Machine learning based detection of genetic and drug class variant impact on functionally conserved protein binding dynamics

Gregory A. Babbitt[1]*, Ernest P. Fokoue[3], Joshua R. Evans[1], Kyle I. Diller[1,2], Lily E. Adams[1]

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester NY, USA 14623
[2]Golisano College for Computing and Information Science, Rochester NY, USA 14623
[3]Applied Statistics, Rochester Institute of Technology, Rochester NY, USA 14623

*email address for primary correspondence: gabsbi@rit.edu

Contact Information –Gregory A. Babbitt (gabsbi@rit.edu).

Abstract

The application of statistical methods to comparatively framed questions about protein dynamics can potentially enable investigations of biomolecular function beyond the current sequence and structure based methods in comparative genomics. However, addressing this problem requires proper statistical inferences obtained from large ensembles of individual molecular dynamic (MD) simulations that represent the comparative functional states of a given protein under investigation. Meaningful interpretation of such large and temporally rich forms of data poses serious challenges to users of MD. Here, we announce the release of DROIDS v3.0, an Amber18/Chimera-based software package for comparative protein dynamics, incorporating many new features including maxDemon v1.0, a multi-method machine learning application that trains on large ensembles of MD comparisons generated by DROIDS and deploys learned classifications of dynamic states to newly generated protein simulations. Up to seven different machine learners can be deployed on the dynamics of each amino acid, and local canonical correlations in learning patterns generated from self-similar MD runs are used to identify regions of functionally conserved protein dynamics. The subsequent impacts of genetic and drug class variants on conserved dynamics can also be analyzed by deploying the trained classifiers on new MD runs. Still and moving images of comparative dynamics allow users to see both when and where a protein dynamic simulation displays a specific functional state defined by the functional protein comparison. Our software allows quantitative visualization of complex changes in functional dynamics caused by temperature changes, chemical mutations, or binding interactions with nucleic acids or a variety of small molecules. Here, we demonstrate the utility of DROIDS 3.0 + maxDemon in four case studies of the impact of genetic and drug class variation on functionally conserved protein dynamics in ubiquitin, TATA binding protein and Hsp90.

Introduction

The physicist Richard Feynman is said to have once famously quipped, 'all biology is ultimately due to the wiggling and jiggling of atoms'. Stated with more precision, Feynman's conjecture would imply that all biological function at the molecular level can actually be understood by analyzing rapid atom motions in biomolecular structures as they alter their functional state(s). Many decades later, functional shifts in molecular dynamics are now being illuminated by structural and computational biology. Examples of such functional dynamics include the destabilization of inter-residue contacts, in both disease malfunction and signal activation during phosphorylation, as well as stabilization of inter-residue contacts during protein folding, the formation of larger complexes, and various other binding interactions to small molecules. And while the functional role of rapid vibrations revealed by short term dynamic simulations has been debated in the past, more recent empirical and computational studies have clearly demonstrated that differences in rapid and directed vibrations can drive longer term functional conformational change (1, 2). From a broader perspective, if Feynman's conjecture is true, then the specific details of a given protein systems biomolecular dynamics will represent a potentially large source of latent variability in our functional understanding of the genome; a problem largely ignored by the considerable amount of current analyses of static forms of 'omic' data (i.e. DNA sequence, transcript level, and protein structure)(3). However, in the last decade, simultaneous advances in the development of gaming graphics hardware and biomolecular force fields has elevated our ability to computationally simulate molecular dynamics (MD) long enough to capture the most biologically relevant timescales for protein function (4, 5) (i.e. ns to μs). And now, the application of proper statistical methods to comparatively framed questions about protein dynamics can potentially enable meaningful comparisons of large ensembles of shorter timed framed MD simulation (6). But despite the rich visual complexity of data and movie images generated by MD software, the functional interpretation of these large complex data sets created by modern MD simulations poses a serious challenge to current users, especially where large ensembles of runs are collected and compared. A potential solution to this problem exists with the application of machine learning to the feature extraction and classification of the dynamic differences between ensembles of MD runs representing the functional states of biomolecular systems (e.g. before/after chemical mutation or binding interaction). In other words, high performance computation to generate simulated protein motions for comparison can be effectively partnered with high performance methods for optimally extracting and learning what the underlying dynamic feature differences are that define the functional protein states. Although machine learning has recently been applied to individual MD studies for a variety of specific tasks (7–9), there is no current software platform for the general application of machine learning to comparative protein dynamics.

In 2018, we released DROIDS v1.2 and v2.0 (Detecting Relative Outlier Impacts from molecular Dynamic Simulation), a GPU accelerated software pipeline designed for calculating and visualizing statistical comparisons of protein dynamics drawn from large repeated ensembles of short dynamic simulations representing two protein states (6). This application allowed simple visual and statistical comparison of protein MD ensembles set up in any way the user wanted to define them. Here, we announce the release of DROIDS v3.0, which comprises multiple pipelines specifically tailored for specific functional comparisons. These include different temperatures, different binding states (i.e. to DNA, drugs, toxins or natural ligands), or divergent genetic mutant states. Many other new improvements to DROIDS are listed below, however we also include a major new machine learning tool, maxDemon v1.0, a multi-machine learning post-processing application capable of training on the data representing the normal functionally divergent dynamic states compared in DROIDS, and subsequently identifying the effects on these learned dynamics upon the deployment of new MD simulations representing genetic or drug class variants of interest. Thus, much like James Clerk Maxwell's mythical creature (10), maxDemon derives important information from all atom resolution observation of dynamic motion. The three primary features/aims of our newly expanded software is to (A) improve user experience in comparative protein dynamics, (B) enable the local detection of functionally conserved protein dynamics, and to (C) enable the assessment of the local dynamic impacts of both genetic and drug class variants within the functional context of protein system of interest. Functionally conserved dynamics is defined as 'sequence-dependent dynamics' discovered after training machine learners on the functional state ensembles derived with DROIDS. Conserved dynamics are detected through the deployment of learners on new MD simulation runs that were setup identically to the dynamic state recognized or queried by the MD ensemble training sets. We expect that functionally conserved dynamics will be sequence encoded and therefore should display a repeated position dependent signature in our learned pattern profiles whenever MD runs are set up identically to MD upon which learners were trained. Therefore, a significant local canonical correlation (i.e. Wilk's lambda) between learning performance profiles of self-similar MD runs can be used to detect local regions of conserved protein dynamics. By extension, mutational impacts of genetic or drug class variants on the functionally conserved dynamics can be quantified by their effects that range significantly beyond that observed in the self-similar runs. Thus when canonical correlations of variants differ significantly from the self-correlation observed, according to a bootstrap or z-test, we plot the relative entropy or magnitude of impact defining how the variant's dynamics differs from the self-similar dynamics of the normal functioning protein. Because the machine learning algorithms are ultimately trained on MD data representing this normal functioning state(s), this metric of impact is potentially very specific to how a given mutation or drug impacts a specific protein. Thus, it potentially gives considerably more functional relevance to our results when compared to more general database-derived metrics of mutational tolerance (e.g. SIFT, PolyPhen2, MegaMD etc.).

In Table 1, we list five primary methods pipelines in available in DROIDS 3.0+maxDemon to address functional questions in comparative protein dynamics. In our results and discussion here, we present data on four case studies of functional protein dynamics that include feature extraction and classification of (A) a simple temperature shift in ubiquitin dynamics, (B) mutational tolerance in ubiquitin dynamics, (C) mutation specific impacts of DNA binding of TATA binding protein, and (D) comparison of binding dynamics of drug class variants that mimic ATP binding in Hsp90.

