ATOMDANCE v1.3.0 - statistical machine learning post-processor for comparative biomolecular dynamics

ATOMDANCE software containing DROIDS 5.0/maxDemon 4.0/Choreograph 2.0 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 program students at the Rochester Institute of Technology in 2017-2023.

to start type 'python3 ATOMDANCE.py' (control= python3 ATOMDANCE_ctlNEG.py or ATOMDANCE_ctlPOS.py)

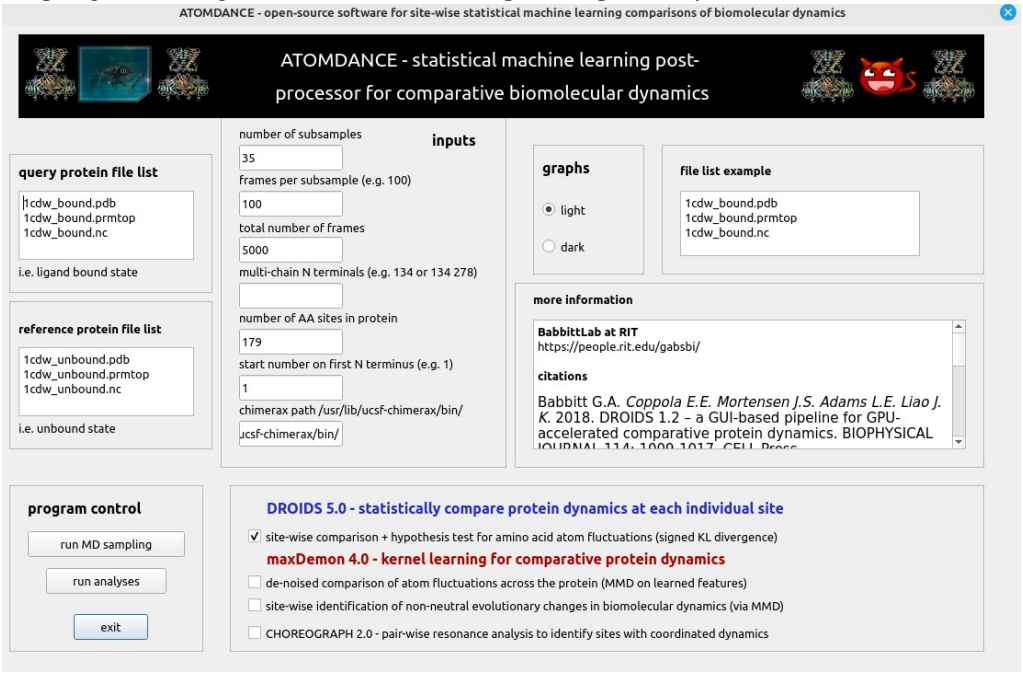
ATOMDANCE combines 3 main programs

DROIDS 5.0 - (Detecting Relative Outlier Impacts in Dynamic Simulation) providing site-wise comparisons (i.e. divergence metrics) and hypothesis testing for amino acid fluctuations. This analysis is appropriate for initial comparisons of dynamic states in which a proteins response to thermal noise is important to consider.

maxDemon 4.0 - (kernel-based macine learning for comparative protein dynamics) This provides (A) de-noised site-wise comparisons of atom fluctuations in molecular dynamics simulations utilizing max mean discrepancy (MMD) on learned features. This is comparison is very useful for finding functional binding sites that are obscured by thermal noise caused by the solvent in the dynamic system (B) site-wise identification of non-neutral evolutionary changes in molecular dynamics (also via MMD). When comparing ortholog proteins in identical functional states, this is useful for identifying amino acid replacements that have caused larger adaptive changes or smaller functionally conserved changes in dynamics than is expected due to chance alone (i.e. neutral evolution or genetic drift)

Choreograph 2.0 – This analysis provides heatmaps and network graph community detection of coordinated site dynamics or resonance (i.e. groups of amino acid sites that move in a coordinated fashion over time in either the reference or query states of a proteins dynamics). It maps all pair-wise site comparisons on a given protein using a mixed-model ANOVA comparing atom fluctuation of site i to site j with atom fluctuation as fixed effect and time as a random effect in the model. The p-values of the interaction between the difference in atom fluctuation between two sites and their changes over time is used to detect site resonance or coordination. Communities of sites that resonate together are detected using Louvain community detection algorithm and presented as a network map. Heatmaps and community graphs of non-significant difference in atom fluctuation (i.e. fixed effect in the model) are used to detect adjacent and non-adjacent sites of points of contact in the protein structure as well.

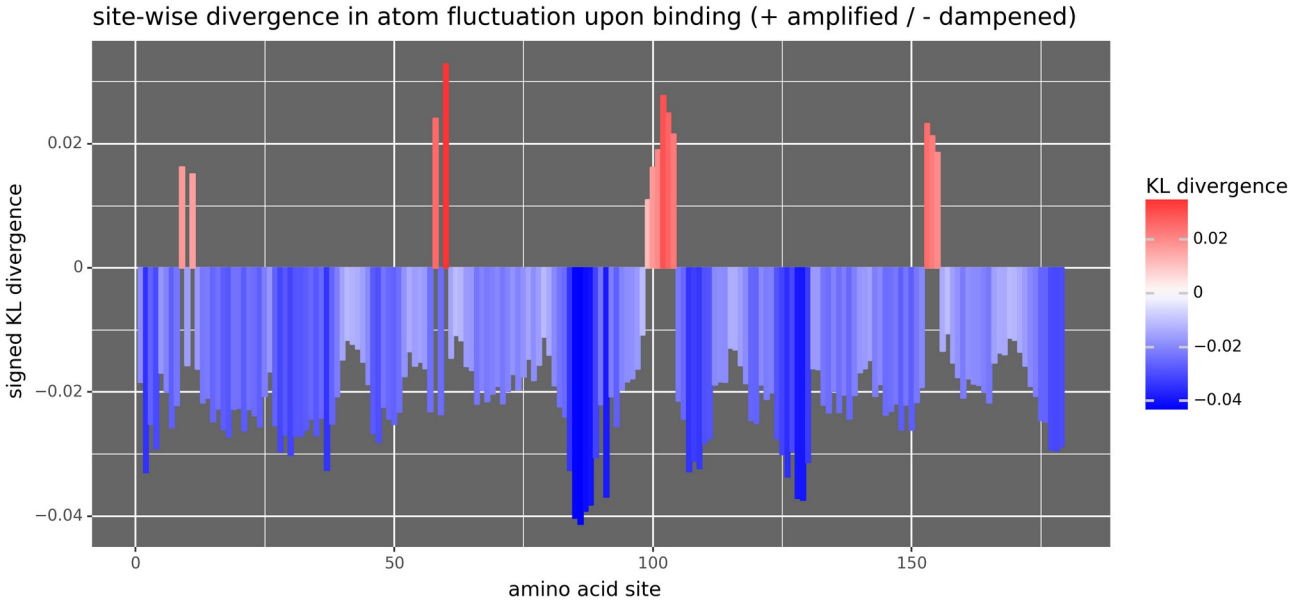
GUI layout - (https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/atomdance_gui.png)

