ATOMDANCE v1.0.0 - statistical machine learning post-processor for comparative biomolecular dynamics

ATOMDANCE software containing DROIDS 5.0/maxDemon 4.0/Choreograph 1.0 is a python-based suite of machine learning assisted statistical methods for comparing molecular dynamic trajectories of proteins in two functional states (e.g. unbound vs. bound to something or wildtype vs mutated or hot vs. cold). It was developed on a python 3 science stack and only additionally requires the cpptraj library software and UCSF Chimerax molecular visualization software to be installed. The methods and software is offered freely (without guarantee) under GPL 3.0 and was developed by Dr. Gragory A. Babbitt and bioinformatics students at the Rochester Institute of Technology in 2017-2023.

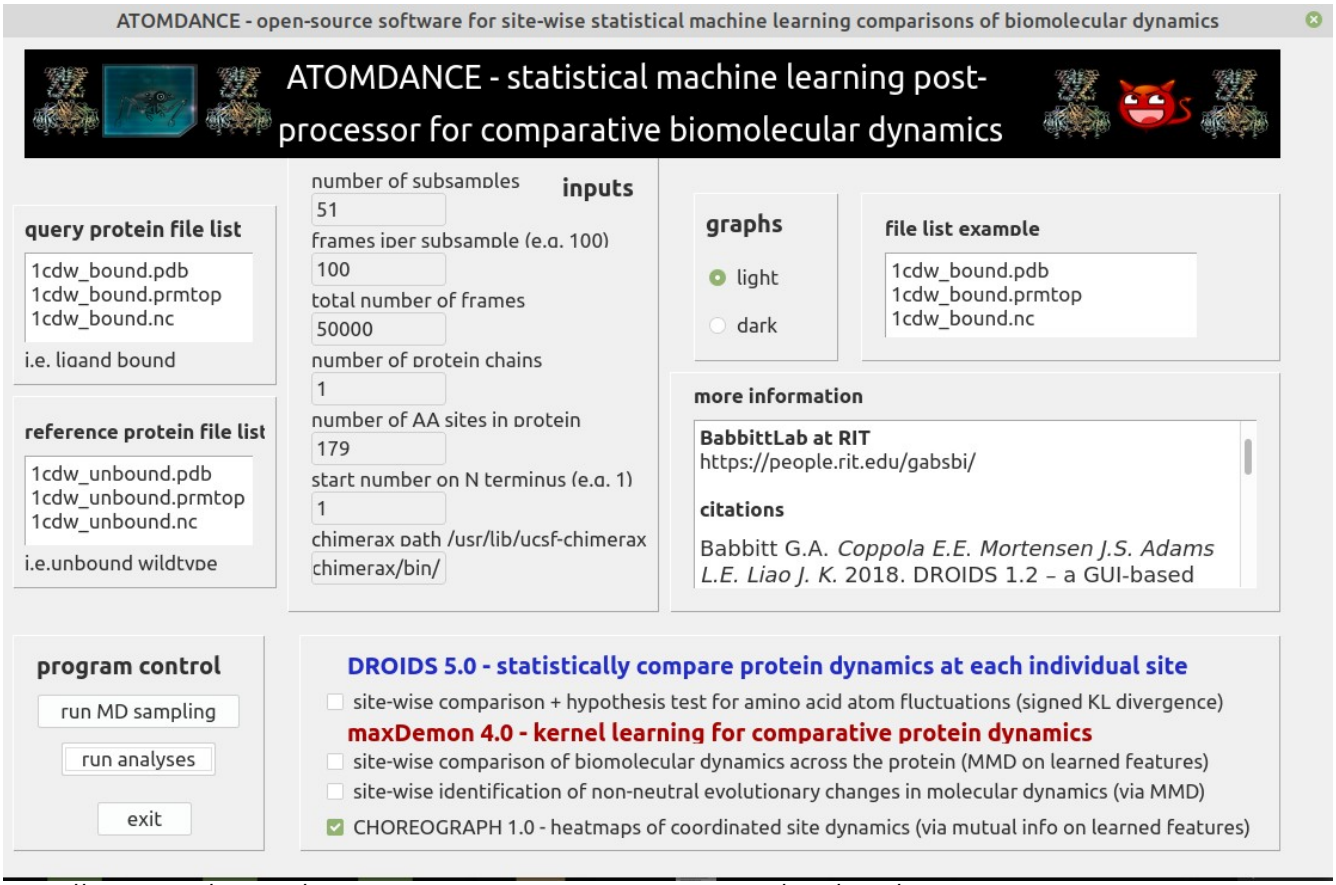
ATOMDANCE combines 3 main programs

DROIDS 5.0 - (Detecting Relative Outlier Impacts in Dynamic Simulation) providing stie-wise comparisons (i.e. divergence metrics) and hypothesis testing for amino acid fluctuations