Materials and Methods

Overview

The machine learning-based detection of variant impacts on functional protein dynamics presented here is outlined schematically in Figure 1. Upon launch, the DROIDS graphical user interface (GUI) will help the user write a control file for required working path directories on their system (first use only) and then proceeds to a main GUI outlining the various types of comparisons that can be generated (as detailed in Table 1) and the number of GPU available on the system. The next step is provides a user-friendly GUI to control and schedule Amber16/18 GPU-accelerated MD simulation to generate ensembles of short MD runs representing two functional protein states wanting to be compared. These functional comparisons are not limited, but would typically entail the impact of mutation (comparing dynamics before and after one or more amino acid replacements), the impact of an environmental change (comparing two states of temperature of solvent set up), or the impact of a molecular interaction (comparing bound to an unbound state). The DROIDS GUI will lead users through the building of a structural alignment file using UCSF Chimera's MatchMaker and Match-Align tools. This will be needed later by the graphics components of DROIDS to make sure that only homologous regions of structures are being compared and analyzed. In this application, where the user is primarily interested in genetic or drug class variant impacts on an interactive signaling function, the typical training ensembles generated by DROIDS for further analysis with maxDemon should represent the normal binding function of the wild-type protein and therefore the bound vs unbound comparison would typically be used. A PDB file of the bound state can be the starting point and an unbound PDB model can be saved after deleting chains in the original file. If a small molecule ligand interaction is under study and requires application of an additional force fields such as GAFF, than an additional file representing only the ligand should also be generated and saved for preparation with antechamber software prior to building the solvent models using teLeAP. The GUI will pop open the .bat files that control more details of the simulation setup allowing advanced users to write more lines into the teLeAP modeling prep (e.g. to alter the water box dimensions, the water model itself, or to add additional ions beyond simple charge neutralization). The user should read all warnings provided to the terminal at this stage by the Amber software. Our GUI script will also double check the sizes of the files generated at this stage and will supply a warning if teLeap failed altogether to set up the complete model system for simulation. Upon successful setup the user can launch all the MD runs from the GUI. The requested jobs are automatically scheduled to each GPU one at a time by our software. When finished, the user can easily generate rmsf data by using the GUI to setup and launch cpptraj software provided in Ambertools. Thus the total process from file preparation, MD production and post-processing for DROIDS analysis by simply working down the buttons on each GUI from top to bottom and subsequently following the directions on the main terminal. After MD simulation and post-processing, DROIDS will take users to a second GUI for generating R plots and analyses for statistically comparing the dynamics, and then to a third GIU for visualization and movie generation. We refer users to our user manual and previous publication for more details. This third GUI has buttons to optionally launch our new machine learning application maxDemon if users wish to go beyond simple comparative protein dynamics and investigate novel simulations utilizing the DROIDS MD ensembles as a training set for subsequent machine learning. Users can select any of seven different machine learning classification algorithms including K-nearest neighbors, naïve Bayes, linear discriminant analysis, quadratic discriminant analysis, random forest, adaptive boosting and support vector machine (kernels include linear, tuned polynomial, laplace and tuned radial basis function). We generally recommend users select all of them. The random forest and adaboost algorithms are programmed to use all available CPU cores found on the system.

To detect functionally conserved dynamics after training on two MD ensembles representing a query and reference functional state of a protein (e.g. bound and unbound), the machine learners are deployed on two new MD simulations across space and time (i.e. fluctuations backbone atoms of individual amino acids over subdivided time intervals) that are identical to the original query state. Any sequence dependent or 'functionally conserved' dynamics can be recognized through a significant canonical correlation in the profile of the overall learning performance along the amino acid positions for the two similar state runs. In effect, this metric defines dynamics that are functionally conserved by capturing a signal of significant self-similarity in dynamics that localizes to a specific part of the protein backbone.

Eqn. 1

$$conserved_{dynamics} = significant\ (CC_{self})$$

The impacts of dissimilar states caused by altered amino acid sequence or different binding partners can be subsequently assessed through their local effect on this canonical correlation (i.e. conserved dynamics). Here, we introduce a metric of relative entropy relating the canonical correlations in both the self-similar and altered state. In essence, this is a metric of the 'impact' of a given genetic or drug class variant within the context of normal functioning dynamics. For example, when trained on a natural binding interaction (e.g. DROIDS analysis comparing a DNA binding protein in its bound and unbound states), novel MD simulations with a variety of amino acid replacements can be deployed to see whether the learners can recognize the functional dynamics in the mutant forms. In this case, functionally tolerated mutations will result in dynamics that do not vary outside of the normal bounds of the self-similar runs, whereas functionally intolerant mutations will result in significant deviations from self-similarity of motion.

Eqn. 2

$$variant_{impact} = CC_{self} * log \frac{CC_{variant}}{CC_{self}}$$

Variant impacts are only plotted when they differ significantly the self-similar MD runs according to a simple Z test.

More detailed instructions to users are included with our DROIDS 3.0 user manual available in the GitHub repository.

The main repository for DROIDS 3.0 and maxDemon 1.0 can be found here. Please follow the link to "Releases" and download the latest release as .tar.gz or .zip file

https://github.com/gbabbitt/DROIDS-3.0-comparative-protein-dynamics

and DOI: 10.5281/zenodo.3358976 concurrent with this publication

https://zenodo.org/record/3358976#.XURVkOhKiiM

We also post various videos of examples using DROIDS, video tutorials, and ongoing projects here

https://www.youtube.com/channel/UCJTBqGq01pBCMDQikn566Kw

To enhance the user experience and scientific utility, DROIDS v3.0 offers many new features beyond earlier major release versions 1.2 and 2.0.   These are summarized below.

- New GUI organization directs users to specific comparative tasks/applications in Table 1
- A new control file builder for managing path dependencies in Linux is included
- Amber16/18 support has been beta tested and is defined via paths.ctl file
- Single or dual GPU user options are available for faster analyses
- Automated structure prep (dry and reduce) via pdb4amber is now included in the GUI. The 'reduce' variable is optional allowing users to either setup their own protonation states ahead of DROIDS, or simply allow DROIDS to hydrogenate the input structures entirely.
- Program/package dependency installer script named 'DROIDSinstaller.pl' is included. It will lead users through all dependencies required after a fresh Linux build, including CUDA libraries and tools required for Nvidia GPU accelerated Amber in the Linux environment
- KL divergence (= relative entropy) definition of dFLUX is now included as an option providing a richer color mapping of dFLUX in images and movies than the simple averaging algorithm offered in earlier DROIDS versions

dFLUX collected as an angstrom average

$$dFLUX_{aa} = \left(\sum_{i=1}^{4} FLUX_{atom}\right)_{query} - \left(\sum_{i=1}^{4} FLUX_{atom}\right)_{reference}$$

$$dFLUX_{chain} = \sum_{i=1}^{L} |dFLUX_{aa}|$$

where L = number of structurally homologous amino acids in reference chain

$i = 4$ *or avg atom flux on N, O, CA and C backbone atoms*

dFLUX collected as symmetric KL divergence

$$dFLUX_{aa} = \begin{bmatrix} D_{KL}(FLUX_{query}|FLUX_{reference}) \\ + D_{KL}(FLUX_{reference}|FLUX_{query}) \end{bmatrix} \Big/ 2$$

where

$$D_{KL}(FLUX_{query}|FLUX_{reference}) = \sum_i FLUX_{query}(i) log \frac{FLUX_{query}(i)}{FLUX_{reference}(i)}$$

$$D_{KL}(FLUX_{reference}|FLUX_{query}) = \sum_i FLUX_{reference}(i) log \frac{FLUX_{reference}(i)}{FLUX_{query}(i)}$$

and

$i = 4$ *or avg atom flux on N, O, CA and C backbone atoms*

- Binding interaction analysis for both protein-DNA and protein-ligand systems is now offered with dedicated GUI for these comparisons. Protein-ligand system setup includes QMMM preprocessing in Antechamber and SQM.
- LeAP control files for explicit solvent runs are now presented for advanced user modifications (e.g. changing ion concentration, water model, water box dimension of volume).
- Dedicated GUI allowing genetic mutation placement (on DNA or AA) are included for setting up variants to analyze
- Self-stability and temperature shift analysis has its own dedicated GUI, allowing users to copy the input pdb file to compare MD ensembles generated on identical structures at the same of at different temperatures
- MaxDemon 1.0 - machine learning based detection of functionally conserved dynamic regions
- MaxDemon 1.0 - machine learning based impact assessment of variants (genetic, structural or binding)
- Dynamic visualization and movie rendering of machine learning classification performance
- Virtual reality and ChimeraX compatibility is also supported (additional information and download code can be found here
  https://cxtoolshed.rbvi.ucsf.edu/apps/moleculardynamicsviewer
  https://github.com/kdiller713/ChimeraX_MolecularDynamicViewer

To demonstrate the performance and utility of DROIDS 3.0 + maxDemon 1.0, we ran the following four comparative case studies using the PDB IDs mentioned below. Bound and unbound files were created by deleting binding partners in UCSF Chimera and resaving PDBs (e.g. 3t0z_bound.pdb, 3t0z_unbound and 3t0z_ligand). Each MD run ensemble consisted of 200 production runs at 0.5ns each explicitly solvated in a size 12 octahedral water box using TIP3P solvent model with constant temperature under an Anderson thermostat. The models were charge neutralized with both Na+ and Cl- ions. The heating and equilibration runs prior to production were 0.3ns and 10ns respectively. Prior to heating 2000 steps of energy minimization were also performed.   All seven available machine learning classifiers were trained on the functional MD ensembles and deployed upon new 5 ns production runs for each variant analyzed.

