ATOMDANCE: kernel-based de-noising and choreographic analysis for protein dynamic comparison

Gregory A. Babbitt[1,*], Madhusudan Rajendran[1] , Miranda L. Lynch[2], Richmond Asare-Bediako[1], Leora T. Mouli[1], Cameron J. Ryan[3], Harsh Srivastava[4], Patrick Rynkiewicz[1], Kavya Phadke[1], Makayla L. Reed[1], Nadia Moore[1], Maureen C. Ferran[1] and Ernest P. Fokoue[5*]

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.

[2]Hauptmann Woodward Medical Research Institute, Buffalo, NY , USA.

[3]McQuaid Jesuit High School Computer Club, Rochester, NY 14618, USA.

[4]New York University, Rochester, NY 14618, USA.

[5]School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.

*Correspondence: gabsbi@rit.edu, epfsms@rit.edu

Running Title – comparative protein dynamics software

Keywords – comparative method, machine learning, molecular dynamics simulation, molecular evolution, coordinated dynamics, allostery

Abstract –

Comparative methods in molecular evolution and structural biology rely heavily upon the site-wise analysis of DNA sequence and protein structure, both static forms of information. However, it is widely accepted that protein function results from nanoscale non-random machine-like motions induced by evolutionarily conserved molecular interactions. Comparisons of molecular dynamics (MD) simulations conducted between homologous sites representative of different functional or mutational states can potentially identify local effects on binding interaction and protein evolution. Additionally, comparisons of different (i.e. non-homologous) sites within MD simulations could be employed to identify functional shifts in local time-coordinated dynamics indicative of logic-gating within proteins. However, comparative MD analysis is challenged by the large fraction of protein motion caused by random thermal noise in the surrounding solvent. Therefore, properly de-noised MD comparisons could reveal functional sites involving these machine-like dynamics with good accuracy. Here, we introduce ATOMDANCE, a user-interfaced suite of comparative machine learning based de-noising tools designed for identifying functional sites and the patterns of coordinated motion they can create within MD simulations. ATOMDANCE-maxDemon4.0 employs Gaussian kernel functions to compute site-wise maximum mean discrepancy (MMD) between learned features of motion, thereby assessing de-noised differences in the non-random motions between functional or evolutionary states (e.g. ligand bound vs. unbound, wild-type vs. mutant). ATOMDANCE-maxDemon4.0 also employs MMD to analyze potential random amino-acid replacements allowing for a site-wise test of neutral vs. non-neutral evolution on the divergence of dynamic function in protein homologs. Lastly, ATOMDANCE-Choreograph2.0 employs mixed-model ANOVA and graph network to detect regions where time synchronized shifts in dynamics occur. Here, we demonstrate ATOMDANCE's utility for identifying key sites involved in dynamic responses during functional binding interactions involving DNA, small molecule drugs, and virus-host recognition, as well as understanding shifts in global and local site coordination occurring during allosteric activation of a pathogenic protease.

Statement of Significance – ATOMDANCE is a suite of software pipelines controlled by a graphical user interface and designed to comprehensively simulate, calculate and compare protein motions between two functional or evolutionary states while controlling for random thermally-induced noise. ATOMDANCE is useful for finding amino acid sites on a given protein that are functionally important in interaction with other proteins, DNA, or other small molecules. It can also be used to assess the protein site-specific effects of genetic mutation that affect protein interaction. Lastly, ATOMDANCE identifies regions of proteins that coordinate shifts in motions in potentially complex ways as a single choreographed unit or community, allowing investigators to identify what sites are coordinating biophysical changes in proteins during changes in functionally logic-gated states.

Introduction –

The complex functioning of many protein pathway systems often relies upon allosteric protein-ligand interactions (PLI) chained together by upstream/downstream protein-protein interactions (PPI). Traditionally, protein function has been studied from a largely structural perspective, with logic-gating often portrayed as the result of the specificity of certain structural conformations in their ability to form PLI or PPI. However, these 'lock-key', 'puzzle piece', or 'induced fit' generalizations of protein function are incapable of fully capturing the soft matter biophysics that are probably involved in most PLI and PPI (1–3). Much in the same way that a description of the state of a light switch as on/off fails to capture the functional details of electron behavior in the wiring underneath the wall, we argue that the description of allosteric activation and PPI as a binary state controlled only by binding interaction (4, 5) similarly fails to fully examine how both the complex coordinated motions of disordered regions such as linkers and loops, as well as the vibrational resonance of more ordered regions within proteins, are measurably altered to affect switching between the 'tensed' vs 'relaxed' logic states likely involved in allostery and PPI (4, 6). In these logic states, the complex protein dynamics both within and between binding partners must be more fully analyzed at single-site resolution so as to better understand underlying mechanisms and their subsequent biological evolution. Often proteins are described as analogous to nanoscale-sized machines (7–9), where non-random repetitive motions, are key characteristics of protein function. In this more dynamic view, the function and evolution of the protein-based regulatory pathways must depend heavily upon how protein structures are able to alter or shift their dynamics during important logic gating functions required when proteins and other biological macromolecules collectively assemble to form larger complexes (10, 11).

Comparative methods of molecular analysis are well developed for protein sequences and structures in the disciplines of phylogenetics (12), molecular evolution (13) and structural biology (14), as well as the history of molecular biology and modern statistics (15). In molecular evolution and comparative genomics, comparative methods are applied in a site-wise manner because genetic mutations tend to act independently at individual sites over time. Site-wise analyses of root mean square deviation (RMSD) of superimposed homologous protein chains are also very common in structural biology. However, site-wise comparative methods for application to molecular dynamics are still only now beginning to be developed (16, 17). Important types of homologous site-wise comparisons derived from two different molecular dynamics (MD) simulations (Figure 1) might include site-wise comparisons of proteins in (A) two functional states (e.g. binding to drugs, toxins, nucleic acids, or other proteins), (B) two different temperatures (e.g. thermostability), (C) two different evolutionary lineages (e.g. before and after some significant mutation events), or (D) two different epigenetic states (e.g. involving phosphorylation or methylation). Being able to make site-wise determinations of similarities and differences in protein dynamics has significant potential application to the fields of computational pharmacology and vaccine development as they interface with the basic

science of molecular evolution (18–22), with specific applications towards identifying single protein sites with large impacts upon the evolution of vaccine (20) or drug escape (19).

A major challenge with functionally analyzing the dynamic trajectories of atoms in MD simulations is caused by the large fraction of motion in the system that is due to random thermal noise. This noise can obscure both harmonic and anharmonic vibrational frequencies (23) as well as other more complex non-random machine-like motions connected to protein function. This is more problematic in explicit solvent based MD approaches, which can more accurately replicate protein dynamics than other methods (24–26). Thermal noise in explicit solvent MD simulation is caused by the random collision of molecules in the solvent with the protein chains. Our past methods of comparative molecular dynamics analysis have relied upon a large amount of resampling of the atom trajectories to be able to resolve site-wise functional differences in dynamics caused by binding interactions and mutations (16–22). Here, we will present a novel and efficient machine learning-based approach to site-wise comparative MD analysis that is robust to the effect of thermal noise in the raw trajectory data. Because a machine learning algorithms cannot learn from noise, they can theoretically provide useful methods of detecting signal from noise (i.e. de-noising) in MD comparisons. Despite their promise for filtering random from non-random motion in site-wise MD comparisons, they remain largely unexplored for this purpose. As the distributions of atom fluctuations over short time scales in MD simulation is largely Gaussian, we propose that noise filtering can be addressed by Gaussian kernel-based approaches to machine learning. This approach also allows differences in learned features to mathematically lend themselves to a very useful comparative representation known as the maximum mean discrepancy (MMD) in the reproducing kernel Hilbert space (RKHS). A Gaussian process kernel also has an advantage in that it is more interpretable than black box methods such as support vector machine and neural networks when applied to many physical systems (27). This interpretability might be very important to future biomedical researchers when navigating a rapidly evolving policy landscape regarding the application of machine learning to drug discovery.