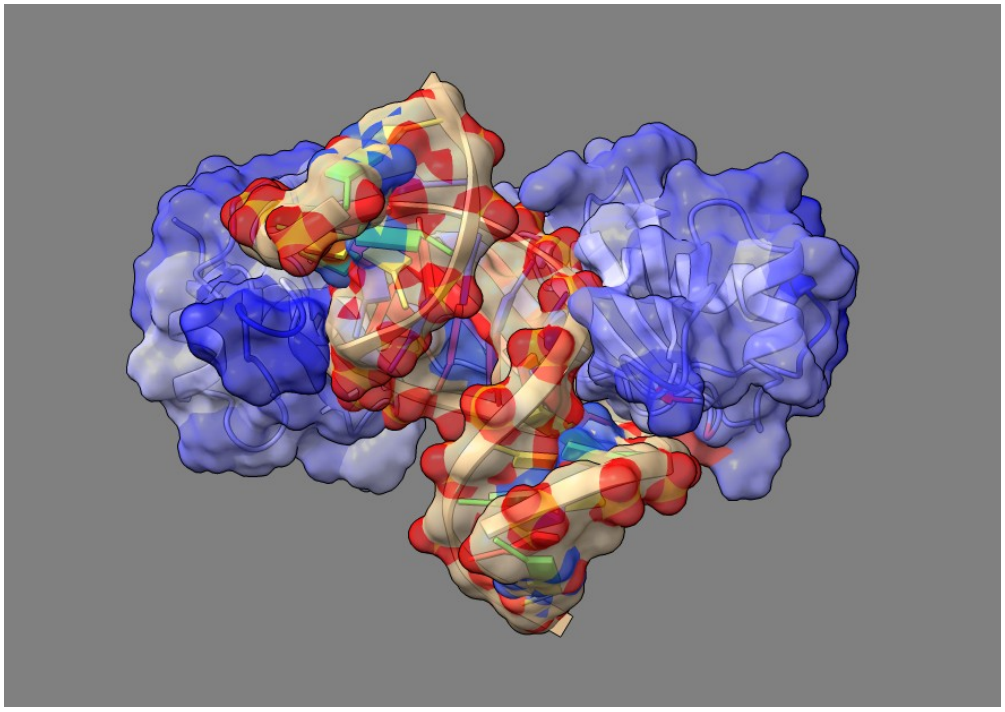


Analyses generally consist of two steps (A) random sampling of the MD trajectories (via GUI launch of cpptraj_sampler.py) and (B) statistical/machine learning post-processing (via GUI launch of chimerax_analyzer.py)

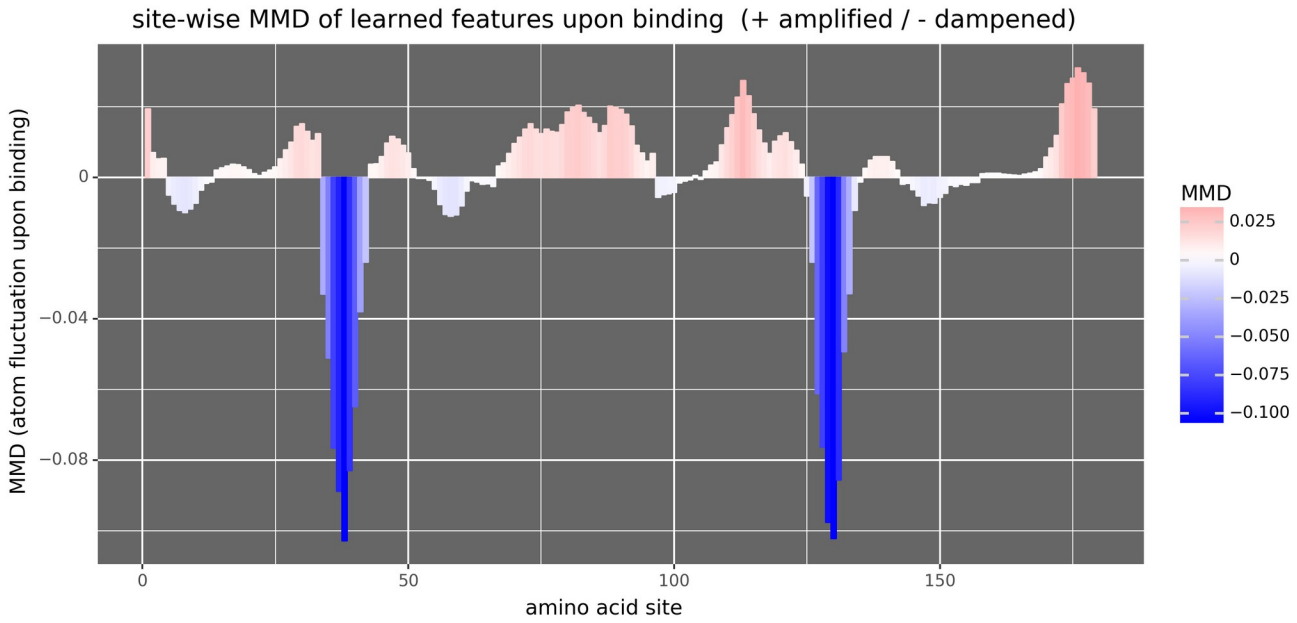
SOME EXAMPLES (blue indicates dampened atom motion while red indicate amplified atom motion)

DROIDS 5.0 comparison of atom dampening when TATA binding protein interacts with DNA



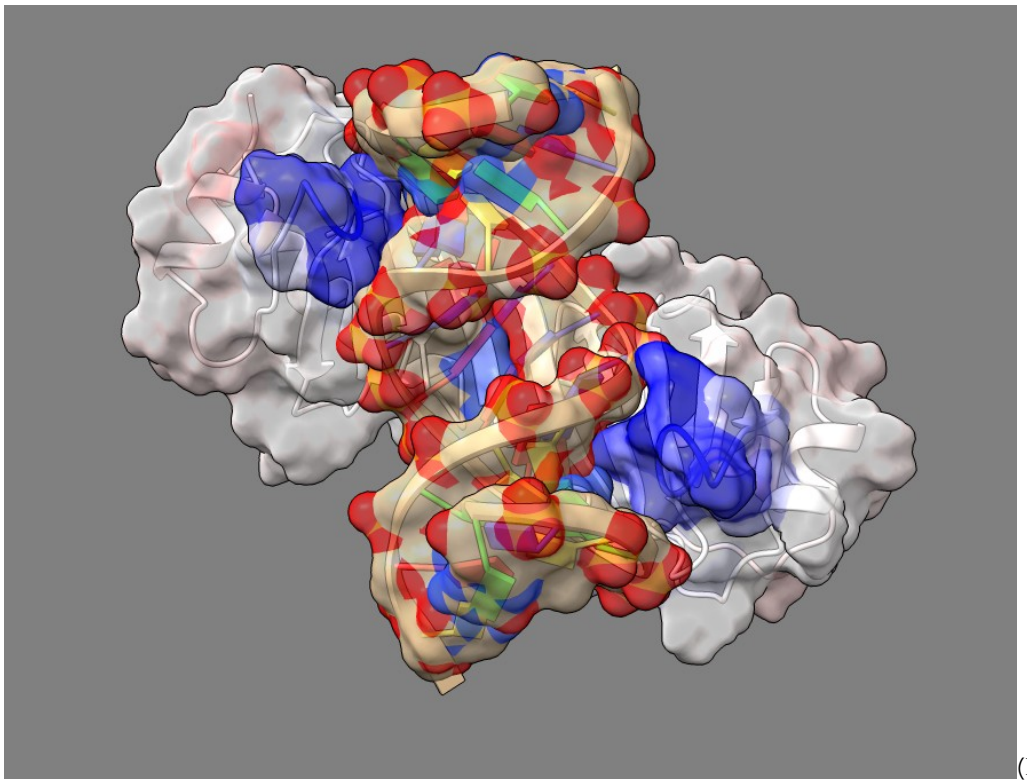


maxDemon 4.0 analysis of denoised machine learning identification of functional DNA binding sites in TATA binding protein (via MMD)