Case study 1 (figure 2) – PDB ID = 1ubq – to analyze self-stability and effect of temperature shift in ubiquitin

Case study 2 (figure 3)– PDB ID = 2oob – to analyze functional binding of ubiquitin to ubiquitin ligase and impacts of several tolerance pre-classified genetic variants

Case study 3 (figure 4) – PDB ID = 1cdw – to analyze functional binding of TATA binding protein to DNA and impacts of several genetic variants

Case study 4 (figure 5) – PDB ID = 3t0z – to analyze functional ATP binding in Hsp90 and subsequent impacts of six inhibitor drug variants

Results and Discussion

To demonstrate the variety of comparative analyses that can be addressed with the new release of DROIDS 3.0 and maxDemon 1.0, we chose four different case studies of comparative protein dynamics. These included (A) an analysis of self-stability and temperature effects in free ubiquitin, (B) a functional genetic variant analysis of ubiquitin and ubiquitin ligase binding interaction, (C) a functional genetic variant analysis of DNA binding in TATA binding protein, and (D) a drug class variant analysis of compounds targeting ATP binding region of Hsp90 heat shock protein.

Machine learning analysis of impacts due to simple environmental temperature shift

We first ran a null comparison as a 'sanity check' by running a query and reference ubiquitin (11) MD at the same temperatures (both 300K) and same solvent conditions. The DROIDS analysis (Figure 2A-C) showed identical atom fluctuation profiles along the backbone and a random dFLUX profile indicative of nonsignificant differences due to small random local thermal differences in the training sets. The machine learning classification plots on new MD runs vary randomly around 0.5 reflecting the fact that the learning algorithms had no features to train on (Figure 2D). As expected, no significantly conserved dynamics were identified either (Figure 2E). By contrast, a protein dynamic comparison run with a 50K temperature difference (Figure 2 F-H) shows a much higher machine learner performance upon deployment (i.e. 70-80% successful classification – Figure 2I). Because environmental temperature shifts are not expected to reflect evolutionary conserved dynamics (i.e. are not position dependent), they subsequently do not result in canonical correlations in the learning profiles (Figure J). Representative time slices of the positional classifications in each of these experiments are shown in K and L resp and indicate that our machine learning is capable of extracting and identifying simple differences in dynamics due to temperature. Another interesting observation here was the slightly higher learning performance of the simpler machine learning methods QDA and LDA over others at all sites in the temperature shifted example. We interpret this to be related to the fact that underlying rmsf distributions are probably Gaussian, a critical assumption of these two models, with unequal variances caused by steric hindrances on the backbone. This would predict that QDA might outperform other learners in this situation and it appears that it does. We note that where more complex functional dynamics are concerned, the more sophisticated learning methods such as support vector machine and adaboost often perform slightly better than others. However, we also note that these performance differences are usually quite small and that all learning methods generally come to similar local conclusions about functional dynamics. We now move on to demonstrate and examine machine learning performance regarding more functional binding dynamics in ubiquitin and two other protein systems.

Machine learning analysis of impacts of genetic variants on functional binding interactions

To examine functional dynamics in ubiquitin, we conducted a DROIDS analysis comparing its two functional states, bound and unbound to the ubiquitin associated binding domain of ubiquitin ligase (12)(Figure 3A-D). This binding domain is highly conserved among many other proteins that interact directly with ubiquitin. The binding interaction greatly reduces the atom fluctuation in ubiquitin at 3 characteristic positions, two loop structures centered at LEU 8 and ALA 46 and a portion of beta sheet at the C terminus (Figure 3C). These three regions also drive significant differences in dynamics across the whole protein. In novel self-similar MD runs on the bound state, we successfully detect significant canonical correlations indicating conserved dynamics in these three regions with a broad expanse in conserved dynamics (Figure 3E and F) across the UBA region (Figure 3G). We tested a set of 24 mutations that included sites with the most and least tolerated effects on growth rate in vivo in yeast according to a study by Roscoe et al. (13).   In this study in vivo, nearly all mutations at E18 and G53 are tolerated while nearly all mutations at K48 and R72 are not. Ultimately, the causes of tolerance in these variants are not known, and do not necessarily invoke functional problems in dynamics. However, the impacts that we did observed in simulation were on average twice as strong in the intolerant backgrounds when compared to the mutation tolerant backgrounds. And, while we did not see large differences in mutational impacts on dynamics between tolerated and non-tolerated mutant groups, the 24 mutations analyzed all show a general trend of dynamic impact falling outside of most of the functional binding region (Figure 3H-K), suggesting that ubiquitin may have evolved a tertiary structure that allosterically translates dynamic impacts to less functional regions of the protein. One exception to this rule was demonstrated by the very large impact of R72D, centered squarely in the functionally conserved binding region of ubiquitin, and would obviously heavily disrupt electrostatic charge interactions as well.

We also conducted a DROIDS analysis comparing human TATA binding protein (TBP) (14) in its functionally bound and unbound states (Figure 4A-C). TBP exhibits a characteristic signature of dampening of atom fluctuation throughout its entire structure with most pronounced effects in two loop regions that interact with the minor groove of DNA (arrows in Figure 4A and 4C). Canonical correlations in new self-similar MD runs marking increased performance in classification were observed in these regions (Figure 4D) along with corresponding regions of conserved dynamics identified by significant Wilk's lamda (Figure 4E). Conserved dynamics from these loop areas are connected through the chains in the beta sheet region of TBP spanning the DNA major groove contact. Mutational impacts of four variants affecting the binding loop most proximal to the C terminal exhibited followed our expectation of increasing impact ordering from R192Q, R192K, R192polyD, and R192polyW (Figure 4G and 4H). The polyD and polyW mutations incorporated 5 sequential ASP or TRP residues centered at R192, both causing the loop region to become more rigid (causing increased negative dFLUX). We expected the strong functional binding affect observed across nearly all residues in this system would make it relatively highly tolerant to single amino acid substitutions, even when located in the most functional binding loop. In accordance with this idea, we found the most impactful multiple mutation (i.e. R192polyW) significantly affected the dynamics of nearly 6 times more local residues than the least impactful single substitution (i.e. R192Q).

Machine learning analysis of impacts of drug class variants targeting the ATP binding region of Hsp90

Lastly, we conducted a DROIDS analysis comparing the dynamics of Hsp90 chaperone, a common drug target for inhibitors in many cancer therapies, in both its ATP bound and unbound states. The binding of ATP was discovered to significantly destabilize three co-localized alpha helical regions of the protein adjacent to and extending from the ATP binding site (Figure 5A-D). MaxDemon analysis confirmed the dynamics of this region to be highly conserved in new MD runs (Figure 5D-G). We also analyzed the impacts of the six drug class variants targeting the ATP site (15–19), but interacting differently with residues in this region (Figure 5H). The contacts in the ATP binding site are shown in Figure 5I. While the localized patterns of impacts of the drug variants were all quite similar to ATP (Figure 5J), the drug variants that most closely mimicked the contacts of ATP (i.e. geldanamycin) had far less impact on dynamics than variants that interacted very differently with the binding pocket (i.e. benzamide SNX1321 and inhibitor FJ1(Figure H-I). We feel that this finding demonstrates that while it is important to be able to target a druggable protein binding site (20), researchers should also consider how these various chemicals might alter the natural dynamics of the system. In situations where a drug might too closely mimic the dynamic effects of a natural activator like ATP, a hyperactivation response might occur in non-tumor cells leading to secondary cancer (21–23). Alternatively, other situations may require drug targeting that does not alter the natural dynamic behavior too much, potentially activating proteolytic systems in the cell. Our software allows more detailed investigations of these potential dynamic impacts of drug class variants.

