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 structural methods in bioinformatics. However, chaotic behavior in single protein trajectories requires statistical inference be obtained from large ensembles of molecular dynamic (MD) simulations representing the comparative functional states of a given protein. Meaningful interpretation of such a complex form of big data poses serious challenges to users of MD. Here, we announce DROIDS v3.0, a molecular dynamic (MD) method + software package for comparative protein dynamics, incorporating many new features including maxDemon v1.0, a multi-method machine learning application that trains on large ensemble comparisons of concerted protein motions in opposing functional states and deploys learned classifications of these states onto newly generated protein dynamic simulations. Local canonical correlations in learning patterns generated from self-similar MD runs are used to identify regions of functionally conserved protein dynamics. Subsequent impacts of genetic and drug class variants on conserved dynamics can also be analyzed by deploying the classifiers on variant MD runs and quantifying how often these altered protein systems display the opposing functional states. Here, we present several case studies of complex changes in functional protein dynamics caused by temperature, genetic mutation, and binding interaction with nucleic acids and small molecules. We studied the impact of genetic variation on functionally conserved protein dynamics in ubiquitin and TATA binding protein and demonstrate that our learning algorithm can properly identify regions of conserved dynamics. We also report impacts to dynamics that correspond well with predicted disruptive effects of a variety of genetic mutations. In addition, we studied the impact of drug class variation on the ATP binding region of Hsp90, similarly identifying conserved dynamics and impacts that rank accordingly with how closely various Hsp90 inhibitors mimic natural ATP binding.

Keywords

Molecular dynamics, machine learning, molecular evolution, pharmaceuticals, genetic variation, binding interaction

Statement of significance

We propose a statistical method as well as offer a user-friendly graphical interfaced software pipeline for comparing simulations of the complex motions (i.e. dynamics) of proteins in different functional states. We also provide both method and software to apply artificial intelligence (i.e. machine learning methods) that enable the computer to recognize complex functional differences in protein dynamics on new simulations and report them to the user. This method can identify dynamics important for protein function, as well as to quantify how the motions of molecular variants differ from these important functional dynamic states. For the first time, this method of analysis allows the impacts of different genetic backgrounds or drug classes to be examined within the context of functional motions of the specific protein system under investigation.

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 can ultimately be understood by analyzing rapid molecular motions in biomolecular structures as they alter or shift their functional state(s). Many decades later, these functional shifts in molecular dynamics are being illuminated by structural and computational biology. Examples of functionally altered dynamics include the destabilization of inter-residue contacts, in both disease malfunction and normal signal activation, as well as the 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 molecular dynamic (MD) simulations has been debated in the past, more recent empirical and computational studies have clearly demonstrated that differences in both 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 system's biomolecular dynamics will represent a potentially large source of latent variability in our functional understanding of the genome; a problem largely ignored by those disciplines currently generating the vast amounts of static forms of 'omic' type data (i.e. DNA sequence, transcript level, and protein structure)(3). However, in the last decade, simultaneous advances in the development of graphics hardware and biomolecular force fields has elevated our ability to computationally simulate MD long enough to capture ns to µs timescales for moderately-sized proteins (4, 5), and finally 'see' some of their functionally relevant motions. And now, the application of proper statistical comparisons of ensembles of short-timed framed MD simulation can potentially enable meaningful interpretations of comparative questions about protein dynamics (6). But due to the richly complex structure of data underlying the moving images generated by MD software, functional interpretation of modern MD simulations poses a serious challenge to current users, especially with comparatively-framed questions, where large ensembles of many production runs need to be generated and subsequently analyzed. 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. These ensembles can be designed to represent pair-wise functional states of biomolecular systems (e.g. before/after chemical mutation or binding). Therefore, the high performance accelerated computation used to generate simulated protein motions for comparison can be effectively partnered with high performance methods for optimally extracting and learning the underlying dynamic feature differences defining the different functional states of proteins. 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 now offers multiple pipelines tailored for specific functional comparisons of systems comprised of combinations of proteins, nucleic acids, and small ligand molecules. Comparisons can include different temperatures, different protein binding states (i.e. to DNA, drugs, toxins or natural ligands), or divergent genetic/epigenetic mutant states. We also include a major new machine learning tool, maxDemon v1.0, a multi-machine learning post-processing application for DROIDS that trains on the data representing the comparatively divergent functional dynamic states,

and subsequently identifies functionally conserved dynamics and genetic and/or drug class binding variant effects when deployed on new MD simulations representing these 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. Because the machine learning model we employ is trained on MD data representing normal functioning dynamic states of a protein, this metric of impact is highly context dependent to how a given mutation or drug impacts a specific protein. Thus, it potentially gives considerably more functional relevance to the analysis of variants when compared to more general database-derived metrics of mutational tolerance (e.g. SIFT, PolyPhen2 etc.). In Table 1, we list five primary methodological 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 of comparative dynamics and visualization with DROIDS v3.0

Our DROIDS method/software leverages several important key concepts when making comparisons between MD. The method utilizes structural alignment to restrict comparison of dynamics between individual homologous amino acids. The method also restricts comparison averaged over atoms common to all amino acids (i.e. backbone C, N, O and C_α). The method also employs statistical ensembling to make a robust comparison between protein dynamics in different functional states (6). While this is computationally intensive, it is necessary because of the inherent chaotic nature and unpredictability of single protein trajectory projections. This logic is analogous to the many 'storm tracks' repeatedly modeled by meteorologists to gain statistical confidence in a weather forecast, where an ensemble of model runs all with slightly different initial conditions has far more predictive power than any single simulation. In DROIDS, the user can decide how large the MD ensembles need to be based upon the inherent stability of the protein under investigation. Generally, an ensemble size of 200 to 300 MD runs at 0.5-1 ns will suffice for most proteins. The dynamics is summarized by calculation of root mean square fluctuations (*rmsf*) over constant time intervals represented by a constant number of image frames defined by the user (thus allowing *rmsf* values to be on comparable scales). The default number of frames (*i*) in the software for a given time slice is n=50 representing 0.01 ns of simulation time. The rmsf value is thus

$$rmsf = \frac{1}{4}\sum_{i=1}^{4}\sqrt{\frac{1}{n} * \sum_{j=1}^{n}\left(v_{jx} - w_x\right)^2 + \left(v_{jy} - w_y\right)^2\left(v_{jz} - w_z\right)^2}$$

where *v* represents the set of XYZ atom coordinates for *i* backbone atoms (C, N, O, and C_α) for a given amino acid residue over *j* time points and *w* represents the reference coordinate structure at the beginning of each MD production run for a given ensemble. Therefore, *rmsf* values as defined here represent molecular dynamics at the resolution of a single amino acid, the same resolution at which fine scale protein-level molecular evolution operates via amino acid replacement, insertion and deletion. The *rmsf* is the most also basic underlying functional quantity to extract from MD simulation as its underpins all hierarchical levels of motion (1). Two ensembles of *rmsf* values (a query set and a reference set) are compared to calculate average delta *rmsf* or *dRMSF*. The user can choose to see the average angstrom difference between sets of values, or more preferably the user can calculate the symmetric Kullback-Leibler divergence (i.e. relative entropy) between the two empirical statistical distributions of *rmsf*. The KL divergence generally provides a richer more informative view of dynamic differences with less loss of information than simple averaging. Thus *dRMSF* comparing *rmsf* values for two ensembles of size *m* for a given amino acid is

$$dRMSF_{avg} = \left(\sum_{i=1}^{m} rmsf\right)_{query} - \left(\sum_{i=1}^{m} rmsf\right)_{reference}$$

or

$$dRMSF_{KL} = \frac{1}{2} * \left[\left(\sum_{i=1}^{m} p\left(rmsf_{query}\right) * log\frac{p\left(rmsf_{query}\right)}{p\left(rmsf_{reference}\right)}\right) + \left(\sum_{i=1}^{m} p\left(rmsf_{reference}\right) * log\frac{p\left(rmsf_{reference}\right)}{p\left(rmsf_{query}\right)}\right)\right]$$

The resulting dRMSF values are color mapped to either still structures or movie images of the dynamics according to either 'temperature' scale where (+) dRMSF is red and (–) dRMSF is blue, or 'stoplight' scale where (+) dRMSF is green and (–) dRMSF is red. On both scale neutral values are shaded towards white.