Besides homologous site-wise comparisons of dynamics, there are also non-homologous site-wise comparisons that can be made with MD simulations (Figure 1). This involves comparing the dynamics of two nearby adjacent or even more distant non-adjacent amino acid sites over time in order to ascertain how coordinated or ‘choreographed’ they are in their dynamic behavior, possibly indicating either resonance due to strong native contact or possibly even longer range interactions created by allosteric effects. A common traditional approach to the analysis of coordination of protein dynamics is via the property of resonance, observed through site-wise correlation or covariance matrices derived from MD trajectories. Hybrid approaches to resonance analysis often combine nuclear magnetic resonance (NMR) and MD simulation studies. Less computationally expensive methods such as normal mode analysis (NMA), often coarse-grained using elastic networks (28) are commonly conducted as well. These approaches have some limitations both practical (e.g. access to NMR equipment) and theoretical (e.g. NMA’s assumption that all protein motion is only harmonic (28)). Where protein allostery is

investigated via MD simulations (29, 30), current state-of the-art methods often also apply graph network-based methods to covariance matrices to define optimal/suboptimal pathways connecting two chosen sites (31) or to define regional or network community boundaries inclusive of important sites that may influence each other (32). However, the covariance of MD trajectories between different sites on the protein are often statistically weakened, again by the effect of thermal noise. Covariance in trajectories can also be rather ineffective in capturing allosteric effects involving more complex motions of disordered regions such as loops and linker regions often involved in important dynamic protein function (33). Also, simple resonance patterns can sometimes be artifacts in MD simulations as well (34). A proper comparative analysis of functionally dynamic shifts involving more disordered regions of proteins will require identifying non-resonant but still time synchronous controlled changes in overall magnitudes of relative motions across non-homologous sites. Inspired by recent advances in the computational analysis of human dance itself (35), where more complex and shifting forms of coordination involve spatial patterning, tessellation, repetitive sequences, and variations on themes, we introduce a broader concept and statistical method of 'choreographic' analysis for studying coordinated protein dynamics. Choreographic analysis can still capture simple molecular resonance if performed on highly resolved time scales, but on longer time scales, it can also capture non-resonant allosteric regulatory influences on protein loop and linker dynamics as well. Choreographic analysis of MD simulations takes advantage of a mixed model analysis of variance (ANOVA) approach to capture coarse-grained time synchronous differences in atom fluctuation between sites and thus it is insensitive to shorter time-scales upon which thermal noise in occurs. Choreographic analysis also employs a graph network-based community detection algorithm to define the boundaries of regions of coordinated motions in the MD simulation.

Here, we introduce the ATOMDANCE statistical machine learning post-processor for comparative molecular dynamics performed at individual site-wise resolution. ATOMDANCE is a python-based graphical interfaced software suite for machine learning both direct and de-noised comparison, as well as choreographic analysis of functional protein dynamics. ATOMDANCE is the first software suite that provides researchers with a user-friendly computational platform for supplementing comparative sequence/structure analyses with important information about the functional motion and functional evolution of proteins undergoing complex interactions with DNA/RNA, drugs, toxins, natural ligands, or other proteins. It offers three analytical pipelines for comparative analysis of MD as well as choreographic analysis for detecting how local regions of proteins are coordinate shifts in dynamics during mutation and/or functional interaction with other molecules in the cell (Figure 1). In addition to static plots, ATOMDANCE interfaces automatically with UCSF ChimeraX to produce color-mapped structural images and movies as well.

Methods –

ATOMDANCE.py is a PyQt5 GUI designed for post-processing comparative molecular dynamics and delivering information about important protein site differences between the dynamics of proteins in two different functional states. It also can be used to investigate potential site-wise evolutionary changes in protein dynamics and to investigate where sites share coordinated dynamics states as well. After randomly subsampling the atom trajectory files and calculating amino acid site atom fluctuations using the atomicfluct functions from the cpptraj library, ATOMDANCE.py runs 4 types of analyses listed below. The types of typical comparisons of protein dynamics and the methods that can be used are summarized in Figure 1.

(A) DROIDS 5.0 for direct site-wise comparison of protein dynamics

This option is an acronym for Detecting Relative Outlier Impacts in Dynamics Simulations and calculates both the average differences and KL divergences in the atom fluctuation at every protein site. Fluctuations are averaged by residue for each amino acid. Significant differences in dynamics of the two protein states are determined by a two sample Kolmogorov-Smirnov test corrected for the number of sites in the protein corrected for the false discovery rate (i.e. Benjamini-Hochberg method) caused by the total number of sites on the protein. This method is described and published previously in DROIDS v1.0-4.0 (16, 17). The only difference in v5.0 is that the subsampling is taken from random window positions along a single long MD production run, rather than multiple short MD production runs. The site-wise average differences and Kullback-Leibler (KL) divergences in atom fluctuation are reported similarly to the DROIDS 4.0 method previously developed by our lab group (16, 17). Atom fluctuation is defined as the root mean square fluctuation (eqn 1) taken over a user defined number of image frames acquired in a given time interval. The MD interface software we provide generates 5000 image frames per 1ns of MD simulation. Thus a typical 10ns comparative analysis of 50,000 image frames might randomly resample the rmsf values for 100 randomly positioned windows taken along the MD trajectory, each with 200-300 frames used for each rmsf calculation,

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

where v represents the set of XYZ atom coordinates for m atoms for a given amino acid residue over n time points, and w represents the average coordinate structure for each MD production run for a given ensemble (using the ‘‘atomicfluct’’ function from cpptraj software (36). Thus, a signed and symmetric Kullback-Leibler (KL) divergence metric (eqns 2-3) for comparing rmsf values for two ensembles of size n (i.e. number of resamples) for a given amino acid is

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

$$signed\ KL = \{KL;\ if\ rmsf_{query} > rmsf_{reference}\ ,$$

$$\{-KL;\ if\ rmsf_{query} < rmsf_{reference} \quad (3)$$

Note that the sign is simply determined by the relative value of the site specific rmsf of the query state of the protein compared to the site specific rmsf of the reference state of the protein. Therefore, a positive sign indicates an amplification of motion during binding or mutation while a negative sign indicates a dampening of motion.

In ATOMDANCE, we offer the same analysis in a completely python-based application as DROIDS 5.0. Sites with significantly different dynamics are identified with a multiple test corrected two sample Kolmogorov-Smirnov test. While older versions of DROIDS were developed as a perl plus R language pipeline that interfaced directly with licensed Amber software, DROIDS 5.0, like all of ATOMDANCE runs exclusively in python code, and now executes independently of MD simulation software.

(B) maxDemon 4.0 for de-noised ***functional*** comparisons of protein dynamics

The maxDemon 4.0 program in ATOMDANCE is trained on feature vectors of local atom fluctuations derived from the molecular dynamics trajectories of proteins in two functional states (e.g. bound vs. unbound or wild-type vs. mutant). The comparison between MD simulations at given sites are reported as maximum mean discrepancy (MMD) in the reproducing kernel Hilbert space (RKHS). Hypothesis tests for significance of functional dynamic changes reported via MMD are also provided using a bootstrapping approach. A more detailed graphical summary of this method is shown in Supplemental Figure 1.

This analysis option uses site-wise training of Gaussian processes machine learners with tuned radial basis kernel functions in order to specify a maximum mean discrepancy (MMD) in reproducing kernel Hilbert space (RKHS) that describes the distance in learned features between the two protein dynamic states at all given sites on the protein.

Thus the kernel function (eqn 4) describing the mapping of the data points x_i and x_j being compared is

$$k(x_i\ , x_j\ ) = exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (4)$$

Note that σ is sampled derived. And the empirical estimations of MMD, or distance between feature means (eqn 5) is given by

$$MMD^2(X,Y) = \frac{1}{m(m-1)}\sum_i\sum_{j\neq i} k(x_i\ , x_j\ ) - 2\frac{1}{m(m-1)}\sum_i\sum_j k(x_i\ , y_i\ ) + \frac{1}{m(m-1)}\sum_i\sum_{j\neq i} k(y_i\ , y_j\ ) \quad (5)$$

where x's are the data points we have and y's are generated examples evaluated on the kernel.