Conclusion

We provide a well demonstrated method and user-friendly software pipeline for conducting statistically sound comparative studies of large ensembles of comparative protein dynamics. The method/software also now provides machine learning based extrapolations of effects on novel MD simulations representing various functional variants of interest to the user. While there currently is at least one other software allowing users to connect sequence-based evolutionary metrics to protein dynamics (24), our method/software is unique in that regions of functional conservation are identified by analyzing self-similar features of dynamics themselves rather than traditional sequence-based approaches, which do not necessarily assume that a protein function has a strong dynamic component. Another advantage to our method/software is that our functional impacts (i.e. mutational tolerance) are defined solely within the context of protein dynamic system being simulated. This provides a much deeper look into protein specific function than current genomic and proteomic database methods of predicting mutational tolerance (25, 26) currently allow. As GPU technology continues to advance at a rapid pace over the next few years, our method/software may have profound potential application to the development of precision and personalized medicine, where understanding the detailed interaction between genetic and drug class variants within the context of specific protein dynamic systems will be greatly needed.

Supporting Material

The main repository for DROIDS 3.0 and maxDemon 1.0 can be found at the GitHub repository link below. Please follow the link to “Releases” and download the latest release as .tar.gz or .zip file https://github.com/gbabbitt/DROIDS-3.0-comparative-protein-dynamics

We also post various videos of examples using DROIDS, video tutorials, and ongoing projects here

https://www.youtube.com/channel/UCJTBqGq01pBCMDQikn566Kw

A 'live version' of the figures in this manuscript is also available on our YouTube channel.

Author Contribution

GAB and EPF conceived the project and method. All authors contributed to the code base. GAB and LEA worked on beta testing and debugging.

Acknowledgements

We acknowledge the lab of Dr. Andre O. Hudson for beta testing our software during its early development. We acknowledge Dr. Miranda Lynch and her colleagues at the Hauptman-Woodward Medical Research Institute for many helpful comments.

References

1. Henzler-Wildman, K.A., M. Lei, V. Thai, S.J. Kerns, M. Karplus, and D. Kern. 2007. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature. 450: 913–916.

2. Niessen, K.A., M. Xu, A. Paciaroni, A. Orecchini, E.H. Snell, and A.G. Markelz. 2017. Moving in the Right Direction: Protein Vibrations Steering Function. Biophys. J. 112: 933–942.

3. Babbitt, G.A., E.E. Coppola, M.A. Alawad, and A.O. Hudson. 2016. Can all heritable biology really be reduced to a single dimension? Gene. 578: 162–168.

4. Götz, A.W., M.J. Williamson, D. Xu, D. Poole, S. Le Grand, and R.C. Walker. 2012. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. 8: 1542–1555.

5. Salomon-Ferrer, R., A.W. Götz, D. Poole, S. Le Grand, and R.C. Walker. 2013. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9: 3878–3888.

6. Babbitt, G.A., J.S. Mortensen, E.E. Coppola, L.E. Adams, and J.K. Liao. 2018. DROIDS 1.20: A GUI-Based Pipeline for GPU-Accelerated Comparative Protein Dynamics. Biophys. J. 114: 1009–1017.

7. Díaz, Ó., J.A.R. Dalton, and J. Giraldo. 2019. Artificial Intelligence: A Novel Approach for Drug Discovery. Trends Pharmacol. Sci. .

8. Plante, A., D.M. Shore, G. Morra, G. Khelashvili, and H. Weinstein. 2019. A Machine Learning Approach for the Discovery of Ligand-Specific Functional Mechanisms of GPCRs. Mol. Basel Switz. 24.

9. Terayama, K., H. Iwata, M. Araki, Y. Okuno, and K. Tsuda. 2018. Machine learning accelerates MD-based binding pose prediction between ligands and proteins. Bioinforma. Oxf. Engl. 34: 770–778.

10. Maxwell, J.C. 2001. Theory of Heat. 9 Reprint edition. Mineola, NY: Dover Publications.

11. Vijay-Kumar, S., C.E. Bugg, and W.J. Cook. 1987. Structure of ubiquitin refined at 1.8 A resolution. J. Mol. Biol. 194: 531–544.

12. Peschard, P., G. Kozlov, T. Lin, I.A. Mirza, A.M. Berghuis, S. Lipkowitz, M. Park, and K. Gehring. 2007. Structural basis for ubiquitin-mediated dimerization and activation of the ubiquitin protein ligase Cbl-b. Mol. Cell. 27: 474–485.

13. Roscoe, B.P., K.M. Thayer, K.B. Zeldovich, D. Fushman, and D.N.A. Bolon. 2013. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J. Mol. Biol. 425: 1363–1377.

14. Nikolov, D.B., H. Chen, E.D. Halay, A. Hoffman, R.G. Roeder, and S.K. Burley. 1996. Crystal structure of a human TATA box-binding protein/TATA element complex. Proc. Natl. Acad. Sci. U. S. A. 93: 4862–4867.

15. Stebbins, C.E., A.A. Russo, C. Schneider, N. Rosen, F.U. Hartl, and N.P. Pavletich. 1997. Crystal structure of an Hsp90-geldanamycin complex: targeting of a protein chaperone by an antitumor agent. Cell. 89: 239–250.

16. Li, J., L. Sun, C. Xu, F. Yu, H. Zhou, Y. Zhao, J. Zhang, J. Cai, C. Mao, L. Tang, Y. Xu, and J. He. 2012. Structure insights into mechanisms of ATP hydrolysis and the activation of human heat-shock protein 90. Acta Biochim. Biophys. Sin. 44: 300–306.

17. Immormino, R.M., Y. Kang, G. Chiosis, and D.T. Gewirth. 2006. Structural and quantum chemical studies of 8-aryl-sulfanyl adenine class Hsp90 inhibitors. J. Med. Chem. 49: 4953–4960.

18. Fadden, P., K.H. Huang, J.M. Veal, P.M. Steed, A.F. Barabasz, B. Foley, M. Hu, J.M. Partridge, J. Rice, A. Scott, L.G. Dubois, T.A. Freed, M.A.R. Silinski, T.E. Barta, P.F. Hughes, A. Ommen, W. Ma, E.D. Smith, A.W. Spangenberg, J. Eaves, G.J. Hanson, L. Hinkley, M. Jenks, M. Lewis, J. Otto, G.J. Pronk, K. Verleysen, T.A. Haystead, and S.E. Hall. 2010. Application of chemoproteomics to drug discovery: identification of a clinical candidate targeting hsp90. Chem. Biol. 17: 686–694.

19. Austin, C., S.N. Pettit, S.K. Magnolo, J. Sanvoisin, W. Chen, S.P. Wood, L.D. Freeman, R.J. Pengelly, and D.E. Hughes. 2012. Fragment screening using capillary electrophoresis (CEfrag) for hit identification of heat shock protein 90 ATPase inhibitors. J. Biomol. Screen. 17: 868–876.

20. Vajda, S., D. Beglov, A.E. Wakefield, M. Egbert, and A. Whitty. 2018. Cryptic binding sites on proteins: definition, detection, and druggability. Curr. Opin. Chem. Biol. 44: 1–8.

21. Poulikakos, P.I., C. Zhang, G. Bollag, K.M. Shokat, and N. Rosen. 2010. RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. Nature. 464: 427–430.

22. Hatzivassiliou, G., K. Song, I. Yen, B.J. Brandhuber, D.J. Anderson, R. Alvarado, M.J.C. Ludlam, D. Stokoe, S.L. Gloor, G. Vigers, T. Morales, I. Aliagas, B. Liu, S. Sideris, K.P. Hoeflich, B.S. Jaiswal, S. Seshagiri, H. Koeppen, M. Belvin, L.S. Friedman, and S. Malek. 2010. RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth. Nature. 464: 431–435.

23. Cichowski, K., and P.A. Jänne. 2010. Drug discovery: inhibitors that activate. Nature. 464: 358–359.

24. Bakan, A., A. Dutta, W. Mao, Y. Liu, C. Chennubhotla, T.R. Lezon, and I. Bahar. 2014. Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics. 30: 2681–2683.