maxDemon 4.0 - (kernel-based macine learning for comparative protein dynamics) This provides (A) site-wise comparisons of atom fluctuations and atom correlations in molecular dynamics simulations utilizing max mean discrepancy (MMD) on learned features (B) site-wise identification of non-neutral evolutionary changes in molecular dynamics (also via MMD).

Choreograph 1.0 - provides heatmaps of coordinated site dynamics (via mutual information of learned feature classifications obtained by a support vector machine)

GUI layout -

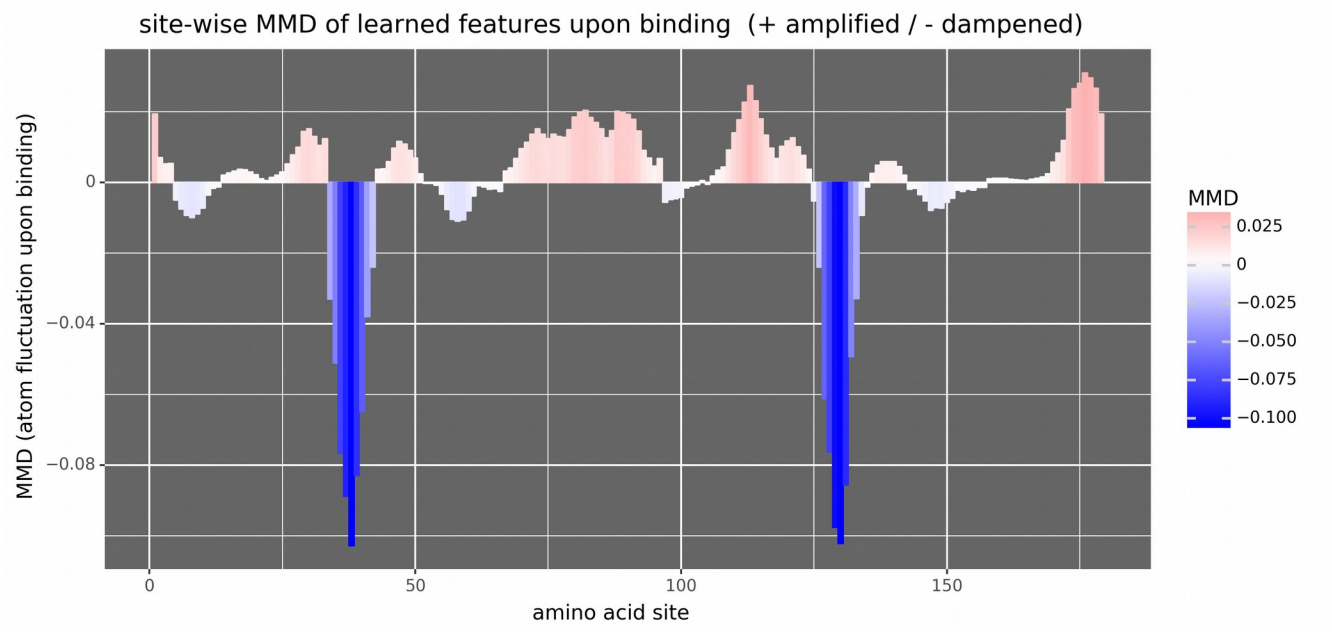


(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/atomdance_gui.png)

Analyses consist of two steps (A) random sampling of the MD trajectories (via GUI launch of cpptraj_sampler.py) and (B) statistical/machine learning post-processing (via GUI launch of chimerax_analyzer.py)

