Machine learning based detection of genetic and drug class variant impact on functionally conserved protein dynamics

Dr Gregory A. Babbitt – T.H. Gosnell School of Life Sciences, Rochester Inst. Technol. (RIT) NY USA
Dr Ernest P. Fokoue – School of Mathematical Sciences, Rochester Inst. Technol. (RIT) NY USA

Machine learning based detection of genetic and drug class variant impact on functionally conserved protein dynamics

OUTLINE

1. **Statement of the Problem** – Traditional bioinformatics favor static representations of data over dynamics. Dynamics of soft matter is where biological function lies. We need comparative methods for study of molecular dynamics (MD) much as we would compare sequence and structural data.

2. **Conceptual Solution to Problem** – DROIDS+maxDemon. We apply statistical ensembling (aka weather forecasting) to MD simulations of comparative functional states of proteins. We make statistical comparisons of dynamics that are inherently robust to model assumptions. We utilize machine learning to detect functional transitions in dynamics on new MD simulations. We utilize signatures of sequence-dependent (i.e. encoded) canonical self-correlation in machine learning to identify regions of functionally conserved dynamics. We now can compare variant vs self-correlation in functional dynamics with an entropy-based metric variant impact.

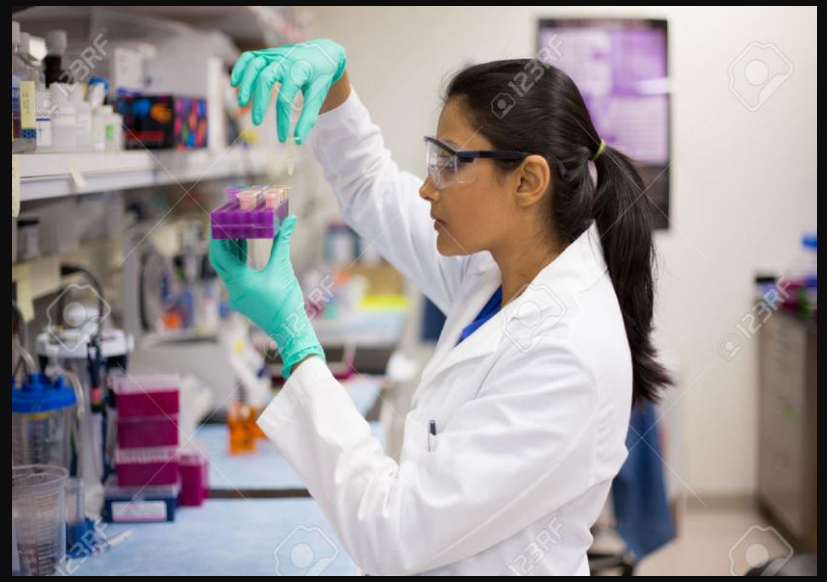
3. **Case Studies (Proof of Concept)** – temperature shifts and genetic mutations in ubiquitin, DNA binding in TBP, protein-ligand interactions targeting hemoglobin, Hsp90 and BRAF

Statement of the problem

Since all real science involves much failure, where would you prefer to fail?

expensive wet bench biology

Not an exact science / results not always reproducible
Achieving an atomic resolution is nearly impossible



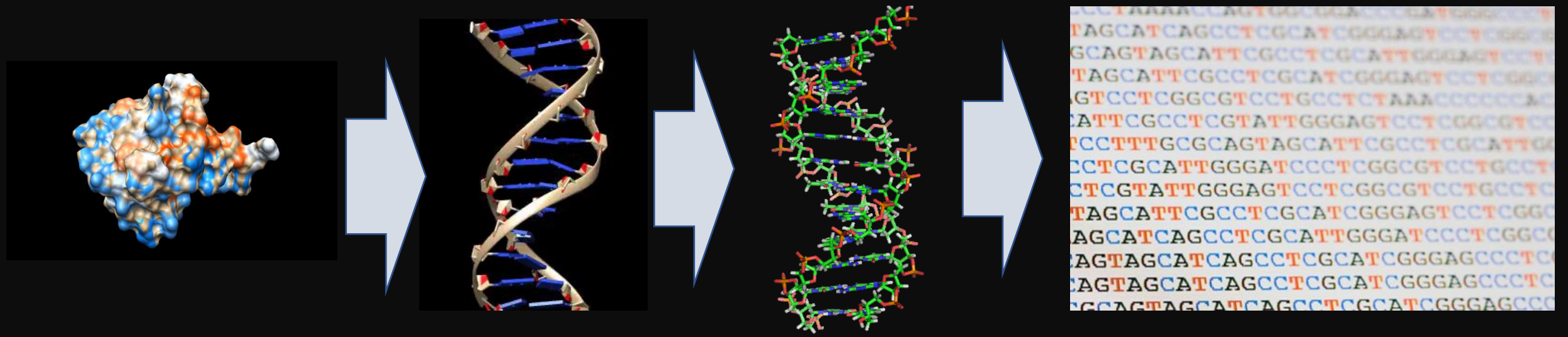
versus

inexpensive dry bench biology

Results highly reproducible at atomic resolution
But always beware GIGO

All of biology can be understood through the jiggling and wiggling of atoms – Richard Feynman

All of biology can be understood through the jiggling and wiggling of atoms – Richard Feynman

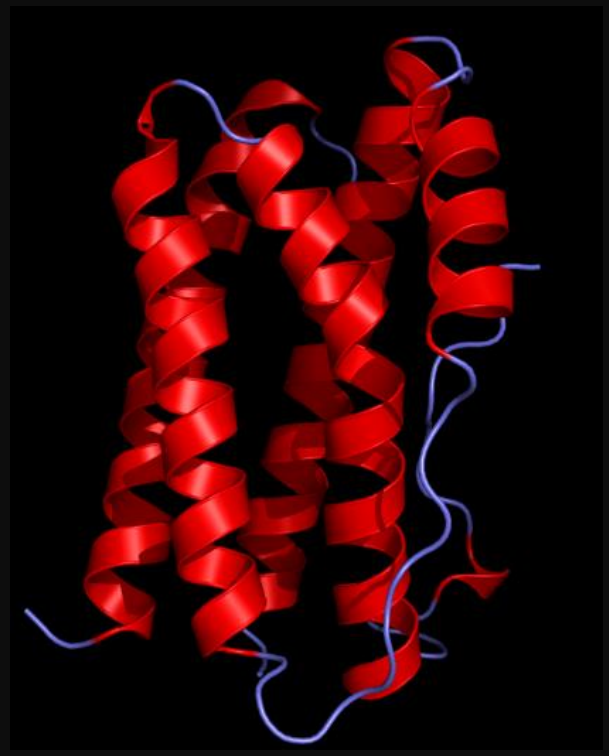


What is lost when we reduce biology to static forms of data?

“Genotype encodes its own phenotype” – Tom Tullius, Boston University

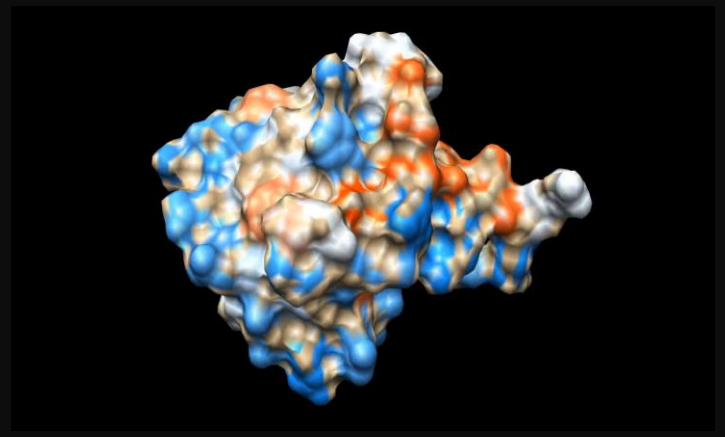
“Can all heritable biology really be reduced to a single dimension?” – G.A. Babbitt, RIT

Theodius Dobzhansky
-“nothing in biology makes any sense except in the light of evolution”

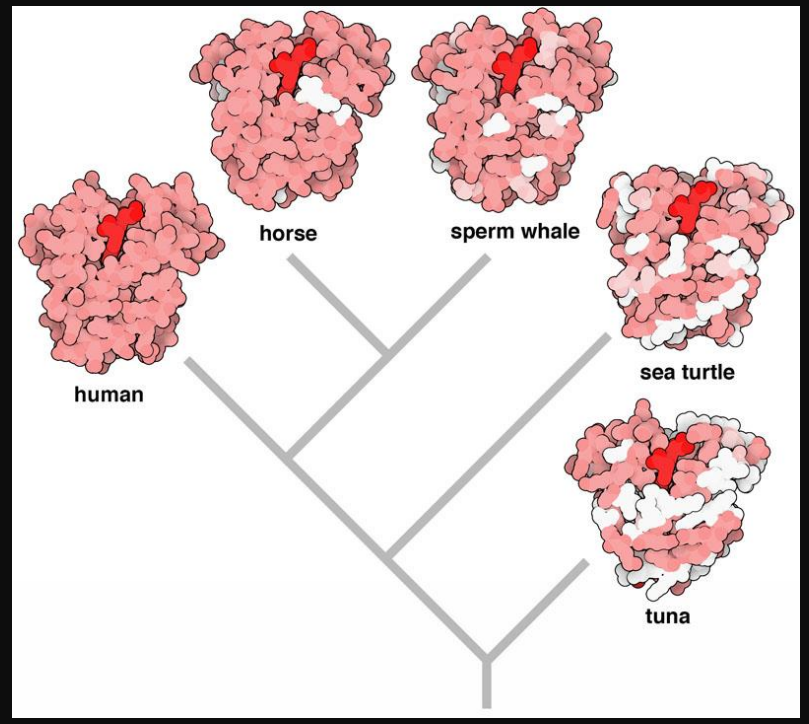


I believe this applies equally to functional traits of molecules as it does to functional traits of organisms

We aim to connect one of the shortest atom event timescales in the universe with one of the longest timescales



How do we measure and visualize this relationship?



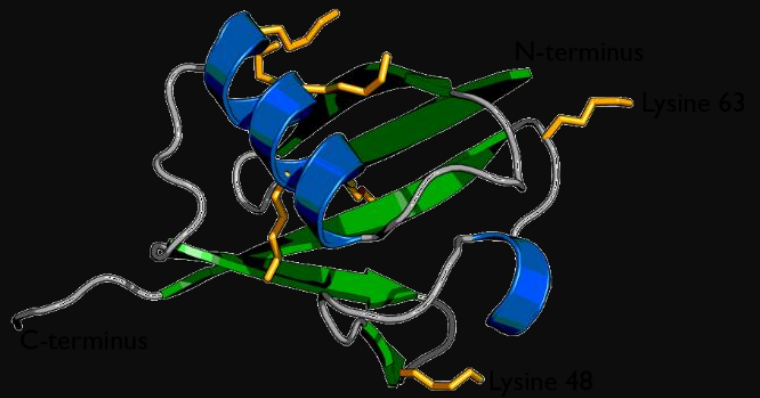
Molecular Function & Motion over femto/pico/nanoseconds

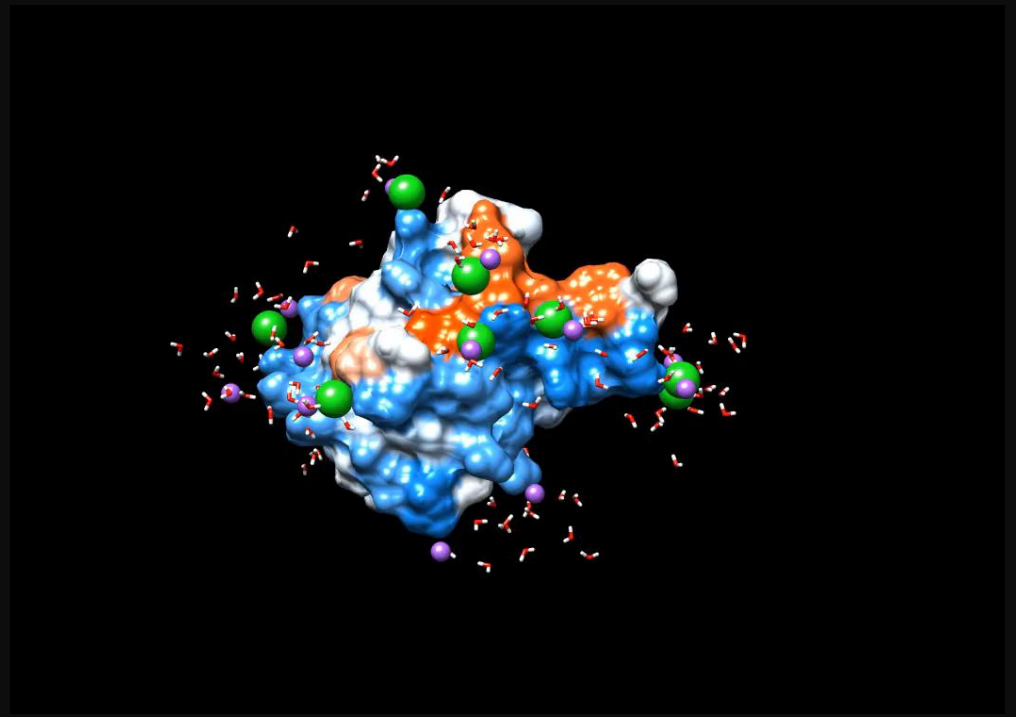
Molecular Evolution over 100s million/billions of years

What sort of information lies in ubiquitin structure versus explicit solvent MD simulation?

Given its inherent complexity, how can we make any sense of it?