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

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

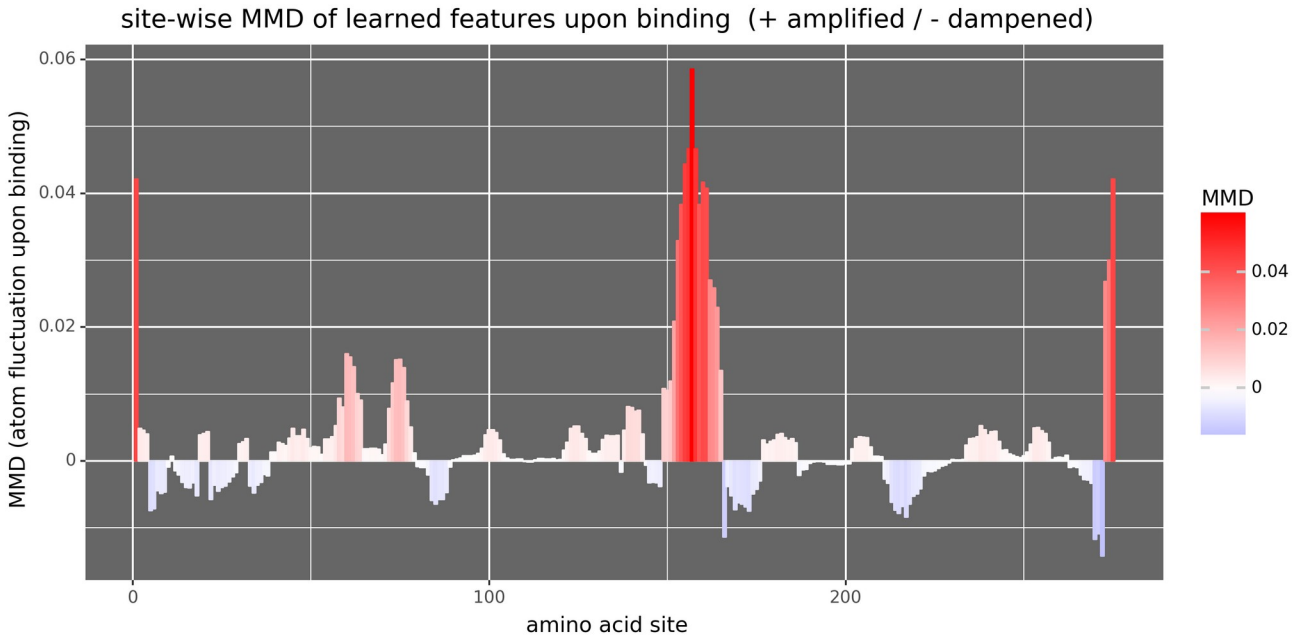


(https://github.com/gbabbitt/DROIDS-5.0-comparative-protein-dynamics/blob/main/TBPmap.png)

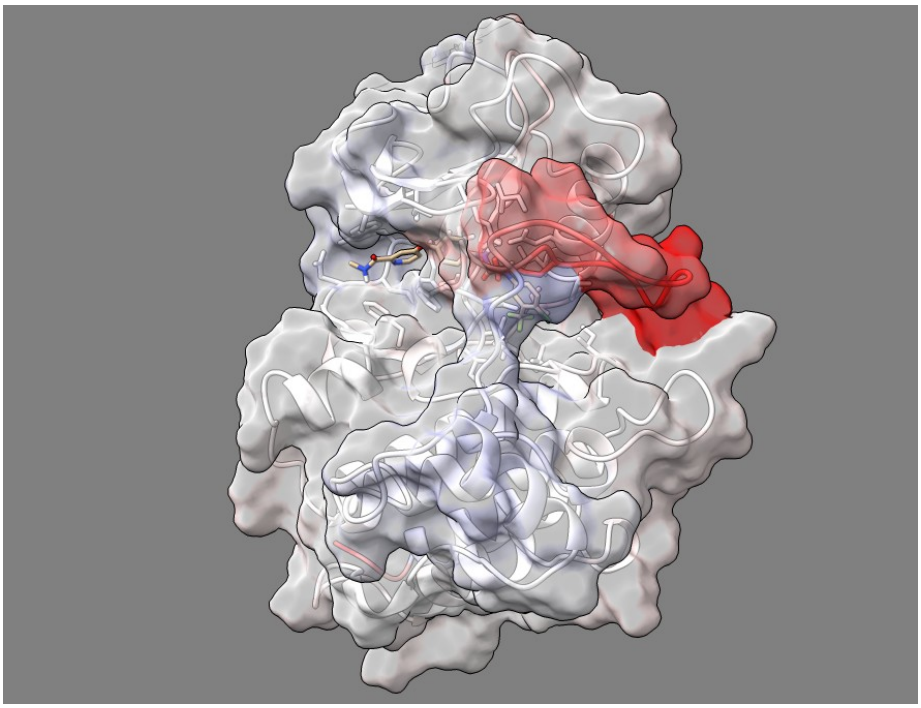
ANOTHER EXAMPLE
denoised machine learning identification of BRAF activation loop during drug binding of ATP pocket

(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)

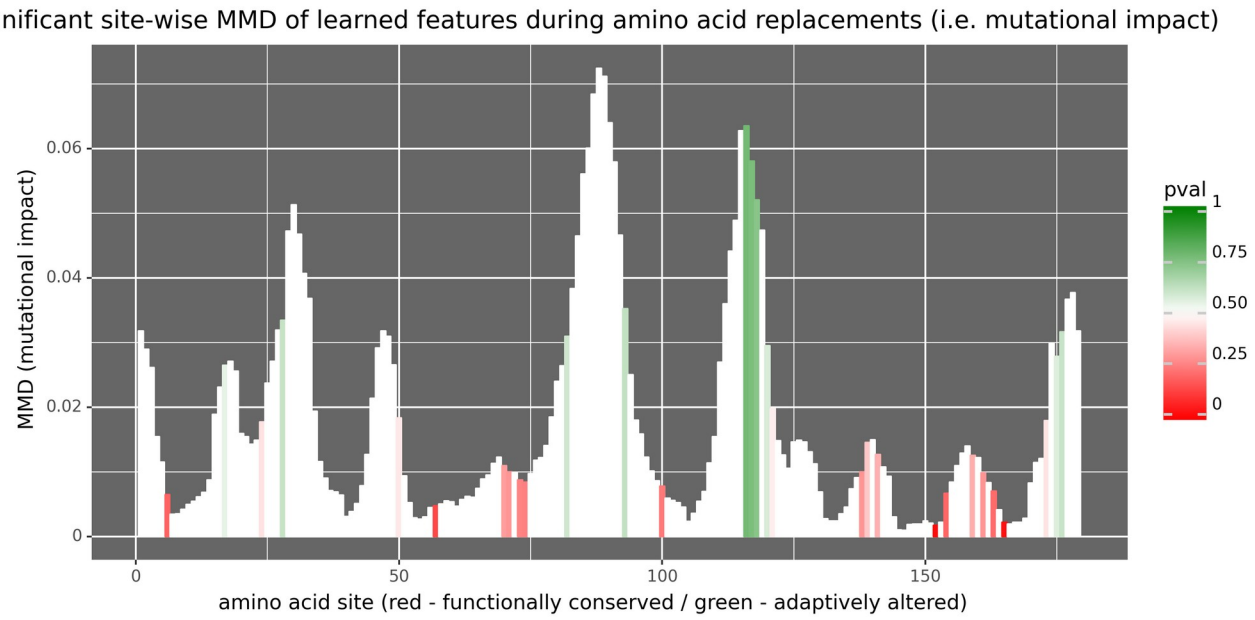


In these above examples, a Gaussian process kernel using a radial basis function is trained upon local atom fluctuations and each amino acid site. The feature vector for training includes atom fluctuation at the local site and that of its two flanking neighbors on the protein backbone chain. Fluctuations 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