25. Kumar, P., S. Henikoff, and P.C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4: 1073–1081.

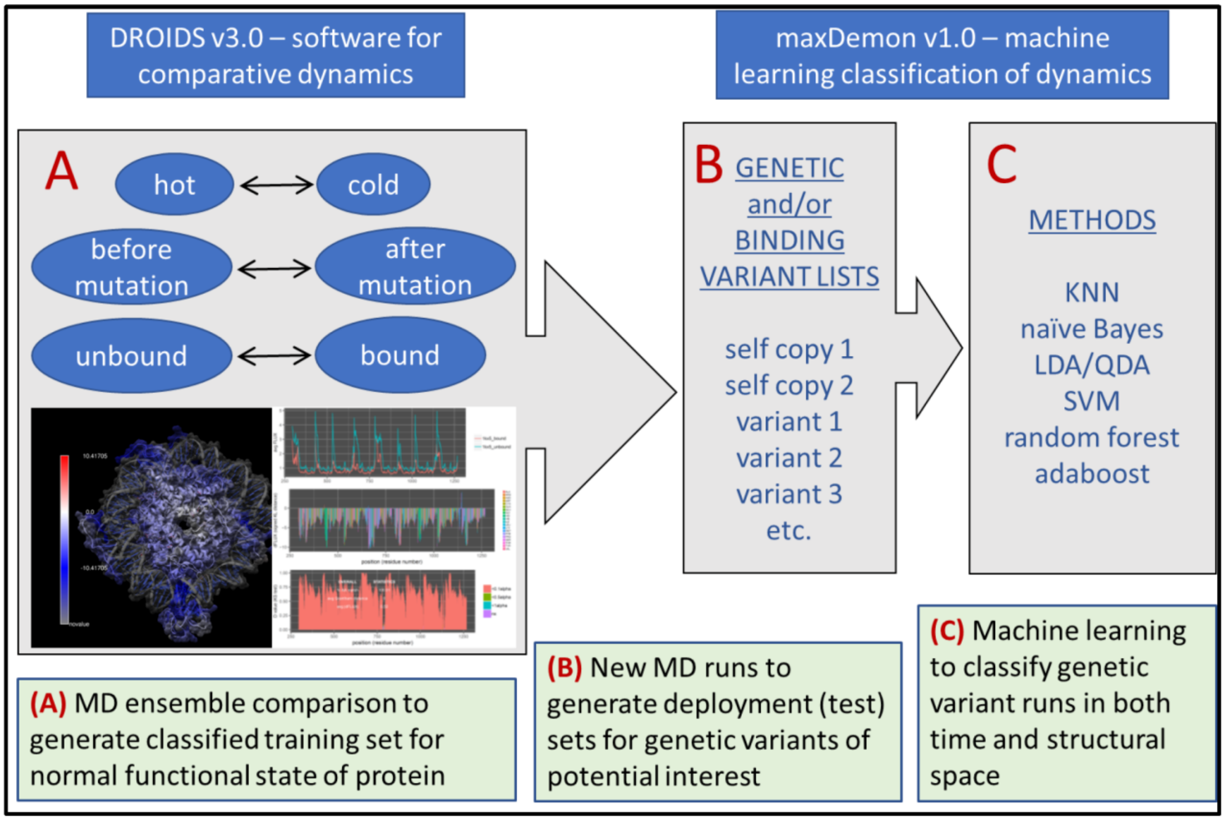
26. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Chapter 7: Unit7.20.

Table 1. Common learner assisted comparative protein dynamic investigations enabled by DROIDS 3.0 + maxDemon 1.0.

QUESTION	DROIDS 3.0 training comparison	Deployment of learners in maxDemon	Important notes
Measure dynamic tolerances of single protein to various genetic mutations	Two sets (ensembles) of MD on the same protein at the same temperature	MD run on one or more genetic mutant structures	Isolates MD impacts of mutation(s) from natural variability in self-similar dynamics
Measure dynamic tolerances of DNA binding interaction to genetic mutation(s)	MD ensembles comparing both the unbound and DNA bound protein	MD run on one or more unbound genetic mutant structures	Isolates MD impacts of mutation from natural binding function of the system
Measure dynamic tolerances of individual genetic differences to a given drug	MD ensembles comparing both the unbound and drug bound protein	MD run on one or more drug-bound genetic mutant structures	Isolates MD impacts of mutation from novel drug binding function of the system
Measure dynamic similarities of different drug candidates to	MD ensembles comparing both the unbound and ligand bound protein	MD run on one or more drug variant bound structures	Isolates MD impacts of drug candidates from the natural binding function of the ligand

natural ligand binding interaction			
Measure evolution of novel dynamics in paralog genes	MD ensembles comparing two ortholog proteins (i.e. same gene different species)	MD runs on one or more paralogs (i.e. duplicated genes in same species)	Isolates potential MD novelty in duplicated gene product from nonfunctional or neutral changes in different species

Figures Captions



(A) MD ensemble comparison to generate classified training set for normal functional state of protein

(B) New MD runs to generate deployment (test) sets for genetic variants of potential interest

(C) Machine learning to classify genetic variant runs in both time and structural space

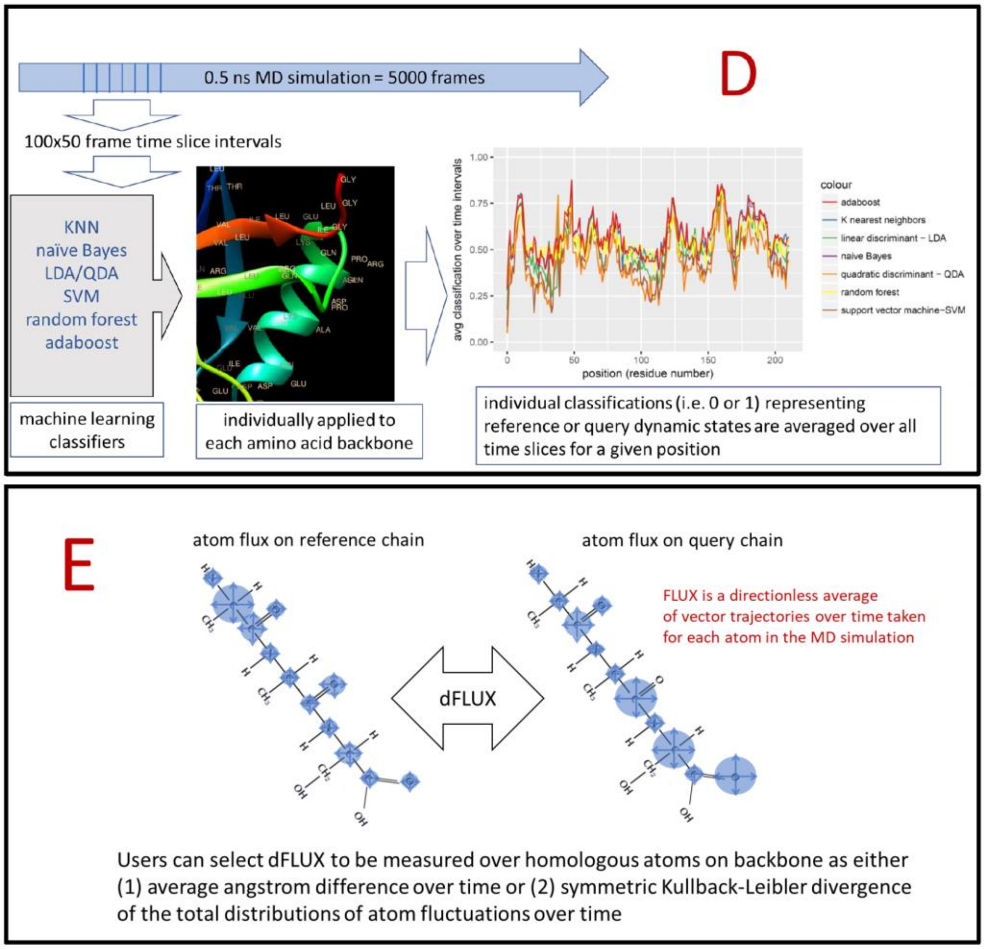

D
0.5 ns MD simulation = 5000 frames
100x50 frame time slice intervals
KNN
naïve Bayes
LDA/QDA
SVM
random forest
adaboost
machine learning classifiers
individually applied to each amino acid backbone
avg classification over time intervals
position (residue number)
colour
adaboost
K nearest neighbors
linear discriminant - LDA
naive Bayes
quadratic discriminant - QDA
random forest
support vector machine-SVM
individual classifications (i.e. 0 or 1) representing reference or query dynamic states are averaged over all time slices for a given position
E
atom flux on reference chain
atom flux on query chain
FLUX is a directionless average of vector trajectories over time taken for each atom in the MD simulation
dFLUX
Users can select dFLUX to be measured over homologous atoms on backbone as either (1) average angstrom difference over time or (2) symmetric Kullback-Leibler divergence of the total distributions of atom fluctuations over time

