User documentation for DROIDS 3.0+maxDemon 1.0 – a machine intelligent GUI-based pipeline for comparative protein dynamics

Gregory A. Babbitt T.H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester NY USA

author email address: gabsbi@rit.edu

please report bugs to this email

System Requirements – Debian Linux desktop OS with 1 or more GPUs. Linux Mint 18/19 is recommended with Nvidia GTX 1080 or larger. Be sure to also check the Linux Mint 'Driver Manager' after initial build and install all recommended Nvidia drivers.

NOTE – software can be most easily installed by running 'perl DROIDS+AMBERinstaller.pl' a perl installer script included with our download folder or GitHub repo. After installation, the software runs as 'perl DROIDS.pl on the Linux terminal opened from within the DROIDS folder'

Software – Amber16/18, AmberTools 16/18, UCSF Chimera 1.11 or 1.13 (additionally ChimeraX optional), CUDA 8.0/9.0, CUDA toolkit, perl-tk, python-tk, and R. (Note: Amber18 install on Linux Mint 19 will likely require setting up older versions of the gcc, g++, and gfortran compilers. Version 5.0 works well. Our installer will lead the user through this process if needed. Do not use CUDA 7.0 or earlier nor version 10.0 or later.

Debian and python packages – gedit, grace, gdebi, gparted, evince, perl-tk, python-tk, python-gi, gstreamer (and dependencies), and these Amber dependencies (csh flex patch gfortran g++ make xorg-dev bison libbz2-dev). If you plan to use VR, install Steam, SteamVR and vulkan library.

Perl packages – Statistics::Descriptive

R packages – ggplot2, gridExtra, dplyr, caret, FNN, e1071, kernlab, class, MASS, ada, randomForest, CCA, CCP, parallel, foreach and doParallel.

NOTE: We supply an installer script (DROIDSinstaller.pl) with download. It can setup the whole system (including CUDA, Amber18, UCSF Chimera, and R) on a fresh Linux Mint install or Virtual Machine. It can also skip over these if already installed (i.e. only will install package dependencies).

INSTRUCTIONS FOR DROIDS v3.0 ON WINDOWS and/or GOOGLE CLOUD PLATFORM

Option 1: running a Linux Mint Virtual Machine (VM) on Windows PC

1. Download and install VirtualBox from this website

   https://www.virtualbox.org/

2. Get a .vdi file (for Linux Mint) from this website

   https://www.osboxes.org/virtualbox-images/

3. Build a Linux Mint VM following instructions from VirtualBox
4. Copy the required files to the VM (i.e. DROIDS-3.0.tar.gz, Amber18.tar.gz, AmberTools18/19.tar.gz, Chimera-1.14-linux_x86_64.bin). These can be obtained from the following websites
   https://people.rit.edu/gabsbi/
   https://ambermd.org/
   https://www.cgl.ucsf.edu/chimera/download.html
   and you'll probably want Modeller enabled on Chimera
   https://salilab.org/modeller/
5. Run the DROIDS+AMBERinstaller.pl script on the VM's desktop terminal

   perl DROIDS+AMBER_installer.pl

   NOTE: your PC must already have Nvidia GPU hardware, CUDA and Nvidia graphics drivers properly installed.

Option 2: running an Ubuntu Linux VM on GCP (Google Cloud)

https://cloud.google.com/

1. Open GCP account, go to your console and request resource quota to enable building GPU VM instances. (e.g. use 8CPUs and add 1 V100 GPU)
2. Upload the required files to a Google Cloud Storage Bucket (i.e. DROIDS-3.0.tar.gz, Amber18.tar.gz, AmberTools18/19.tar.gz, cuda-repo-ubuntu1704-9-0-local_9.0.176.deb (or an equivalent version of cuda), Chimera-1.14-linux_x86_64.bin). These can be obtained from the following websites
   https://people.rit.edu/gabsbi/
   https://ambermd.org/
   https://developer.nvidia.com/cuda-downloads
   https://www.cgl.ucsf.edu/chimera/download.html
   and you'll probably want Modeller enabled on Chimera
   https://salilab.org/modeller/
3. On your Dashboard, go to 'Compute Engine' and build your VM instance. IMPORTANT: Be sure to add an Nvidia GPU and use Ubuntu 16.04LTS or 18.04 LTS so that you can install and link from a remote desktop application.
4. Once the VM appears, connect to it via the SSH link and run the following commands.

sudo passwd (to reset root password)
sudo passwd yourusername (to reset your user password)
sudo apt-get install xrdp
sudo apt-get install xfce4
sudo service xrdp restart

5. Leave this terminal open and go to your Windows RDP (Remote Desktop) application and enter the external IP address from your VM to open the new xfce desktop you installed with your new passwords.
6. At this point you can transfer your files from your cloud bucket or your own computer to the VM home folder (using the 'gear' icon on the ssh terminal of your VM). Open the home folder and move your files to your desktop and then proceed with the installation using the DROIDS+AMBERinstaller.pl script at either the desktop or SSH terminal.

   perl DROIDS+AMBERinstaller.pl

   NOTE: when the script pauses and asks for information typed into secondary terminals, these terminals may need to be opened manually at the VM instance SSH link or else on the remote desktop. If R package installations fail after installing R, open R manually at the command line (i.e. type 'R') and then paste or type the following

   >install.packages(c('ggplot2', 'gridExtra', 'dplyr', 'caret', 'FNN', 'e1071', 'kernlab', 'class', 'MASS', 'ada', 'randomForest', 'CCA', 'CCP', 'doParallel', 'foreach', 'rpsychi'))

   then to exit R

   >q()

Improvements and upgrades over previous versions

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
- 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

Current bugs –

1. In some systems, when the maxDemon machine learning deployment button is used for the first time, an error 'can't open temp test file' is thrown. If the user closes the maxDemon GUI and reopens it at the Linux terminal ('perl GUI_ML_DROIDS.pl'), the button will work as intended and learners can then be deployed.

Basic implementation

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.

Scientific overview

DROIDS 3.0 (**D**etecting **R**elative **O**utlier **I**mpacts in molecular **D**ynamic **S**imulation) is an open source software project enabling statistical comparison of large ensembles of molecular dynamic (MD) simulation that represent changes in functional states such as before/after genetic or epigenetic mutation and/or before/after binding of DNA/small molecule/ligand. The software returns both traditional plots of MD comparison along amino acid sequence, as well as color mapped images and movies of functional impacts on protein structure. DROIDS 3.0 is bundled with a new backend application ‘maxDemon’, allowing users to train combinations of different machine learning algorithms on the functional changes extracted from the original comparative MD ensembles and subsequently map what is learned on new MD runs. The selected learners are spatially applied individually to each amino acid in the structure and temporally applied to every 50 frame time slice of MD simulation. Sequence-dependent canonical self-correlation is used to identify regions of functionally conserved dynamics. The impacts of genetic and/or

binding variants are also able to be statistically determined and compared. Thus maxDemon greatly assists the user interpretation of MD simulation by returning analyses, images and movies summarizing 'when and where' functionally important dynamics have occurred. DROIDS 3.0 with maxDemon is designed to allow for statistical and visual exploration of different genetic and drug binding variants with reference to natural dynamic function (see Table 1).