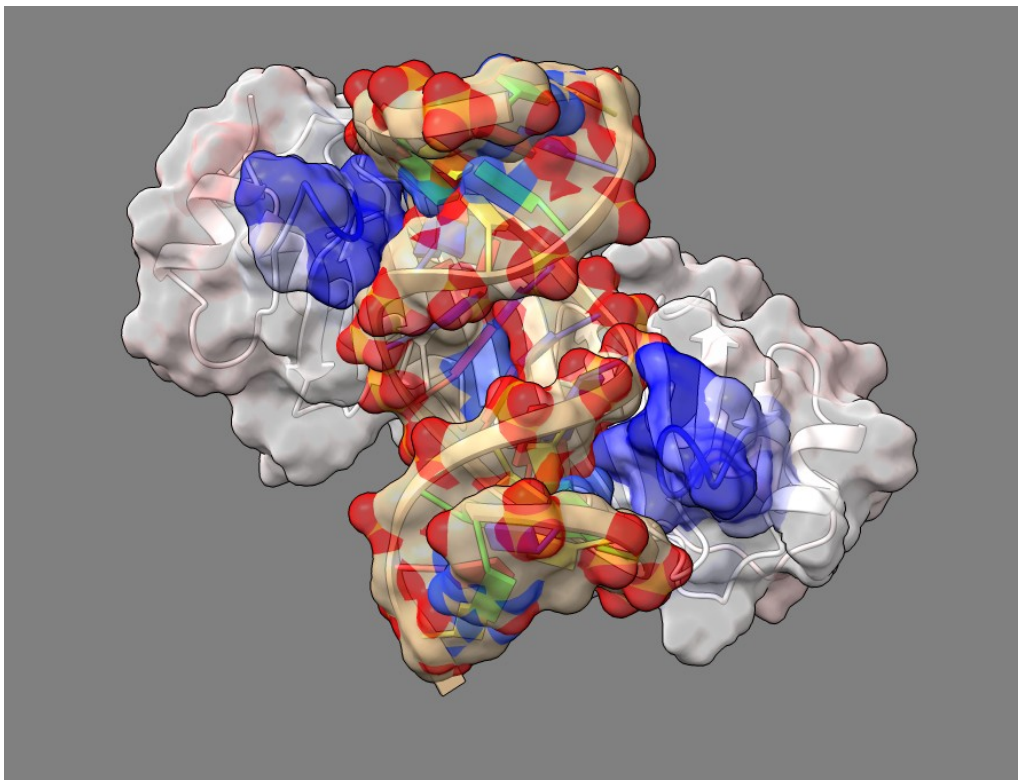
SOME EXAMPLES (blue indicates dampened atom motion while red indicate amplified atom motion)

machine learning identification of functional DNA binding sites in TATA binding protein (via MMD)



(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/TBPplot.png)