Explicitly, how can we (A) identify functional dynamics and (B) employ a comparative framework to identify different functional/malfunctional states





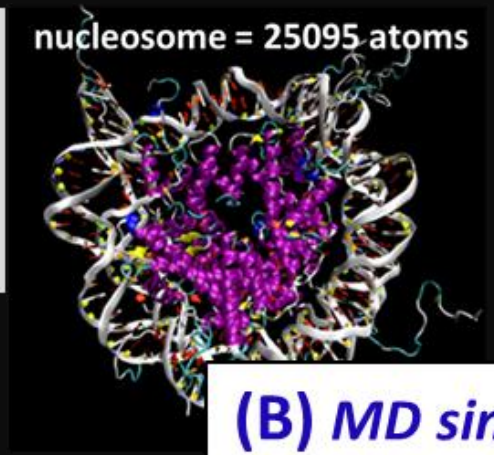
Conceptual Solution to Problem

BEFORE GPU

AFTER GPU

(A) *example game images*

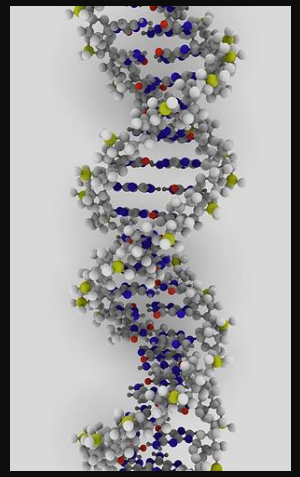
Xeon CPU
20 cores
0.08 ns/day
2fs time step



4 GPU cards
GTX-Titan-X
12208 cores
21.54 ns/day
2fs time step

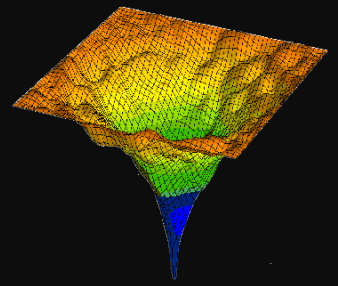
(B) *MD sim benchmarks*

Anatomy of a molecular dynamic simulation



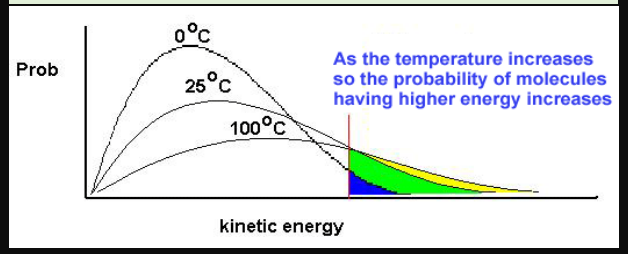
structure = ideal or PDB

minimization of potential energy surface (relaxation)



How might one statistically compare results between MD production runs?

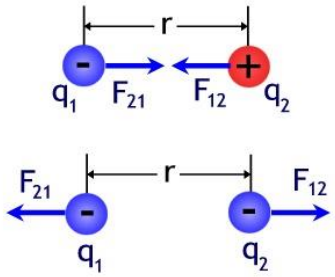
heating applied to all bonds



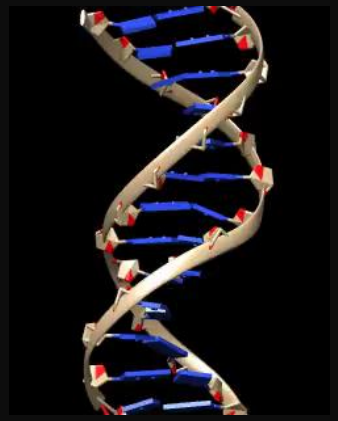
movies and statistics rendered from atomic vector trajectories

bonded atoms

non-bonded atom



MD equilibration and production run(s) track Newtonian mechanics stepped out in femtoseconds



Biophysical Journal
Computational Tool
Biophysical Society

DROIDS 1.20: A GUI-Based Pipeline for GPU-Accelerated Comparative Protein Dynamics

Gregory A. Babbitt,[1,*] Jamie S. Mortensen,[2] Erin E. Coppola,[2] Lily E. Adams,[1] and Justin K. Liao[2]
[1]T.H. Gosnell School of Life Sciences and [2]Department of Biomedical Engineering, Rochester Institute of Technology, Rochester, New York

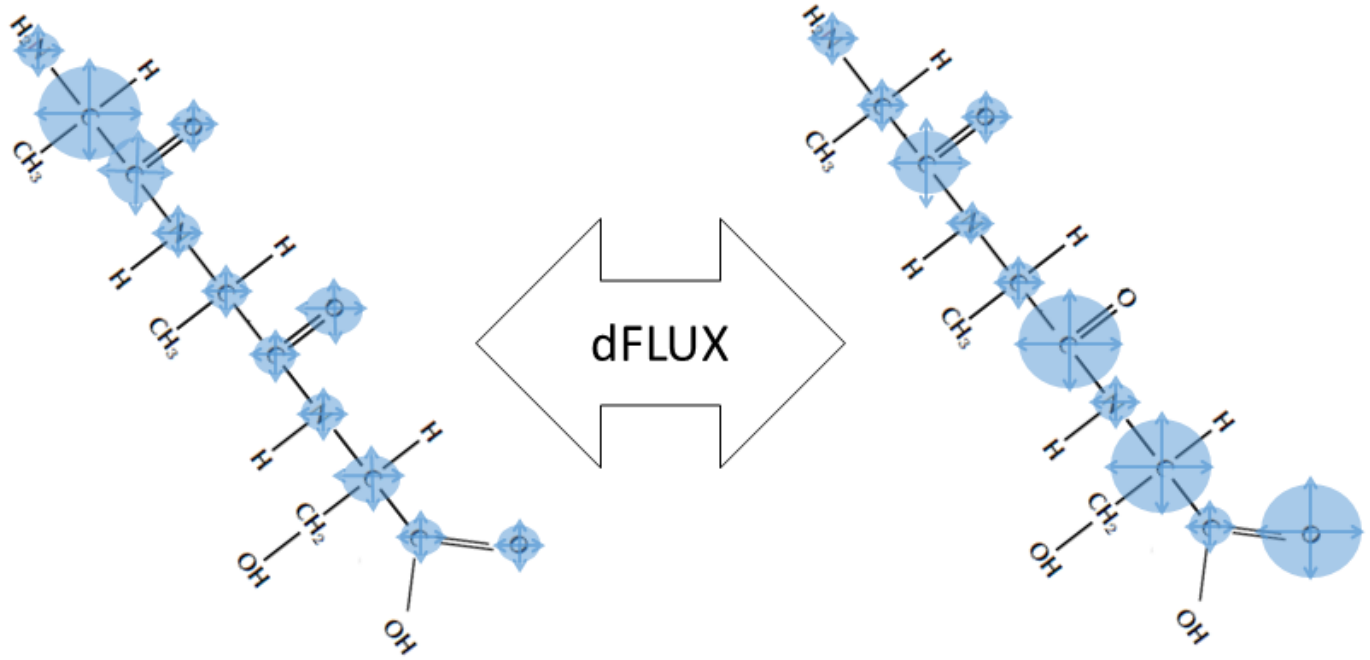
Detecting Relative Outlier Impacts in molecular Dynamic Simulations

primary metric of interest – residue backbone specific root mean square fluctuation or 'rmsf'
i.e. rapid, mostly harmonic bond vibrations measured in femtoseconds

when compared across ensembles we compare local distributions of rmsf using KL divergence (i.e. dFLUX)

atom flux on reference chain

atom flux on query chain



$$dFLUX_{aa} = \left(\sum_{i=1}^{4} FLUX_{atom}\right)_{query} - \left(\sum_{i=1}^{4} FLUX_{atom}\right)_{reference}$$

$$dFLUX_{chain} = \sum_{i=1}^{L} |dFLUX_{aa}|$$

where N = number of structurally homologous amino acids in reference chain

Thermodynamic Relationships

dFLUX = average difference in rapid small amplitude motions measured in angstroms over 100s ps to 10s ns on each chain

...thus
dFLUX ≈ Δq (change in heat)

where U = q + w and H = U + pv

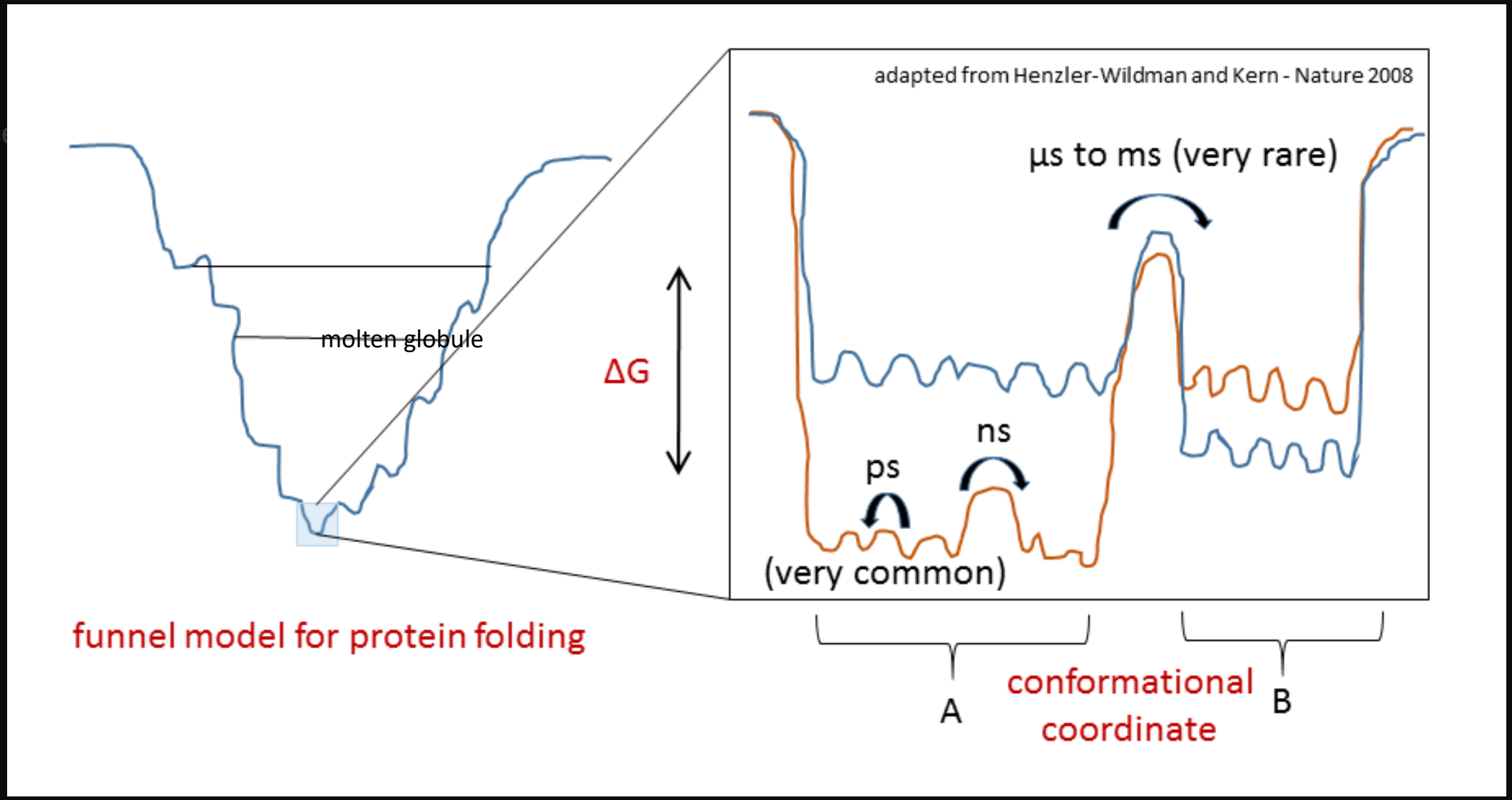
p = pressure, v = volume, w = work
H = enthalpy and U = internal energy

...as p and v are negligible under physiological conditions and heat and work are set only initially during MD simulation ....

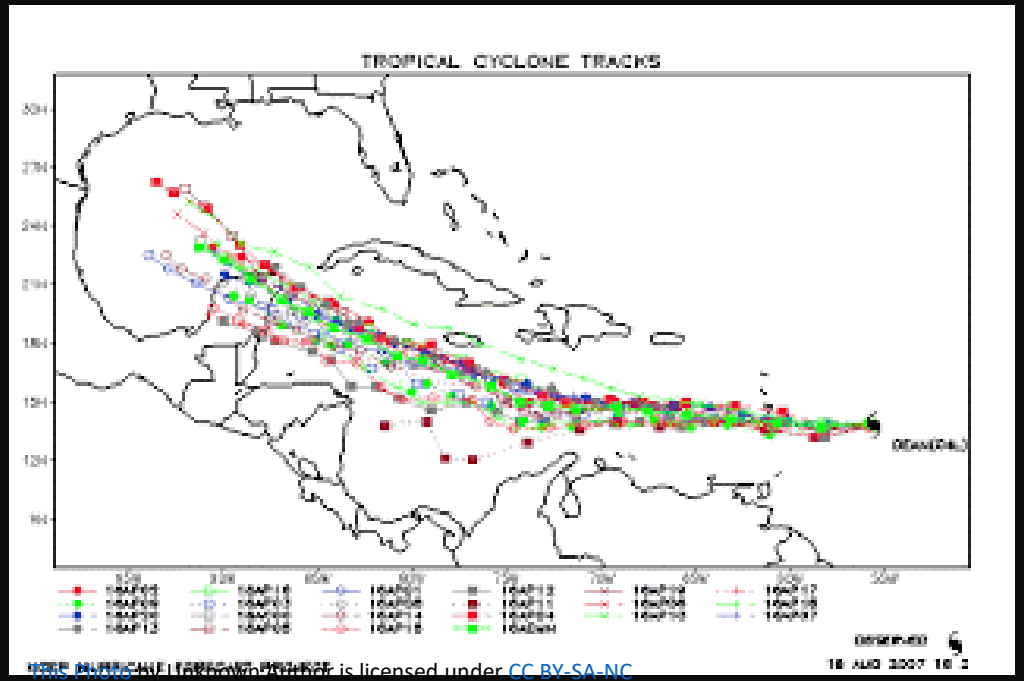
dFLUX ≈ Δq ≈ ΔH ≈ ΔG + TΔS

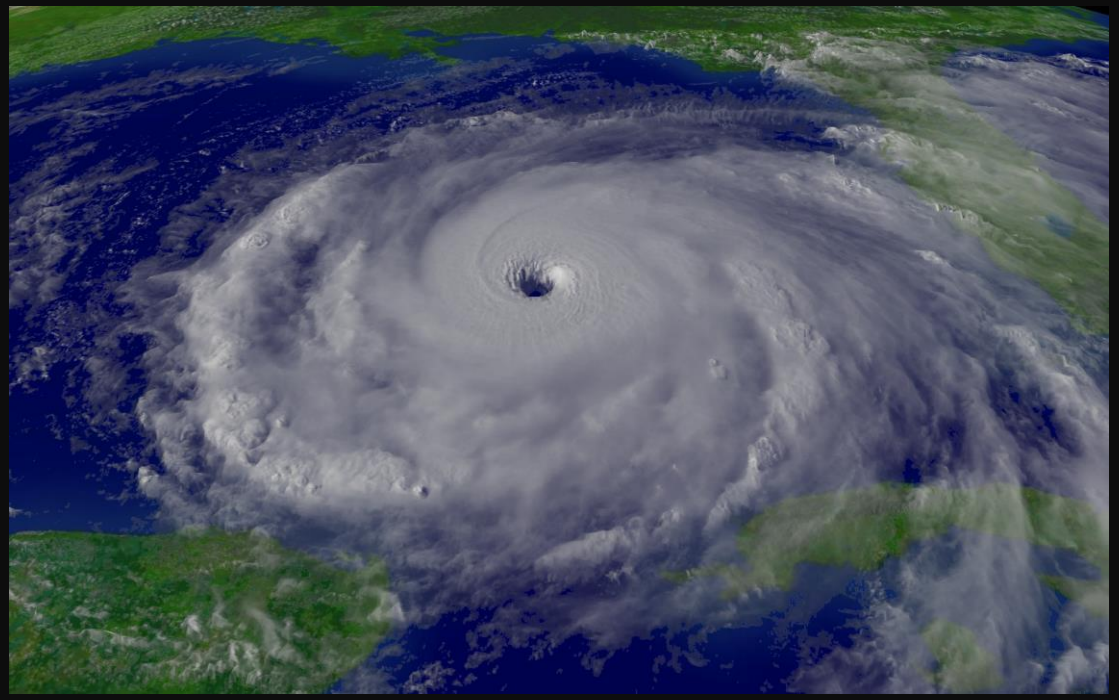
capturing differences due to both change in entropy and free energy on the query and reference structures