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 DNA bound 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 functional evolution of novel dynamics in homolog genes	MD ensembles comparing two functional states of the protein	MD runs on one or more homologs	Isolates potential MD novelty in evolved gene product

maxDemon, is a multi-method machine learning application that trains on the comparative protein dynamics, identifies functionally conserved dynamics, and deploys classifications of functional dynamic states to newly generated protein simulations. Nine different types of machine learners can be deployed on the dynamics of each amino acid, then the resulting classifications are rendered upon movie images of the novel MD runs. This results in movies of protein dynamics where the conserved functional states are identified in real time by color mapping, allowing users to see both when and where a novel MD simulation displays a specific functional state defined by the comparative training. Examples of the functional dynamic effects of solvent temperature change, genetic mutation, and drug binding interaction on protein dynamics will be demonstrated in a future software note. Thus, much like James Maxwell's mythical demon of thermodynamics from 150 years ago, maxDemon software derives potentially important spatiotemporal information from the observation of dynamic motion at all-atom resolution.

More broadly, the DROIDS+maxDemon software project aims to visualize and quantify the impact of one of the longest time scale processes in the universe (i.e. molecular evolution) on one of the shortest time scale processes in the universe (i.e. molecular motion). Specifically, we want to know how molecular evolution over 100s of millions of years impacts the functional molecular motions that play out over a few femtoseconds in real time. A primary motivation of this project is to combine GPU accelerated biophysical simulations and GPU graphics to design a gaming PC into a 'computational microscope' that is capable seeing how mutations and other molecular events like binding, bending and bonding affect the functioning of proteins and nucleic acids. DROIDS-1.20 is a GUI-based pipeline that works with AMBER16/18 (Assisted Model Building with Energy Refinement), Chimera 1.11, R and CPPTRAJ to analyze and visualize comparative protein dynamics on GPU accelerated Linux graphics workstations. DROIDS employs a robust and nonparametric statistical method (multiple test corrected KS tests on root mean square fluctuation or RMSF of all backbone atoms of each amino acid) to detect significant changes in molecular dynamics simulated on two homologous PDB structures. Quantitative KL divergence in atom fluctuation (i.e. calculated from vector trajectories) are displayed graphically and mapped onto movie images of the protein dynamics at the level of individual residues. P values indicating significant changes are also able to be similarly mapped. DROIDS is useful for examining how mutations, epigenetic changes, or binding interactions affect protein dynamics. DROIDS was produced by student effort at the Rochester Institute of Technology under the direction of Dr. Gregory A. Babbitt as a collaborative project between the Gosnell School of Life Sciences and the Biomedical Engineering Dept.

Visit our lab website (https://people.rit.edu/gabsbi/) and download DROIDS from Github at https://github.com/gbabbitt/DROIDS-2.0---free-software-for-comparative-protein-dynamics

We will be posting video results periodically on our YouTube channel https://www.youtube.com/channel/UCJTBqGq01pBCMDQikn566Kw

Implementation

It is strongly advised that users be comfortable with how to prepare PDB files for molecular dynamic (MD) simulation using GPU accelerated AMBER 16/18 (pmemd.cuda). DROIDS assists with modifying .pdb files named in the GUI for AMBER simulation, however the user should become very familiar with the programs running at these steps (i.e. antechamber, pdb4amber, and teLeap) and read through all output at the DROIDS terminal to ensure that the structures are properly prepared for MD simulation. You must consult the AMBER documentation for this knowledge. The DROIDS GUI provides automation of teLeap, a program for pdb file setup, but care must be taken to read output on the Linux terminal for any errors. The programs 'antechamber' and 'pbd4amber' are used by DROIDS in modifying files for MD and are generally prior to starting teLeap in DROIDS. Please consult the Amber16 user manual for more details. Typically preparation includes (A) removing mirrored images and other chemical artifacts (done manually in Chimera prior to DROIDS), (B) performing a structural alignment (using Chimera MatchMaker and Match->Align when prompted by DROIDS) followed by subsequent saving of a Clustal format file (.aln), (C) adding H atoms and removing crystallographic waters (use pdb4amber button in DROIDS to dry and reduce), (D) estimating and loading force field parameterization regarding important ligands if a protein-ligand interaction is modeled (use antechamber button). Then finally (E) run teLeap button in DROIDS to setup topology and coordinate files for simulation. For v2.0 we have added script to check the file sizes of teLeap output files and recommend whether the process likely failed of succeeded at this step. teLeap is nicely verbose, so warnings on terminal when running teLeap button is very helpful for any indications of problems specific to your structural models. For many at this stage of model prep, it is not unusual to go back to modify the original .pdb file and run through the prep stages again. Be sure to view your models in Chimera using the 'all atom' preset so that you do not miss small molecules that might trip up the MD setup. Amber is designed not to run unless all atoms in your system can be properly parametrized by the force field you have chosen. Many force fields are available to try in the amber16/dat/leap/cmd folder. Many are appropriate only for certain macromolecules, and analysis of binding interaction will require several are loaded. ALSO NOTE: AMBER 16/18 software must be licensed from the University of California. More details about purchasing and installation can be found at http://ambermd.org/. DROIDS is tested on Linux Mint 18.1 and Ubuntu 16.04 and is offered freely under the GPL 3.0 license and is available on GitHub https://github.com/gbabbitt/DROIDS-1.0

DROIDS is activated by entering 'perl DROIDS.pl' at the Linux terminal opened from within the DROIDS folder. DROIDS v3.0 initially starts with a small GUI requesting user to add paths to Chimera and

Amber's force field data files (e.g. amber16/dat/leap/cmd). As Amber16/18 is typically installed to the Desktop, this path will be different on different machines. Make sure you edit the path appropriately before attempting to run DROIDS. The GUI will create a paths.ctl file. Once this file is created for your individual machine, it can be saved and dropped into DROIDS folders prior to each run. The typical bashrc file can be used similarly, but this GUI was added to make this initial setup simpler for less experience Linux users. Once the paths GUI is closed, the main DROIDS v3.0 GUI will appear. Here the user is directed to choose one of the various types of comparative analysis that can be done, choose MD sim software, and indicate whether the machine is running a single or dual GPU. Upon clicking 'run DROIDS' the user is taken to the first main GUI for setup, running MD, and parsing of MD simulation output. The second main GUI controls the DROIDS statistical analyses and the last main GUI controls the image color-mapping and movie rendering and viewing options.

