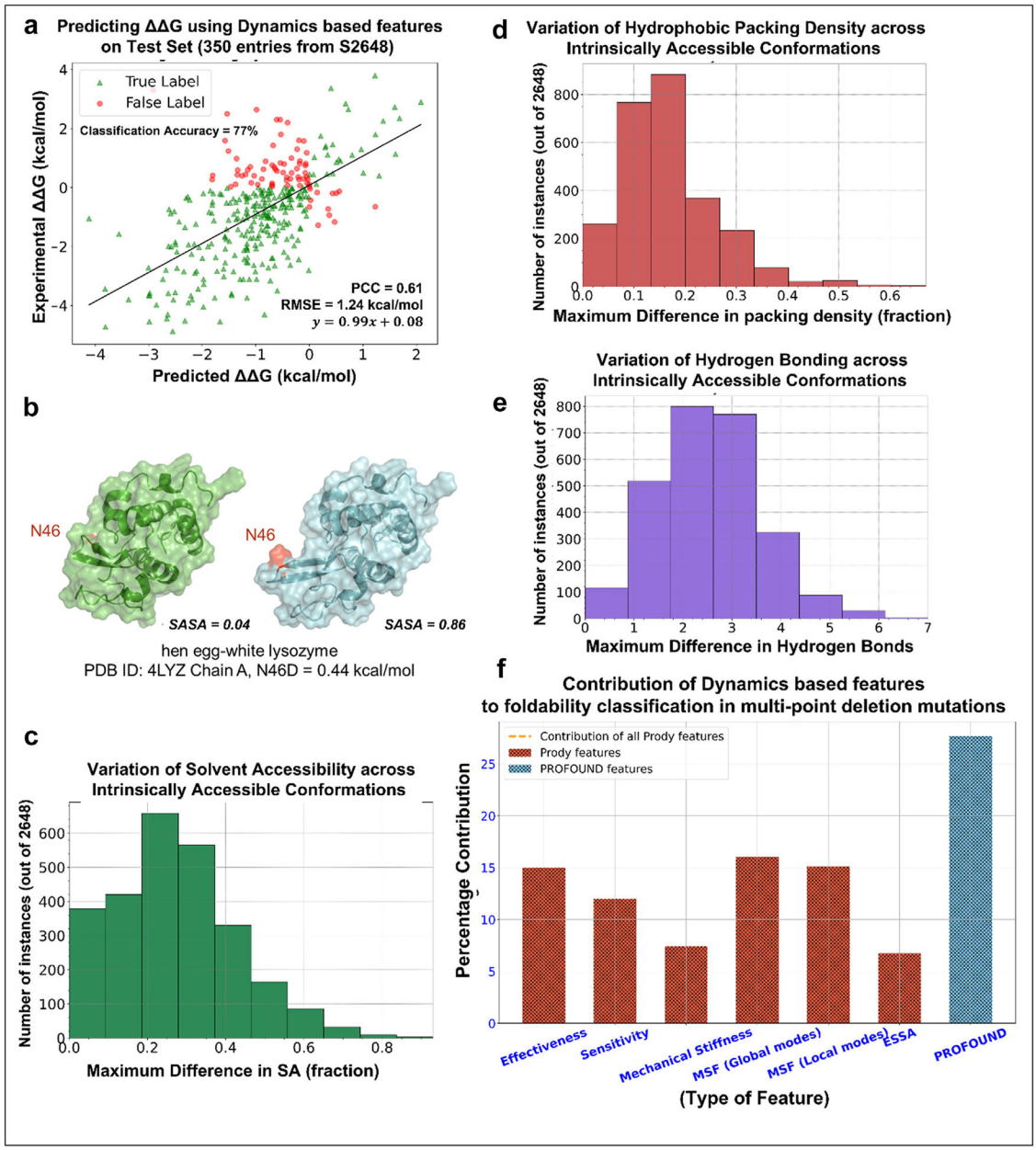


**Figure 3. Significance of equilibrium dynamics and variation of physicochemical attributes across ensembles of conformers in developing ML-predictors of stability.**

**(a)** A gradient boosting regressor (scikit-learn package with n_estimators = 1500, subsample = 0.7, max_depth = 5, max_features = 7) trained on dynamics-based attributes (ESSA score (for wt and substituted residues and their difference); MSFs in the softest 2% of modes, MSFs in the stiffest 5% modes, allosteric signaling sensitivity and effectiveness; mechanical stiffness) The regressor yielded a PCC of 0.61 and a RMSE of 1.24 kcal/mol on a widely benchmarked test set consisting of 350 SAVs. **(b)** Hen egg-white lysozyme N46 corresponding to two intrinsically accessible conformations show a widely varying SASA, from 0.04 to 0.86. **(c)** The distribution of the maximum difference in solvent accessibility, **(d)** distribution of hydrophobic packing density (the ratio of hydrophobic residues within

5 Å sphere radius of the mutated residue to the total number of residues within the same radius), **(e)** distribution of the changes in the number of hydrogen bonds near the investigated residue, **(f)** Contribution of indicated (abscissa labels) dynamics-based attributes and the 39 additional attributes from PROFOUND) to foldability prediction.