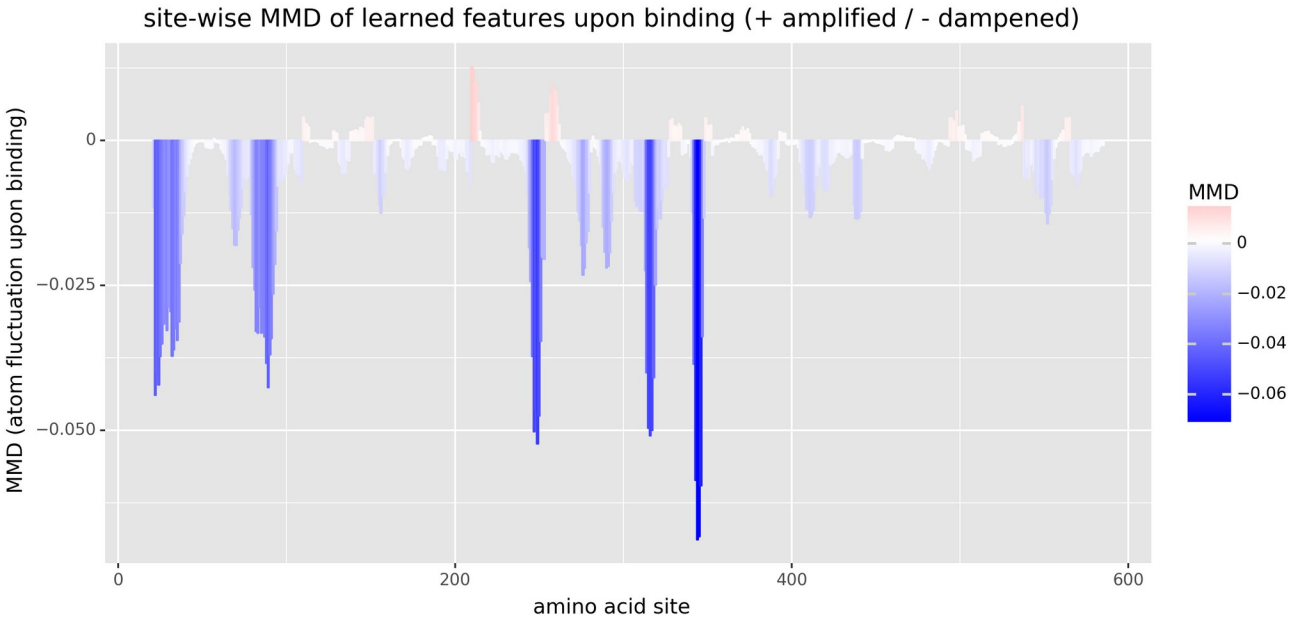
different samples collected from the reference dynamic state (i.e. unbound motions) and then subsequent bootstrapping of MMD calculated between these two reference samples.

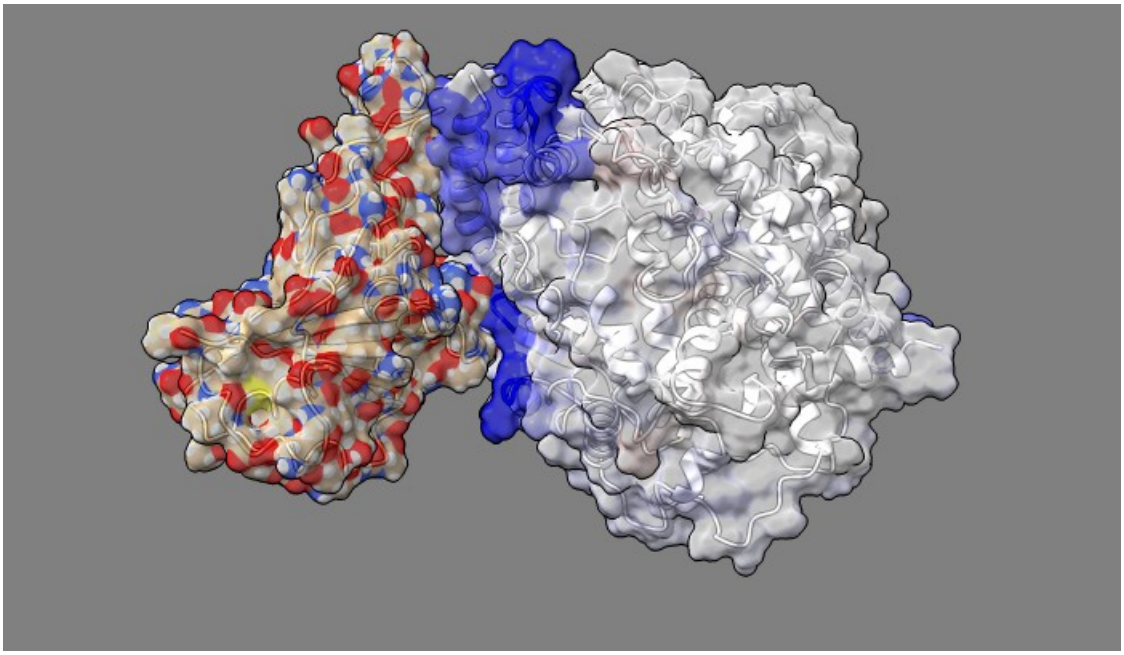
ATOMDANCE can also be used to compare the functional binding dynamics of genetic variants (e.g. orthologs, polymorphisms or mutants). In this analysis, the MMD derived from learned features (including both atom fluctuation and atom correlation) is calculated at amino acid sites that differ between the 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).