The learners in ATOMDANCE are trained using a local atom fluctuation feature vector comprised of fluctuations from sites -2, -1, 0, 1, 2 positions on the protein chain relative to the site being analyzed. The observed site-wise MMD values are further subjected to hypothesis testing using a bootstrap derived empirical p-value whereby the observed MMD values between the functional dynamic states at any given site are compared to 500 bootstrapped MMD values for that site when derived from resampling the dynamics in the same dynamic state. A key concept here when comparing this output to the site-wise KL divergence metrics generated by DROIDS 5.0 is that because the learner cannot optimize on random differences in atom fluctuation caused by thermal noise it acts as a noise filter, thus eliminating motion dampening that is not directly due to non-random differences in atom fluctuation between the sites being compared (i.e. functional aspects of molecular interactions directly involved in the binding interaction)

(C) maxDemon 4.0 for de-noised *evolutionary* comparisons of protein dynamics

For a test of neutral evolution on protein dynamics, the MMD of amino acid replacements observed on orthologs is compared to a neutral model of the MMD of random pairs of differing amino acids at different sites on the two protein simulations being compared. This allows for the identification of potential sites where natural selection has either functionally conserved or adaptively altered the local molecular dynamics of the protein. This analysis option is only appropriate when comparing two homologous proteins in two states of molecular evolution, whereby mutations have accrued over time and the user would like to determine whether the dynamics at a given site of amino acid replacement has likely been functionally conserved, evolved neutrally or evolved adaptively (i.e. under purifying, neutral or adaptive evolution). Comparisons of dynamics between the same protein in two different species (i.e. orthologs) or two related proteins in the same species (i.e. paralogs) are both enabled through this method of analysis. In this case, the MMD in dynamics between each site of amino acid replacement between the homologous proteins is compared to a model of neutral evolution represented by a distribution of MMD taken from the dynamics of 500 random pairs of dissimilar amino acids at different sites on the homologous proteins. If the MMD for an observed amino acid replacement is in the lower or upper extremities of the distribution of neutral MMD (two-tailed level of significance = 0.05), then natural selection acting upon the dynamics can likely be inferred.

(D) Choreograph 2.0 – classical statistical identification and comparison of dynamics that are coordinated across different sites on a given protein

The ChoreoGraph 2.0 program in ATOMDANCE offers a site-wise mixed-effects model ANOVA and graph network community detection analysis of time synchronous differences in atom fluctuation between sites indicative of complex choreographed motions in protein dynamics. It is used for the identification of regions or communities of amino acid sites with high degree of coordination in the shifting of their respective dynamic states. The network communities detected in both functional protein states are compared using bootstrapped resampling of

graph network connectivity and two methods of defining graph network non-randomness. These are the (A) probability of two sites being in the same resonance community and the (B) distance of the graph network from a random graph defined by the Erdos-Renyi model. This analysis option examines the reference and query dynamic state simulations of the proteins compared above and produces (A) site-wise heat maps and community level graph networks identifying groups of amino acid sites where atom fluctuation values are resonating over time in a coordinated fashion and (B) site-wise heat maps and community level graph networks identifying groups of amino acid sites where overall atom fluctuation values are not significantly different from each other (i.e. potentially in contact). In each dynamic state simulation every site *i* on the protein is compared to every site *j* using a mixed effects model ANOVA where atom fluctuation represents a fixed effect (α) in the model and a time sample represents a random effect (β) in the model (with μ = mean fluctuation and ε= error or residual term). Thus the general linear model (eqn 6) becomes

$$Y_{st} = \mu + \alpha_s + \beta_t + \alpha\beta_{st} + \varepsilon_{st} \quad (6)$$

where *s* represents the site class (*i* or *j*) and *t* represents the random time sampling group collected by the cpptraj program.

For the choreographic analysis, the p-value of interaction between atom fluctuation levels between site *i* and site *j* and the time subsamples in the MD simulation (i.e. $\alpha\beta_{st}$) indicates the significance of an interaction of fixed differences in atom fluctuation between the two separate sites over time (i.e. a coordinated or choreographed physical motion). The p-values are corrected for false discovery rate via Benjamini Hochberg method and shown as a heat map representing significance of areas of synchronously shifting site dynamics on the protein. In the second step of the analysis, intended to define communities of coordinated regions of protein dynamics, the strongest interaction p-values for all site *i* to site *j* comparisons are represented by a graph network (eqn 7) where (k_n), the degree of node n is

$$k_n = A_{nm} \sum nm \quad (7)$$

where *n* and *m* are the interaction p values for sites *i* and *j,* and A_{nm} is the adjacency matrix connecting nodes *n* and *m*. The user can choose to build this network using either a user-defined fixed p-value cutoff (e.g. $p < 0.003$) or an autotuned cutoff that collects the 5% strongest interaction p-values (e.g. p='auto'). The reader should note that here the mathematic term 'adjacency' is not defined in terms of structural proximity of sites i and j, but instead it defines a form of adjacency in shifts in atom fluctuation that occur at the similarly in time. Most importantly, while the structural proximity of two given sites i and j can facilitate resonances in dynamic state, our definition of adjacency can also capture synchronous changes of dynamic states even when the sites i and j are quite far apart within the protein. Therefore, this allows us to define communities of time coordinated dynamics that can potentially involve both proximal and distant effects that when color mapped to the protein structure, can show how two distant sites can ultimately influence each other via allostery. The Louvain community

detection algorithm (37) iterates a two-step process of modularity optimization followed by community aggregation until community identities of all nodes are stable. It is implemented in our code by the python package networkx (38).

ATOMDANCE is intuitive and user-friendly, providing a simple graphical user interface (GUI) that only requires structure, topology, and trajectory files (.pdb, .prmtop, .nc) for the two molecular dynamics simulations being compared (Supplemental Figure 2). It is entirely python-based and outside of this it only requires UCSF ChimeraX for molecular visualization and the popular cpptraj library for resampling calculations (36, 39, 40). ATOMDANCE is also supplemented with an optional GUI (Supplemental Figure 3) for generating simulations via open-source tools (i.e. openMM and AmberTools) (41, 42). However it can also potentially be used with files generated using NAMD (qwikMD), CHARMM, or the licensed version of Amber (41, 43–45). We provide an optional GUI for generating multiframe PDB file movies using UCSF ChimeraX (40) where the motions of the protein system are colored and augmented in accordance with the MMD. Examples of these movies can be seen in an introductory video available at https://people.rit.edu/gabsbi/img/videos/MMDmovie.mp4

ATOMDANCE is available at GitHub/GitHub pages

https://github.com/gbabbitt/ATOMDANCE-comparative-protein-dynamics

https://gbabbitt.github.io/ATOMDANCE-comparative-protein-dynamics/

and as a docker container here

https://github.com/patrynk/atomdance-docker

Examples presented in this manuscript were generated from structure, topology, and trajectory files deposited here

https://zenodo.org/record/7679282#.Y_wIK9LMJ9A

DOI 10.5281/zenodo.7679282

See the Supplemental Methods file for more detailed descriptions about the MD preparation and analyses on the specific examples presented in the Results section.

Results –

To demonstrate the utility of ATOMDANCE, we present a comparison of the unfiltered (i.e. noisy) site-wise divergences in atom fluctuation (Figure 2) to the denoised site-wise discrepancy in learned features of local atom fluctuation (Figure 3), presented in four examples of functional binding interactions. These four examples include (A) DNA-bound vs. unbound TATA binding protein (PDB: 1cdw)(46), (B) sorafenib-bound vs. unbound B-Raf kinase domain (PDB: 1uwh)(47), (C) SARS-CoV-2 viral bound vs. unbound angiotensin-converting enzyme 2 (ACE2)

protein (PDB: 6m17)(48), and (D) the allosteric activated (i.e. InsP6 bound) vs inactivated (i.e. unbound) Vibrio cholera toxin RTX cysteine protease domain (PDB: 3eeb)(49). The root mean square fluctuation plots for these comparisons are shown in Supplemental Figure 4.