SEE OUR TUTORIAL (PDF) FOT MORE EXPLICIT INSTRUCTIONS ON RUNNING DROIDS

IMPORTANT NOTE: When running DROIDS on many protein comparisons, we find that explicitly solvated systems (i.e. PME method) tend to yield better and more conservative results regarding the significance of the KS test when compared to implicitly solvated comparisons (i.e. GB method). This is likely expected due to the many more degrees of freedom under the PME option as well as its better approximation to reality. A three point solvent model (tip3p) is default method in DROIDS. This is for sake of efficiency. If a more accurate solvent is needed we recommend the users edit the .bat files that pop open when running teLeap from the DROIDS GUI. The user can manually change the references to tip3p to the tip4p, tip5p or tip6p models. Another default state of our software is to charge neutralize the protein. The .bat files can also be edited by experienced users to alter the ion concentrations in the simulation. For more complicated setups, the numbers of ions needed for given box size and salt concentration can be determined using the method and tool cited below.

SLTCAP: A simple method for calculating the number of ions needed for MD simulation

Jeremy D. Schmit*,†, **Nilusha L. Kariyawasam**‡, **Vince Needham**†, and **Paul E. Smith**‡
†Department of Physics, Kansas State University, Manhattan, KS 66506, USA
‡Department of Chemistry, Kansas State University, Manhattan, KS 66506,

J Chem Theory Comput. 2018 April 10; 14(4): 1823–1827. doi:10.1021/acs.jctc.7b01254.

We recommend that users explore many methods of solvation when using DROIDS. Implicitly solvated protein comparisons run relatively fast and may be useful for an initial investigation of a large system,

however comparison of explicitly solvated systems may yield more realistic local variation in mutational impacts.

IMPORTANT: Given that MD simulations are well known to exhibit complex and often chaotic behavior, we also strongly recommend that users of DROIDS repeat analyses of given systems in order to determine best parameter settings for ensemble size, lengths of production runs and overall reproducibility of the final results.

Specific analyses now offered in DROIDS v2.0 and v3.0

DROIDS v2.0 now offers 10 different pipelines intended for specific types of comparative analysis. Examples with .pdb files are provided. First time users should run the examples provided in the exampleFiles folder first, to get a sense of what setup required and what output is delivered. The 8 main types of comparative analysis are listed here.

1. ***Analysis of self-stability of dynamics on a single protein*** – this compares MD of a protein to itself and is useful for finding regions of protein that are less stable. Try it on 1ubq.pdb and notice the lack of stable dynamics near the c-terminal tail, where ubiquitin is ligated to 'tag' proteins for degradation. In this
2. ***Analysis of mutational impacts on a protein*** – here the user can create mutant versions of a given protein by replacing one or several AAs using automatically optimized selections from the Dunbrack rotamer library and comparatively quantify the local mutational impacts on MD using KL divergence in atom fluctuation. This option is great for simulating studies of site-directed mutagenesis.
3. ***Analysis of evolutionary or functional divergence in MD on a protein*** – here the user can analyze divergence in MD using PDB files for an ortholog pair. This option is interesting when applied to questions of thermostability. For example, compare thermostable Taq DNA polymerase (4n56.pdb) to its less stable cousin in E. coli (1kfd.pdb). Epigenetic changes (i.e. post-translational modifications) can also be compared as well as genetic-based divergences.
4. ***Analysis of impact of DNA protein interaction upon binding*** – This option allows users to identify and visualize where DNA binding in the system occurs by comparing the dynamics of protein in the bound and unbound states. Binding is identified via dampened atom fluctuation in the bound model. Try the TATA binding protein example using 1ytb_bound and 1ytb_unbound (where the DNA chains were removed).
5. ***Analysis of the impact of mutation(s) on DNA-protein interaction*** – Here site-directed mutagenesis in both cis and/or trans can be simulated on the DNA bound protein system and mutational impacts on binding observed. This is particularly useful for questions around gene regulatory evolution on a given transcription factor. Try mutating 1ytb_bound in regions where strong binding is indicated in analysis #5.

6. ***Analysis of the comparison of two DNA-protein interactions*** - Like analysis #3, this option allows comparison of DNA-binding homologs directly from two PDB files. This is useful for analyzing more distant evolutionary divergences in transcription factors.
7. ***Analysis of the impact of protein-ligand interaction with a drug, toxin or activator*** – Like analysis #5, this option allows user to examine how a given protein binds a particular ligand by comparing bound and unbound protein dynamic states. Binding mechanisms are identified by reduced atom fluctuation in the results. Try examples provided to demonstrate binding of HIV drug sustiva (efavirenz) to the drug target of viral reverse transcriptase (1fk9.pdb). Note: requires three files (1fk9_bound, 1fk9_unbound and 1fk9_ligand).
8. ***Analysis of the impact of mutation(s) on protein-ligand interaction*** – Like #6 this option allows user to put mutations onto the protein-ligand system and analyze the effects on MD. This is potentially useful for examining how genetic backgrounds can influence the working of a drug or toxin.

NOTE: If you use DROIDS for published work please use the following citation

Babbitt et al., DROIDS 1.20: A GUI-Based Pipeline for GPU-Accelerated Comparative Protein Dynamics, Biophysical Journal (2018), https://doi.org/10.1016/j.bpj.2018.01.020

The DROIDS pipeline

The DROIDS+maxDemon pipeline is run as a series of linked Perl-Tk scripts that are controlled at the Linux terminal command line. The analysis steps are shown schematically in Figure 1 and 2. The user starts the pipeline by placing the two PDB files to be compared in the DROIDS main folder, opening a terminal, and typing 'perl DROIDS.pl. After the paths.ctl file is created, the main GUI opens allowing choice of analysis, and specification of hardware and software. After this, the user is guided through four main GUI's each for (1) Amber MD simulation, vector trajectory analysis and file preparation and parsing for DROIDS, (2) DROIDS comparative statistical analysis of protein dynamics and graphical plotting in R, (3) PDB structure color-mapping and movie rendering in Chimera and subsequent movie viewing in the DROIDS movie viewer, and (4) functional machine learning interpretation with maxDemon. We now offer GUI for computer builds with either single or dual GPU cards (note: multiple cards connected via SLI are treated as single GPU). Dual GPU systems will run MD on both homologous protein structures at the same time. This GUI interface is designed to control and run all stages of the MD simulations of both the query and reference PDB structures that will be needed for later DROIDS analysis. This includes typical teLeap setup of the PDB file, structural alignment of the query and reference proteins, and an energy minimization, heating and equilibration run on each PDB. These runs are followed by N number of sampling runs with N specified by the user. Random spacer runs precede each sampling run so as to minimize the impact of initial conditions on the MD sampling (i.e. minimize differences merely due to chaos in the MD runs). Afterwards, MD is run, users will collect atom info and flux data using buttons that run typical cpptraj commands that loop through each sampling run. The last step includes the parsing of

the vector trajectory output to the structurally-based sequence alignment in performed earlier in Chimera. Some analyses in DROIDS call for choice of 'strict' vs 'loose' homology (which determines upon which amino acids the DROIDS statistics will be applied). Loose homology should be chosen when evolutionary distances between the PDB files are large. Strict homology should be chosen when sequences are nearly identical (e.g. examination of one or several specific mutations). After parsing, a second GUI will pop up and lead users through DROIDS statistical analysis and graphical output. Here users run the statistical comparisons and choose method of multiple test correction. At this point a third GUI will pop up and allow color-mapping and graphics options to be applied to the static and moving images of the reference PDB. The statistical test employed by DROIDS is a KS test applied specifically to the collective backbone MD of each amino acid residue (i.e. atoms N, CA, C and O masked during cpptraj). A fourth GUI runs the maxDemon application described in Figure 2.