…and mapped to structure (PDB: 1cdw) in ChimeraX

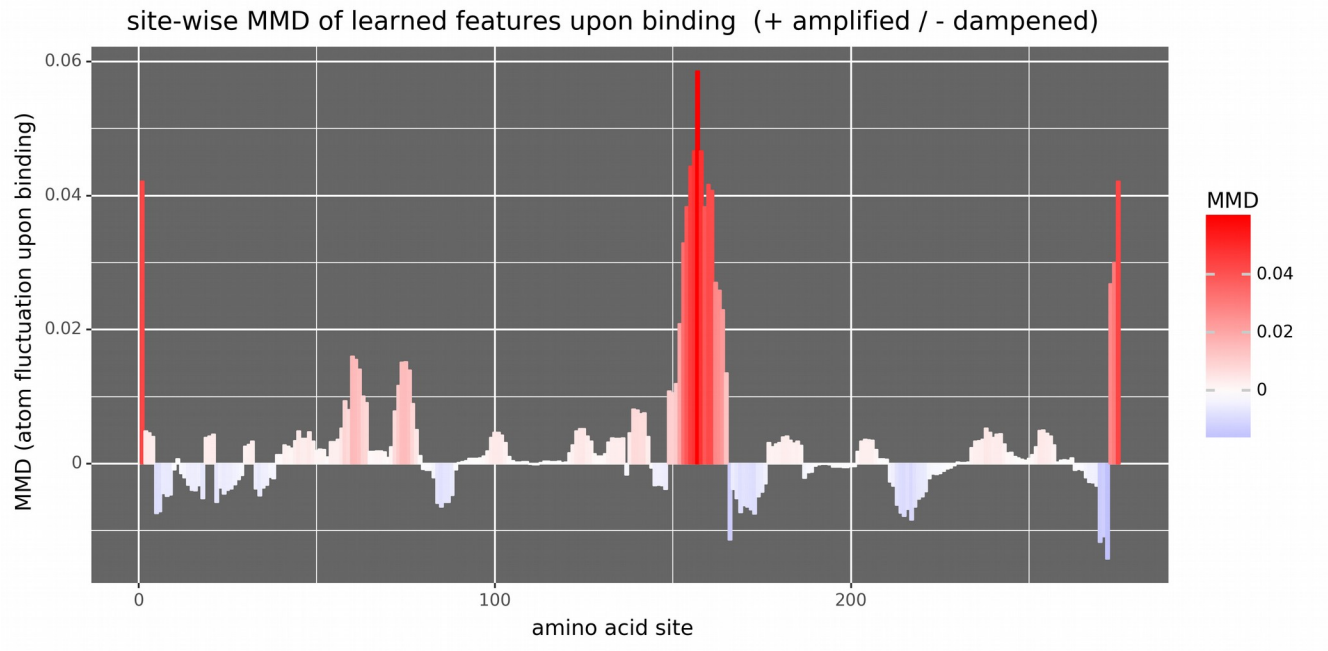


(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/TBPmap.png)

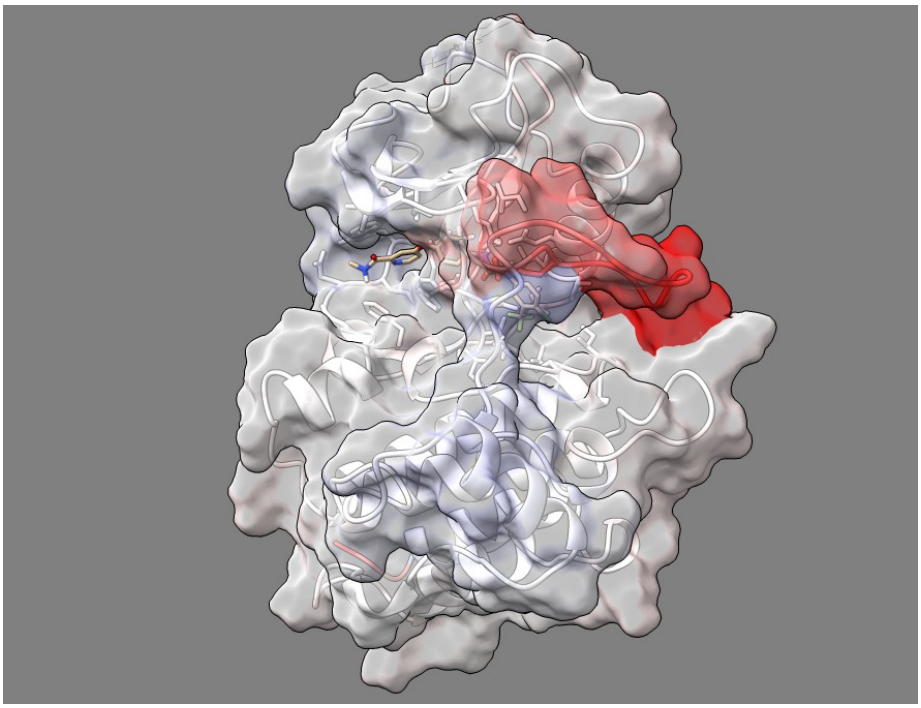
ANOTHER EXAMPLE
machine learning identification of BRAF activation loop during drug binding of ATP pocket

(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/BRAFplot.png)