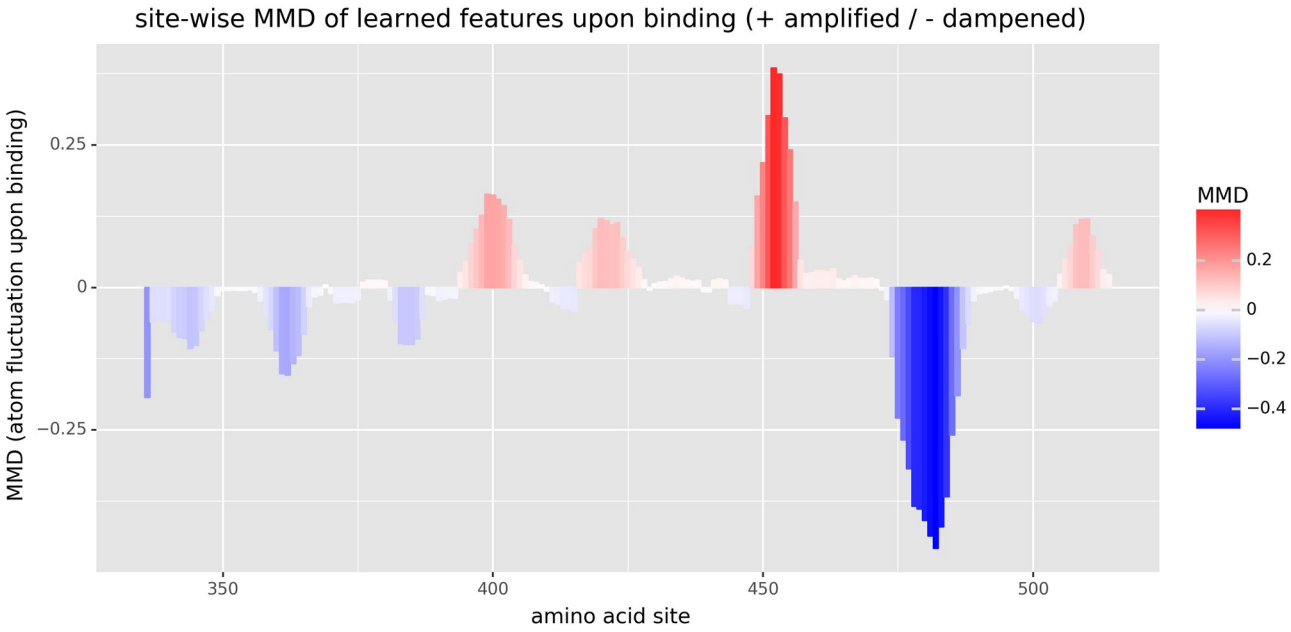


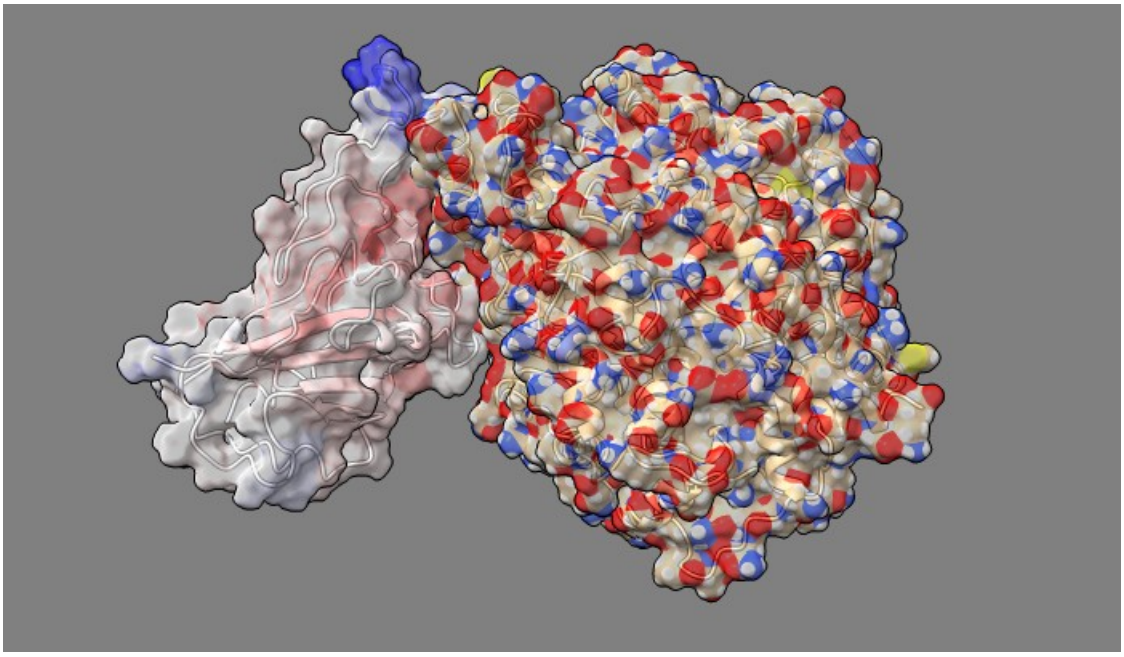
Below is an analysis functional sites of protein-protein interaction between the SARS-CoV-2 receptor binding domain (RBD) and the human angiotensin converting enzyme ACE2. This shows the atom dampening in motion at key binding site on ACE2 during binding by the viral RBD



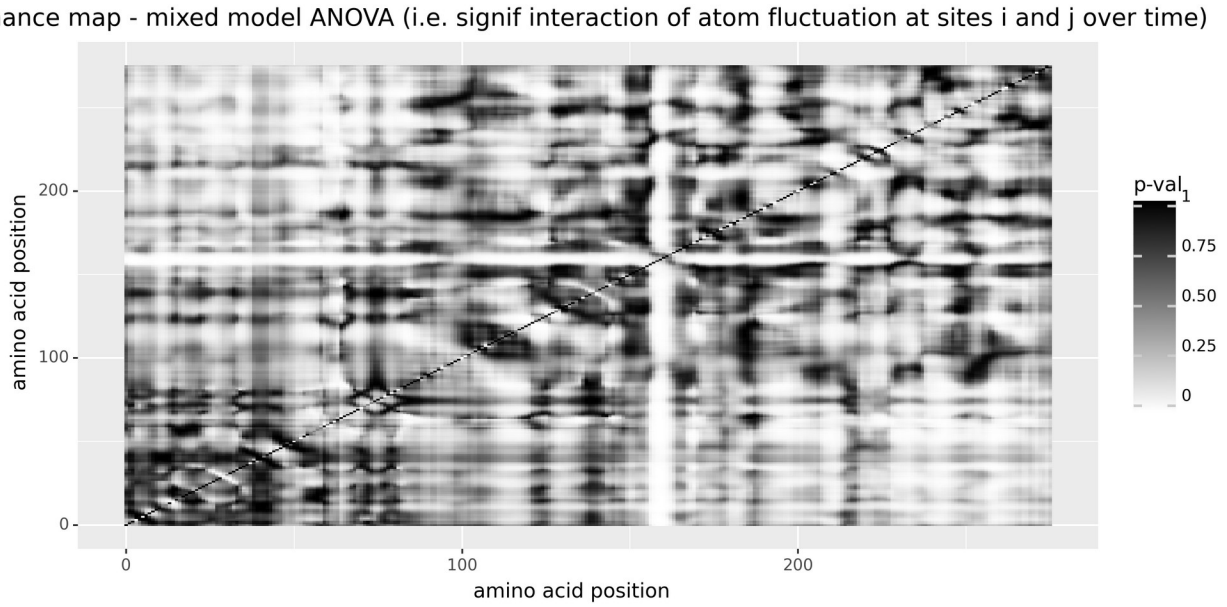


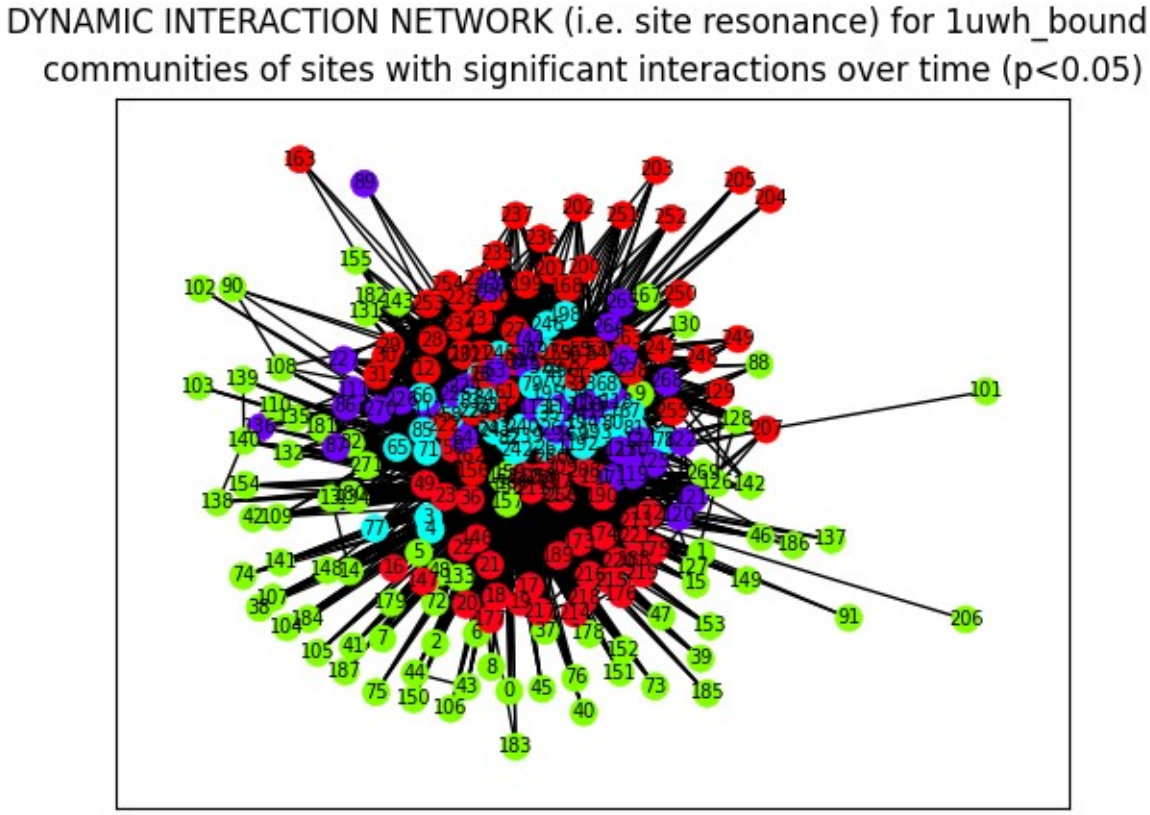
Conversely, this shows the key sites on the viral RBD during binding of human ACE2