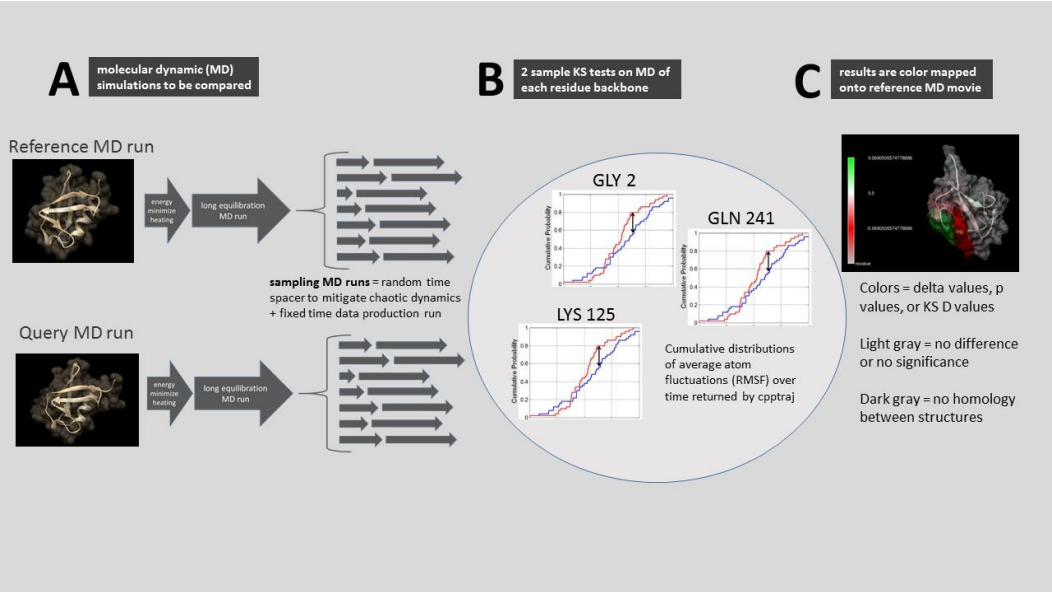

Figure 1. A schematic representation of DROIDS comparative molecular dynamic analysis software. DROIDS is a software tool for multiple test corrected amino acid-level pairwise comparison of molecular dynamics of two comparable PDB structures. The three main phases of analysis include (A) MD sampling runs and vector trajectory analysis, (B) statistical comparison via multiple test corrected KS tests, and (C) visualization results on static and moving images.

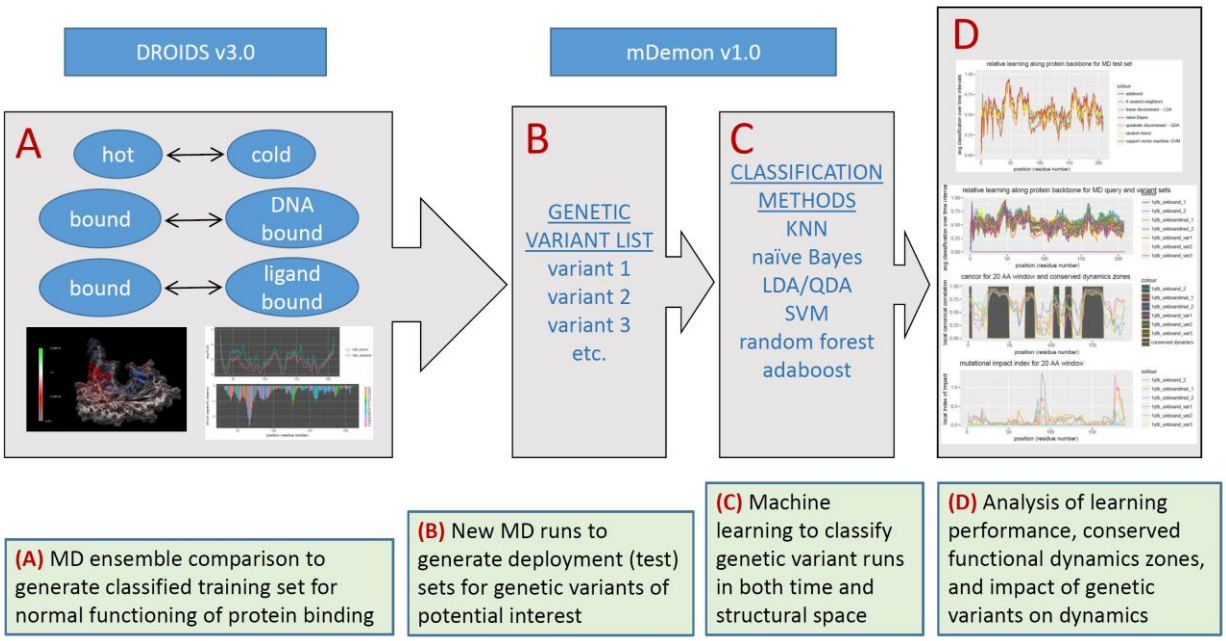

Figure 2. A schematic representation of DROIDS+maxDemon comparative molecular dynamic analysis software. The addition of maxDemon allows for machine learning classification of comparisons trained in DROIDS, to be deployed on new MD runs that represent genetic or drug binding variants. maxDemon reports regions where protein dynamics is functionally conserved and where significant impacts on conserved dynamics is created by each variant(s).