Functional classification of new MD simulation with maxDemon v1.0

While users can easily employ DROIDS 3.0 to examine ensemble differences between functional genetic or binding states, the application of this knowledge to new MD simulation is nearly impossible due to the inherent complexity of the moving protein behavior. Our new post-processing software, maxDemon 1.0, uses machine learning to label or classify the differences learned by a previous DROIDS query/reference state comparison when applied to one or more new MD runs. The machine learning-based detection of variant impacts on functional protein dynamics presented here is outlined schematically in Figure 1. Similar to the statistics for comparative dynamics, the learning algorithms are also applied individually to each amino acid backbone's ensemble of *rmsf* values allowing for single residue resolution in the results. Learners are applied within the same user defined time slices of *rmsf* allowing for visualization of time resolution of classification of functional dynamic behaviors as well. The learning performance is summarized by tallying the average classification over all time slices for each amino acid. Thus, an average performance of 0.5 would indicate that the learners are not finding the functional states defined by and trained by the initial DROIDS comparative analysis. Canonical correlations in the positional performance plots are key in detecting sequence encoded functionally

conserved dynamics regions, as well as genetic and drug class variant impacts to these functional regions as well. This is described with more formality below.

Machine learning training and validation

The feature vectors (*X*) for machine learning are collections of *rmsf* values (x_i) labeled according to a query (q) and reference state (r) defined by the DROIDS MD comparison (i.e. where labels y_i are *q = 1* and *r = 0*).

$$X = \{(x_i, y_i)\}_{i=1}^{N}$$

The length of the vector (N) is defined by the length of the MD production run chosen by the user and the size of the ensemble of MD production runs taken. Thus if the user chooses an ensemble of 200 MD production runs each at a time length of 0.5 ns (= 2500 frames) and uses the default time interval of 50 frames to calculate any given interval of *rmsf* then the resulting feature vector will contain 20,000 data values for training (i.e. 10,000 values each for *q* and *r*).

Users create a 'stacked model' or meta-model containing up to 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 options include tuned linear, polynomial, laplace and radial basis function). R packages employed here are KNN, MASS, kernlab, randomForest and ada. We restricted machine learning to 'shallow' learning methods due to the relatively small datasets created when resolving dynamics of protein systems to short slices of time over single amino acids and also because of the robustness of the R packages when applied sequentially over time and structural space. Therefore, we do not support implementation of deep neural networks. For methodologically robust results on small proteins, we generally recommend users select all seven available methods. As real features of dynamics should be detectable by any method of learning, the agreement of classification obtained by the creation of a stacked model utilizing different methods makes the learning less sensitive to methodological artifacts. Depending upon system resources, users can choose to include or omit methods from four categories of learning (i.e. instance-based = KNN, probabilistic = NB / LDA / QDA, black box = SVM, and ensemble learning = randomForest / adaboost. Users will want to use as many as their system resources can handle, however for faster processing, a minimum of three of the seven learning methods can be chosen. Currently, most methods run on single CPU cores, however, the more CPU intensive methods of random forest and adaboost algorithms are programmed to use all available CPU cores found on the system. Omitting SVM is recommended for larger protein systems.

After learners are trained on the query and reference ensembles, they are validated on a new MD run that matches the state of the reference MD runs during training. For example, when analyzing a binding interaction where the reference ensemble of training runs are conducted in the unbound protein state, a new run will be conducted in the unbound state and a line plot of the machine learning performance (i.e. precision, recall and accuracy) will be generated for all positions on the protein. It would be expected that if comparative differences in dynamics observed in the training set have a genuine relation to function(s) defined during training, they will display repeated behavior in the new reference run and be identified by the stacked learning model generating local peaks in learning performance (i.e. accuracy) at functional regions (Figure 1D). Learner performance (*PERF*) for a given machine learning method is defined as

227 $$PERF = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + 0}{TP + 0 + FP + 0}$$

228 Where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative
229 classifications resp. The zero value terms arise because the validation is conducted on simulation
230 representing just the reference state of the DROIDS comparison (where $y_i = 0$). Therefore, accuracy,
231 and precision are algebraically collapsed to a single equivalent performance metric while recall is always
232 equal to 1.

233 ***Identifying regions of conserved dynamics***

234 Functionally conserved dynamics are defined as 'repeated or self-similar and sequence-dependent
235 dynamics' discovered after training machine learners on the functional state ensembles derived with
236 DROIDS. Conserved dynamics are detected via significant canonical correlations in position specific
237 learning performance patterns after the deployment of learners on new MD simulation runs that were
238 setup identically to the reference dynamic state defined by the MD ensemble training set. We expect
239 that functionally conserved dynamics will be sequence encoded and therefore should display a repeated
240 position dependent signature in our learned pattern profiles whenever MD runs are set up identically to
241 MD upon which learners were trained. Therefore, a significant local canonical correlation (i.e. Wilk's
242 lambda) between learning performance profiles of self-similar MD runs can be used to detect local
243 regions of conserved protein dynamics.

244 To detect functionally conserved dynamics after training and validation, an additional new MD
245 run matching the functional reference state is created (i.e. matching the MD validation run). The
246 learning performance of this run is compared across all selected learners and across space and time (i.e.
247 fluctuations backbone atoms of individual amino acids over subdivided time intervals) to the same
248 results from the MD validation run using a canonical correlation analysis. Any sequence dependent or
249 'functionally conserved' dynamics can be recognized through a significant canonical correlation in the
250 profile of the overall learning performance along the amino acid positions for the two similar state runs.
251 In effect, this metric defines dynamics that are functionally conserved by capturing a signal of significant
252 self-similarity in dynamics that co-localizes to a specific part of the protein backbone.

253 $$conserved_{dynamics} = significant\ (CC_{self})$$

254 Significantly conserved regions can be plotted upon both the positional local correlational value profile
255 (i.e. R value) and the reference structure of the protein, colored in dark gray on a light background.