Dynamic transitions in structure can be chaotic (i.e. unpredictable and sensitive to initial conditions)



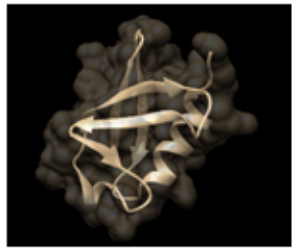
Dynamic transitions in structure can be chaotic (i.e. unpredictable and sensitive to initial conditions)





We take advantage of model simulation ensembling to buffer impact of chaotic behavior on prediction

Reference MD run

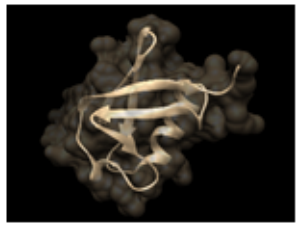


energy minimize heating

long equilibration MD run

sampling MD runs = random time spacer to mitigate chaotic dynamics + fixed time data production run

Query MD run

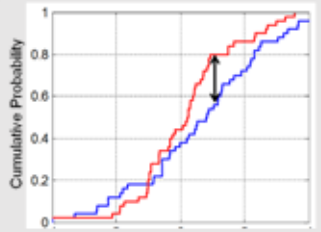


energy minimize heating

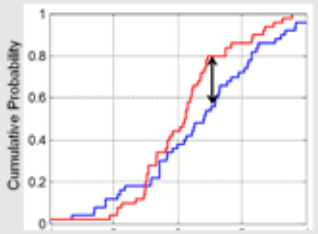
long equilibration MD run

Two sets of molecular dynamic (MD) simulations to be compared

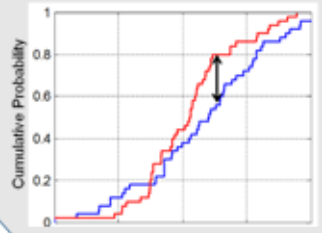
GLY 2



GLN 241



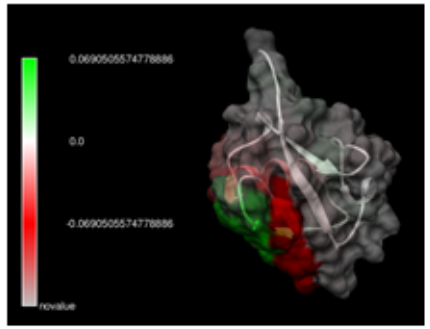
LYS 125



Cumulative distributions of average atom fluctuations (RMSF) over time returned by CPPTRAJ

2 sample KS tests on MD of each residue backbone

Colors = delta values, p values, or KS D values



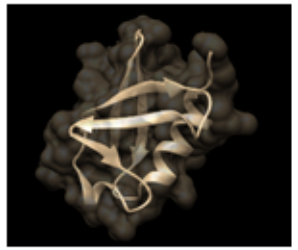
Light gray = no difference or no significance

Dark gray = no homology between structures

results are color mapped onto reference MD movie

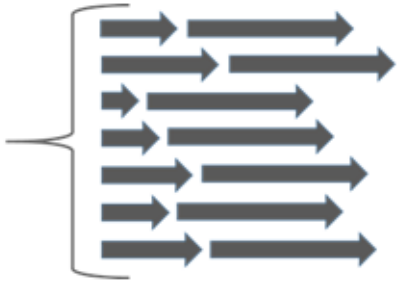
NOTE: KL divergence and two sample KS test make no stat model assumptions i.e. allow for wide range of comparison between complex rmsf patterns driven by many potential latent variables

Reference MD run



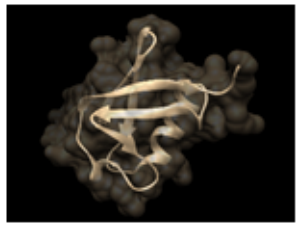
energy minimize heating

long equilibration MD run



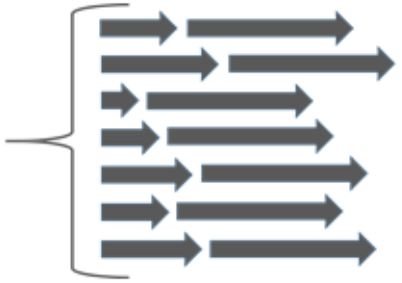
sampling MD runs = random time spacer to mitigate chaotic dynamics + fixed time data production run

Query MD run



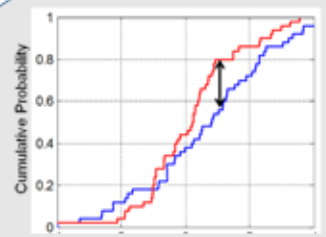
energy minimize heating

long equilibration MD run

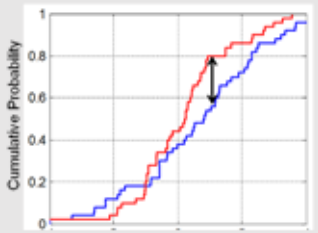


Two sets of molecular dynamic (MD) simulations to be compared

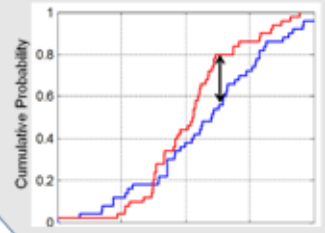
GLY 2



GLN 241



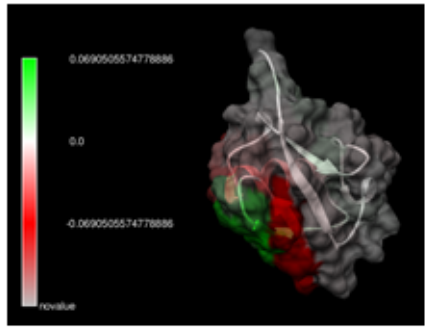
LYS 125



Cumulative distributions of average atom fluctuations (RMSF) over time returned by CPPTRAJ

2 sample KS tests on MD of each residue backbone

Colors = delta values, p values, or KS D values



Light gray = no difference or no significance

Dark gray = no homology between structures

results are color mapped onto reference MD movie

DROIDS v3.0 – software for comparative dynamics

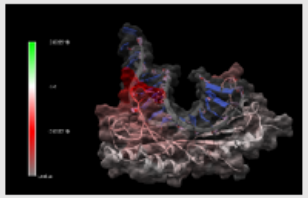
maxDemon v1.0 – machine learning classification of dynamics

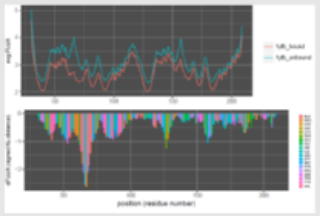
A

hot ↔ cold

before mutation ↔ after mutation

unbound ↔ bound





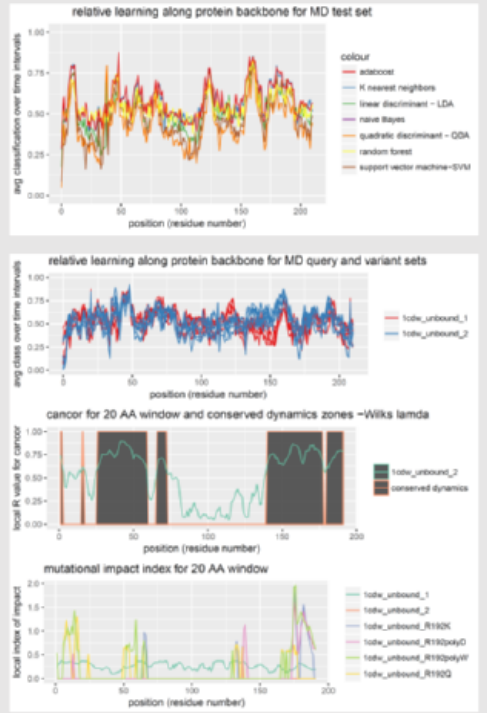
B GENETIC and/or BINDING VARIANT LIST

self copy
variant 1
variant 2
variant 3
etc.

C METHODS

KNN
naïve Bayes
LDA/QDA
SVM
random forest
adaboost

D STATISTICAL ANALYSIS

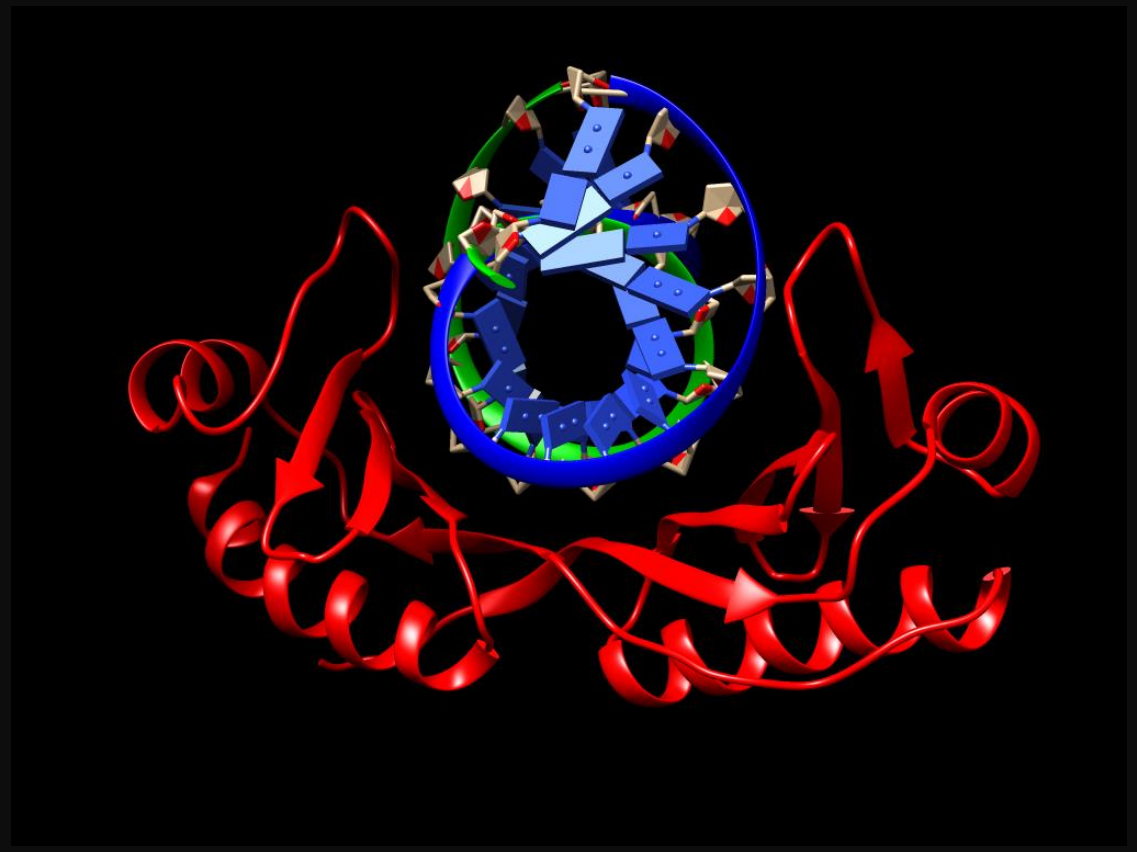


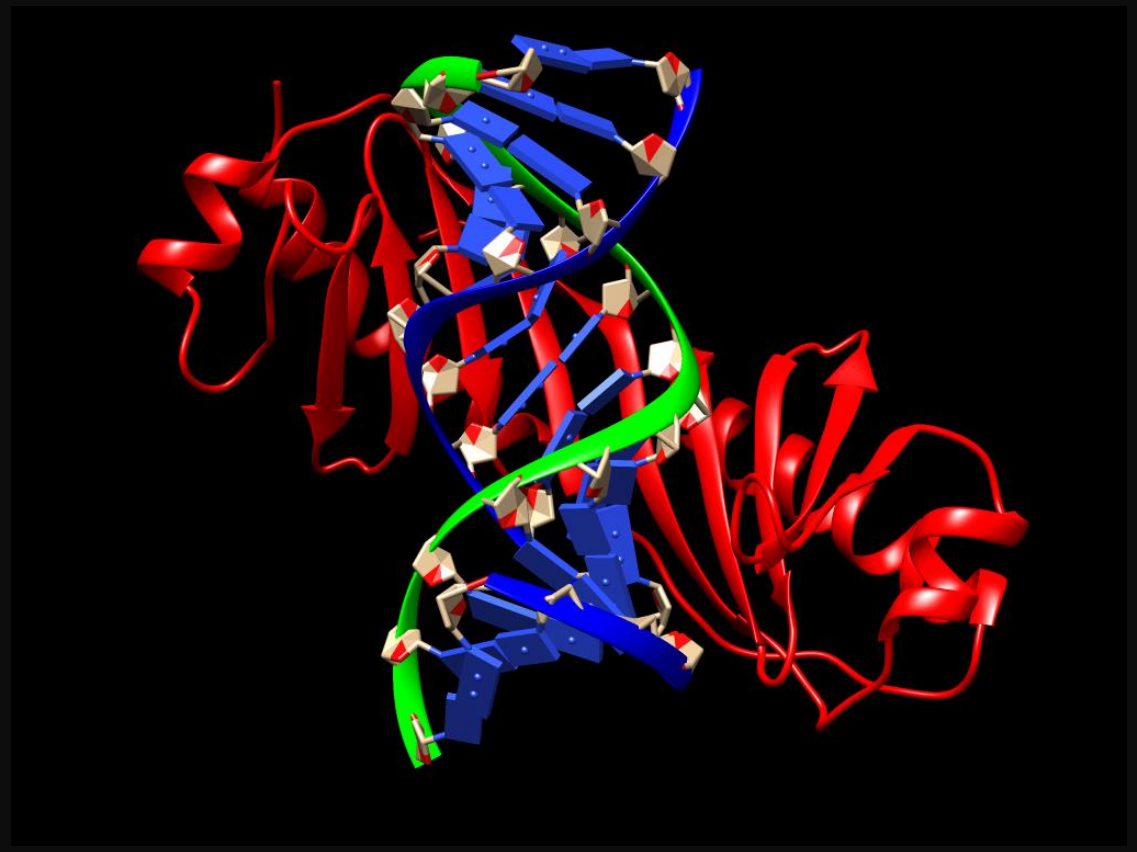
(A) MD ensemble comparison to generate classified training set for normal functional state of protein