…and mapped to structure (PDB: 1uwh) in ChimeraX



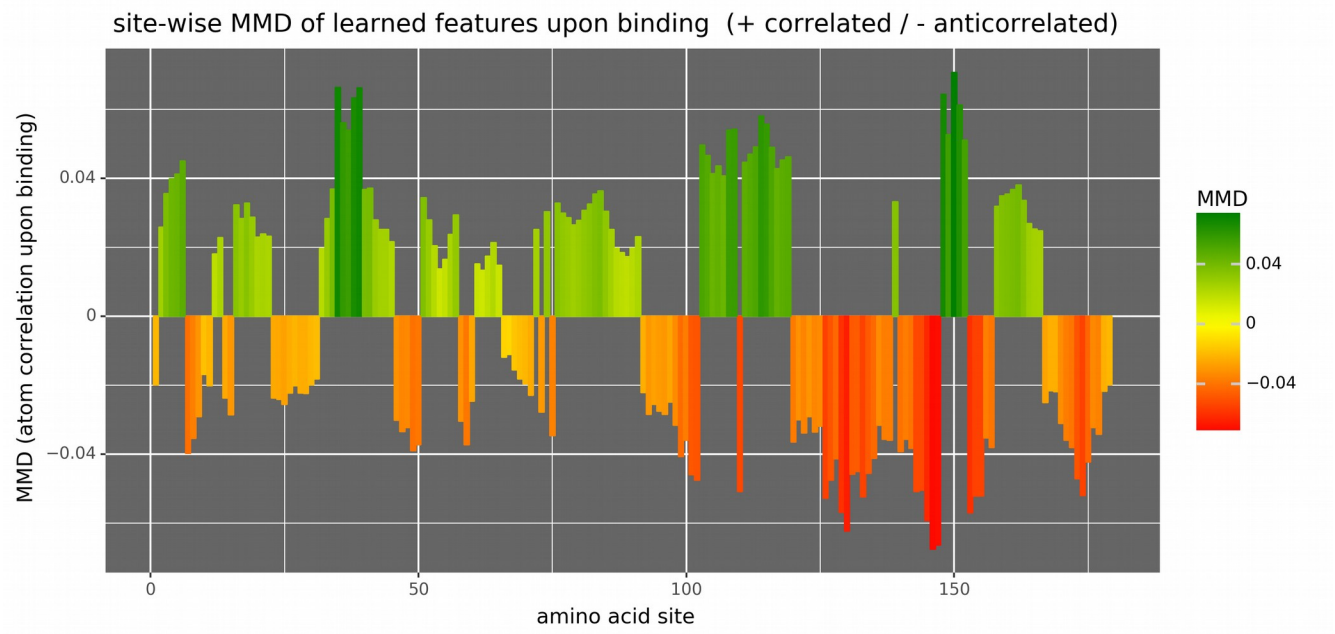
(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/BRAFmap.png)