The first example of comparative protein dynamics analyses conducted with ATOMDANCE investigated the functional effect of DNA binding to TATA binding protein (TBP; PDB 1cdw) by the site-wise comparison of atom fluctuation of TBP in both its DNA bound and unbound state. Figure 2A shows both color-mapped protein surface and site-wise plot of the KL divergence in fluctuation (i.e. DROIDS 5.0). This protein binds quite strongly as is evidenced by a general dampening of fluctuation across the protein (in blue). Supplemental Figure 5 demonstrates alternative plots of the TBP results generated by ATOMDANCE showing site-wise atom fluctuation profiles and average differences in fluctuation colored by amino acid type. The comparison of machine learning derived MMD (i.e. maxDemon 4.0; Figure 3A) clearly captures the key sites of the functional interaction; two loops of the protein that interact directly with the major groove of the DNA (in blue) (46). Close-up views of all MMD color-mapped structures are given in Supplemental Figure 6 with TBP in panel A. Supplemental Figure 7 shows the TBP MMD plot colored by amino acid type and bootstrapped empirical p-values. To examine the consistency of our comparative methods and machine learning application when different types of MD integration and acceleration are used on the TBP structure, we compare the site-wise dampening of atom fluctuation measured via signed KL divergence (eqn 3; Supplemental Figure 8) and the key binding sites identified via MMD (eqns 4 nd 5; Supplemental Figure 9) across four methods of MD simulation using two different softwares packages (41, 42). These include (A) GPU-accelerated Verlet integration with an Andersen thermostat (50) in OpenMM, (B) GPU-accelerated Langevin integration in OpenMM, (C) GPU-accelerated particle-mesh Ewald aMD in Amber20 (i.e. pmemd.cuda; see (51, 52)), and (D) GPU-accelerated aMD integration in OpenMM (53, 54) . While the site-wise divergence patterns do differ slightly (Supplementary Figure 8), the MMD correctly identifies the TBP binding site loops in all cases regardless of the type of integration or acceleration used (Supplementary Figure 9).

The second example of the application of MMD captures the functional amplification of atom fluctuation by the activation loop (shown in red) of BRAF kinase upon the binding of the drug sorafenib in the ATP binding pocket of the kinase domain (PDB 1uwh; Figure 3). In this example the interaction of ATP or ATP-competitive antagonists like the cancer drug sorafenib clearly amplify the motion in the activation loop as can be observed in both the unfiltered and de-noised dynamics (Figure 2B and Figure 3B) (47). While sorafenib binds the site stronger than ATP, thus interrupting the MAPK pathway triggering cell proliferation in tumorogenesis, this amplification of the activation loop by the drug may be functionally related to the hyperactivation of MAPK in surrounding normal cells, leading to cancer recurrence (18).

A third example also demonstrates the utility of the DROIDS 5.0 KL divergence and maxDemon 4.0 MMD to investigate the protein-protein interaction between the viral SARS-CoV-2 receptor binding domain (RBD) and its human protein target angiotensin converting enzyme (ACE2)(PDB

6m17) (48). While the unfiltered divergence in viral-bound vs. unbound dynamics exhibits general dampening of ACE2 target protein's motions (Figure 2C) the key functional sites of ACE2 that are recognized by the viral RBD are quite clearly and dramatically revealed by the MMD (Figure 3C). These include two sites on the N-terminal helices of ACE2 including two well documented additional sites identified at Q325 and K353 identified in previous studies of functional dynamics and evolution (20–22).

In the fourth example, the DROIDS 5.0 KL divergence demonstrates a large allosteric effect of the eukaryotic specific InsP6 signaling molecule in triggering general amplification of dynamics across the whole of the Vibrio cholera RTX cysteine protease (Figure 2D), effectively activating the toxin only when it is present within host tissues thus preventing proteolytic destruction of the bacteria itself (49). The maxDemon 4.0 MMD analysis also captures this effect and additionally reveals the key cysteine and other possible sites that drive this change in dynamics (Figure 3D).

To demonstrate the utility of MMD in a comparative evolutionary analysis of human vs. bacterial TBP (Figure 4), maxDemon 4.0 derived a neutral model distribution of MMD in dynamics for randomly selected pairs of differing amino acid sites on the human and bacterial orthologs (Figure 4A) with the tails indicating non-neutral evolution colored red for functionally conserved dynamics and green for adaptively altered dynamics. The two TBP ortholog structures are nearly identical (Figure 4B), and yet (Figure 4C) two regions of altered dynamics (i.e. high MMD) appear related amino acid replacements that have shifted the protein dynamics related to the TBP central hinge and one of the two loop binding regions highlighted earlier in Figure 4A. Most of the rest of the majority of the amino acid replacements (red bars) have occurred under the selective pressure to functionally conserve the TBP dynamics keeping the MMD low between the two orthologs.

The last ATOMDANCE method demonstrates the utility of a mixed effects model ANOVA combined with network community detection algorithms in ChoreoGraph 2.0 for identifying regions with time coordinated dynamics (Figure 5, Supplemental Figure 10). Here we analyzed our fourth case example above comparing the unbound and InsP6-bound dynamics of the RTX cysteine protease domain. We demonstrate a profound loss of regions of coordinated motions, most likely due to the loss of some site resonance as well as the activation of triggered loop regions during the transition from a tensed to a relaxed state during allosteric transition invoked by the InsP6 ligand. Upon allosteric activation of the protease, the dynamic shift from a tensed to a relaxed state mainly serves to remove most of the coordinated motions between amino acid sites. This was particularly exemplified by enhanced dynamics at the loop or flap regions at the top of the protease. This presumably would allow for more facile interactions with other protein substrates in the cell upon infection by the *V. cholera* pathogen. We have recently investigated protease flap dynamics in another protease system in HIV-1 (19). In this allosterically activated protease, we demonstrate another potentially valuable computational

approach to analyzing allostery and other forms of protein logic gating in metabolic and regulatory pathways in the cell.

Discussion -

Molecular dynamics simulation is a powerful tool for estimating physicochemical properties of systems in modern protein science. However, its utility has been limited by the lack of statistically sound methods that allow site-wise comparative functional and evolutionary analyses of protein dynamics. Unlike protein sequence and structural data, both static forms of data, capturing protein motion via molecular dynamics simulations creates a large component of variation that is induced by solvent-induced random thermal noise, subsequently creating a dataset in which non-random functional motions of proteins are potentially obscured. We have described ATOMDANCE, a software suite for comparative protein dynamics that utilizes a powerful and interpretable kernel-based machine learning-based post-processor that allows users to mitigate the effects of noise and to identify functional and evolutionary differences in molecular dynamics at individual sites on proteins during important logic-gated functions of pathways involving PLI and/or PPI. In combination, the ATOMDANCE software provides users with a comprehensive approach to studying how dynamic PPI and PLI partners alter site-wise atom fluctuation when binding, what sites are most responsible for this function and/or how they evolved to achieve this function and how local communities of sites on the protein are organized or choreographed in their motion to achieve this function.

While site-wise differences and divergence metrics can capture meaningful differences in protein function related to overall shifts in thermodynamics, they often have difficulty achieving accurate resolution for identifying the key binding sites without a large amount of time-consuming sampling via MD simulation to mitigate the effect of noise. The TATA binding protein used in our validations, is a perfect example of such a case, as it utilizes two functional binding recognition loops, but nevertheless binds DNA so strongly so as to even bend the rigid DNA molecule and dampen atom fluctuation across nearly the whole of the TATA binding protein. While divergence metrics in DROIDS 5.0 capture this overall dampening at nearly all protein sites very well, our kernel learner in maxDemon 4.0 clearly identifies the functional binding sites hidden within the thermal noise (i.e. the ends of the two loops of the protein with amino acid side chains that intercalate with nucleotides in the DNA double helix).