Figure 1. Schematic overview of DROIDS 3.0 + maxDemon 1.0 software for machine learning-based detection of variant impacts on functionally conserved protein dynamics. The pipeline starts with (A) generation of two large ensembles of molecular dynamic (MD) simulations that represent a functional comparison of protein states (e.g. mutation, binding or environmental change). The root mean square fluctuations (rmsf) of protein backbone atoms in these ensembles are comparatively analyzed/visualized (i.e. using DROIDS) and are also later used as pre-classified training data sets for machine learning (i.e. using maxDemon). Note: in the pictured DROIDS analysis of nucleosome shows overall dampening of rmsf in the histone core with maximal dampening where the histone tails cross the DNA helix (B) New MD simulations are generated on two structures self-similar to the query state of training as well as a list of functional variants, and (C) up to seven machine learning methods are employed to classify the MD in the self-similar and variant runs according to the functional comparison defined by the initial training step. (D) The relative efficiency of learning is defined by average value of classification (i.e. 0 or 1) over 50 frame time slices for each amino acid position and regions of functionally conserved dynamics are later identified by significant canonical correlations in this learning efficiency (i.e. Wilk's lamda) in self-similar MD runs. The impacts of variants are defined by relative entropy of variant MD compared to the MD in the self-similar runs and plotted when this entropy is significantly different from the variation in self-similarity (i.e. bootstrapped z-test). (E) A visual representation of the difference in local rmsf (dFLUX) is typically calculated using symmetric Kullback-Leibler (KL) divergence between the two distributions of rmsf in the training MD ensembles.

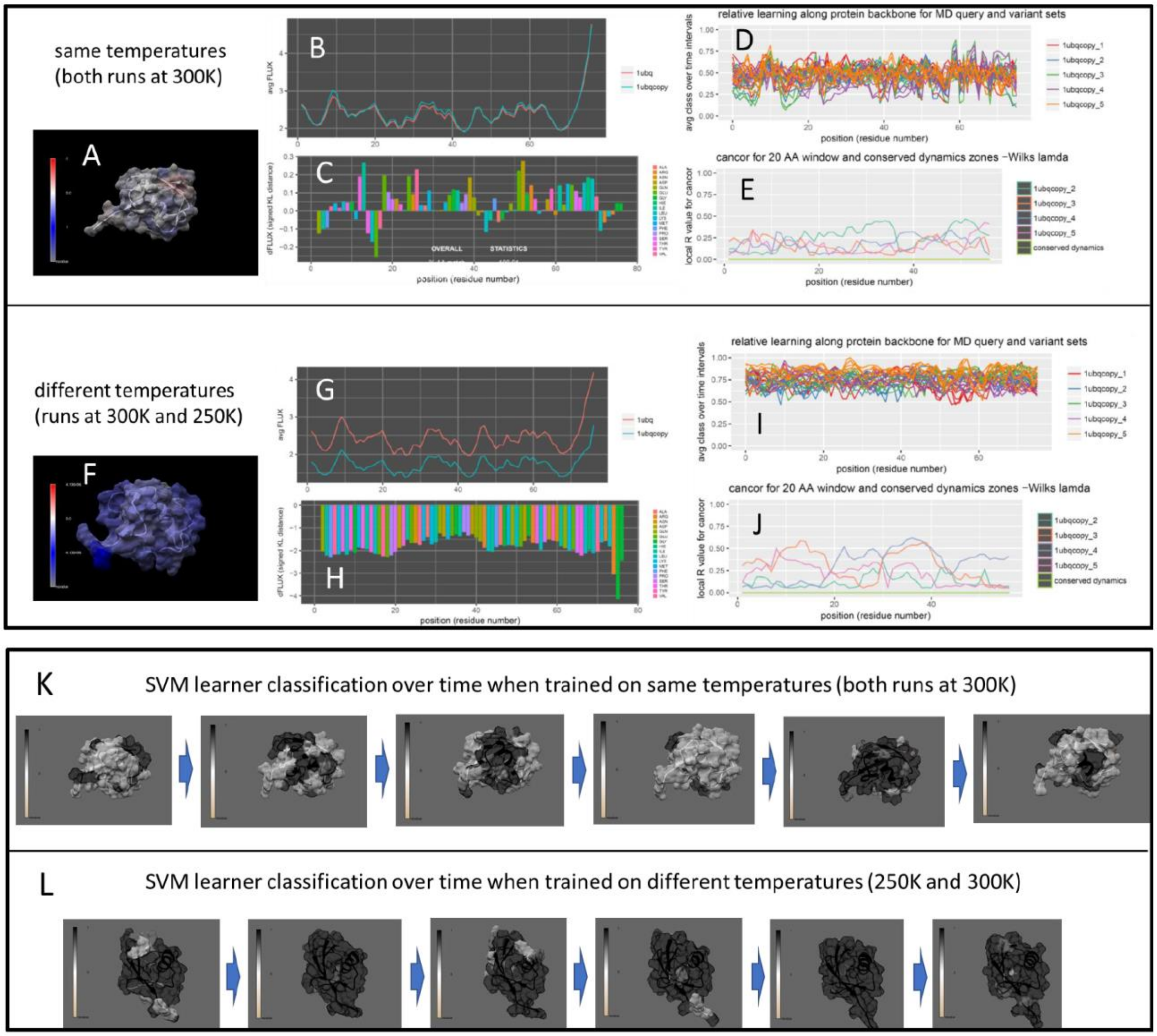


Figure 2. Analysis of environmental temperature change on non-functional ubiquitin dynamics. DROIDS image and analysis of random ubiquitin dynamics compared at the same (A-E) and different (F-J) temperatures. Note: blue color quantifies damped rmsf at temperature lowered by 50K. Note that relative learning is much higher when a temperature difference is modeled (D and I resp), however, as expected, neither comparison offers the machine learners a sequence-dependent profile by which to establish a signal of conserved dynamics (E or J). The learner classifications for the best performing learner in this case (quadratic discriminant function: QDA) is shown imaged on the ubiquitin structure over time in both the (K) random dynamics and (L) temperature dampened dynamics. (Movies of this can be observed in supplemental file A)