256 ***Variant impact assessment***

257 By extension, mutational impacts of genetic or drug class variants on the functionally conserved
258 dynamics can be quantified by their effects that range significantly beyond that observed in the self-
259 similar reference runs. Thus when canonical correlations of variants differ significantly from the self-
260 correlation observed in functionally conserved regions, according to a bootstrap test, we can plot the
261 magnitude of impact defining how the variant's dynamics differs from the normal self-similar dynamics
262 of the in the conserved regions of the normal functioning protein. The impacts of dissimilar states
263 caused by altered amino acid sequence or different binding partners are assessed through their local
264 effect on this canonical correlation identifying conserved dynamics. We introduce a metric of relative
265 entropy relating the canonical correlations in both the self-similar and altered variant state. In essence,

266 this is a metric of the 'impact' of a given genetic or drug class variant within the context of normal
267 functioning dynamics. For example, when trained on a natural binding interaction (e.g. DROIDS analysis
268 comparing a DNA binding protein in its bound and unbound states), novel MD simulations with a variety
269 of amino acid replacements can be deployed to see whether the learners can still recognize the
270 functional dynamics in the mutant forms. In this case, functionally tolerated mutations will result in
271 functionally conserved dynamics that do not vary outside of the normal bounds of the self-similar runs,
272 whereas functionally intolerant mutations will result in significant deviations from self-similarity of
273 motion. An overall impact of a genetic and/or drug class binding variant on the conserved dynamic
274 regions is calculated by

275 $$variant_{impact} = CC_{self} * log \frac{CC_{variant}}{CC_{self}}$$

276 Comparative plots of local variant impacts are also able to be generated as absolute relative declines in
277 learning performance in conserved regions when compared to the validation MD run. Bootstrapped
278 confidence intervals on the plots allow for determining where these variant impacts are most
279 meaningful.

280 ***Four example applications (case studies)***

281 To demonstrate the performance and utility of DROIDS 3.0 + maxDemon 1.0, we ran the following four
282 comparative case studies using the PDB IDs mentioned below. Bound and unbound files were created by
283 deleting binding partners in UCSF Chimera and resaving PDBs (e.g. 3t0z_bound.pdb, 3t0z_unbound and
284 3t0z_ligand). Each MD run ensemble consisted of 200 production runs at 0.5ns each explicitly solvated
285 in a size 12 octahedral water box using TIP3P solvent model with constant temperature under an
286 Anderson thermostat. The models were charge neutralized with both Na+ and Cl- ions. The heating and
287 equilibration runs prior to production were 0.3ns and 10ns respectively. Prior to heating 2000 steps of
288 energy minimization were also performed. All seven available machine learning classifiers were
289 trained on the functional MD ensembles and deployed upon new 5 ns production runs for each variant
290 analyzed.

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

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

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

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

299 ***Improvements and upgrades over previous versions***

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

302 - New GUI organization directs users to specific comparative tasks/applications in Table 1
303 - 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
- 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

Basic implementation of DROIDS and maxDemon

The first step of any DROIDS analysis is to find or create two homologous PDB file format structures that represent the query and reference functional states of the protein system under investigation. Typically, these would represent the same protein in a bound vs. unbound state, or in a mutant vs. wildtype state. If the protein is interacting with a small ligand, and additional 'ligand only' PDB file should also be created for subsequent quantum mechanical optimization and preparation by Ambertools antechamber program. These files should be placed within the DROIDS download folder. Upon implementation via the command 'perl DROIDS.pl' launched from terminal within the DROIDS folder, 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.

More detailed instructions to users are included with our DROIDS 3.0+maxDemon 1.0 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

387

388 # Results and Discussion

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

395 ***Machine learning analysis of impacts due to simple environmental temperature shift***

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

420

421 ***Machine learning analysis of impacts of genetic variants on a functional protein binding interaction***

422 To examine functional dynamics in ubiquitin, we conducted a DROIDS analysis comparing its two
423 functional states, bound and unbound to the ubiquitin associated binding (UBA) domain of ubiquitin
424 ligase (12)(Figure 3A-D). This binding domain is highly conserved among the many other proteins that
425 interact directly with ubiquitin. The binding interaction greatly reduces the atom fluctuation in ubiquitin
426 at 3 characteristic positions, two loop structures centered at LEU 8 and ALA 46 and a portion of beta

427 sheet at the C terminus (Figure 3C). These three regions also drive significant differences in dynamics
428 across the whole protein. In novel self-similar MD runs on the bound state, we successfully detect
429 significant canonical correlations indicating conserved dynamics in these three regions with a broad
430 expanse in conserved dynamics (Figure 3E and F) across the UBA region (Figure 3G). We tested a set of
431 24 mutations that included sites with the most and least tolerated effects on growth rate in vivo in yeast
432 according to a study by Roscoe et al. (13). In this study in vivo, nearly all mutations at E18 and G53 are
433 tolerated while nearly all mutations at K48 and R72 are not. Ultimately, the causes of tolerance in these
434 variants are not known, and do not necessarily invoke functional problems in dynamics. However, the
435 impacts that we did observed in simulation were on average twice as strong in the intolerant
436 backgrounds when compared to the mutation tolerant backgrounds. And, while we did not see large
437 differences in mutational impacts on dynamics between tolerated and non-tolerated mutant groups, the
438 24 mutations analyzed all show a general trend of dynamic impact falling outside of most of the
439 functional binding region (Figure 3H-K), suggesting that ubiquitin may have evolved a tertiary structure
440 that allosterically translates dynamic impacts to less functional regions of the protein. One interesting
441 exception to this rule was demonstrated by the very large impact of R72D, centered squarely in the
442 functionally conserved binding region of ubiquitin, and would obviously heavily disrupt electrostatic
443 charge interactions there as well.

444 ***Machine learning analysis of impacts of genetic variants on DNA binding interaction***

445 TATA binding protein (TBP) is a general transcription factor that binds DNA upstream in most
446 highly regulated eukaryotic gene promoter regions (14). While relatively small, it is a mechanically
447 dynamic protein with a C-clamp like structure that highly distorts the rigid DNA double helix by inserting
448 four phenylalanine side-chains between base pairs. It is thought that this bending allows TBP to be more
449 rapidly released from the TATA element, as opposed to TATA-less promoters, subsequently allowing
450 more highly controlled regulatory responses in TATA box genes (15). Due to its obvious symmetry and
451 ability to impart large forces during binding, we thought that it would represent a good candidate for
452 comparison of its dynamics during its binding interaction with DNA. We conducted a DROIDS analysis
453 comparing human TBP (16) in its functionally bound and unbound states (Figure 4A-C). TBP exhibits a
454 characteristic large signature of dampening of atom fluctuation throughout its entire structure with
455 most pronounced effects in two loop regions that interact with the minor groove of DNA (arrows in
456 Figure 4A and 4C). Canonical correlations in new self-similar MD runs marking increased performance in
457 classification were observed in these regions (Figure 4D) along with corresponding regions of conserved
458 dynamics identified by significant Wilk's lamda (Figure 4E). Conserved dynamics from these loop areas
459 are connected through the chains in the beta sheet region of TBP spanning the DNA major groove
460 contact. Mutational impacts of four variants affecting the binding loop most proximal to the C terminal
461 exhibited followed our expectation of increasing impact ordering from R192Q, R192K, R192polyD, and
462 R192polyW (Figure 4G and 4H). The polyD and polyW mutations incorporated 5 sequential ASP or TRP
463 residues centered at R192, both causing the loop region to become more rigid (causing increased
464 negative dFLUX). We expected the strong functional binding affect observed across nearly all residues in
465 this system would make it relatively highly tolerant to single amino acid substitutions, even when
466 located in the most functional binding loop. In accordance with this idea, we found the most impactful
467 multiple mutation (i.e. R192polyW) significantly affected the dynamics of nearly 6 times more local
468 residues than the least impactful single substitution (i.e. R192Q).