Lastly, we demonstrate that our application of mixed-effects model ANOVA to identify time synchronous differences in atom fluctuations as a pre-processing step prior to building a graph network (ChoreoGraph 2.0), can act effectively to identify dynamically choreographed regions of sites and to determine whether these differences in the magnitude of motion are significantly coordinated over time (i.e. via the interaction term in the model). As the ANOVA approach requires coarse-grained blocks of time to be defined, we suppose that it is likely to be relatively unaffected by thermal noise in the MD simulations, especially if the allosteric effects

on dynamics are happening on a longer time scale when compared to either the thermal noise and/or the molecular resonance in the MD simulation. Our application of mixed-effects model ANOVA combined with network graph Louvain community detection (ChoreoGraph 2.0) offers users a simple GUI interface, that unlike WISP (31), analyzes all pair-wise combinations of sites at once without requiring users to select a predetermined source and sink site on the protein to be connected by the graph network. In our work presented here, ChoreoGraph 2.0 demonstrates a clear role of the coordination of motions of amino acid sites via a combination of local resonance and loop regulation in the allosteric control and logic-gating behavior of a well-studied pathogenic protease (49). We also note that while our choregraphic analysis of this allosteric protease confirms the long hypothesized existence of 'tensed' and 'relaxed' protein states involved in allosteric regulation, it also represents a significant departure from past theory in that it demonstrates that these states are not always dependent upon interactions across separate protein domains that invoke conformational change (4, 5, 55); however see (56). As we observe here, they also occur across communities of sites with coordinated dynamics acting within a single domain and in a relatively short amount of MD simulation (i.e. < 10 ns). Despite the apparent success in our analysis validating many aspects we already know about this pathogenic protease, we would warn users that all of our methods require a well-defined and realistic structure that represents the reference and query state of the protein, along with all the usual caveats needed when carefully applying MD simulation (e.g. sufficient energy minimization and equilibration, adding hydrogens, proper force fields, etc.).

In conclusion, ATOMDANCE software offers suite of analyses useful for (A) identifying single sites with large effect on the function and/or evolution of PLI and PPI and (B) investigating shifts in protein dynamics that are affected by the logic state of these sites (i.e. bound vs. unbound). Here, we have demonstrated the utility of ATOMDANCE for investigating a variety of functional aspects of PLI including TATA-binding and BRAF kinase inhibitor drug induced shifts in dynamics, PPIs involved in infectious disease, as well as the functional evolutionary convergence of dynamic function in human-bacterial TATA-binding protein orthologs. ATOMDANCE offers our traditional comparative metrics applied to molecular dynamics (DROIDS 5.0) as well as a novel kernel-based approach to identifying specific key sites driving either binding interactions and/or protein activation (maxDemon 4.0). ATOMDANCE also offers a choreographic analysis for characterizing regions across the protein where coordinated changes or shifts in dynamics over time may involve multiple sites (ChoreoGraph 2.0). ATOMDANCE is entirely python-based with an easy to use graphical interface with seamless interaction with the open-source cpptraj MD analysis library and the modern UCSF ChimeraX molecular visualization software.

Declaration of Interests -

The authors declare no competing interests

Author Contributions –

GAB, MLL, and EPF contributed to method design, GAB, MLL, and MCF contributed to biological experiment design and interpretation, GAB engineered and released the final code version, HR, KP, NM, LTM, RAB and PR are Bioinformatics BS/MS students/former students from the Rochester Institute of Technology who contributed to software design and development and beta testing. CJR is a McQuaid High School Computer Club student who contributed to software design and development.

Acknowledgements -

We thank Dr. George M. Thurston for insightful feedback and creative suggestions for the early direction of this work.

References

1. Fischer, E. 1894. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch. Chem. Ges.* 27:2985–2993.

2. Koshland, D.E. 1958. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* 44:98–104.

3. Tripathi, A., and V.A. Bankaitis. 2017. Molecular Docking: From Lock and Key to Combination Lock. *J. Mol. Med. Clin. Appl.* 2:10.16966/2575-0305.106.

4. Liu, J., and R. Nussinov. 2016. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* 12:e1004966.

5. Monod, J., J. Wyman, and J.P. Changeux. 1965. ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. *J. Mol. Biol.* 12:88–118.

6. Changeux, J.-P. 2013. The Origins of Allostery: From Personal Memories to Material for the Future. *J. Mol. Biol.* 425:1396–1406.

7. Abendroth, J.M., O.S. Bushuyev, P.S. Weiss, and C.J. Barrett. 2015. Controlling Motion at the Nanoscale: Rise of the Molecular Machines. *ACS Nano*. 9:7746–7768.

8. Flechsig, H., and A.S. Mikhailov. 2019. Simple mechanics of protein machines. *J. R. Soc. Interface*. 16:20190244.

9. Strong, M. 2004. Protein Nanomachines. *PLoS Biol.* 2:e73.

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

11. Morcos, F., and J.N. Onuchic. 2019. The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes. *Curr. Opin. Struct. Biol.* 56:179–186.

12. Cornwell, W., and S. Nakagawa. 2017. Phylogenetic comparative methods. *Curr. Biol.* 27:R333–R336.

13. Suzuki, Y. 2010. Statistical methods for detecting natural selection from genomic data. *Genes Genet. Syst.* 85:359–376.

14. Kufareva, I., and R. Abagyan. 2012. Methods of protein structure comparison. *Methods Mol. Biol. Clifton NJ*. 857:231–257.

15. Parolini, G. 2015. The Emergence of Modern Statistics in Agricultural Science: Analysis of Variance, Experimental Design and the Reshaping of Research at Rothamsted Experimental Station, 1919–1933. *J. Hist. Biol.* 48:301–335.

16. Babbitt, G.A., E.P. Fokoue, J.R. Evans, K.I. Diller, and L.E. Adams. 2020. DROIDS 3.0—Detecting Genetic and Drug Class Variant Impact on Conserved Protein Binding Dynamics. *Biophys. J.* 118:541–551.

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

18. Babbitt, G.A., M.L. Lynch, M. McCoy, E.P. Fokoue, and A.O. Hudson. 2022. Function and evolution of B-Raf loop dynamics relevant to cancer recurrence under drug inhibition. *J. Biomol. Struct. Dyn.* 40:468–483.

19. Rajendran, M., M.C. Ferran, L. Mouli, G.A. Babbitt, and M.L. Lynch. 2023. Evolution of drug resistance drives destabilization of flap region dynamics in HIV-1 protease. *Biophys. Rep.* 3:100121.

20. Rajendran, M., M.C. Ferran, and G.A. Babbitt. 2022. Identifying vaccine escape sites via statistical comparisons of short-term molecular dynamics. *Biophys. Rep.* 2:100056.

21. Rajendran, M., and G.A. Babbitt. 2022. Persistent cross-species SARS-CoV-2 variant infectivity predicted via comparative molecular dynamics simulation. *R. Soc. Open Sci.* 9:220600.

22. Rynkiewicz, P., M.L. Lynch, F. Cui, A.O. Hudson, and G.A. Babbitt. 2021. Functional binding dynamics relevant to the evolution of zoonotic spillovers in endemic and emergent Betacoronavirus strains. *J. Biomol. Struct. Dyn.* 1–19.

23. Wang, S. 2019. Efficiently Calculating Anharmonic Frequencies of Molecular Vibration by Molecular Dynamics Trajectory Analysis. *ACS Omega*. 4:9271–9283.

24. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

25. Petersen, H.G. 1995. Accuracy and efficiency of the particle mesh Ewald method. *J. Chem. Phys.* 103:3668–3679.

26. Wang, H., P. Zhang, and C. Schütte. 2012. On the Numerical Accuracy of Ewald, Smooth Particle Mesh Ewald, and Staggered Mesh Ewald Methods for Correlated Molecular Systems. *J. Chem. Theory Comput.* 8:3243–3256.

27. Ponte, P., and R.G. Melko. 2017. Kernel methods for interpretable machine learning of order parameters. *Phys. Rev. B*. 96:205146.

28. Bauer, J.A., J. Pavlović, and V. Bauerová-Hlinková. 2019. Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules*. 24:3293.

29. Arantes, P.R., A.C. Patel, and G. Palermo. 2022. Emerging Methods and Applications to Decrypt Allostery in Proteins and Nucleic Acids. *J. Mol. Biol.* 434:167518.

30. Verkhivker, G.M., S. Agajanian, G. Hu, and P. Tao. 2020. Allosteric Regulation at the Crossroads of New Technologies: Multiscale Modeling, Networks, and Machine Learning. *Front. Mol. Biosci.* 7.