Running MD with Amber via DROIDS

This MD GUI interfaces (Figure 3 and 4) allow the user to set the most important parameters for the MD (e.g. name the force field, set run times of each phase, choose a solvation method, add salt conc) as well as determine how many sampling MD runs on each protein will be analyzed in later analysis. For most proteins, I often take 50-100 sampling runs at 0.5ns each, after a single equilibration phase of 10-50ns…depending upon how stable the structure behaves. Users are guided through creation of a structurally-based sequence alignment using Chimera MatchMaker and Match->Align, followed by setup of topology and coordinate files using teLeap. Then the script automates the energy minimization, heating, equilibration and MD production sampling runs on the two homologous structures and reports the progress to the Linux terminal. This part of the analysis takes the longest (e.g. the two comparative runs on two typical implicitly solvated systems may take 24-48 hours to run on the GTX 1080 card). Explicit solvated systems may run 2-3X longer. Details about the MD are hard coded into the portion of the script that writes the control file (i.e. the control subroutine). These settings can be easily changed by users with some experience with Amber commands and perl scripting. The default assumes constant temperature (300K) and pressure during production. Note that MD output is produced in the form of binary files (.nc file type extension) rather than text (i.e. .mdcrd file type). This is to allow the saving of hard drive space and proper file type for cpptraj analysis that follows. These files are not ‘readable’ in any

sort of text editor. Jobs are scheduled to the GPU by means of a while loop that periodically pgreps the process ID's produced by pmemd.cuda. The GPU will not automatically control job scheduling the way a CPU will. So we have added a GPU surveillance button that opens terminals that monitor the load on the GPU as well as current running processes. If the user interrupts a script and starts another job, this will not terminate the previous run. If the user sees that two pmemd.cuda processes are running at once, then the data is likely corrupt as the GPU is attempting to run both jobs at the same time. We include a 'kill' button which will pkill all pmemd.cuda jobs. This is handy when restarting DROIDS after previous interruption. It is recommended that user keep surveillance open at all times alongside the main terminal when running then MD wrapping script (GUI_START_DROIDS.pl). See Figure 2 for how this should look on your desktop. Before each sampling MD run, a random time length spacer is generated uniformly distributed between 0 and 0.5 x length of the sampling run. The purpose of this step is to average out the effect of chaotic dynamics that may be observed if the initial starting conditions were always exactly taken after the equilibration step has finished. A typical DROIDS analysis might consist of 0.5ns heating, 10-50ns of equilibration and 50 x 0.5ns of sampling runs on each protein. With this setting, most comparisons of protein dynamics can be achieved in 12-48 hours of run time using a dual GPU machine with GTX 1080.

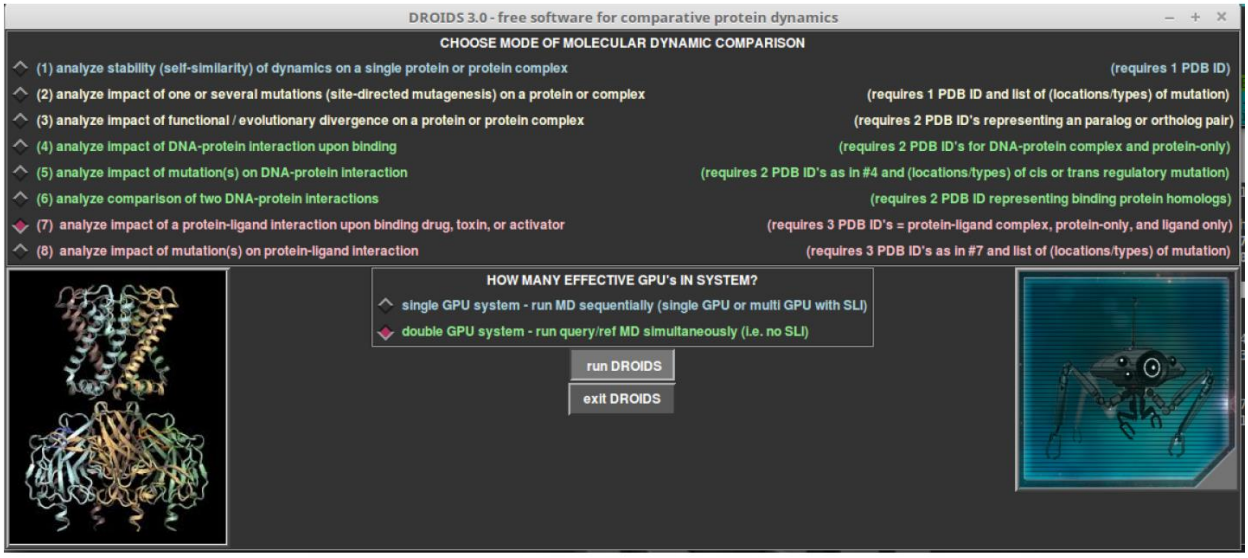

Figure 3. The DROIDS GUI interfaces for controlling molecular dynamic simulations and sampling conditions in Amber16/18 and subsequent cpptraj analysis.

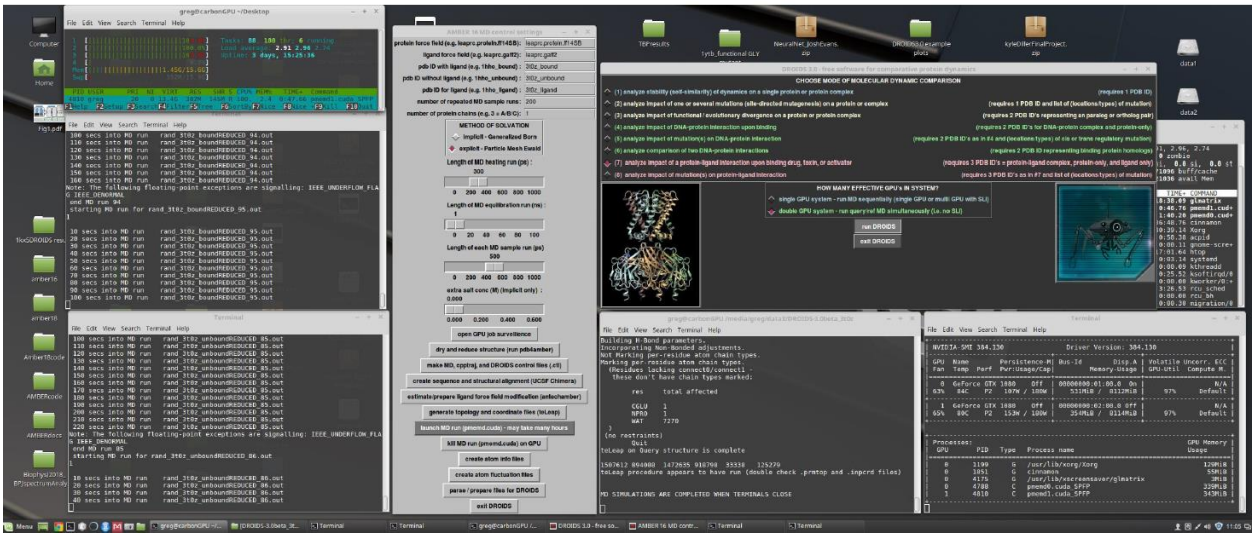

Figure 4. Linux terminal windows showing the progression of the MD simulations as well as general surveillance of GPU loads and process IDs