469

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

471 In contrast to TBP, we wanted to use our method to examine a small molecule binding
472 interaction in a protein with potentially more complex impacts on molecular dynamics. Hsp90 is a well-
473 known chaperone protein that assists the folding of many proteins and thereby mitigating many
474 environmental stresses in the cell. Hsp90 even capacitates the evolutionary process by allowing
475 potential phenotypic variation exhibited under stress to be hidden from natural selection until needed in
476 response to environmental change (17). Hsp90 contains a highly conserved N-terminal domain where
477 ATP binding and activation occurs. The binding of ATP physically changes motions in this region creating
478 a 'lid' that closed during ATP binding and open when conversion to ADP occurs. Due to the role of Hsp90
479 in stress mitigation in most tumors, it is a common drug target for ATP inhibitors in many cancer
480 therapies (18, 19). The amino acid residues that interact with ATP in this region are well known and the
481 inhibitor geldanamycin is known to mimic nearly all the local ATP contacts as well (20). Other more
482 modern inhibitors interact with the ATP binding pocket differently (19, 21, 22), so we thought that this
483 system would be a good candidate for comparative analysis of drug class variants.

484 We conducted a DROIDS analysis comparing the dynamics of Hsp90 chaperone, a common
485 drug target for inhibitors in many cancer therapies, in both its ATP bound and unbound states. The
486 binding of ATP was discovered to significantly destabilize three co-localized alpha helical regions of the
487 protein adjacent to and extending from the ATP binding site (Figure 5A-D). MaxDemon analysis
488 confirmed the dynamics of this region to be highly conserved in new MD runs (Figure 5D-G). We also
489 analyzed the impacts of the six drug class variants targeting the ATP site (20, 22, 21, 23, 24), but
490 interacting differently with residues in this region (Figure 5H). The contacts in the ATP binding site are
491 shown in Figure 5I. While the localized patterns of impacts of the drug variants were all quite similar to
492 ATP (Figure 5J), the drug variants that most closely mimicked the contacts of ATP (i.e. geldanamycin) had
493 far less impact on dynamics than variants that interacted very differently with the binding pocket (i.e.
494 benzamide SNX1321 and inhibitor FJ1(Figure H-I). We feel that this finding demonstrates that while it is
495 important to be able to target a druggable protein binding site (25), researchers should also consider
496 how these various chemicals might alter the natural dynamics of the system. In situations where a drug
497 might too closely mimic the dynamic effects of a natural activator like ATP, a hyperactivation response
498 might occur in non-tumor cells leading to secondary cancer (26–28). Alternatively, other situations may
499 require drug targeting that does not alter the natural dynamic behavior too much, potentially activating
500 proteolytic systems in the cell. Our software allows more detailed investigations of these potential
501 dynamic impacts of drug class variants.

502 ***Conclusion***

503 We provide a well demonstrated method and user-friendly software pipeline for conducting statistically
504 sound comparative studies of large ensembles of comparative protein dynamics. The method/software
505 also now provides machine learning based extrapolations of effects on novel MD simulations
506 representing various functional variants of interest to the user. While there currently is at least one
507 other software allowing users to connect sequence-based evolutionary metrics to protein dynamics (29),
508 our method/software is unique in that regions of functional conservation are identified by analyzing
509 self-similar features of dynamics themselves rather than relying upon marrying dynamics to traditional
510 static sequence-based approaches, which do not necessarily assume that a protein function has a strong
511 dynamic component. By providing a systematic way of comparing protein dynamics at single residue

resolution, our method/software provides an important step beyond traditional bioinformatics, allowing investigators to gain a much more biophysically-grounded view of functional and evolutionary change. 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 (30, 31) 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.

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

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

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

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

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

557 8. Plante, A., D.M. Shore, G. Morra, G. Khelashvili, and H. Weinstein. 2019. A Machine Learning
558 Approach for the Discovery of Ligand-Specific Functional Mechanisms of GPCRs. Mol. Basel Switz.
559 24.

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

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

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

565 12. Peschard, P., G. Kozlov, T. Lin, I.A. Mirza, A.M. Berghuis, S. Lipkowitz, M. Park, and K. Gehring.
566 2007. Structural basis for ubiquitin-mediated dimerization and activation of the ubiquitin protein
567 ligase Cbl-b. Mol. Cell. 27: 474–485.

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

570 14. Kornberg, R.D. 2007. The molecular basis of eukaryotic transcription. Proc. Natl. Acad. Sci. U. S.
571 A. 104: 12955–12961.

572 15. Tora, L., and H.T.M. Timmers. 2010. The TATA box regulates TATA-binding protein (TBP)
573 dynamics in vivo. Trends Biochem. Sci. 35: 309–314.

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

577 17. Rutherford, S.L., and S. Lindquist. 1998. Hsp90 as a capacitor for morphological evolution.
578 Nature. 396: 336–342.

579 18. Neckers, L., and P. Workman. 2012. Hsp90 Molecular Chaperone Inhibitors: Are We There Yet?
580 Clin. Cancer Res. 18: 64–76.

581 19. Yuno, A., M.-J. Lee, S. Lee, Y. Tomita, D. Rekhtman, B. Moore, and J.B. Trepel. 2018. Clinical
582 Evaluation and Biomarker Profiling of Hsp90 Inhibitors. Methods Mol. Biol. Clifton NJ. 1709: 423–
583 441.

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

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

589 22. 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.
590 Structure insights into mechanisms of ATP hydrolysis and the activation of human heat-shock
591 protein 90. Acta Biochim. Biophys. Sin. 44: 300–306.

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

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

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

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

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

608 28. Cichowski, K., and P.A. Jänne. 2010. Drug discovery: inhibitors that activate. Nature. 464: 358–
609 359.

610 29. Bakan, A., A. Dutta, W. Mao, Y. Liu, C. Chennubhotla, T.R. Lezon, and I. Bahar. 2014. Evol and
611 ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics. 30: 2681–
612 2683.

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

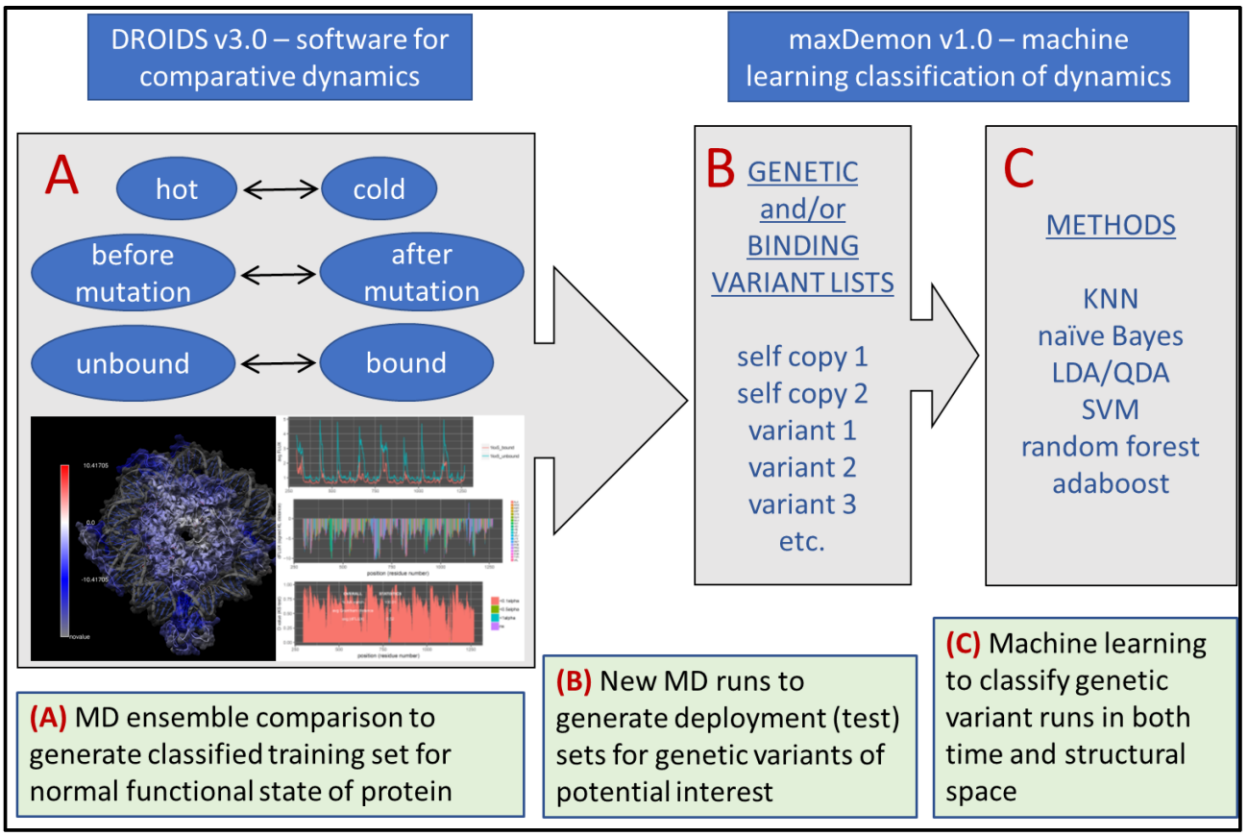
31. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Chapter 7: Unit7.20.