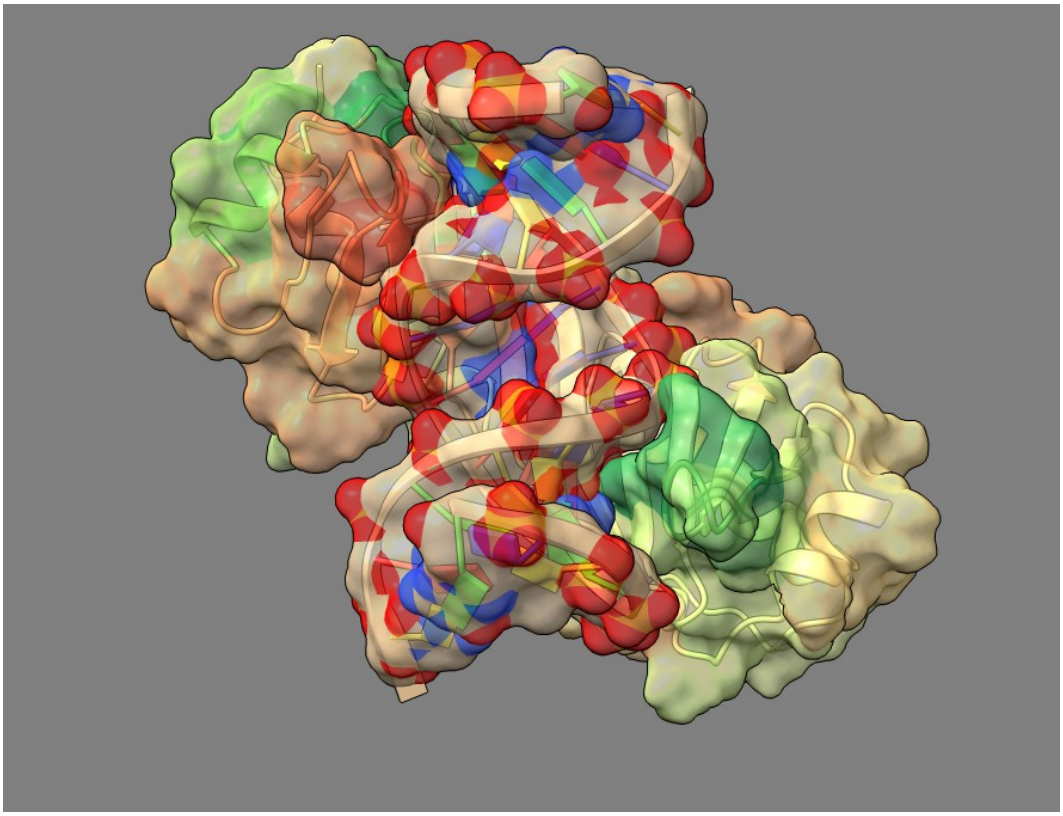


In these above examples, a Gaussian process kernel using a radial basis function is trained upon local atom fluctuations and each amino acid site. The feature vector for training includes atom fluctuation at the local site and that of its two flanking neighbors on the protein backbone chain. Fluctuations are filtered by residue and masked to include only the homologous atoms on the protein backbone (i.e. C, C-alpha, O and N). The maximum mean discrepancy (MMD) in the reproducing kernel Hilbert space is calculated between learned features of the bound vs unbound dynamic states. An empirical p-value for this MMD derived by training the machine learning on two different samples collected from the reference dynamic state (i.e. unbound motions) and then subsequent bootstrapping of MMD calculated between these two reference samples.

In addition to analyzing comparative shifts in atom fluctuation (above), ATOMDANCE also analyzes MMD in atom correlation in a similar fashion. The feature vector here comprises of a reduced correlation matrix created via truncated singular value decomposition representing 80% of the original variance of the unreduced correlation matrix between all amino acid sites on the protein. This analyses can provide comparison of much more subtle functional shifts in molecular dynamics upon binding that are affecting the interactions between amino acid sites.

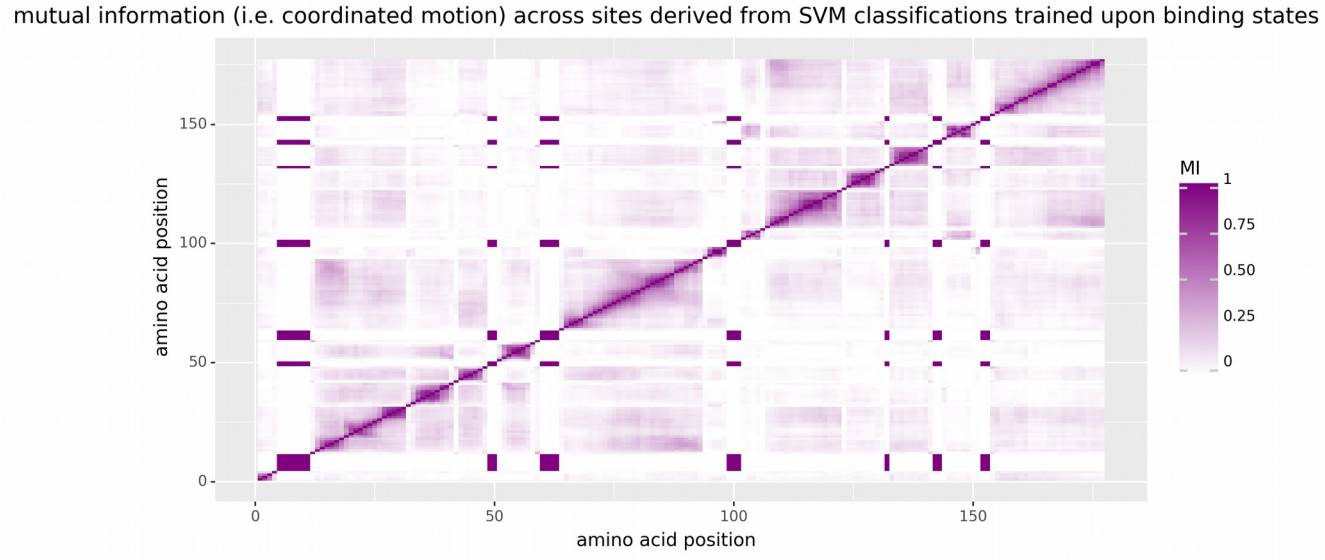
For example, here is output from an analysis of atom correlation during DNA binding of TATA binding protein. You can see that the two sites of atom dampening during binding (shown above) are actually anti-correlated and strongly influenced/connected to the major secondary structures in the rest of the protein.