Calculation and collection of atom fluctuations

After the end of the MD simulations the user is guided through vector trajectory analysis using cpptraj (Ambertools16/17). The buttons are run from top to bottom and include making control files, collecting atom information, calculating atom fluctuations, and lastly, preparing and parsing the cpptraj output for subsequent DROIDS analysis. The setup we use under the hood is designed to return amino acid averaged motions collected only over the backbone of the polypeptide chain (i.e. N, CA, C, O …Figure 5). Fluctuation is very rapid (10-20 femtoseconds on most bonds) and largely harmonic and thus is relevant to comparative studies of protein stability (i.e. evolution of thermostability, functional epigenetic modifications, or disease-related genetic mutations that globally destabilize function. During initial setup (start GUI), the user is also guided from the terminal through the creation of a structural alignment of both protein structures using Chimera's MatchMaker and Match -> Align tools. The user is directed to save the resulting sequence alignment as a Clustal format file (.aln) using the name of the reference PDB ID in the title as follows Nxxx_align.aln (e.g. ubiquitin would be 1ubq_align.aln). Not that it is very important that the user trims the N terminal chains to the same length after alignment so that data is collected correctly from homologous amino acids. In GUI 2, the user is now also asked to specify whether the DROIDS statistics and mapping are to be conducted using 'loose' or 'strict' homology. Strict homology will only conduct MD comparisons on the backbone atoms of the protein when the aligned amino acid residues are identical. Loose homology will compare backbone MD even when residues are different as long as the structural alignment file identifies then as homologous. Note: atoms in sidechains are always excluded from all analyses via a mask used in cpptraj. When pipelines use strict homology on a protein comparison

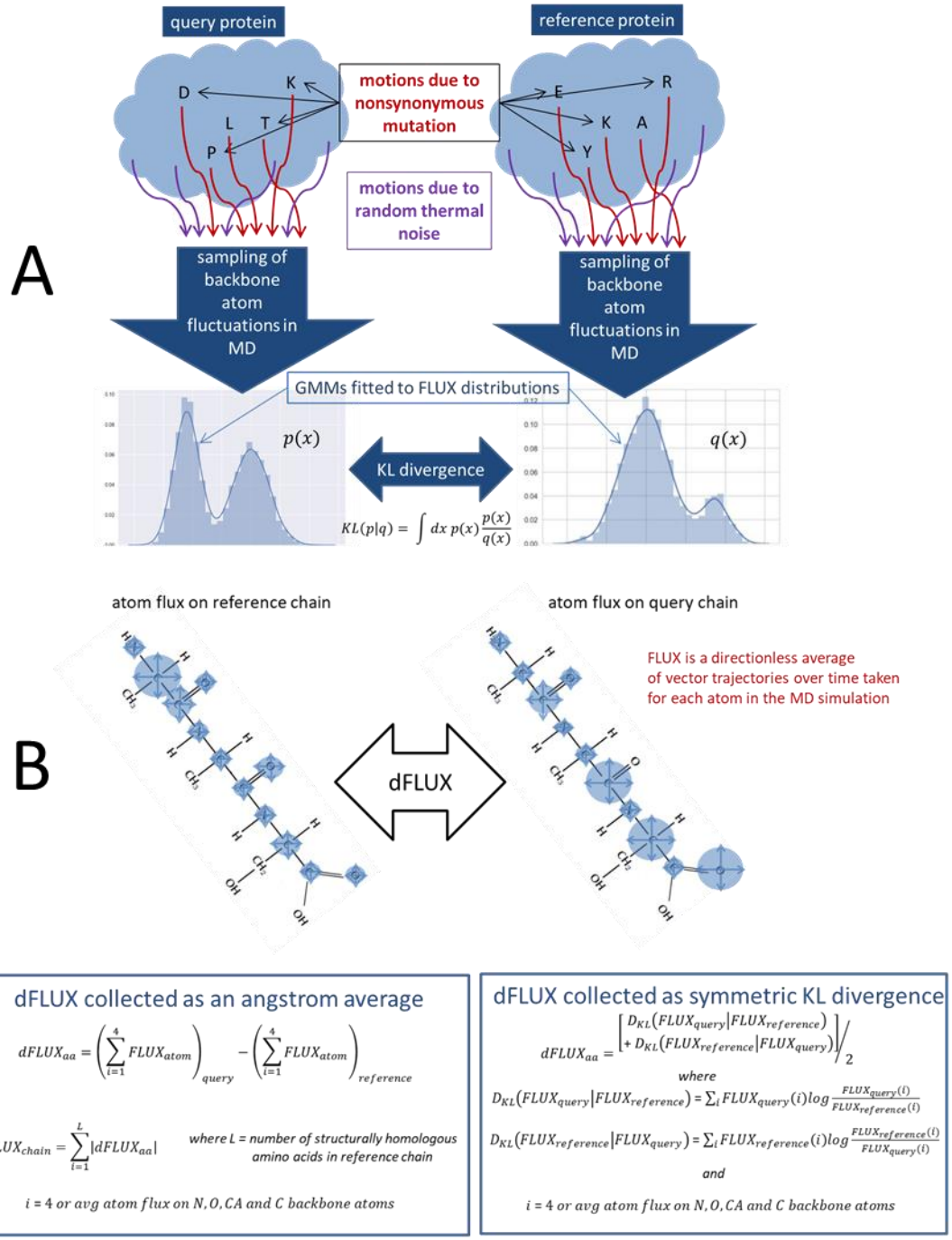

Figure 5. A schematic representation hypothetical differences atom fluctuation (dFLUX). Functional analysis of destabilization due to mutation and or evolution of functional thermostability can be addressed using dFLUX. In the DROIDS color mapping, (A) dFLUX is averaged over the 4 backbone atoms of each amino acid. Global dFLUX for the whole chain is simply the sum of absolute dFLUX over the length of the polypeptide chain. (B) Version 2.0 also allows dFLUX to be defined using symmetric Kullback-Leibler divergence between the distributions of atom fluctuation. This option provides a richer view of differences when color-mapping dFLUX.

without a large evolutionary distance a brighter color selection (i.e. red or yellow) is used for nonhomologous regions as a way to label interesting mutations in the resulting images and movies of the dynamics. Under structural comparisons of greater evolutionary distances, where the underlying protein sequences are likely to be quite different, loose homology will be used along with a less conspicuous color (i.e. usually gray) to mark regions in the protein comparison that lack true homology (i.e. are poorly aligned) MatchMaker provides user ability to choose appropriate substitution matrices and gap penalties to reduce the problem of poor alignment. DROIDS automatically excludes these regions from analysis. NOTE: at the end of parsing, a folder named 'atomflux' should appear with individual files for each comparison per residue. The number of files in this folder should correspond to the number of residues in the reference protein that have homologous residues in the query protein. If there are far fewer files in the atomflux folder than expected, this is most likely due to the fact the sequence at PDB does not exactly match the structure. Occasionally, one will need to trim the alignment file to match the structure, and then rerun the parsing again.