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 natural ligand binding interaction	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
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



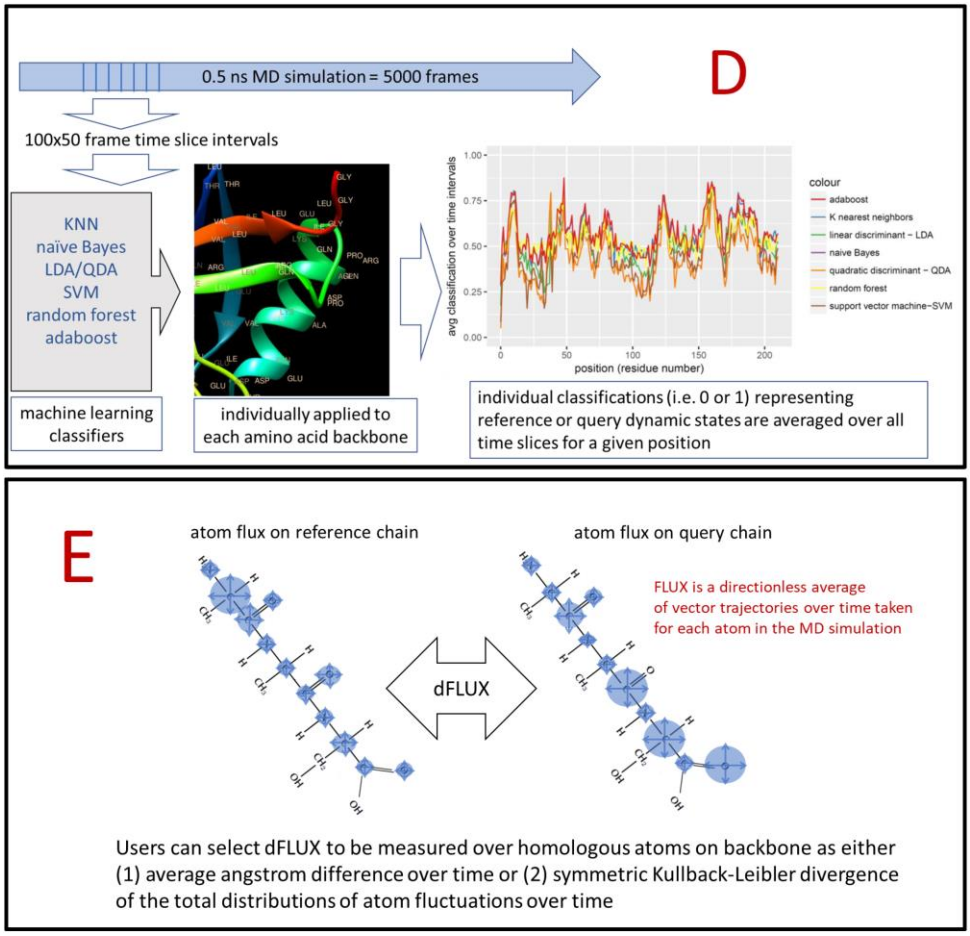


Figure 1. Schematic overview of DROIDS 3.0 + maxDemon 1.0 software for machine learning-based detection of variant impacts on functionally conserved protein dynamics. The pipeline starts with (A) generation of two large ensembles of molecular dynamic (MD) simulations that represent a functional comparison 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.