(B) New MD runs to generate deployment (test) sets for genetic variants of potential interest

(C) Machine learning to classify genetic variant runs in both time and structural space

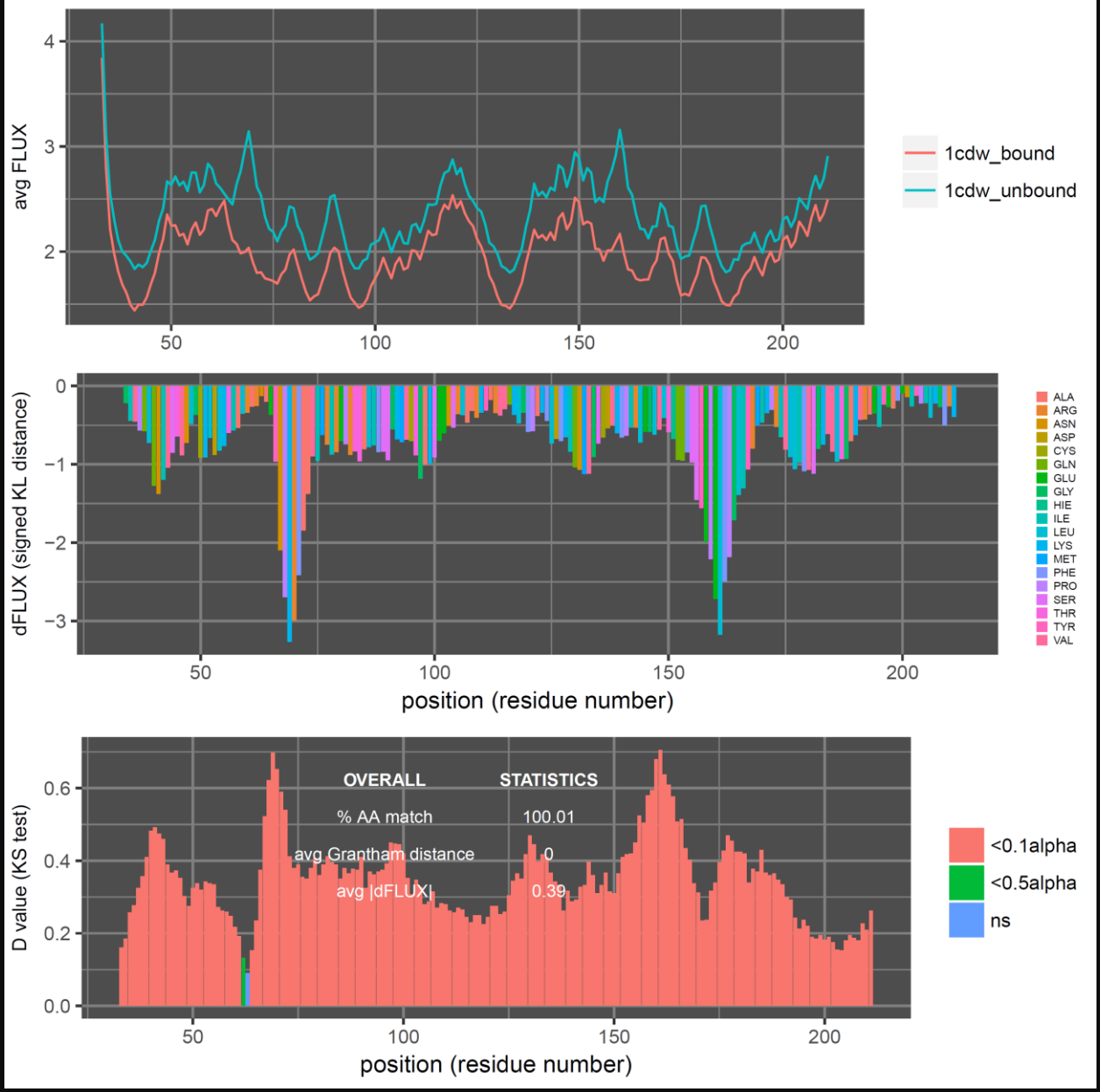
(D) Analysis of learning performance, conserved functional dynamics zones, and impact of genetic variants on dynamics



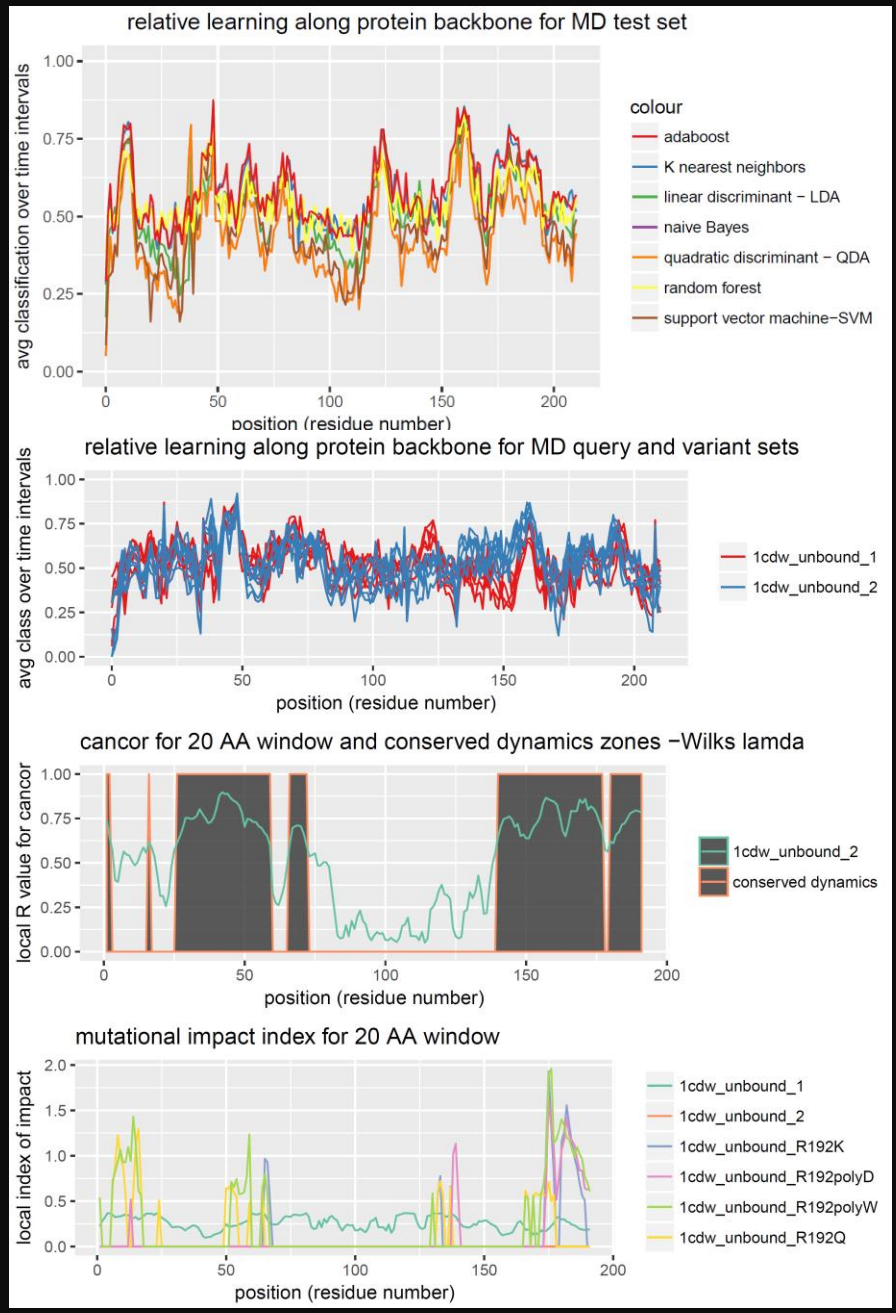


TATA binding protein

DROIDS 3.0 software



maxDemon 1.0 software



Local relative mach learning color by method

Local relative mach learning during two self similar MD runs

Conserved dynamics = local canonical corr (CC) in mach learning during self similar runs

Mutant impact = relative entropy betw CC-variant and CC-self

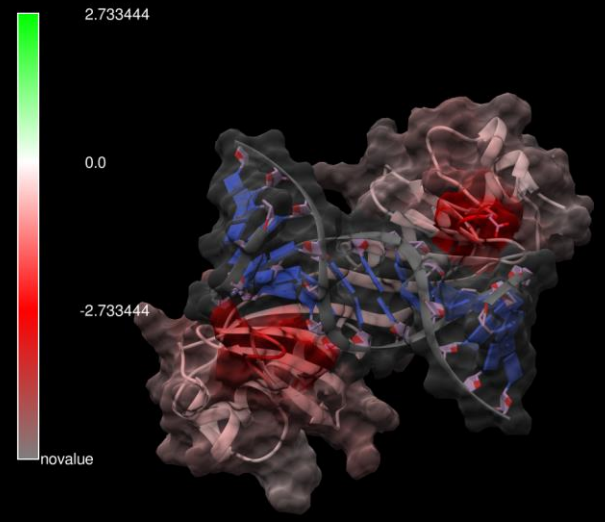
KEY CONCEPTUAL POINTS

- We expect that functionally conserved dynamics to be sequence encoded and therefore should display a repeated sequence-dependent signature in our machine learning profiles when novel MD runs are set up identically to MD upon which learners were trained. A significant local canonical correlation (Wilk's lambda) between 2 self-similar MD runs can be used to detect this.

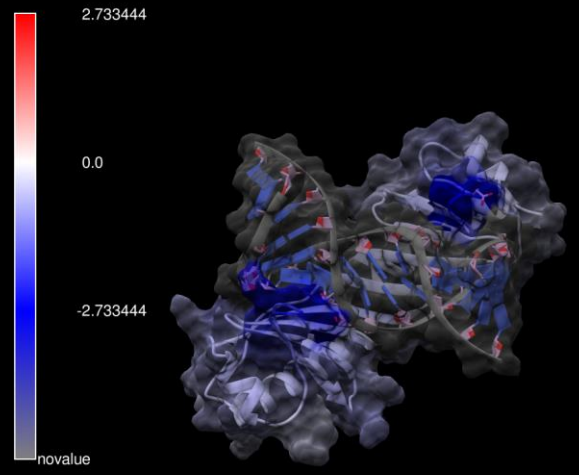
- Mutational impacts of genetic or drug class variants can be quantified by their effects on functional dynamics (can corr = CC) beyond that observed between self-similar runs. Thus when variant CC differs significantly from self CC (bootstrap or z-test), we plot the relative entropy i.e. impact = self CC*log(variant CC / self CC)