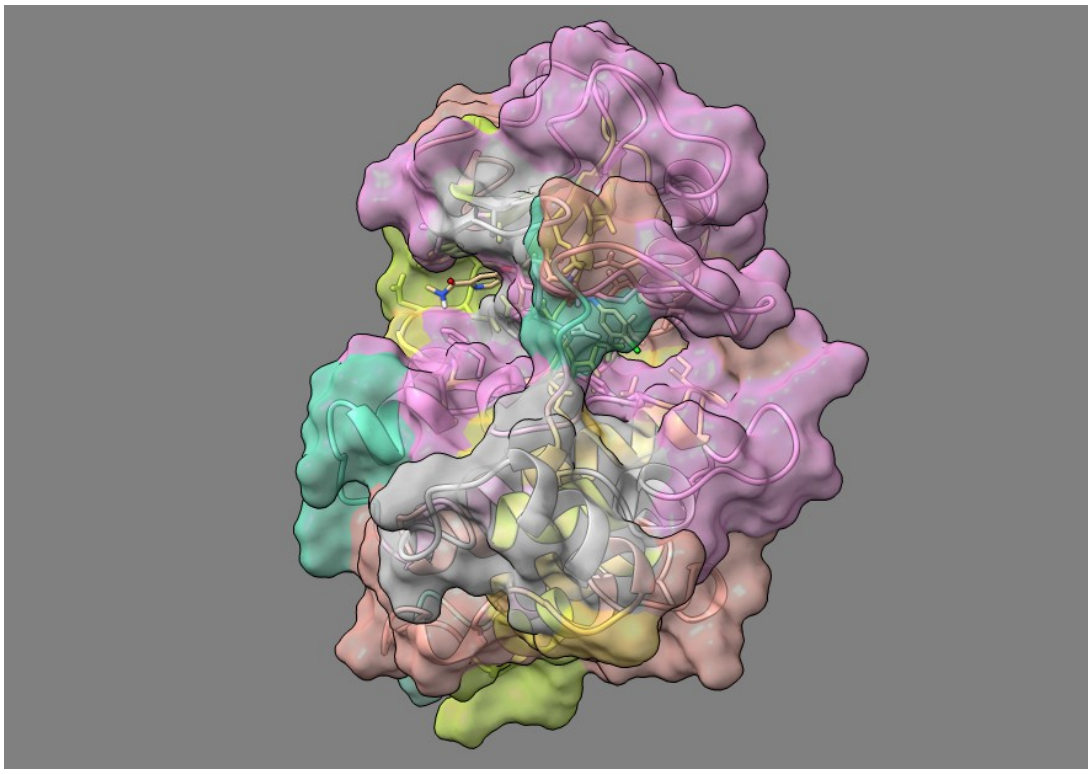


Coordinated site dynamics can also be analyzed using Choreograph 2.0. This is a heatmap and community analysis of pairwise site resonance detected with mixed-model ANOVA for the ATP-bound BRAF kinase protein

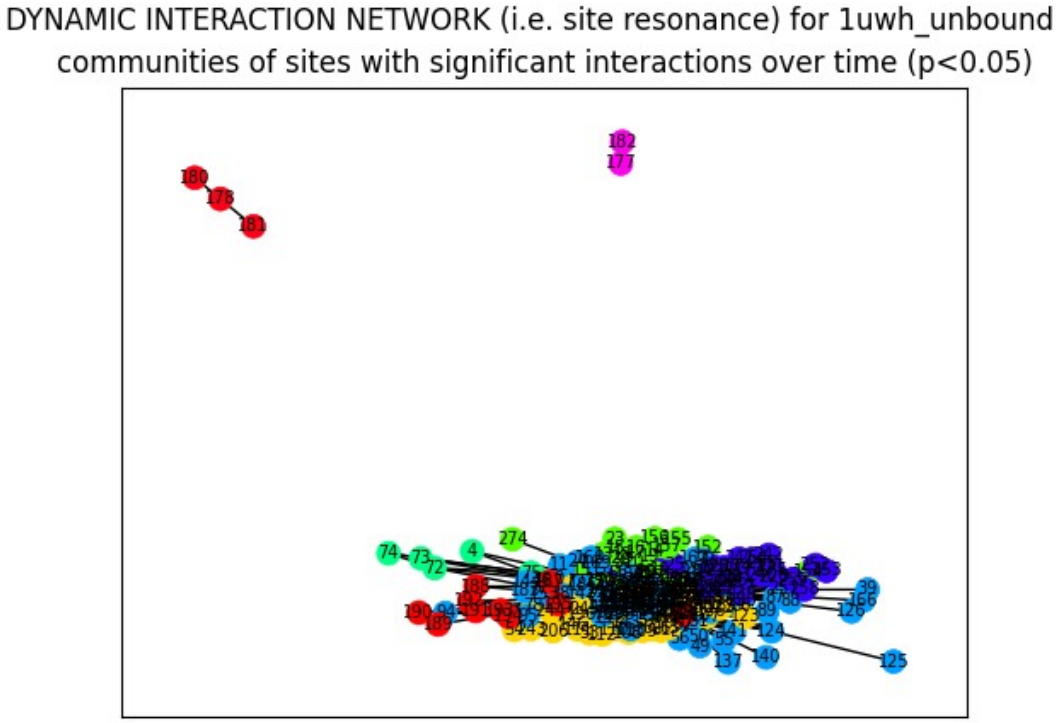


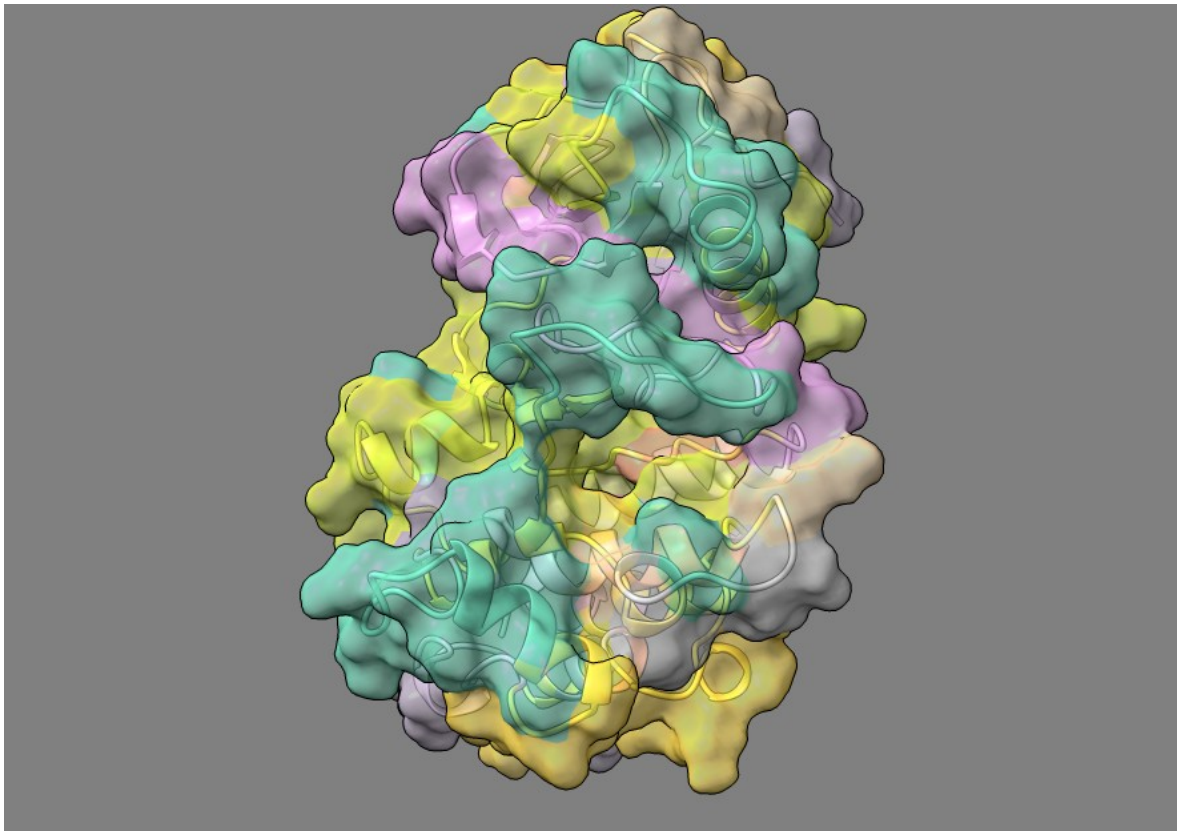


and communities color-mapped to the ATP-bound structure (note how ATP binding site and activation loop it affects are all comprised of a single resonance community – in lavender)