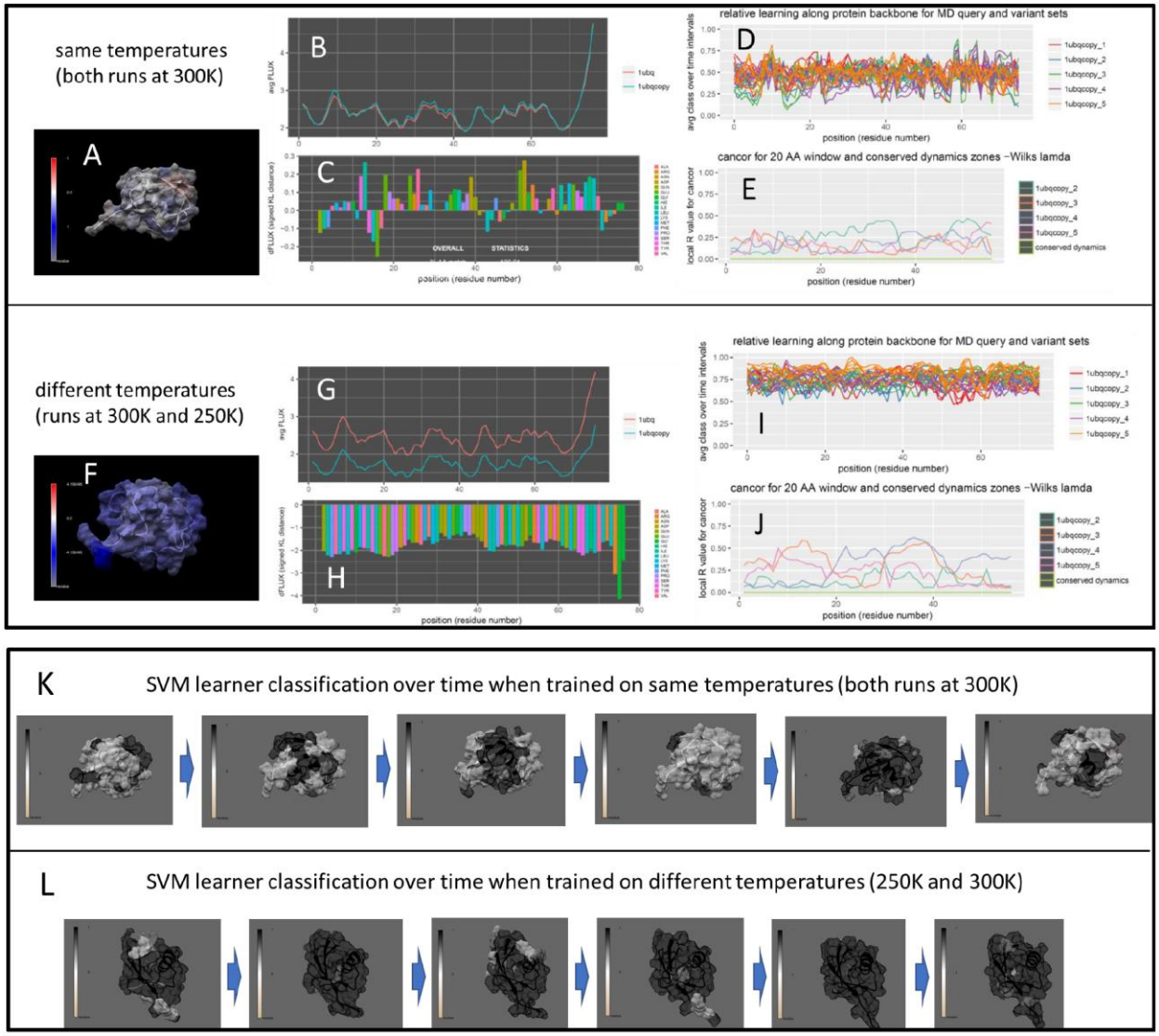


Figure 2. Analysis of environmental temperature change on non-functional ubiquitin dynamics. DROIDS image and analysis of random ubiquitin dynamics compared at the same (A-E) and different (F-J) temperatures. Note: blue color quantifies damped rmsf at temperature lowered by 50K. Note that relative learning 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)

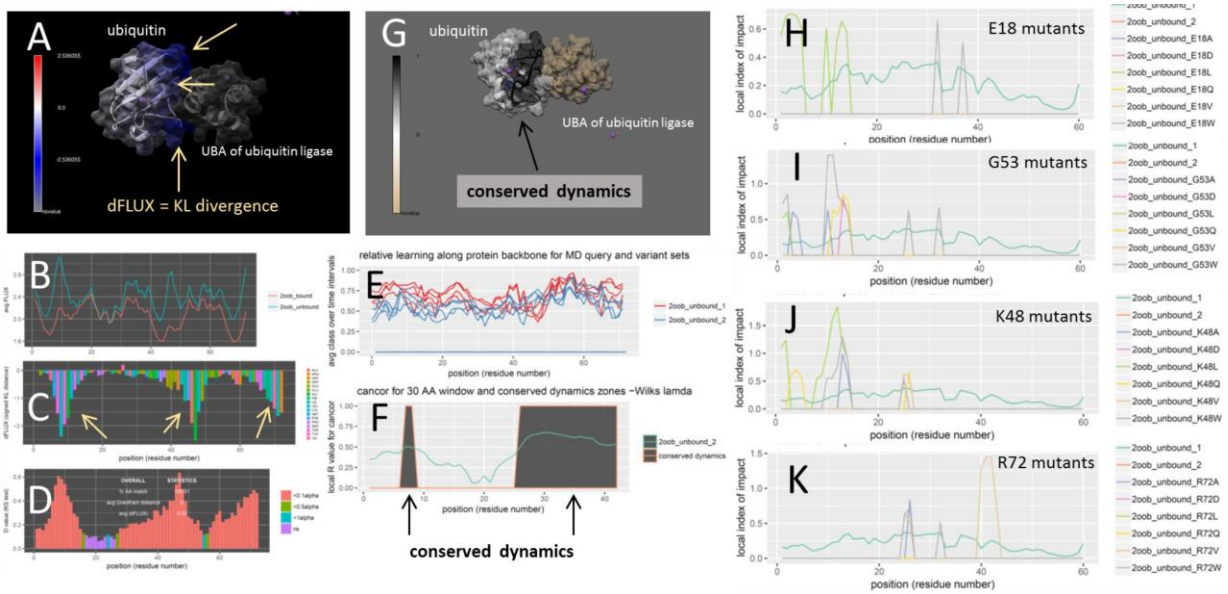


Figure 3. Analysis of mutational impact and tolerance on functional ubiquitin dynamics. (A) DROIDS image and analysis of ubiquitin bound to the ubiquitin associated binding domain (UBA) of ubiquitin ligase. Note: blue color quantifies damped rmsf at binding interface. (i.e. negative dFLUX) also by the (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.

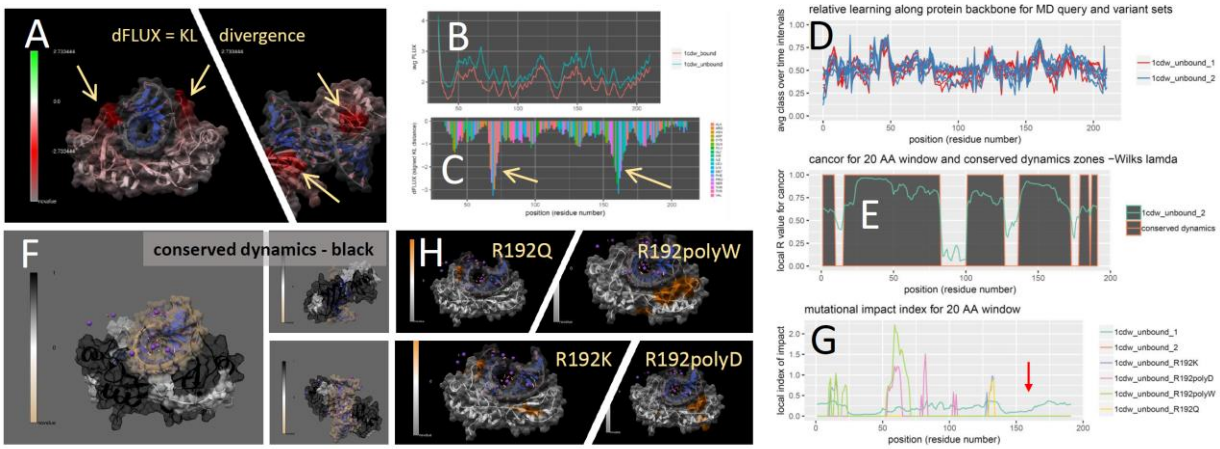
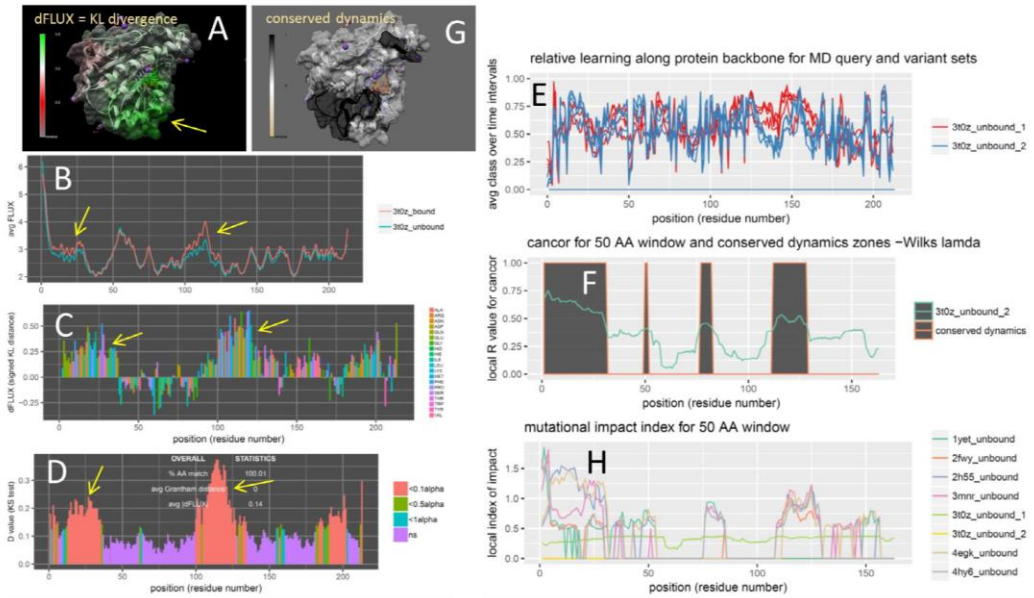
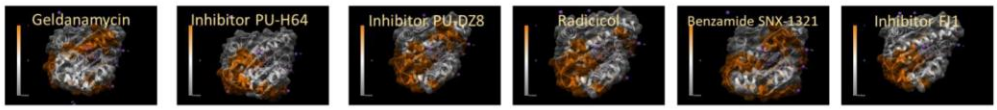