DROIDS 3.0 software

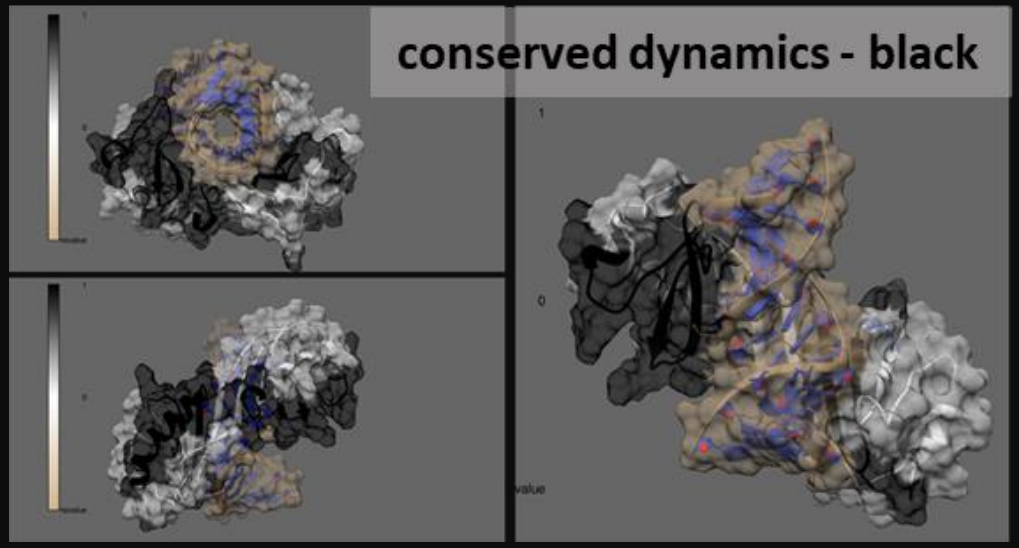
dFLUX = ‘stoplight’ color scheme

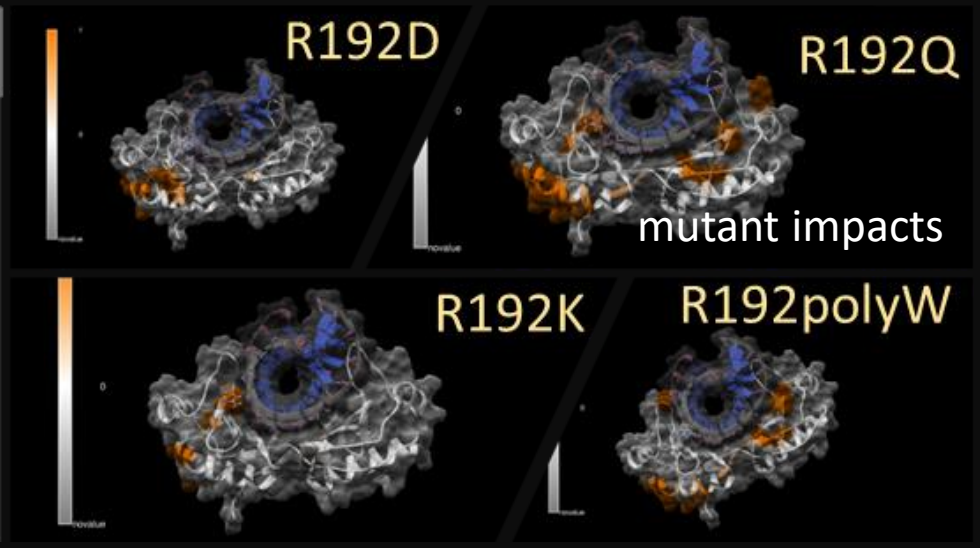


‘temperature’ color scheme

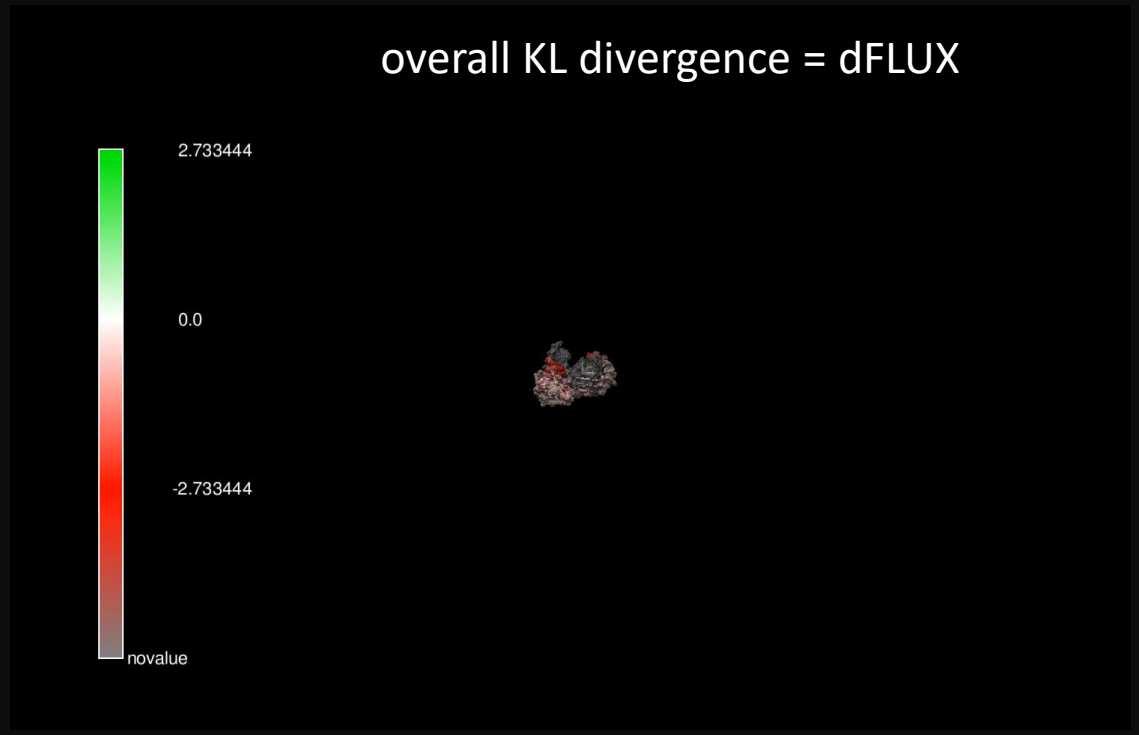


maxDemon 1.0 software

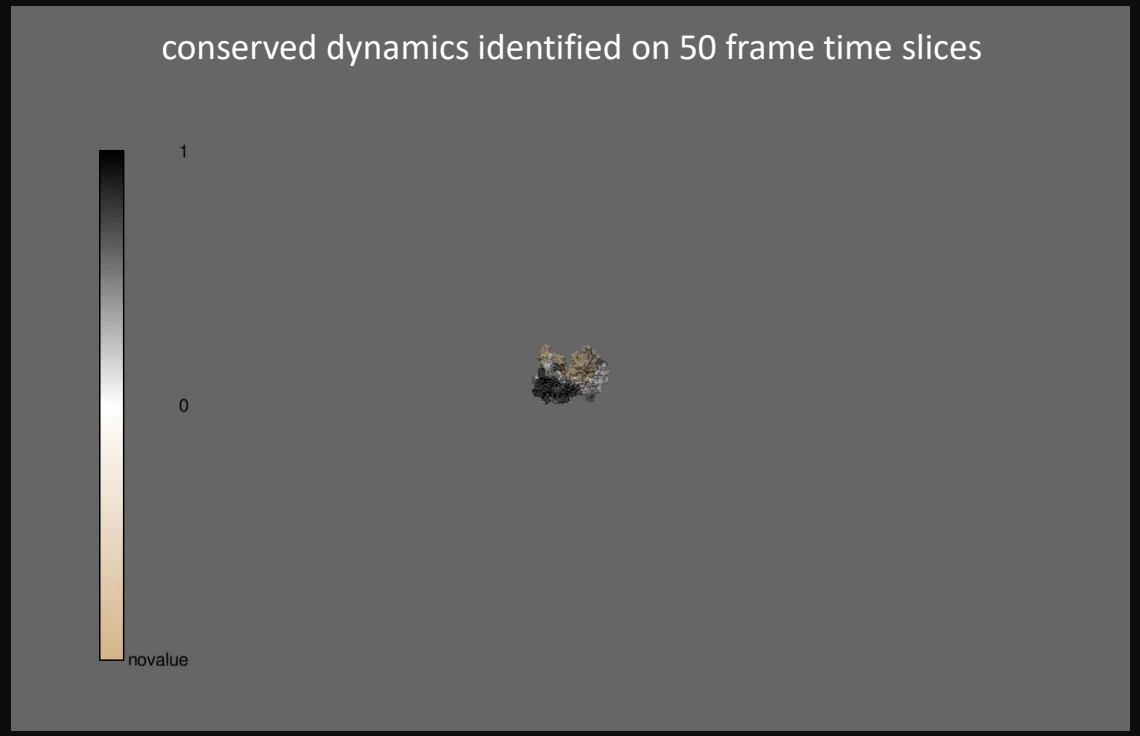




DROIDS 3.0 software



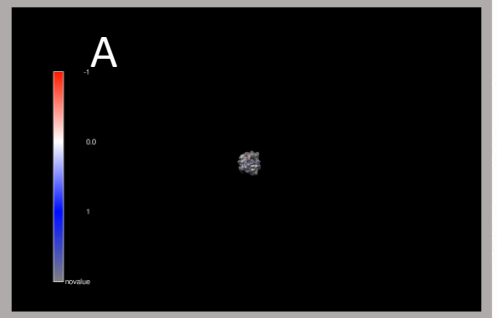
maxDemon 1.0 software

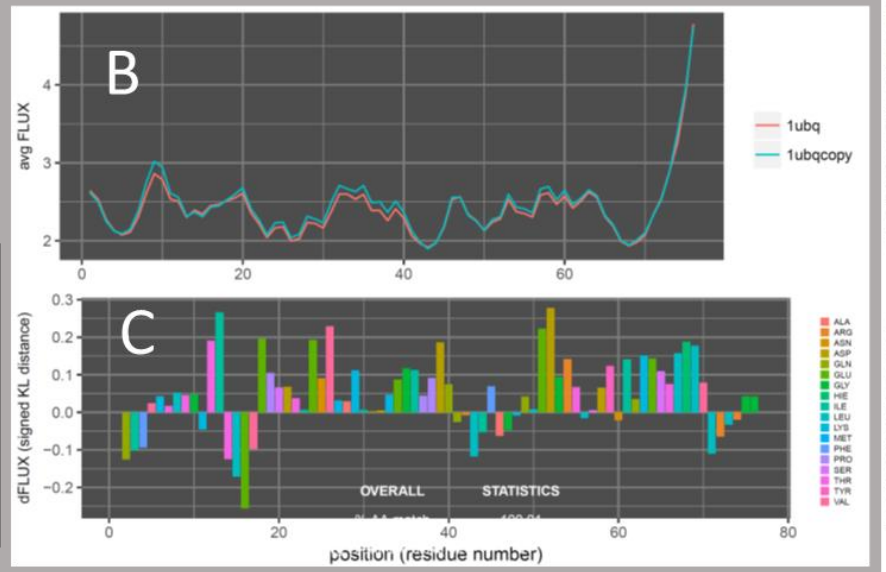


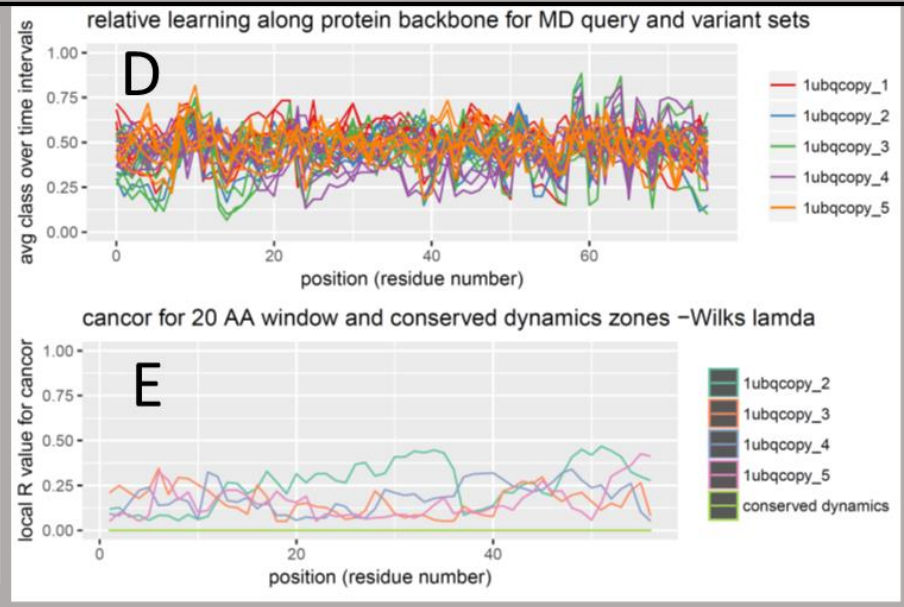
Proof of Concept - Case Studies

Environmental temperature shift in ubiquitin

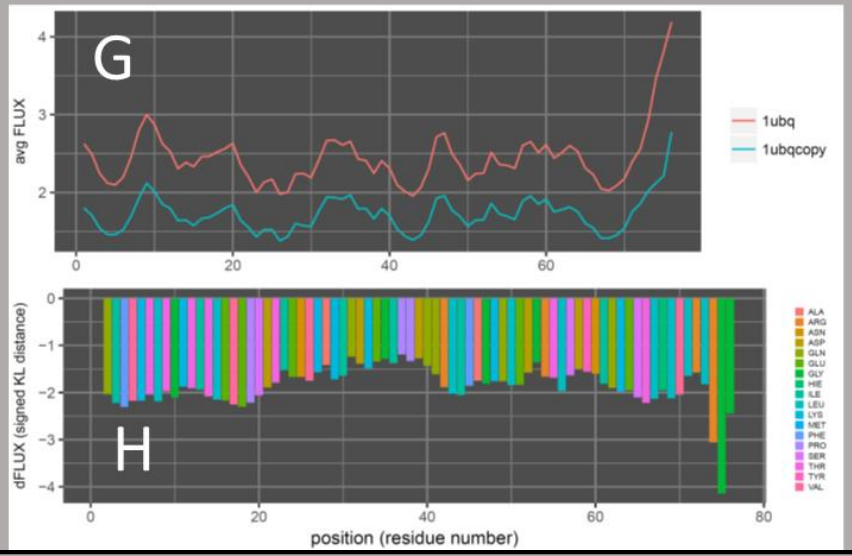
same temperatures
(both runs at 300K)

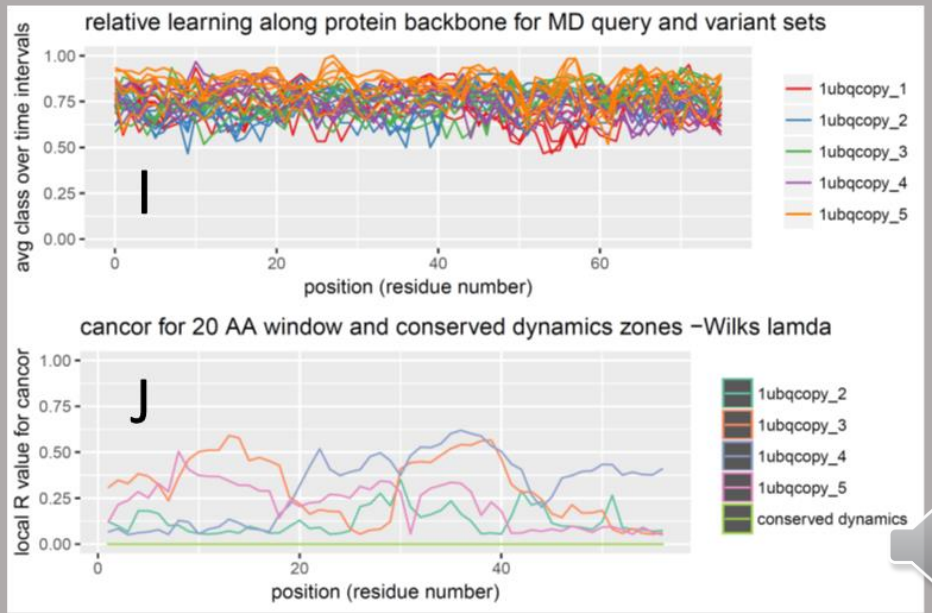




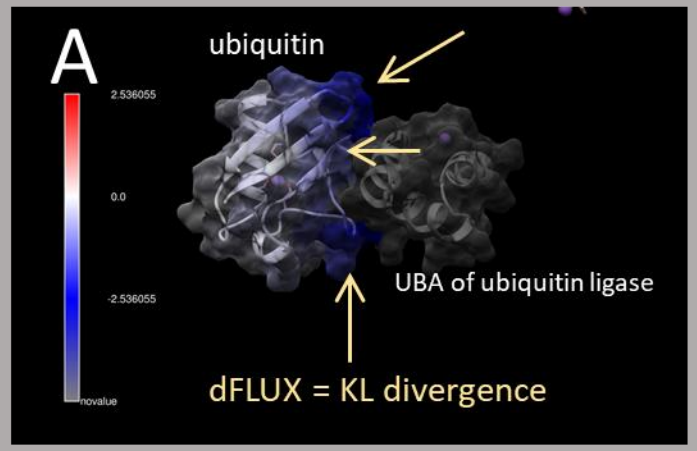


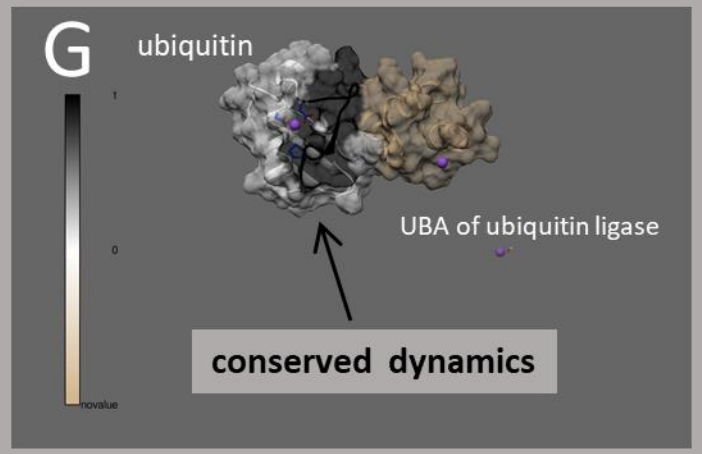
different temperatures
(runs at 300K and 250K)