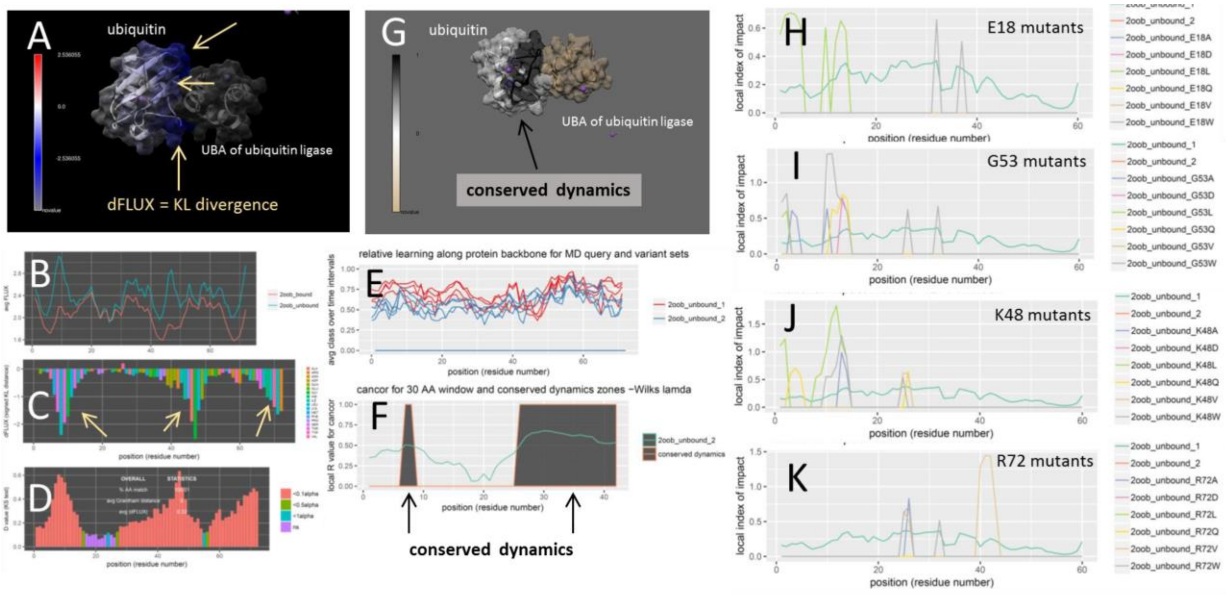


Figure 3. Analysis of mutational impact and tolerance on functional ubiquitin dynamics. (A) DROIDS image and analysis of ubiquitin bound to the ubiquitin associated binding domain (UBA) of ubiquitin ligase. Note: blue color quantifies damped rmsf at binding interface. (i.e. negative dFLUX) also by the (B) respective rmsf profiles of bound and unbound training states and (C) the KL divergence or dFLUX profile colored by residue. Arrows indicate most prominent dampening of rmsf near loops at THR 9, ALA 46 and C terminus. (D) Significant differences in these rmsf profiles is determined by multiple-test corrected two sample KS test. (E) Local relative learning efficiency of each machine learning method in self-similar testing runs are shown color-coded by run and regions of functionally conserved dynamics, determined via significant local canonical correlation are shown in dark gray in both (F) traditional N to C terminal plot as well as (G) structural image. The mutational impacts of 24 genetic variants (H-K: six variants at each or four sites) are shown all demonstrating lack of impact in functionally conserved regions of the binding interaction.

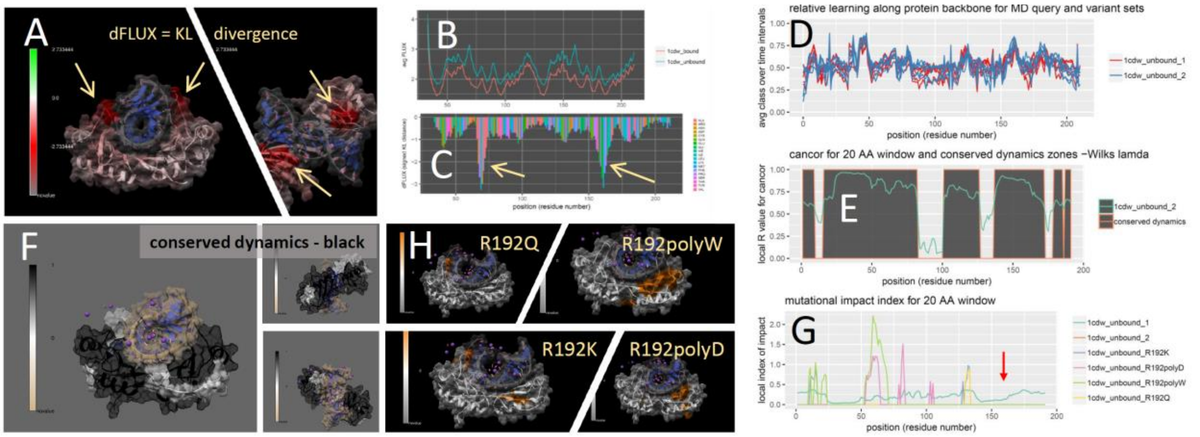
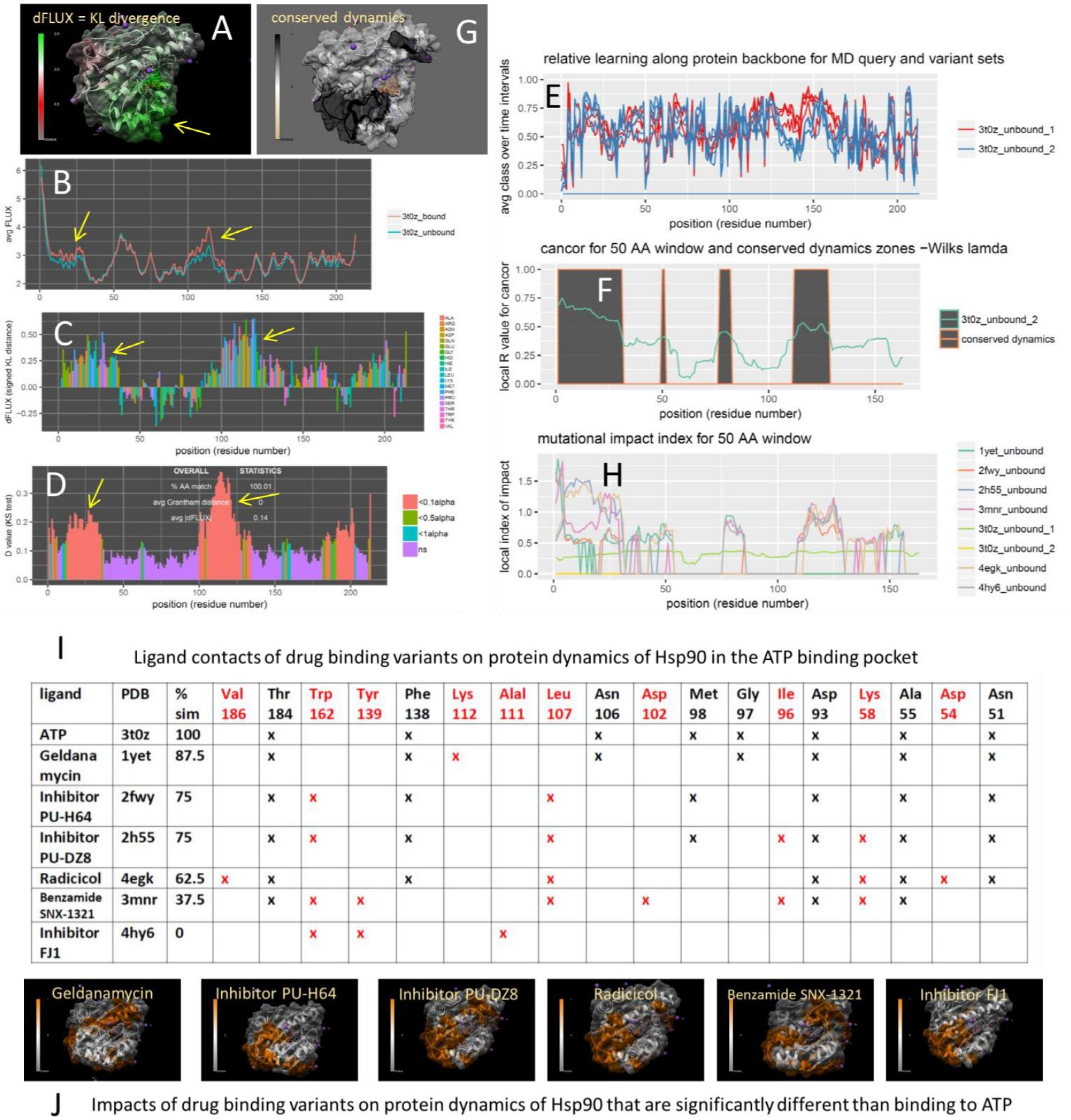


Figure 4. Analysis of mutational impact and tolerance on DNA binding in Tata Binding Protein (TBP). DROIDS image and analysis of TBP in DNA-bound and unbound states showing (A) colored TBP structure, (B) respective rmsf profiles and (C) KL divergence (dFLUX) plot. Note: arrows indicate functional binding loops in the DNA minor groove red color indicates dampened rmsf. maxDemon analysis (D-E) identifying conserved dynamics supporting both minor groove binding loops and (F) connecting them through the central region of the beta sheet in the main body of TBP closest to the DNA. Mutational impacts of 4 genetic variants with increasing impact one of the functional loops are also shown (G) plotted and (H) on the TBP structure. They are R192K, R192D, R192Q and polyW centered at R192 in 1cdw.pdb and and position 161 (red arrow) in plots (Note: 31 position offset is due to DNA in the original file).