Figure 4. Analysis of mutational impact and tolerance on DNA binding in Tata Binding Protein (TBP). DROIDS image and analysis of TBP in DNA-bound and unbound states showing (A) colored TBP structure, (B) respective rmsf profiles and (C) KL divergence (dFLUX) plot. Note: arrows indicate functional binding 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).



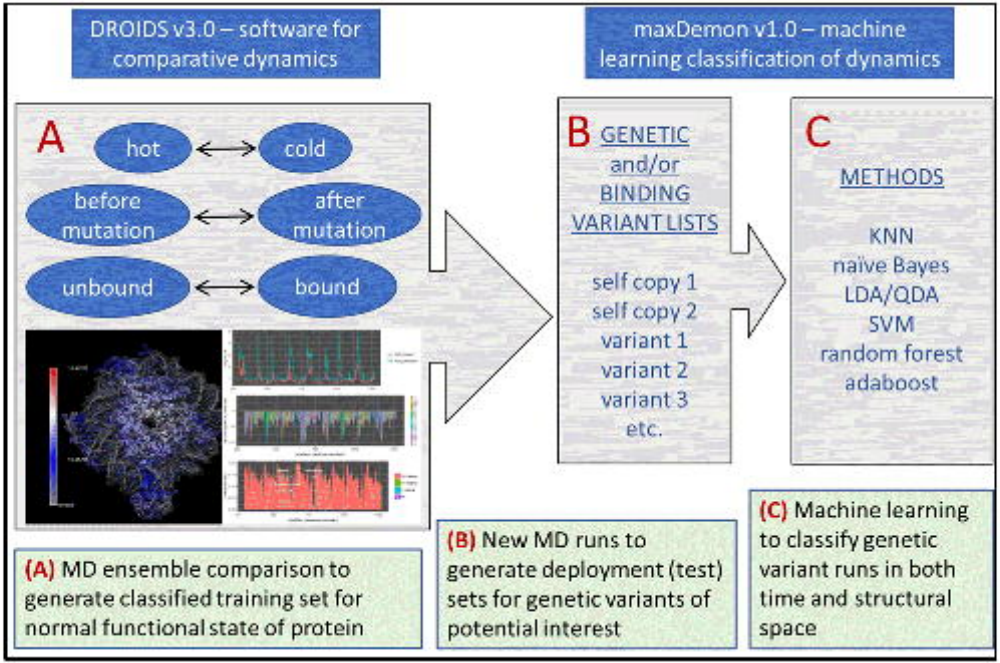
I Ligand contacts of drug binding variants on protein dynamics of Hsp90 in the ATP binding pocket

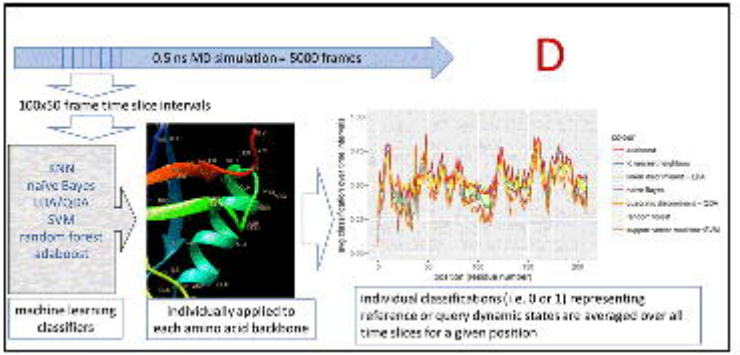
ligand	PDB	% sim	Val 186	Thr 184	Trp 162	Tyr 139	Phe 138	Lys 112	Alal 111	Leu 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 93	Lys 58	Ala 55	Asp 54	Asn 51
ATP	3t0z	100		x			x				x		x	x		x		x		x
Geldanamycin	1yet	87.5		x			x	x			x			x		x		x		x
Inhibitor PU-H64	2fwy	75		x	x		x			x			x			x		x		x
Inhibitor 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											

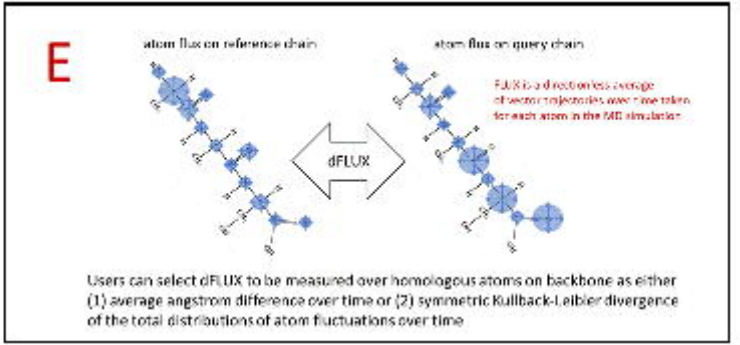


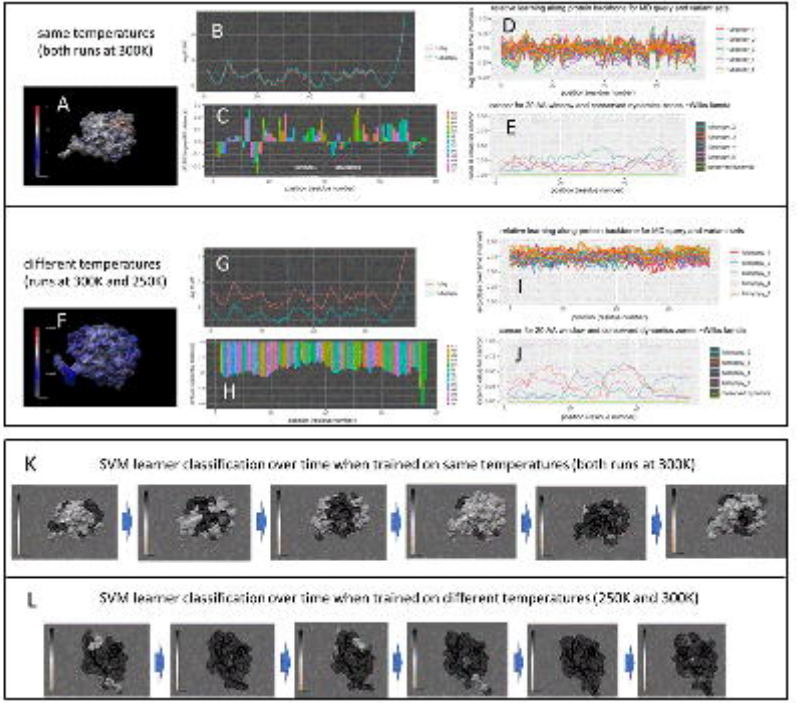
J Impacts of drug binding variants on protein dynamics of Hsp90 that are significantly different than binding to ATP