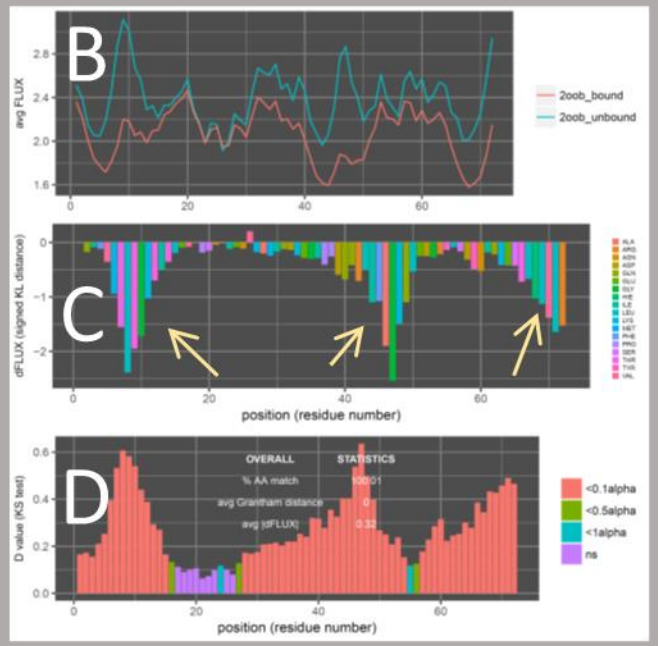


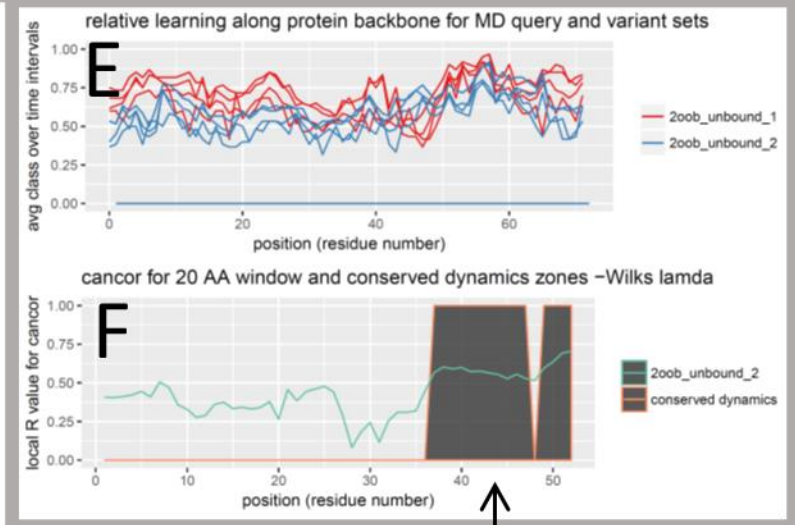


Genetic mutants in ubiquitin

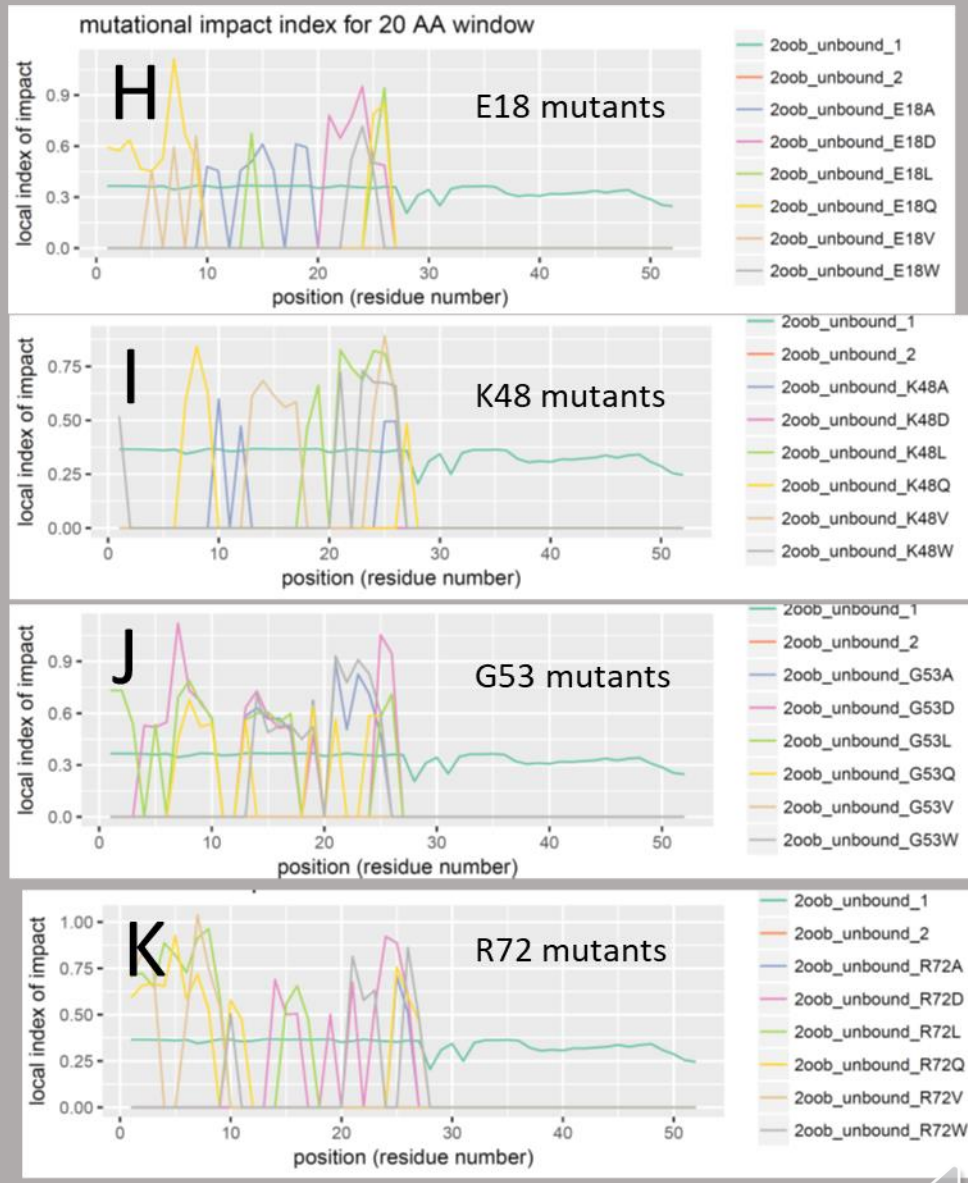




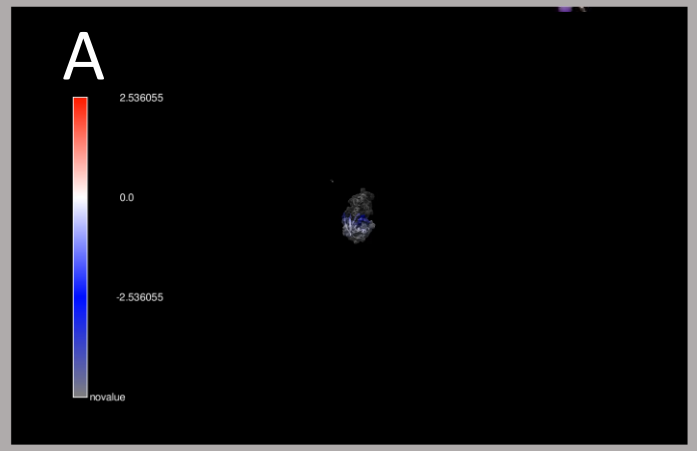


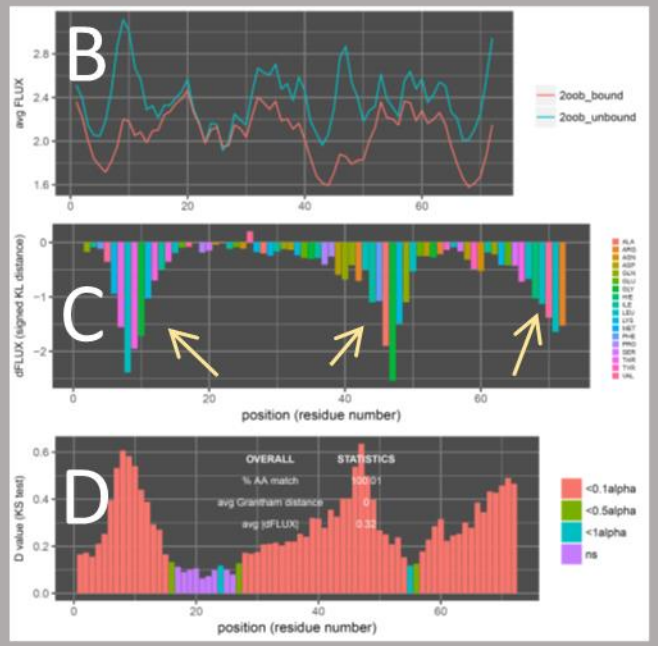


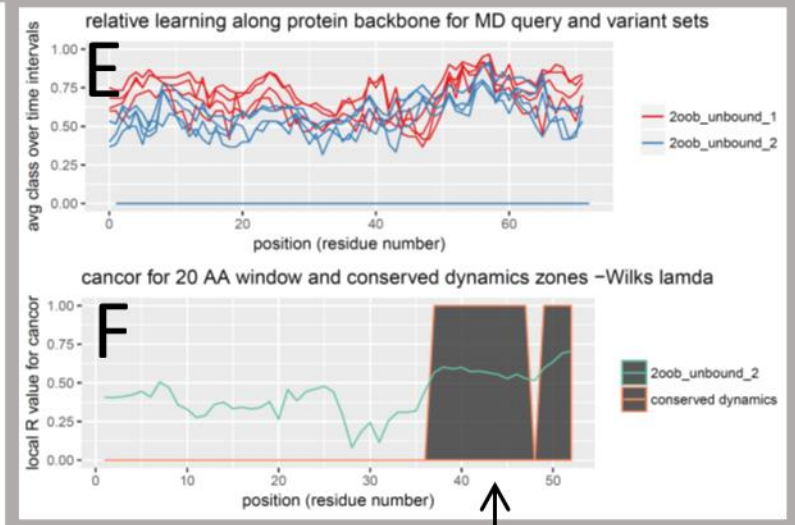
conserved dynamics



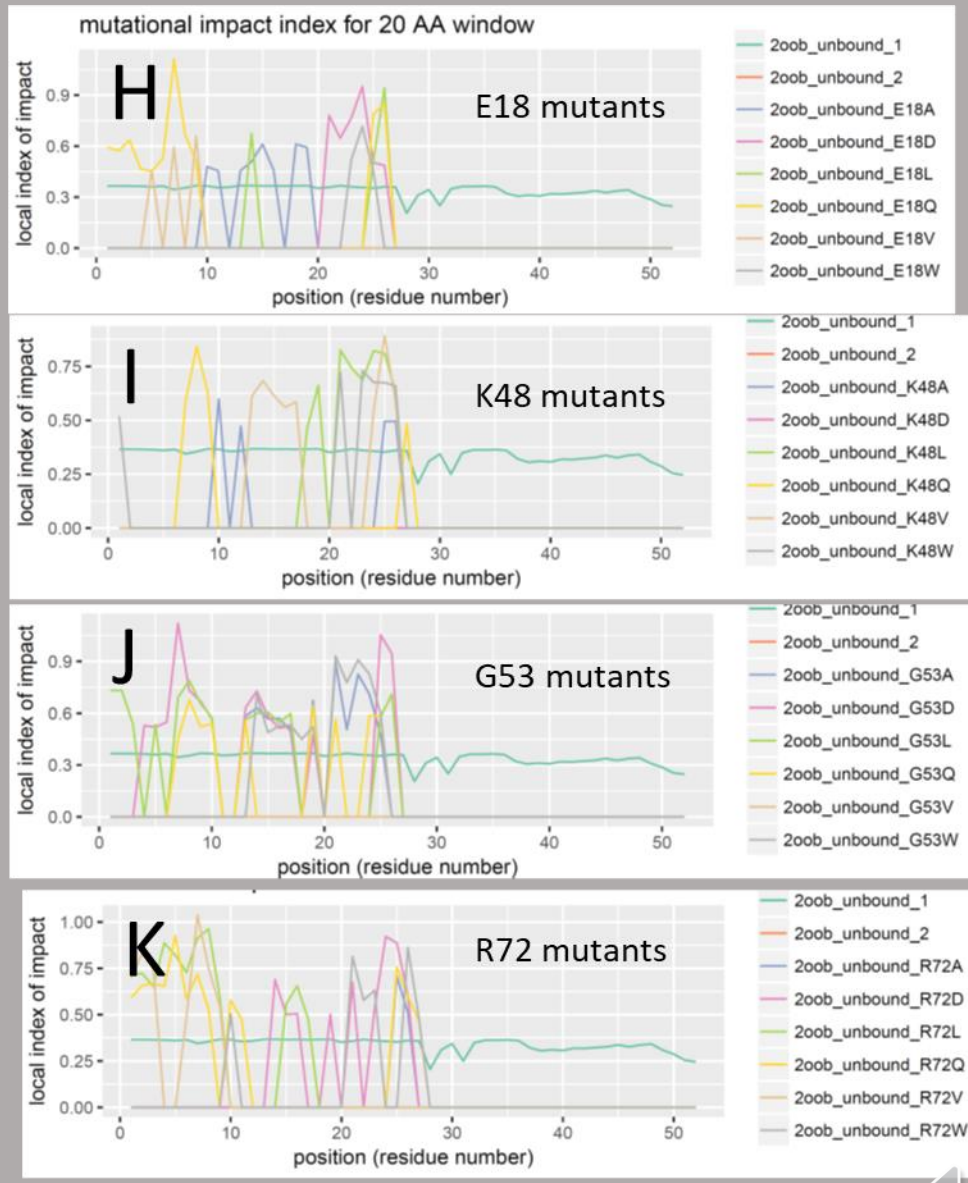
Genetic mutants in ubiquitin



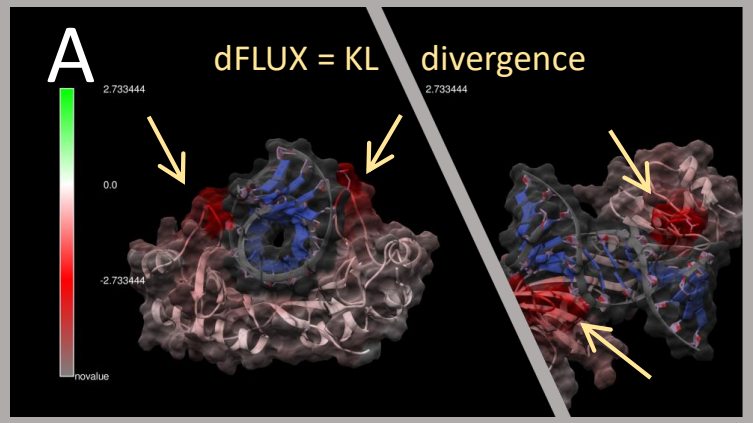


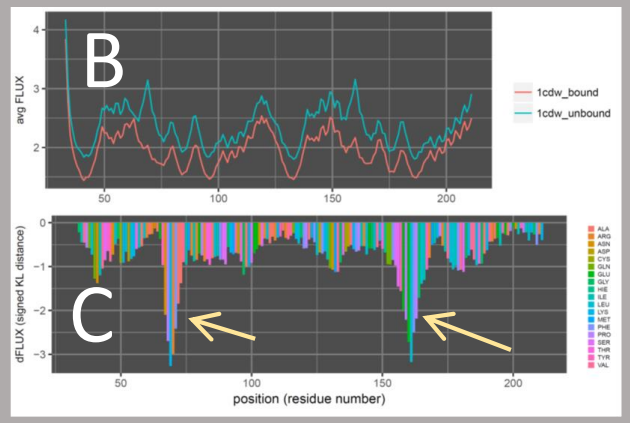


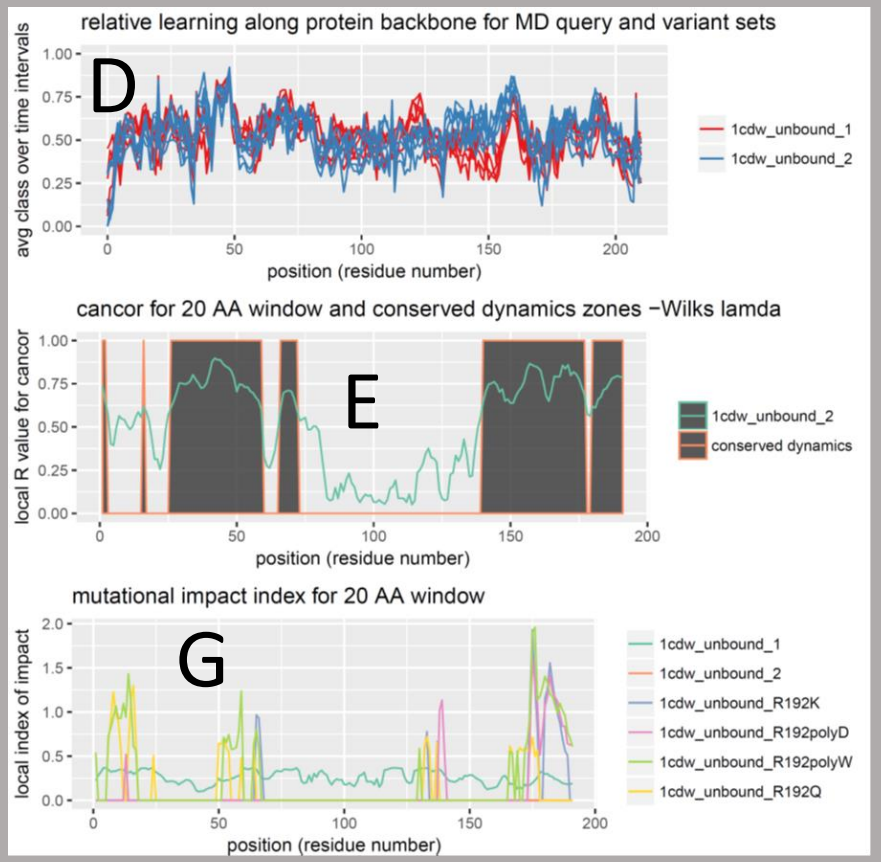