31. Van Wart, A.T., J. Durrant, L. Votapka, and R.E. Amaro. 2014. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* 10:511–517.

32. Bowerman, S., and J. Wereszczynski. 2016. Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods Enzymol.* 578:429–447.

33. Papaleo, E., G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F.L. Gervasio, and R. Nussinov. 2016. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chem. Rev.* 116:6391–6423.

34. Schlick, T., M. Mandziuk, R.D. Skeel, and K. Srinivas. 1998. Nonlinear Resonance Artifacts in Molecular Dynamics Simulations. *J. Comput. Phys.* 140:1–29.

35. Rosa, H.B., Rosemary Candelario, J. Lorenzo Perillo, Cristina Fernandes. 2023. Choreographic Analysis as Dance Studies Methodology: Cases, Expansions, and Critiques. In: Dance Research Methodologies. Routledge.

36. Roe, D.R., and T.E.I. Cheatham. 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* 9:3084–3095.

37. Blondel, V.D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008:P10008.

38. Hagberg, A., D. Schult, and P. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Proceedings of the 7th Python in Science conference. G Varoquaux, T Vaught, J Millman. pp. 11–15.

39. Goddard, T.D., C.C. Huang, E.C. Meng, E.F. Pettersen, G.S. Couch, J.H. Morris, and T.E. Ferrin. 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci. Publ. Protein Soc.* 27:14–25.

40. Pettersen, E.F., T.D. Goddard, C.C. Huang, E.C. Meng, G.S. Couch, T.I. Croll, J.H. Morris, and T.E. Ferrin. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. Publ. Protein Soc.* 30:70–82.

41. Case, D.A., T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. 2005. The Amber biomolecular simulation programs. *J. Comput. Chem.* 26:1668–1688.

42. Eastman, P., J. Swails, J.D. Chodera, R.T. McGibbon, Y. Zhao, K.A. Beauchamp, L.-P. Wang, A.C. Simmonett, M.P. Harrigan, C.D. Stern, R.P. Wiewiora, B.R. Brooks, and V.S. Pande. 2017. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* 13:e1005659.

43. Jo, S., T. Kim, V.G. Iyer, and W. Im. 2008. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* 29:1859–1865.

44. Phillips, J.C., D.J. Hardy, J.D.C. Maia, J.E. Stone, J.V. Ribeiro, R.C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M.C.R. Melo, B.K. Radak, R.D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L.V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid. 2020. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* 153:044130.

45. Ribeiro, J.V., R.C. Bernardi, T. Rudack, J.E. Stone, J.C. Phillips, P.L. Freddolino, and K. Schulten. 2016. QwikMD — Integrative Molecular Dynamics Toolkit for Novices and Experts. *Sci. Rep.* 6:26536.

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

47. Wan, P.T.C., M.J. Garnett, S.M. Roe, S. Lee, D. Niculescu-Duvaz, V.M. Good, C.M. Jones, C.J. Marshall, C.J. Springer, D. Barford, R. Marais, and Cancer Genome Project. 2004. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*. 116:855–867.

48. Yan, R., Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou. 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*. 367:1444–1448.

49. Lupardus, P.J., A. Shen, M. Bogyo, and K.C. Garcia. 2008. Small molecule-induced allosteric activation of the Vibrio cholerae RTX cysteine protease domain. *Science*. 322:265–268.

50. Andersen, H.C. 1980. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* 72:2384–2393.

51. Ewald, P.P. 1921. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* 369:253–287.

52. Pierce, L.C.T., R. Salomon-Ferrer, C. Augusto F. de Oliveira, J.A. McCammon, and R.C. Walker. 2012. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* 8:2997–3002.

53. Hamelberg, D., J. Mongan, and J.A. McCammon. 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120:11919–11929.

54. Hamelberg, D., C.A.F. de Oliveira, and J.A. McCammon. 2007. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J. Chem. Phys.* 127:155102.

55. Guo, J., and H.-X. Zhou. 2016. Protein Allostery and Conformational Dynamics. *Chem. Rev.* 116:6503–6515.

56. Cooper, A., and D.T. Dryden. 1984. Allostery without conformational change. A plausible model. *Eur. Biophys. J. EBJ*. 11:103–109.

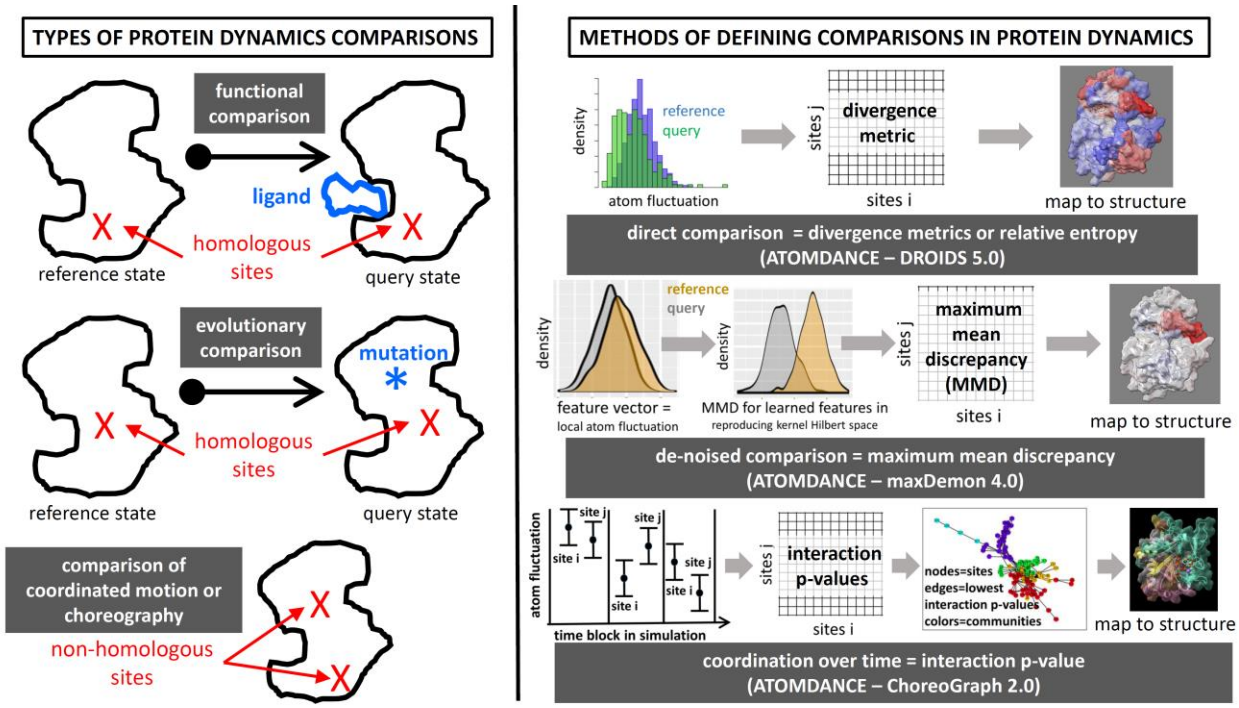

Figure 1 – Overview of the ATOMDANCE statistical machine learning post-processor for comparative protein dynamics. The three main types of site-wise comparisons of protein dynamics (left) and the three main methods of comparative analysis (right) are visually summarized. The types of protein dynamics comparisons shown here include (A) functional comparisons across homologous sites where the reference state and query state of the dynamics is measured from two molecular dynamics (MD) simulations of the unbound protein and its ligand bound state (resp.), (B) evolutionary/genetic comparisons across homologous sites where the reference and query state is measured from dynamics before and after the mutation, and (C) comparisons of dynamics of non-homologous sites over time collected from a single MD simulation. Note that functional comparisons could also include changes in temperature and evolutionary comparisons could also include changes in epigenetics as well. The methods of comparative analysis include (A) a direct comparison of atom fluctuations across all sites computed using divergence metrics (i.e. DROIDS 5.0), (B) a de-noised comparison of learned features of the local atom fluctuations across all sites computed via Gaussian process kernel machine learners (maxDemon 4.0), and (C) a comparison of how

coordinated the atom fluctuations are between all pair-wise combinations of sites computed via a mixed-effects model analysis of variance (ChoreoGraph 2.0).