Coordinated site dynamics can also be analyzed using Choreograph 1.0. This is a heatmap of mutual information derived from a kernel-based classifier (support vector machine) and can identify sites that are shifting in and out of correlated states together over time (indicated by high mutual information).

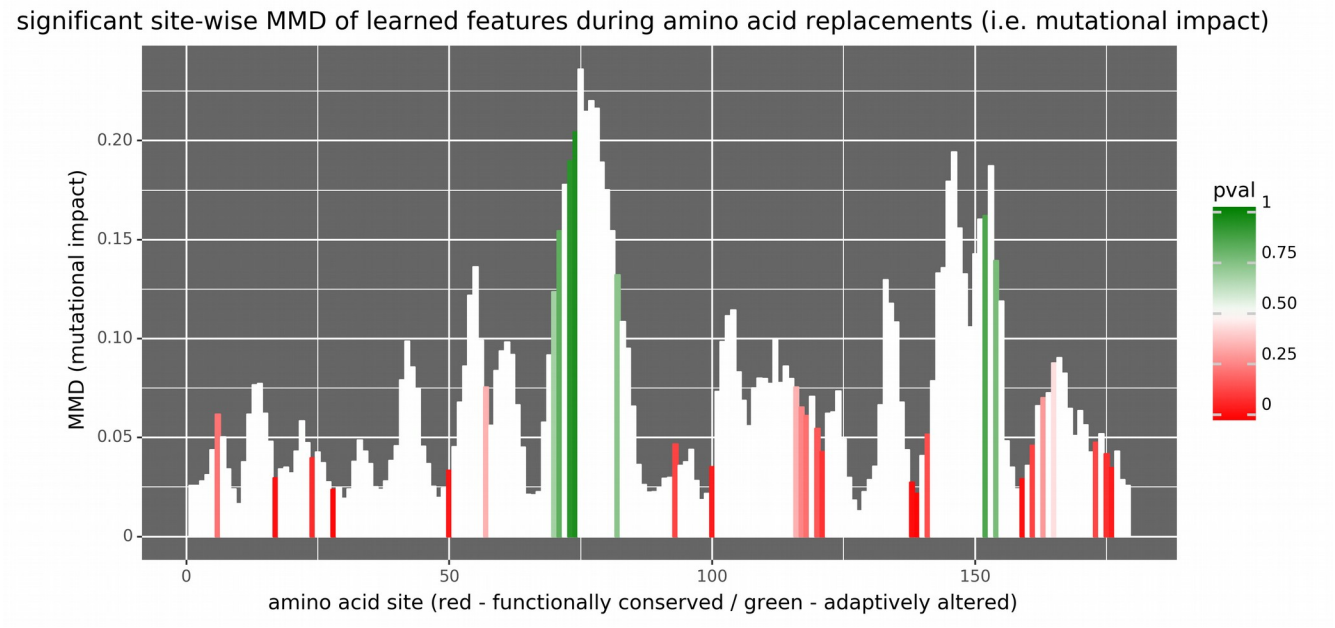
Here is the heatmap of mutual information for DNA-bound TATA binding protein



We can compare this to the same plot derived from a negative control (two MD runs of ubiquitin conducted at the same temperature)

ATOMDANCE can also be used to compare the functional binding dynamics of genetic variants (e.g. orthologs, polymorphisms or mutants). In this analysis, the MMD derived from learned features (including both atom fluctuation and atom correlation) is calculated at amino acid sites that differ between the wildtype and the genetic variant. An empirical p value is derived from a bootstrap distribution of MMD calculated from comparison of dynamics of different amino acid sites across the bound state of the wildtype and variant proteins. This allows the MMD at sites of amino acid replacement to be compared to a proxy for neutral evolution, thus allowing the identification of sites with unusually high MMD (altered dynamics) from unusually low MMD (conserved dynamics).