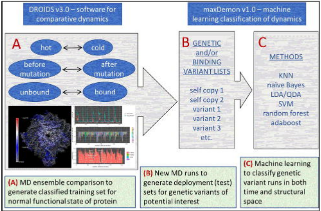


I Ligand contacts of drug binding variants on protein dynamics of Hsp90 in the ATP binding pocket

ligand	PDB	% sim	Val 186	Thr 184	Trp 162	Tyr 139	Phe 138	Lys 112	Alal 111	Leu 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 93	Lys 58	Ala 55	Asp 54	Asn 51
ATP	3t0z	100		x			x				x		x	x		x		x		x
Geldana mycin	1yet	87.5		x			x	x			x			x		x		x		x
Inhibitor PU-H64	2fwy	75		x	x		x			x			x			x		x		x
Inhibitor PU-DZ8	2h55	75		x	x		x			x			x		x	x	x	x		x
Radicicol	4egk	62.5	x	x			x			x						x	x	x	x	x
Benzamide SNX-1321	3mnr	37.5		x	x	x				x		x			x	x	x	x		
Inhibitor FJ1	4hy6	0			x	x			x											

J Impacts of drug binding variants on protein dynamics of Hsp90 that are significantly different than binding to ATP

Figure 5. Analysis of drug class variant binding in the ATP-binding domain of Hsp90. DROIDS image and analysis of Hsp90 in ATP-bound and unbound states showing (A) colored Hsp90 structure, (B) respective rmsf profiles and (C) KL divergence (dFLUX) plot and (D) significant differences in dynamics determined via the KS test. Note: arrows and green color indicate regions where rmsf is amplified in response to ATP binding. maxDemon analysis (E-G) identifying conserved dynamics connecting the ATP binding pocket and region of amplified rmsf. (H) Mutational impacts of 6 drug class variants targeting the ATP binding pocket of Hsp90 are plotted and (I) ordered by number of differences in structural contacts within the binding pocket. (J) Mutational impacts of these variants are demonstrated to predominantly impact the functionally conserved region of amplified rmsf thus mimicking the dynamic effect of functional ATP binding.



DROIDS v3.0 – software for comparative dynamics

maxDemon v1.0 – machine learning classification of dynamics

A

hot ⟷ cold

before mutation ⟷ after mutation

unbound ⟷ bound

B

GENETIC and/or BINDING VARIANT LISTS

self copy 1
self copy 2
variant 1
variant 2
variant 3
etc.

C

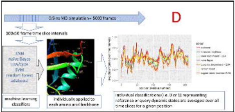
METHODS

KNN
naïve Bayes
LDA/QDA
SVM
random forest
adaboost

(A) MD ensemble comparison to generate classified training set for normal functional state of protein

(B) New MD runs to generate deployment (test) sets for genetic variants of potential interest

(C) Machine learning to classify genetic variant runs in both time and structural space

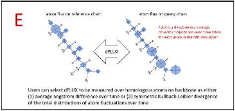


D
0.5 ns MD simulation – 5000 frames
100x50 frame time slice intervals
KNN
naive Bayes
LDA/QDA
SVM
random forest
adaboost
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
position (residue number)
color
adaboost
K nearest neighbors
linear discriminant - LDA
naive Bayes
quadratic discriminant - QDA
random forest
support vector machine (SVM)

machine learning classifiers

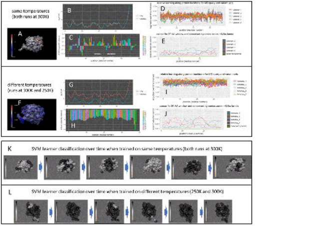
individually applied to each amino acid backbone

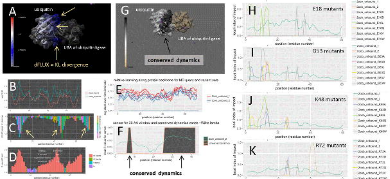
individual classifications (i.e. 0 or 1) representing reference or query dynamic states are averaged over all time slices for a given position

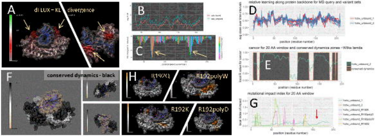


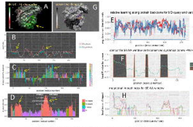
FLUX is a directionless average of vector trajectories over time taken for each atom in the MD simulation

Users can select dFLUX to be measured over homologous atoms on backbone as either (1) average angstrom difference over time or (2) symmetric Kullback-Leibler divergence of the total distributions of atom fluctuations over time


same temperatures
(both runs at 300K)
A
B
C
D
E
different temperatures
(runs at 300K and 250K)
F
G
H
I
J
K
SVM learner classification over time when trained on same temperatures (both runs at 300K)
L
SVM learner classification over time when trained on different temperatures (250K and 300K)


A
ubiquitin
UBA of ubiquitin ligase
dFLUX = KL divergence
G
ubiquitin
UBA of ubiquitin ligase
conserved dynamics
B
C
D
position (residue number)
relative learning along protein backbone for MD query and variant sets
E
2oob_unbound_1
2oob_unbound_2
position (residue number)
cancor for 30 AA window and conserved dynamics zones ~Wilks lambda
F
2oob_unbound_2
conserved dynamics
position (residue number)
conserved dynamics
H
E18 mutants
local index of impact
position (residue number)
2oob_unbound_1
2oob_unbound_2
2oob_unbound_E18A
2oob_unbound_E18D
2oob_unbound_E18L
2oob_unbound_E18Q
2oob_unbound_E18V
2oob_unbound_E18W
I
G53 mutants
local index of impact
position (residue number)
2oob_unbound_1
2oob_unbound_2
2oob_unbound_G53A
2oob_unbound_G53D
2oob_unbound_G53L
2oob_unbound_G53Q
2oob_unbound_G53V
2oob_unbound_G53W
J
K48 mutants
local index of impact
position (residue number)
2oob_unbound_1
2oob_unbound_2
2oob_unbound_K48A
2oob_unbound_K48D
2oob_unbound_K48L
2oob_unbound_K48Q
2oob_unbound_K48V
2oob_unbound_K48W
K
R72 mutants
local index of impact
position (residue number)
2oob_unbound_1
2oob_unbound_2
2oob_unbound_R72A
2oob_unbound_R72D
2oob_unbound_R72L
2oob_unbound_R72Q
2oob_unbound_R72V
2oob_unbound_R72W


A
dl LUX – KL divergence
B
hotlw_bound
hotlw_unbound
C
D
relative learning along protein backbone for MD query and variant sets
avg class over time intervals
position (residue number)
hotlw_unbound_1
hotlw_unbound_2
E
cancor for 20 AA window and conserved dynamics zones - Wilks lamda
local R value for cancor
position (residue number)
hotlw_unbound_2
conserved dynamics
F
conserved dynamics - black
G
mutational impact index for 20 AA window
local index of impact
position (residue number)
hotlw_unbound_1
hotlw_unbound_2
hotlw_unbound_R192K
hotlw_unbound_R192polyD
hotlw_unbound_R192polyW
hotlw_unbound_R192Q
H
R192Q
R192polyW
R192K
R192polyD



I Ligand contacts of drug binding variants on protein dynamics of Hsp90 in the ATP binding pocket

ligand	PDB	% sim	Val 186	Thr 184	Trp 162	Tyr 139	Phe 138	Lys 112	Ala 111	Leu 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 93	Lys 58	Ala 55	Asp 54	Asn 51
ATP	3t0z	100		•			<				•		x	>		<		>		•
Geldanamycin	1yet	87.5		•			•	x			•			x		•		•		•
Inhibitor PU-H64	2fwy	75		•	•		x			•			x			•		•		•
Inhibitor PU-DZ8	2fwz	75		•	•		<			•			x		•	<	x	>		•
Radicicol	4egk	62.5	x	•			<			•						<	x	>	•	•
benzamide inhibitor	3rlr	87.5		•	•	•				•		•			•	•	•	•		
Inhibitor [illegible]	4byo	0			•	x			>											

J Impacts of drug binding variants on protein dynamics of Hsp90 that are significantly different than binding to ATP