conserved dynamics

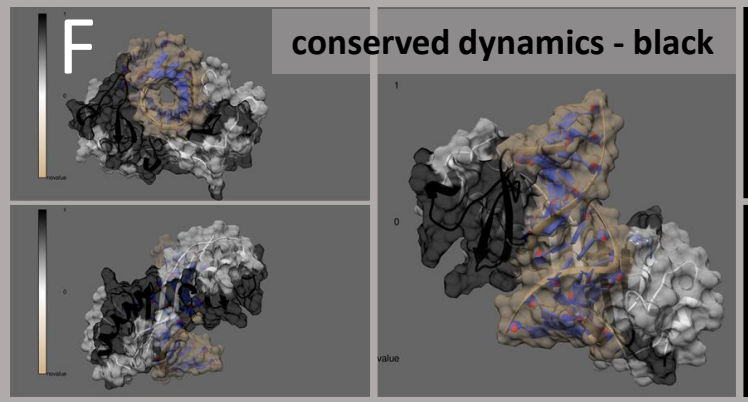


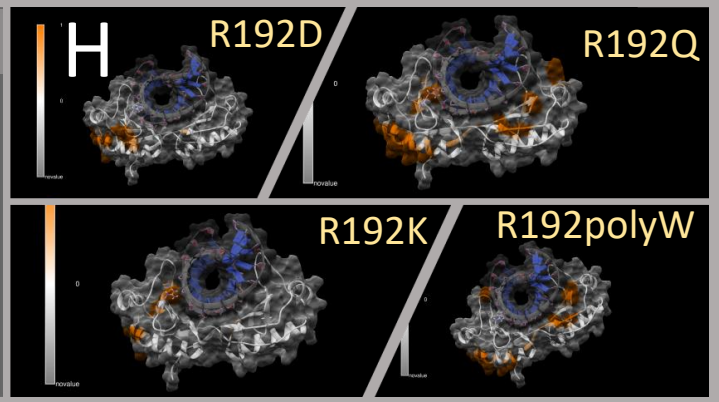
Genetic mutants in TBP



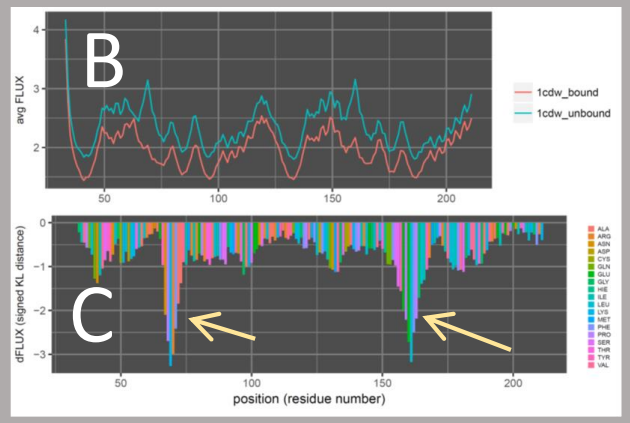


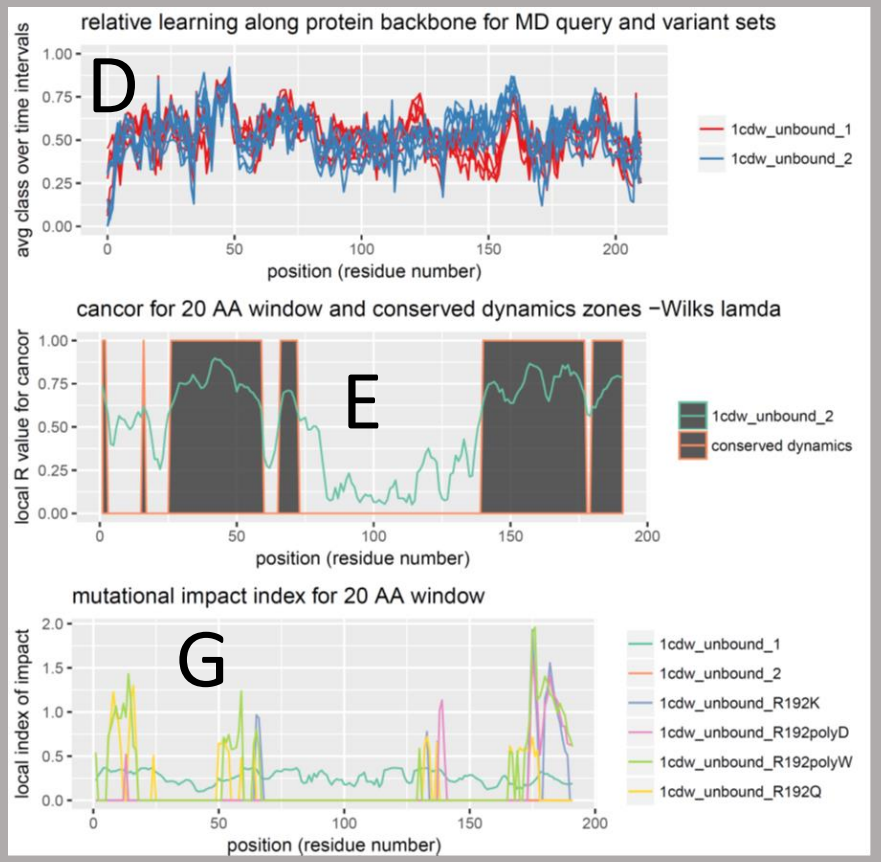


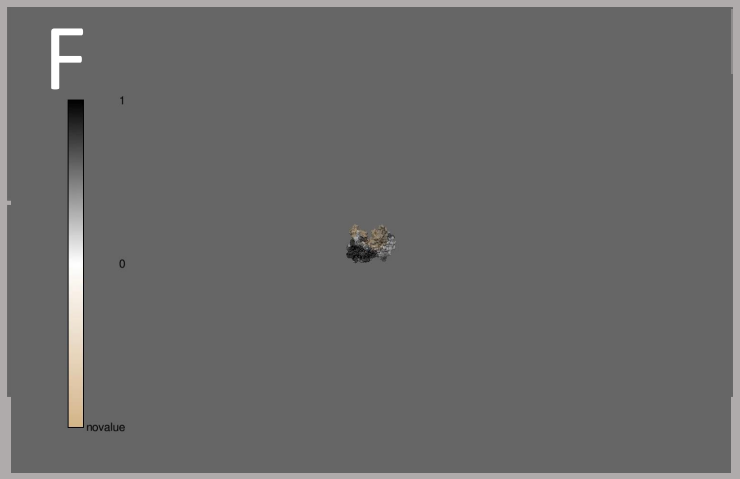


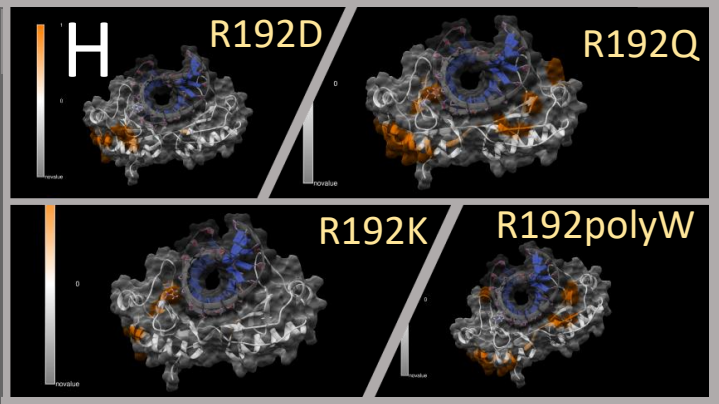


Genetic mutants in TBP

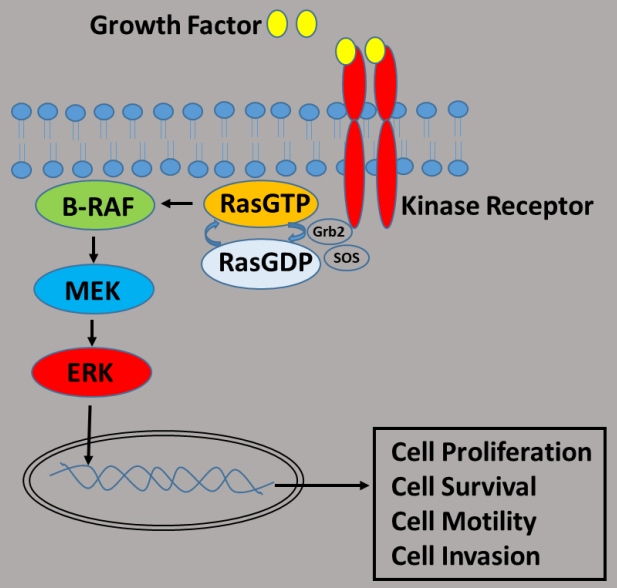


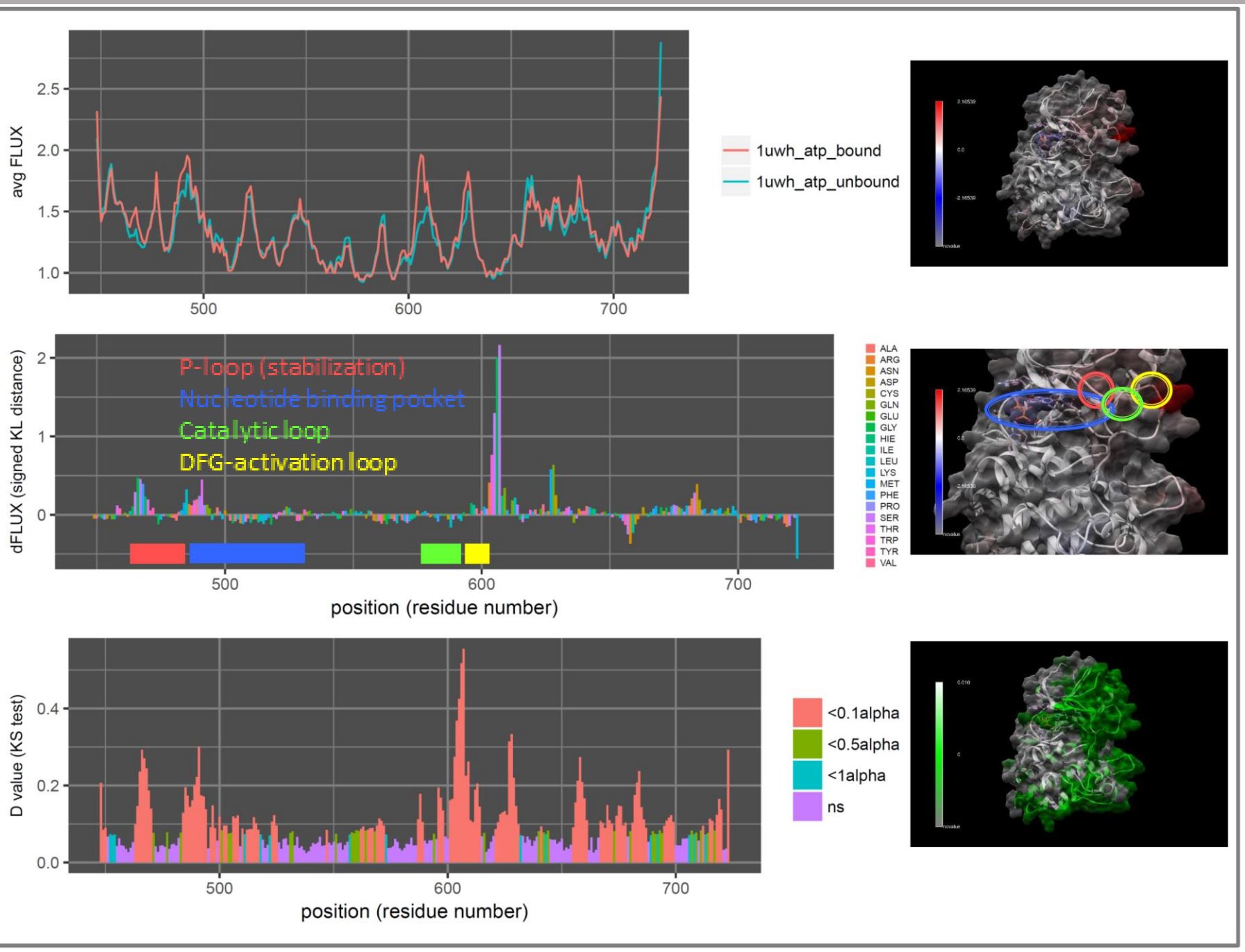






Normal ATP Activation of BRAF

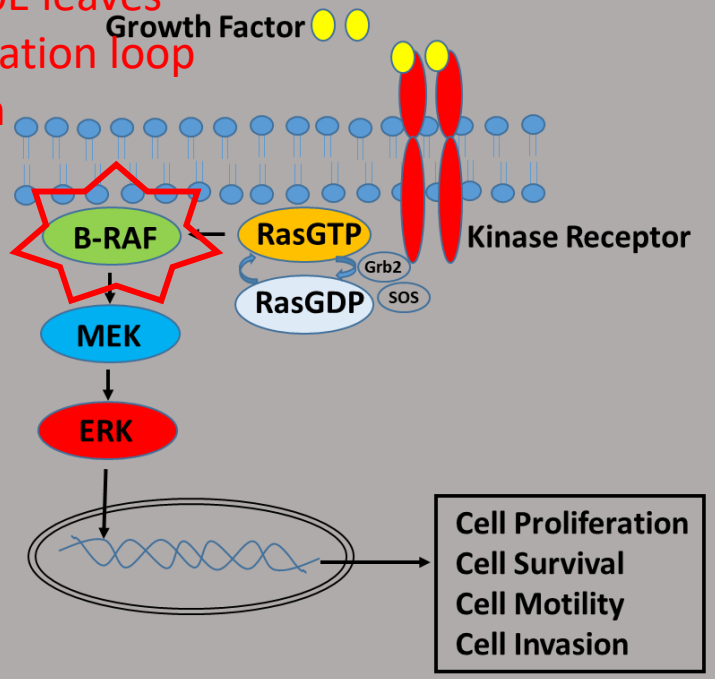