Comparative analysis and visualization of mutational impacts on protein dynamics

The statistical analysis is the heart of comparative protein dynamics using DROIDS (Figure 6-8). The initial steps include making choices about the type of analysis you want, then producing the control files you need. Then you run the KS tests in R on the next button. R graphics will show analyses as a popup in the pdf viewer. After this step the user will generate Chimera 'attribute' files for color mapping. Color mapping generally scales in saturation with the strength of the delta shift in atom motion (fluctuation or correlation) between the two sets of MD runs. Regions lacking homology are darker gray. If you are only changing the mapping options (i.e. data types – delta, p, or D values, color schemes or scaling of plots), you do not need to rerun the statistical tests. If you change statistical test options (i.e. motion type, p value cutoff, or multiple test correction), you will need to rerun the KS tests again. As the number of KS tests equals the number of amino acids on the chain, correction for multiple testing is highly recommended. Multiple test correction methods included as options in DROIDS are the Bonferroni correction or Benjamini-Hochberg estimation of false discovery rate (see Figure 6). The dFLUX values of the query runs can be scaled to the absolute dFLUX values of the reference runs if the user is more interested in relative difference rather than absolute difference. Be sure to choose color schemes that correspond to the data type as indicated on the screen. Color gradients can be auto-scaled (highest to lowest value) or fixed at one of several options. When statistical options are changed (excepting p value corrections) a new DROIDS results folder is generated for each set of tests. After the Chimera attributes are stored for mapping, the user can generate color mapped static structures in Chimera and/or render movies with the appropriate color mapping shown from 6 points of view X1, X2, Y1, Y2, Z1, and Z2…or alternatively with 2 points of view incorporating a smooth vertical and horizontal roll during playback. These movies can be viewed simultaneously in concert in the DROIDS movie viewers (see Figure 7).

While the colors mapped correspond to the overall analysis, the movie dynamics correspond to only the first MD sampling run taken on the reference PDB structure.

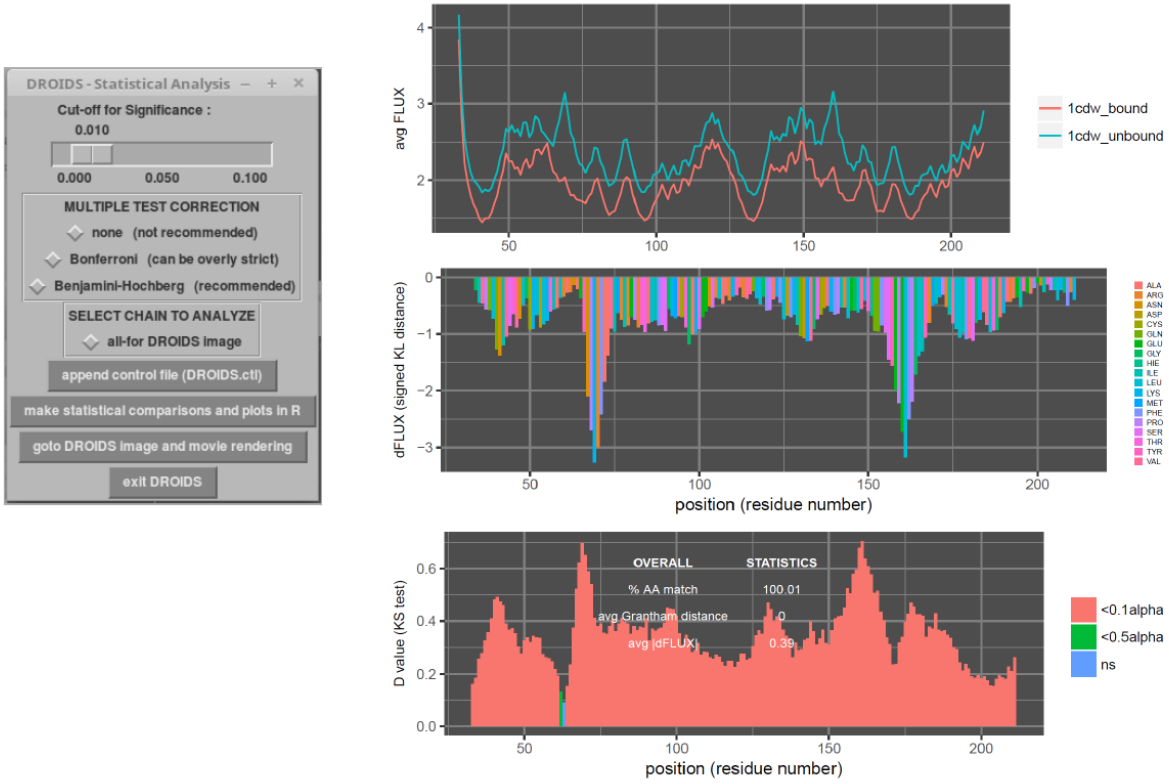

Figure 6. The stats DROIDS GUI controlling the KS statistics, multiple test correction method and graphics options. Negative peaks in dFLUX plot (middle) indicate regions of protein where DNA binding is most pronounced.