Below is a such a comparison of TATA binding protein of human and plants (arabidopsis).



ATOMDANCE utilizes the cpptraj program (Daniel Roe) and UCSF ChimeraX and a minimal number of python libraries. More information about installing these can be read below.

more about the BabbittLab@RIT https://people.rit.edu/gabsbi/

more on cpptraj https://github.com/Amber-MD/cpptraj

GitHub repo for cpptraj https://amber-md.github.io/cpptraj/CPPTRAJ.xhtml

TO INSTALL cpptraj

check/install gcc, g++ and gfortran compilers (e.g. sudo apt install gcc g++ gfortran)
sudo ./configure gcc
make install
NOTE: after installing cpptraj then open bashrc file (e.g. $ gedit .bashrc), then add the following lines to open cpptraj from everywhere.

export CPPTRAJ_HOME=/home/myUserName/Desktop/cpptraj-master export PATH=$PATH:$CPPTRAJ_HOME/bin

To check this, open a terminal and type 'cpptraj'. If the program opens, this has worked. If you get an error message, you'll likely need to correct the bashrc file and try again

NOTE: to use older versions of cpptraj (version 18 and prior) open the three following files (cpptraj_parser.py, cpptraj_ortholog_sampler.py, and chimerax_coordyn.py) and change the line of code in the header part of the script to read 'cpptraj_version = 'old'' instead of 'cpptraj_version = 'new''.

more on UCSF ChimeraX https://www.rbvi.ucsf.edu/chimerax/

FOR OUR CODE: python module dependencies (os, getopt, sys, threading, random, re, chimerax.core.commands) python modules to be installed (PyQt5, numpy, scipy, pandas, sklearn, matplotlib, plotnine, progress) NOTE: for best results, the CPU on the computer should support at least 4-6 cores

Molecular dynamics file inputs to ATOMDANCE include 6 files (3 for each functional state including a .pdb formatted structure file, a .prmtop formatted topology file and a .nc (i.e. NetCDF) formatted trajectory file. To run the program put these input files in the local folder you have downloaded from us, open a terminal or cmd line from that folder and type 'python3 ATOMDANCE.py'. Then follow directions on the graphical interface. These files can be generated on any molecular dynamics engine the user prefers (e.g. QwikMD using NAMD, OpenMM in python, or Amber/Ambertools in Linux). For beginners, we also offer a useful GUI for Amber MD simulations on Linux available here

https://gbabbitt.github.io/amberMDgui/ https://github.com/gbabbitt/amberMDgui

NOTE: before statistical comparison, your MD simulations should be appropriately set up (e.g. PDB should be cleaned up removing crystallographic waters and other stray molecules used in crytallization cocktails), your simulations should be appropriately equilibrated for stability, and your trajectory should be appropriately long enough to allow statsitical resampling of many conformational states. This is very different for various protein systems. However, it can often requires 10-100+ nanoseconds of simulation which can take many days even on the fastest GPU processors. The example files included with the ATOMDANCE software only have been run for relatively shorter periods on relatively stable proteins to allow ease of download from our website. To demonstrate the software, we include a negative control consisting of two MD runs on a small protein ubiquitin (1ubq) in identical function states and a positive control consisting of TATA binding protein simulated in both its DNA-bound and unbound functional state. To run these

python3 ATOMDANCE_ctlNEG.py python3 ATOMDANCE_ctlPOS.py

please cite us (as well as ChimeraX and cpptraj)

Babbitt G.A. Coppola E.E. Mortensen J.S. Adams L.E. Liao J. K. 2018. DROIDS 1.2 – a GUI-based pipeline for GPU-accelerated comparative protein dynamics. BIOPHYSICAL JOURNAL 114: 1009-1017. CELL Press.

Babbitt G.A. Fokoue E. Evans J.R. Diller K.I. Adams L.E. 2020. DROIDS 3.0 - Detection of genetic and drug class variant impact on conserved protein binding dynamics. BIOPHYSICAL JOURNAL 118: 541-551 CELL Press.

Babbitt G.A. Fokoue E.P. Srivastava H.R. Callahan B. Rajendran M. 2022. Statistical machine learning for comparative protein dynamics with the DROIDS/maxDemon software pipeline. In press. STAR Protocols CELL Press.