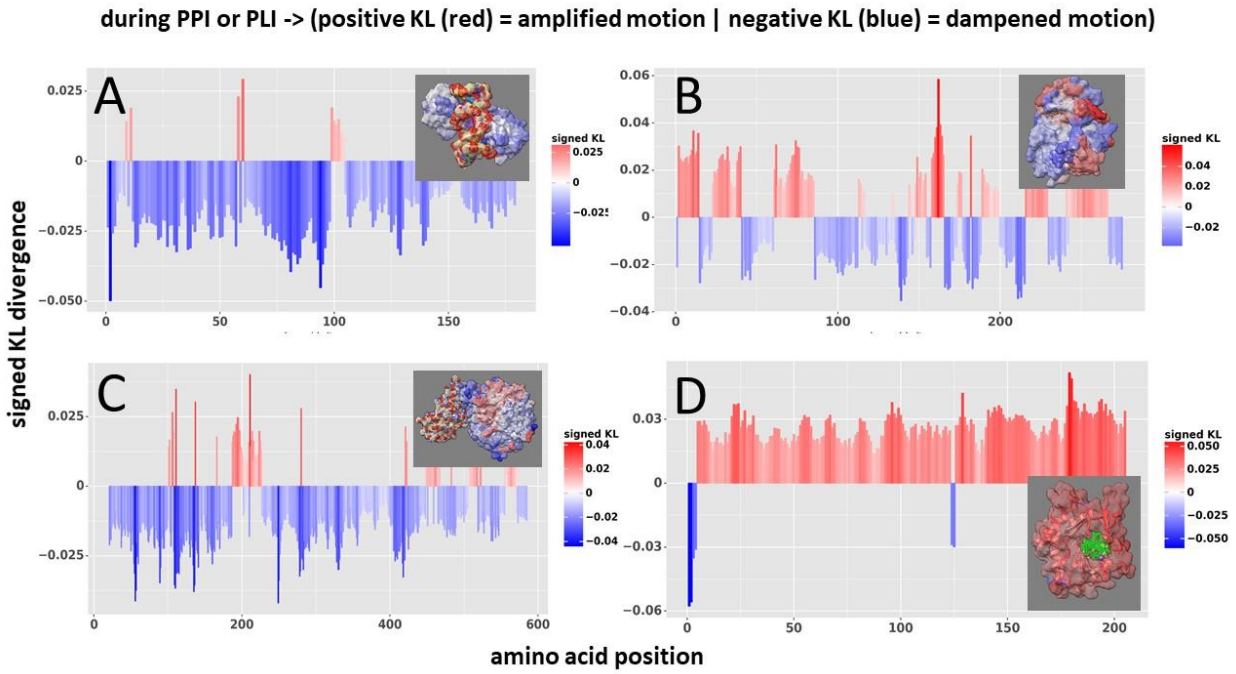

Figure 2 – DROIDS 5.0 analysis of direct site-wise divergence metrics in local atom fluctuation when comparing molecular dynamics simulations of functionally bound vs. unbound target proteins. The comparisons include (A) DNA-bound vs. unbound TATA binding protein (PDB: 1cdw), (B) sorafenib-bound vs. unbound B-Raf kinase domain (PDB: 1uwh), (C) SARS-CoV-2 viral bound vs. unbound angiotensin-converting enzyme 2 (ACE2) protein (PDB: 6m17), and (D) the allosteric activated (i.e. InsP6 bound) vs inactivated (i.e. unbound) Vibrio cholera toxin RTX cysteine protease domain (PDB: 3eeb). Signed symmetric Kullback-Leibler (KL) divergence in atom fluctuation indicates sites where motion is dampened during binding (blue) and where motion is amplified (red). Note that while binding typically dampens atom fluctuation locally or even globally, in the case of this example of allostery (D) it actually amplifies atom fluctuation globally.

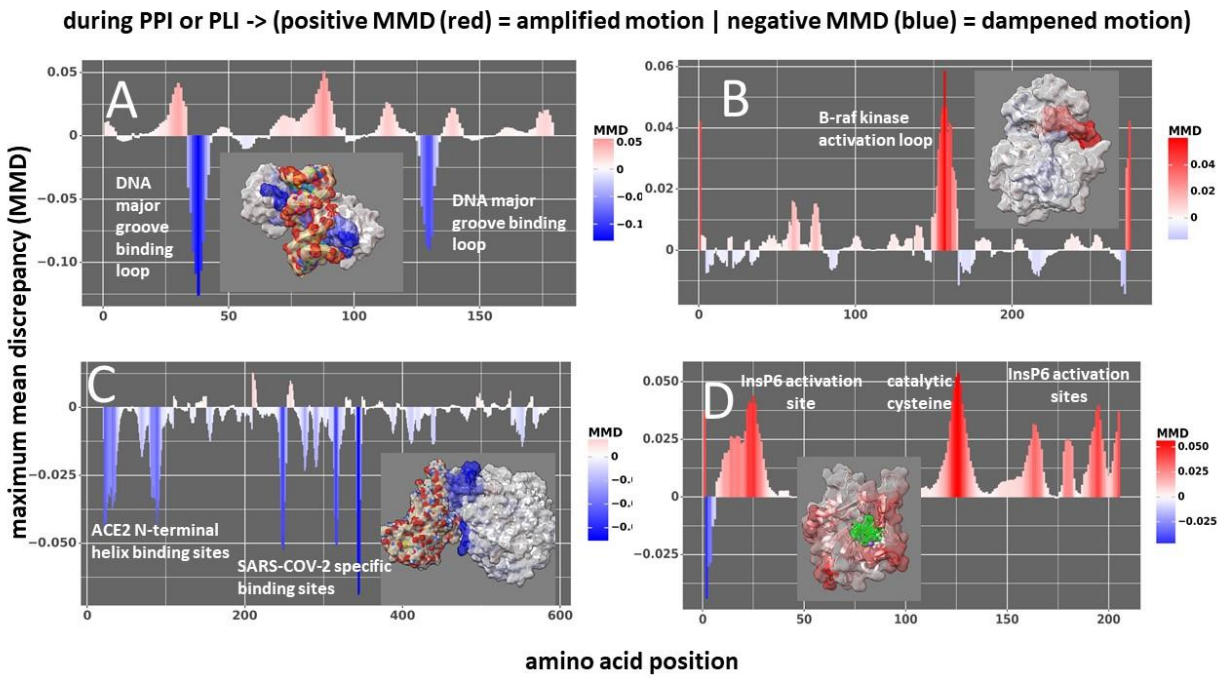

Figure 3 – maxDemon 4.0 analysis of signed maximum mean discrepancy (MMD) learned features regarding local atom fluctuation when comparing molecular dynamics simulations of functionally bound vs. unbound target proteins. The comparisons include (A) DNA-bound vs. unbound TATA binding protein (PDB: 1cdw), (B) sorafenib-bound vs. unbound B-Raf kinase domain (PDB: 1uwh), (C) SARS-CoV-2 viral bound vs. unbound angiotensin-converting enzyme 2 (ACE2) protein (PDB: 6m17), and (D) the allosteric activated (i.e. InsP6 bound) vs inactivated (i.e. unbound) Vibrio cholera toxin RTX cysteine protease domain (PDB: 3eeb). Signed MMD in atom fluctuation indicates sites where motion is dampened during binding (blue) and where motion is amplified (red). Note that the kernel-based learning applied to local atom fluctuation (i.e. signed MMD) removes the overall effect of differences in thermal noise present in divergence metrics (Figure 2) from the functional comparison and so much better isolates the binding sites themselves. Close-up views of the color-mapped structures are shown in Supplemental Figure 6.