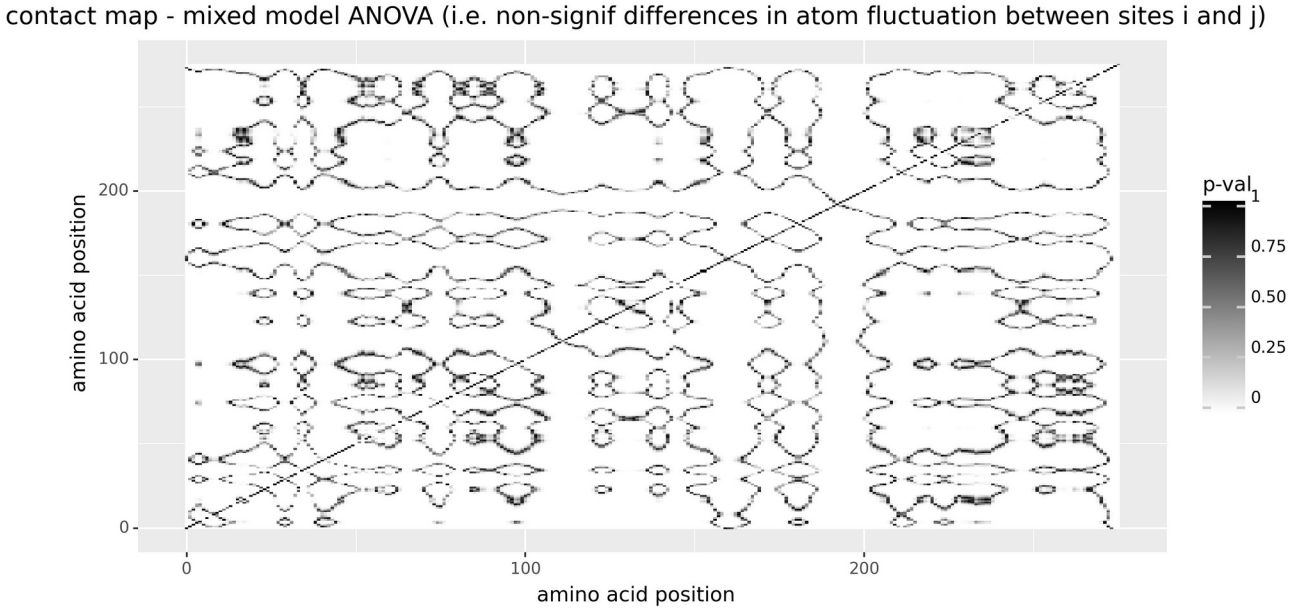
When ATP is not present, this community structure is quite different

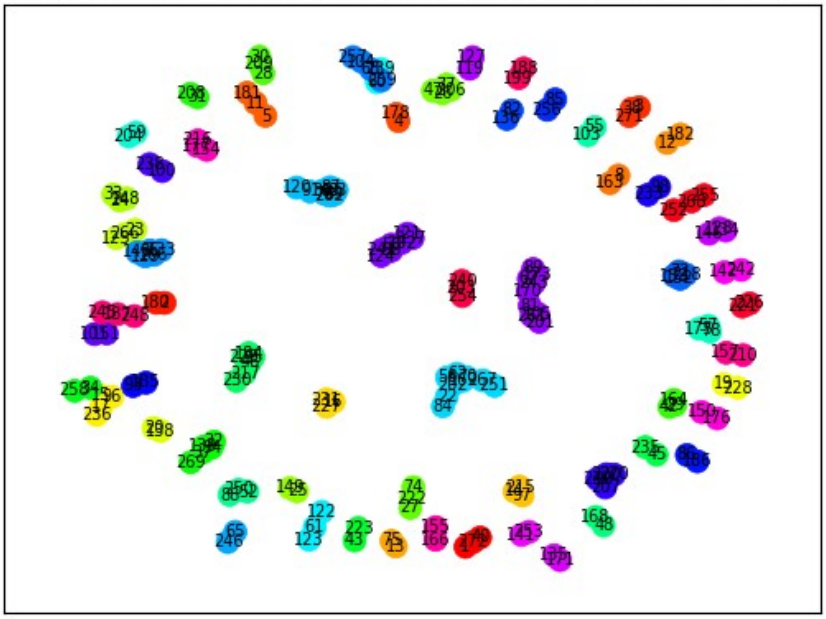




NOTE- turqoise indicates regions with no significant resonance interaction

The overall non-significant differences or similarities in atom fluctuation can also be mapped revealing potential contact interactions between sites. Here the adjacent sites share similar fluctuations and form a snaking pattern on the heatmap. Regions where this pattern folds back onto itself are regions where similar dynamics are being created by non-adjacent contacts between the folded protein chains. The labeling of sites on the community network allows us to identify the positions of the potential contact points that organize the thermodynamics of the protein structure.

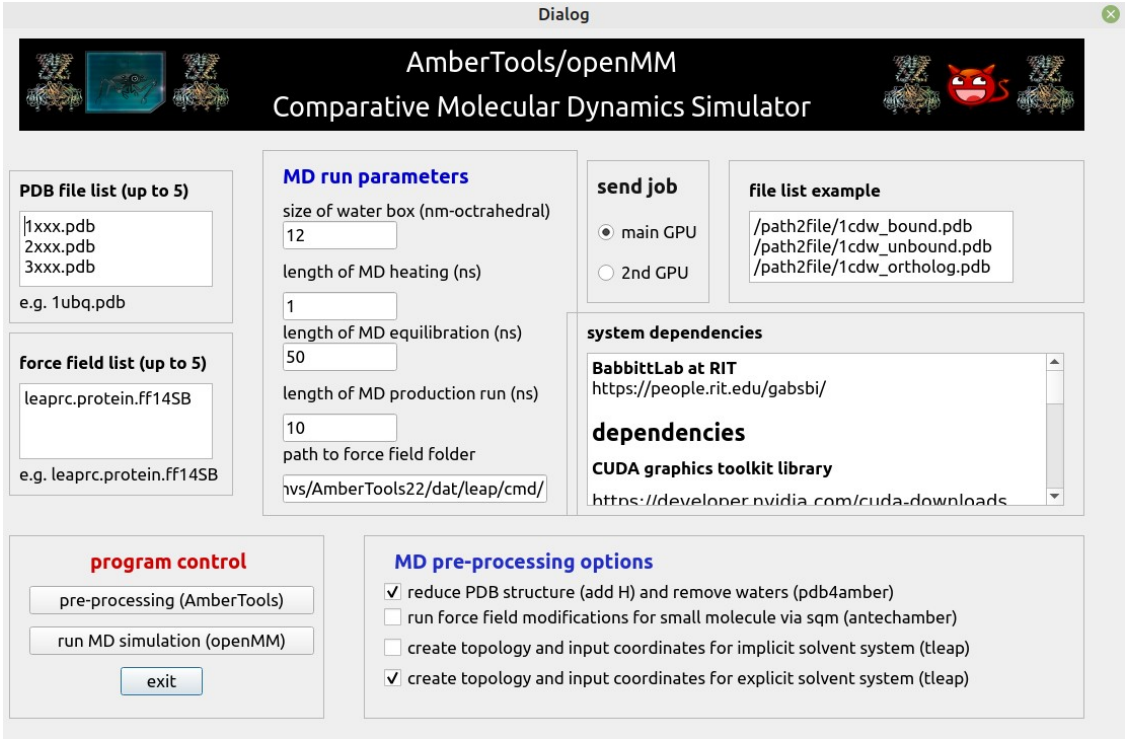




##

MOLECULAR DYNAMICS SIMULATION ENGINE AND GRAPHICAL USER INTERFACE

As a companion to ATOMDANCE, we offer a GUI for running molecular dynamics simulations using the free softwares AmberTools and openMM. The GUI is shown below