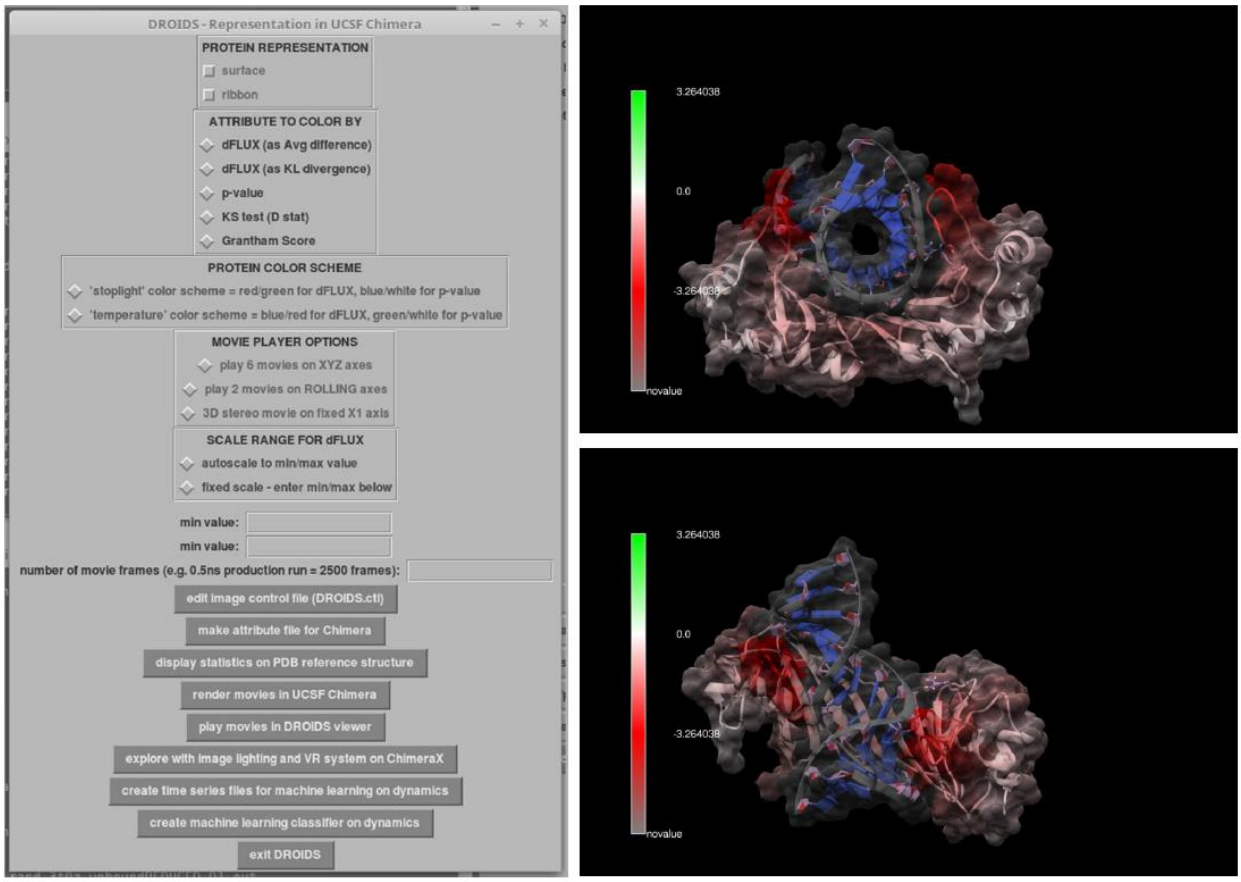

Figure 7. GUI for graphics options. A movie viewer showing six points of view (front, back, left, right, top, bottom) is also provided. Bivariate color option 'stoplight' for dFLUX is shown here. Red indicates dampening of rmsf values during DNA binding. Peaks in Figure 4 have deep red color and indicate loops in the DNA minor groove. Univariate coloring options for p or D values of the KS test are also provided.

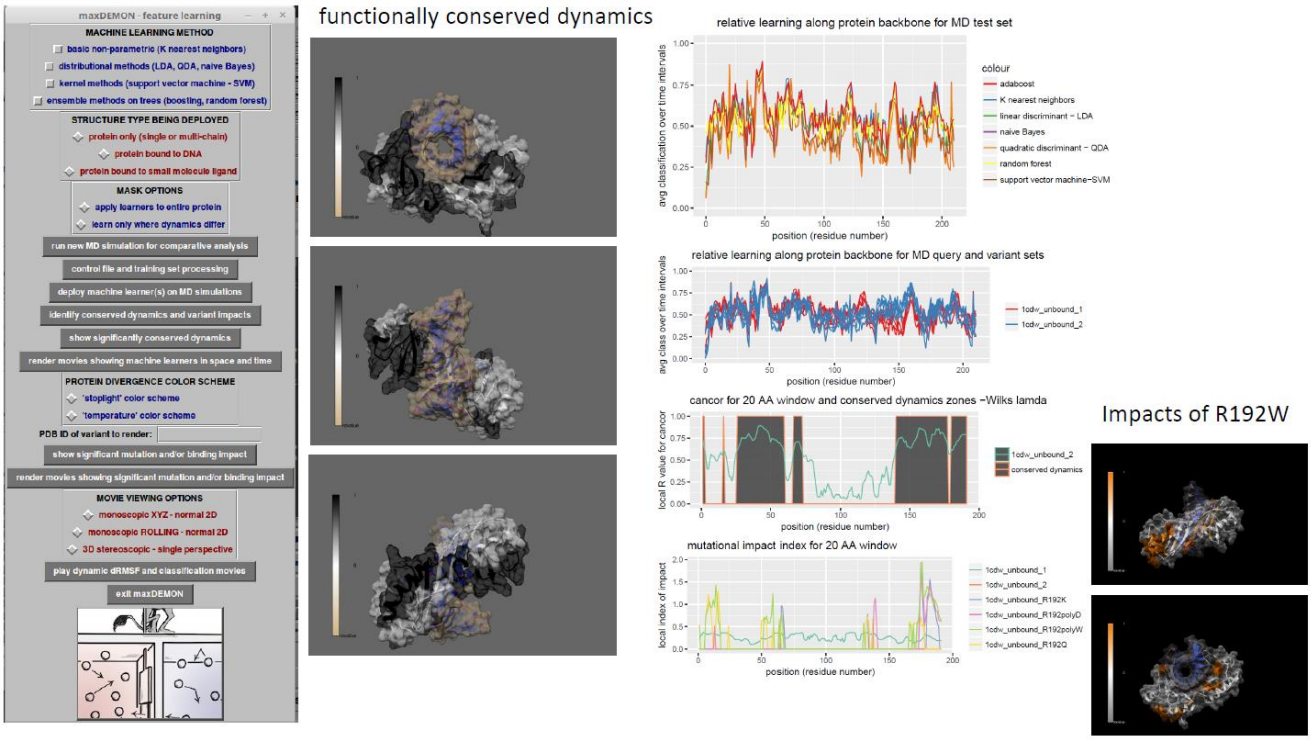

Figure 8. GUI for maxDemon – machine learning assistance for comparative protein dynamics.

The maxDemon application (Figure 8) allows for further machine learning assisted functional interpretation of the DROIDS comparison. The details of this analysis will be described in a future upcoming software note. Essentially, maxDemon trains up to nine different machine learning methods on the original comparative ensembles run earlier in DROIDS. It then allows the user to deploy a list of new runs representing the original query PDB (i.e. copies) and different PDB (i.e. variants). The significance of canonical correlation (i.e. Wilk's lamda) of the positional learning efficiencies of each method deployed on identical MD runs of the original PDB query on which the learners were trained (i.e. copies) is used to define functional conserved dynamics (i.e. dynamic signatures that are repeatable and dependent upon amino acid sequence). These areas are shown in very dark gray in the plots and images above. The impacts of variants are defined where the relative entropy of the variant canonical correlation differs significantly from that of the self comparison. The regions impacted significantly by mutations or drug variants are plotted as peaks in the bottom graph and as orange regions on the PDB structure.