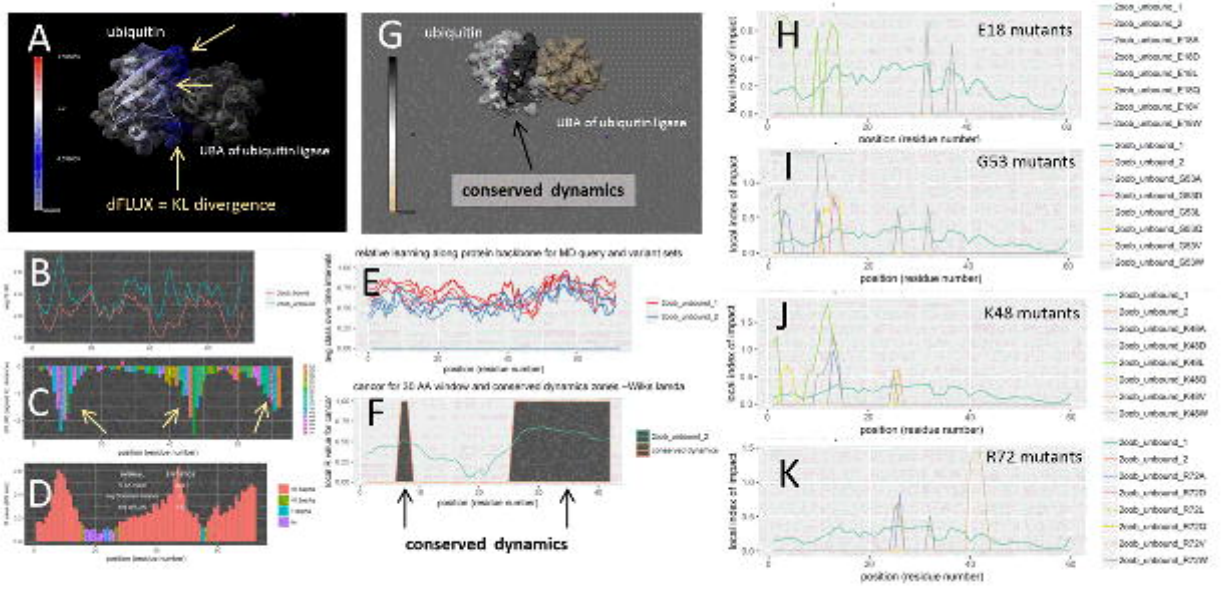
Figure 5. Analysis of drug class variant binding in the ATP-binding domain of Hsp90. DROIDS image and analysis of Hsp90 in ATP-bound and unbound states showing (A) colored Hsp90 structure, (B) respective rmsf profiles and (C) KL divergence (dFLUX) plot and (D) significant differences in dynamics determined 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.

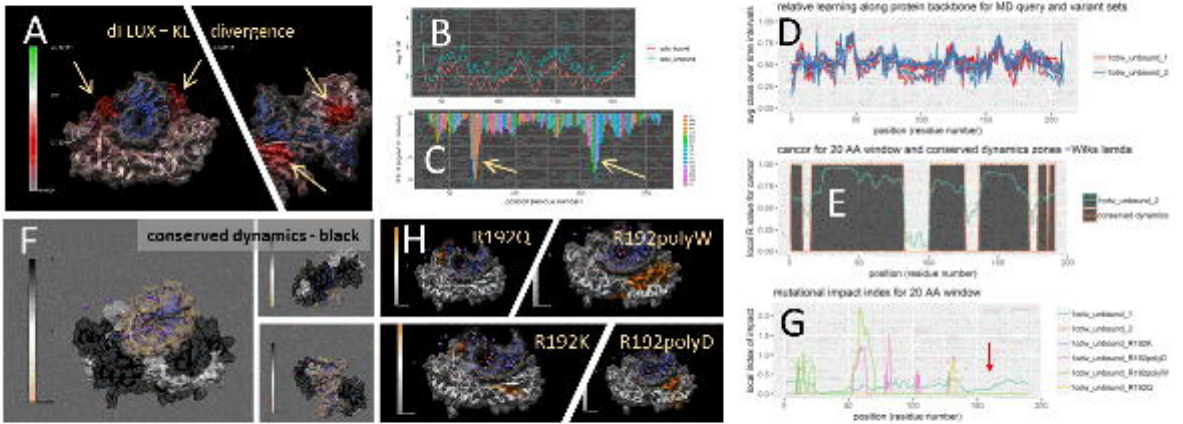

DROIDS v3.0 – software for comparative dynamics
maxDemon v1.0 – machine learning classification of dynamics
A
hot
cold
before mutation
after mutation
unbound
bound
B GENETIC and/or BINDING VARIANT LISTS
self copy 1
self copy 2
variant 1
variant 2
variant 3
etc.
C
METHODS
KNN
naïve Bayes
LDA/QDA
SVM
random 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

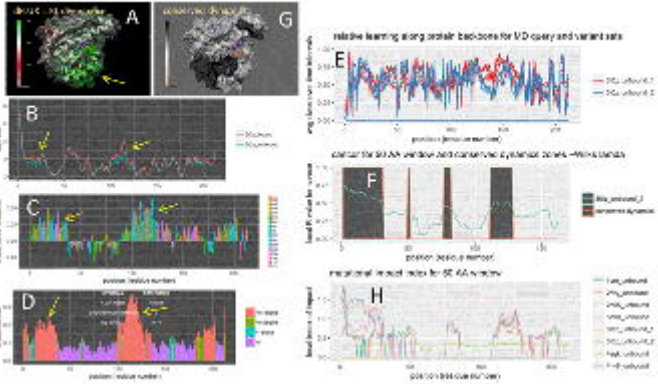

D
0.5 ns MD simulation – 5000 frames
100x50 frame time slice intervals
kNN
naive Bayes
LDA/QDA
SVM
random forest
adaboost
machine learning classifiers
individually applied to each amino acid backbone
individual classifications (i.e. 0 or 1) representing reference or query dynamic states are averaged over all time slices for a given position


E
atom flux on reference chain
atom flux on query chain
FLUX is a dimensionless 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


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
E
F
conserved dynamics
H
E18 mutants
I
G53 mutants
J
K48 mutants
K
R72 mutants
position (residue number)
local index of impact


A
di LUX – KL divergence
B
C
D
relative learning along protein backbone for MD query and variant sets
position (residue number)
cancer for 20 AA window and conserved dynamics zones - Wilks lambda
E
position (residue number)
mutational impact index for 20 AA window
G
position (residue number)
F
conserved dynamics - black
H
R192Q
R192polyW
R192K
R192polyD



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

Ligand	PDB	% size	Val 186	Thr 184	Tyr 192	Tyr 138	Ser 138	Lys 112	Ala 111	Ins 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 95	Lys 58	Ala 55	Asp 54	Asn 51
ATP	2BSE	100		x			x				x		x	x		x		x		x
Reference inhibitor	1yet	82.5		x			x	x			x			x		x		x		x
Inhibitor PU-H54	2hey	75		x	x		x			x			x			x		x		x
Inhibitor PU-H58	2BSE	75		x	x		x			x			x		x	x	x	x		x
Radicicol	4ngb	62.5	x	x			x			x						x	x	x	x	x
Semantic inhibitor	5ewe	52.5		x	x	x				x		x			x	x	x	x		
Inhibitor FLS	4hy6	0			x	x			x											

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