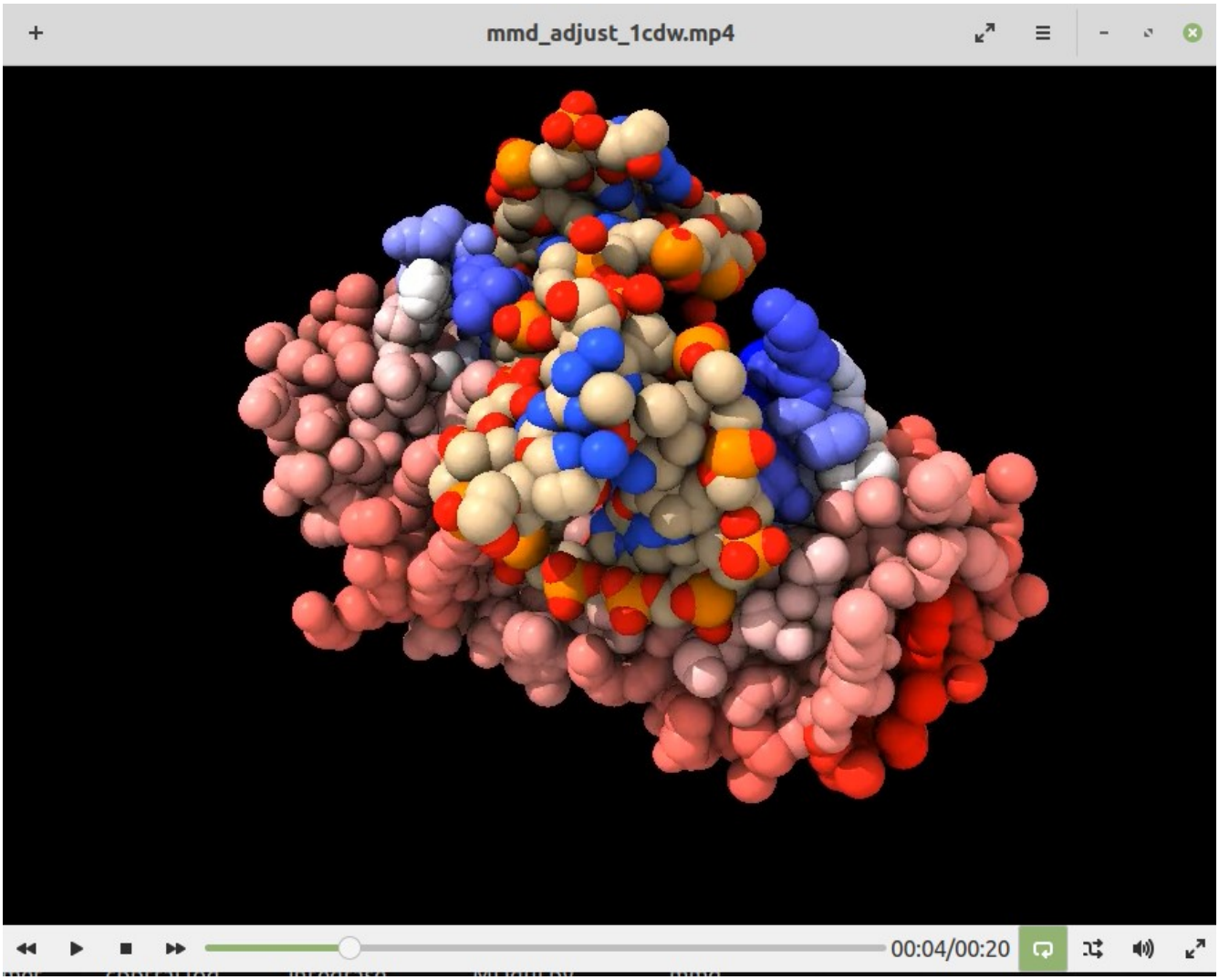
To utilize this feature we require a high end GPU preferably installed on a Linux Mint OS with recent installation of CUDA and CUDA toolkit installed with proper Nvidia graphics driver. An AmberTools and openMM installation

via anaconda or miniconda is also needed. The program terminal will prompt routines for both conda or native installations of these tools. Conda installations are easier and less likely to break existing packages. The python modules 'parmed' and 'netCDF4' are required and should also be installed as well. For later comparisons using ATOMDANCE, this interface allows preparation and simulation of systems with simple protein, protein+small molecule ligand, and protein+DNA or RNA.

To start type 'python3 Mdgui.py'

##

To make movies that have dynamic motions that are weighted to the normalized maximum mean discrepancy of the atom fluctuations, run makeMovie.py and enter into the GUI the .pdb sturcture file, the .prmtop topology file, the .nc trajectory file and the maximum mean discrepancy data file (i.e. maxMeanDiscrepancy_flux.txt) produced by the ATOMDANCE.py software. This will produce a multiframe .pdb file called mfMMD_pdbname.pdb. Close the GUI and open ChimeraX and load the Molecular Dynamics Viewer (from ChimeraX ToolShed). Enter the multiframe PDB file above and the .dat attribute file produced in the chimeraxVis folder to color the movie image by the max mean discrepancy (blue = atom motion dampening; red = atom motion amplification), then click make movie. NOTE: the ribbon representation should be disabled as the noise weighting will fragment the image. It is best to record the movie using all atom view with spacefill enabled to get a movie such as in the image below.



See this movie at https://people.rit.edu/gabsbi/img/videos/MMDmovie.mp4

##
ATOMDANCE INSTALLATION INSTRUCTIONS AND USAGE
##

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 (independent of AmberTools)

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, scikit-learn, matplotlib, patchworklib, plotnine, progress, parmed, netCDF4, pingouin, networkx) 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

IMPORTANT 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

ALSO IMPORTANT: Make sure the chains and amino acid sites in the reference protein are labeled sequentially (chain A:0-183, B:184-224, C:225-307 etc) then make sure the query protein is labeled identically in its homologous regions and includes the additional nonhomologous structures for ligand/nucleic acid/protein chain interaction partners following the homologous portion of the PBD file. To reset the x axis (starting amino acid position) of output plots, the user can change the number of the starting position (i.e. N terminus of chain A) on the GUI interface.

##
PLEASE CITE US (as well as ChimeraX and cpptraj and openMM and AmberTools if used)

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.
##
The naming of things:
DROIDS – Detecting Relative Outlier Impacts in Dynamics Simulations
maxDemon – from Maxwell's Demon, a 19th century thought experiment connecting the concepts of information and entropy in thermodynamics involving a mythical demon watching/assessing the motion of every atom in a system.
ChoreoGraph – from a notion of when motions of atoms at amino acids site 'move together' in a coordinated manner, in much the same way dancers may move together in choreography.
ATOMDANCE – named after a composition by Icelandic singer Bjork Guomundsdottir from her album Vulnicura