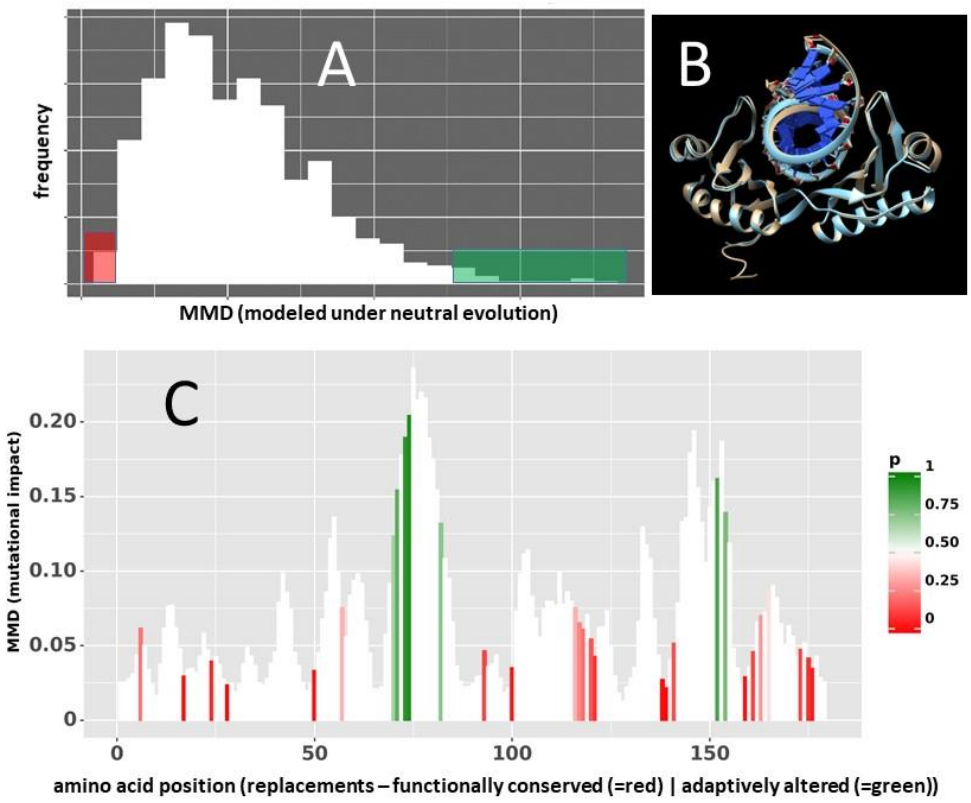

Figure 4 – Site-wise unsigned maximum mean discrepancy MMD in local atom fluctuation and atom correlation comparing DNA-bound models of bacterial and human orthologs of TATA binding protein (PDB: 1qna and PDB 1cdw resp.). As a test of neutral evolution, the MMD between dynamics on randomly chosen differing amino acid sites between the orthologs is used to generate (A) an expected distribution of MMD for the effects of random amino acid replacement on molecular dynamics. The tails of the distribution are used to identify MMD values indicative of functionally conserved dynamics (red) or adaptively altered dynamics (green). (B) The superimposition of the two structures shows that the protein has maintained near perfect structural similarity since the divergence of common ancestor between bacteria and humans despite many amino acid replacements over time. (C) The MMD profile of the dynamic differences between orthologs is the background (in white) with the bootstrap analyses of MMD for the existing amino acid replacements (in color). Red indicates dynamic changes that are significantly smaller than expected under the neutral model (i.e. functionally conserved) while green indicates dynamic changes that are significantly larger under the neutral model (i.e. adaptively altered).

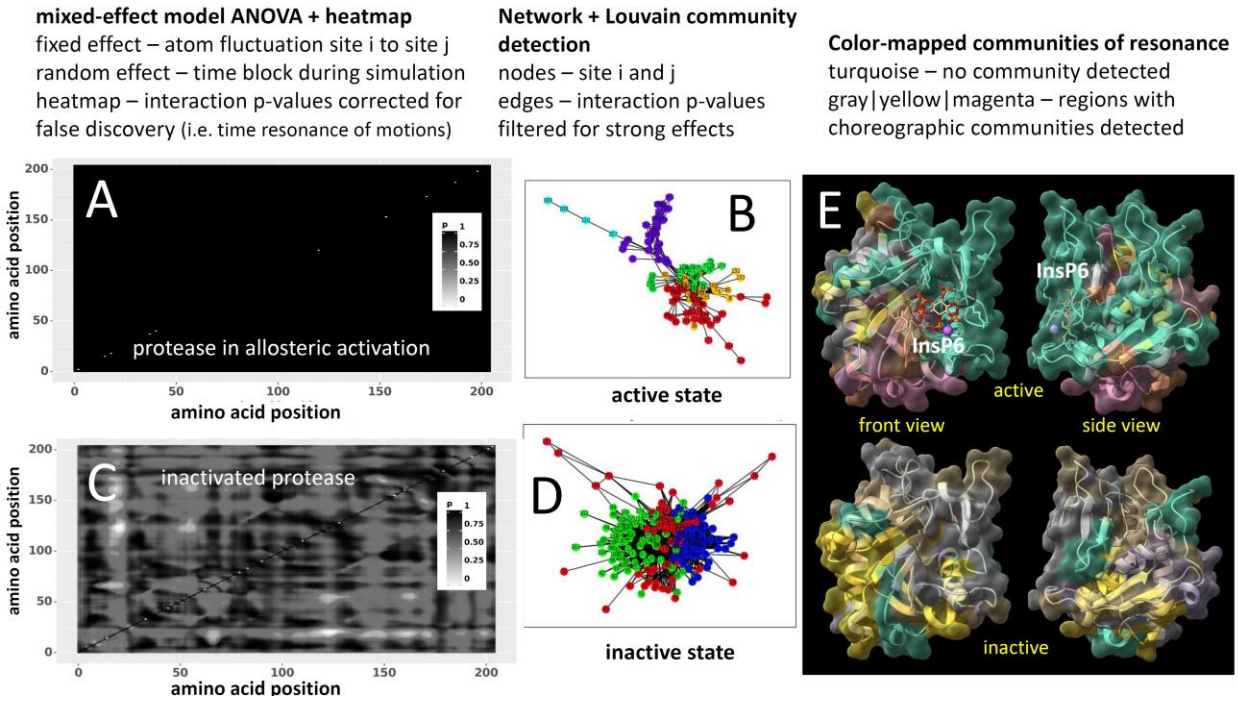

Figure 5 – ChoreoGraph 2.0 - choreographic analysis with interaction p-value heat maps and network-based community detection indicating regions of coordinated protein dynamics over time. The heat maps of the interaction term p-values for all pair-wise comparison of sites I to sites j on the (A) inactivated or unbound and (B) allosteric activated (i.e. InsP6 bound) Vibrio cholera toxin RTX cysteine protease domain (PDB: 3eeb) are shown. Multiple-test corrected interaction P- values for time synchronous differences in atom fluctuation across sites over time (i.e. coordinated motions) are derived from pair-wise mixed effects model ANOVAs where atom fluctuation is the dependent variable and sites i vs site j is the fixed effect and time samples are the random effect). Patterns of coordinated motion across sites are indicated by significant p-value (white). Regions of coordinated motion derived from Louvain community detection applied to graph network analysis are shown for (C) unbound and (D) InsP6-bound V. cholera protease. Choreographed regions (i.e. communities of sites with significant time interactions) are similarly color mapped to the surface of the protein. Regions that fail to form coordinated motions are colored light turquoise green. Note that upon binding (D) several very large resonance communities/regions disappear probably allowing the protease to more easily interact with host protein targets than when in the inactivated state within the bacteria. Connectivity and non-randomness of the interaction p-value network is significantly higher in the inactivated state (connectivity $t=-397.83$, $p<0.0001$ | non-randomness $t=-47.75$, $p<0.0001$). Close-up view of the interaction heat map for the inactivated protease (C) is given in Supplemental Figure 10.

Supporting Material – video overview with dynamics of DNA-bound TATA binding protein and sorafenib drug-bound B-Raf kinase domain weighted in accordance with maximum mean discrepancy in atom fluctuation. https://people.rit.edu/gabsbi/img/videos/MMDmovie.mp4