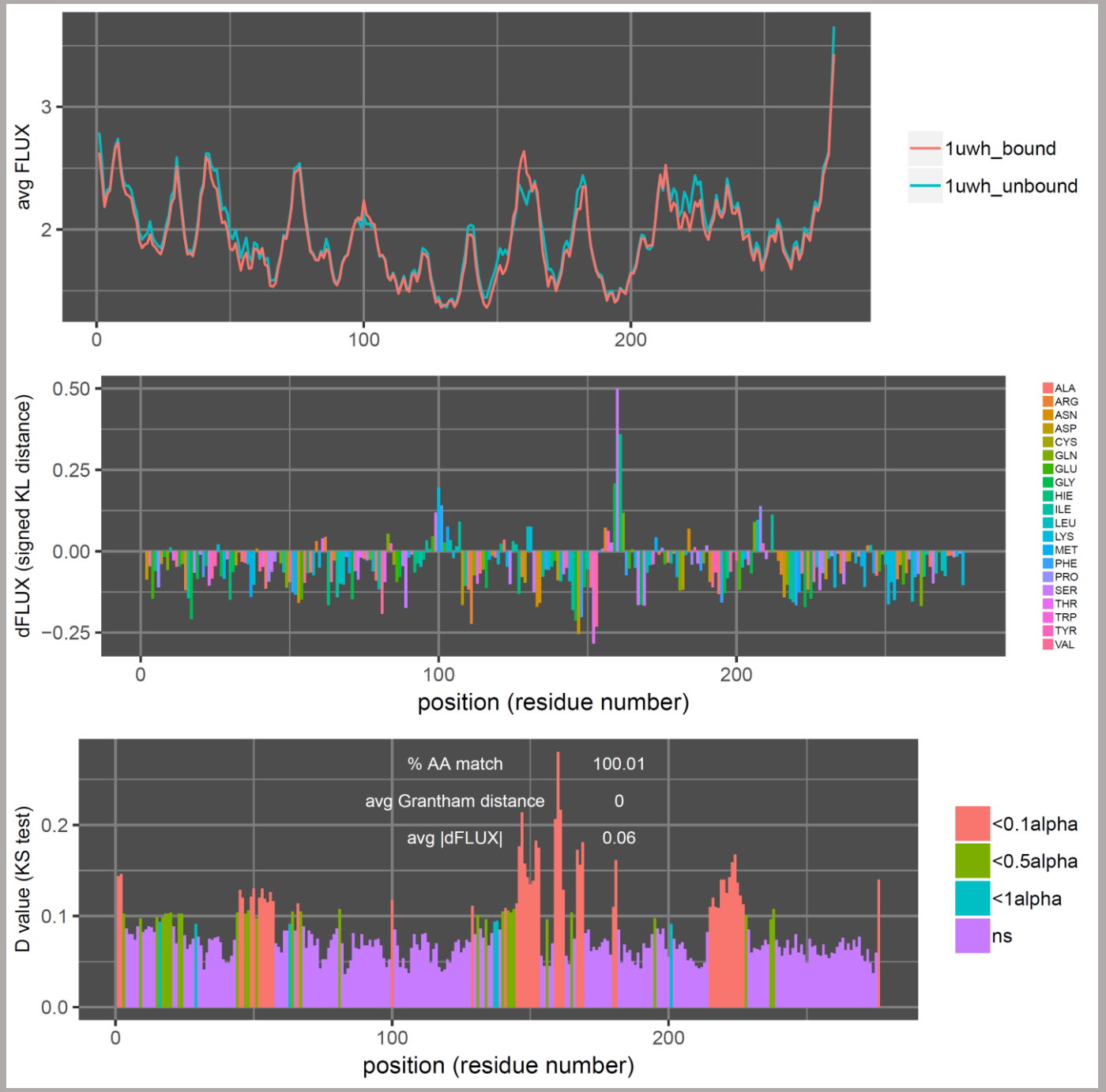


Raf inhibitor Sorafenib targeting V600E mutant BRAF

Note: sorafenib hyperactivates melanoma cancer in wild-type cells in absence of 2nd mutation

V600E leaves activation loop open

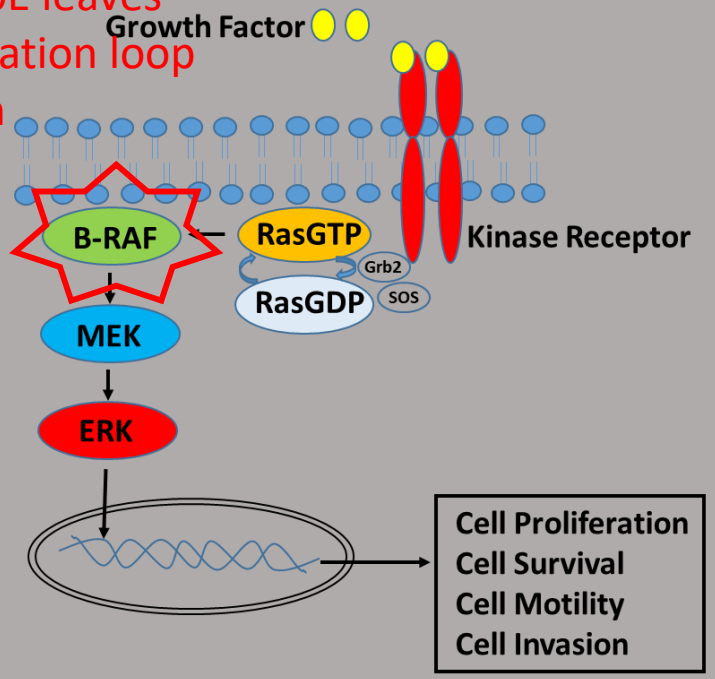


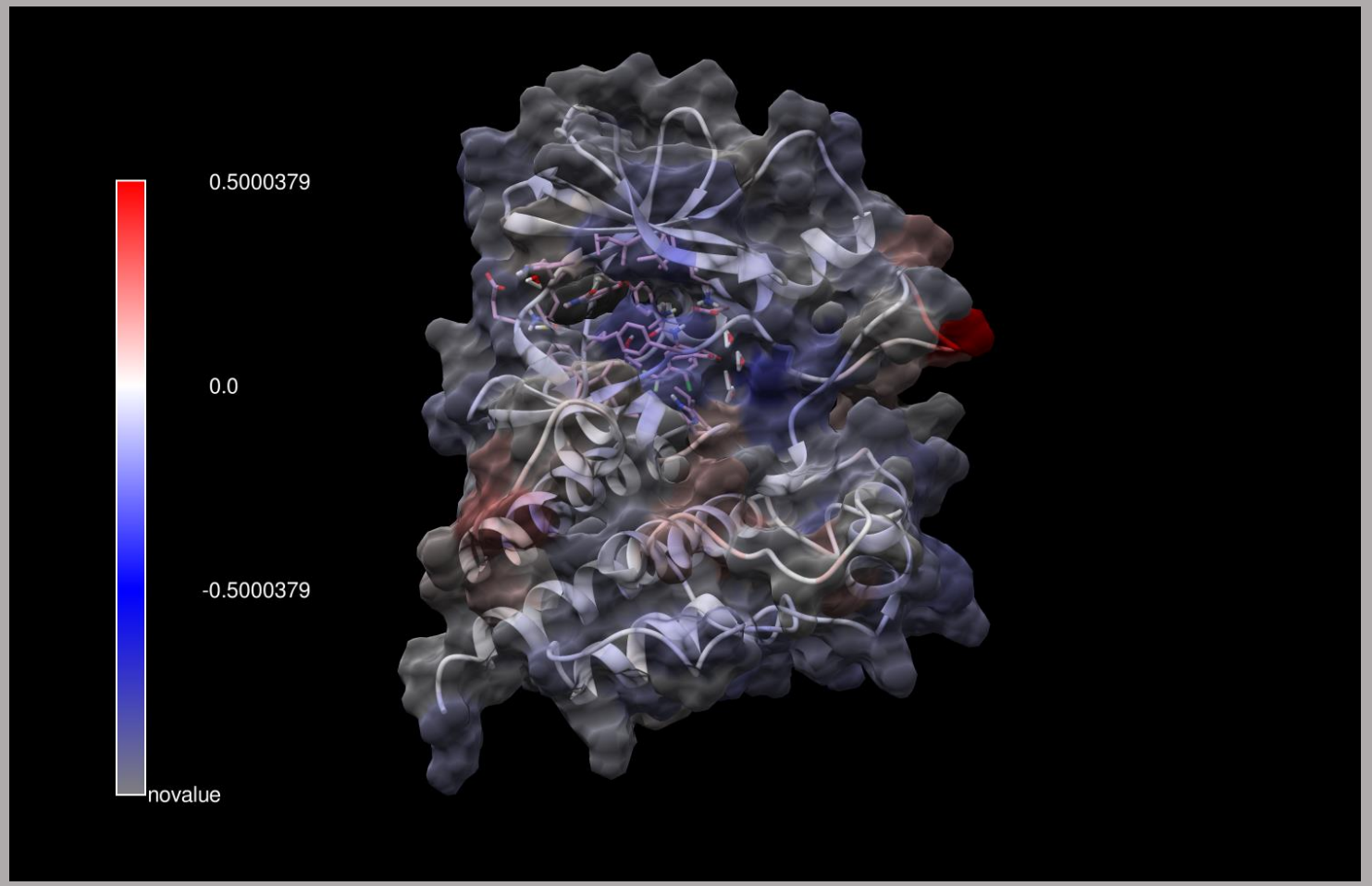


Raf inhibitor Sorafenib targeting V600E mutant BRAF

Note: sorafenib hyperactivates melanoma cancer in wild-type cells in absence of 2nd mutation

V600E leaves activation loop open

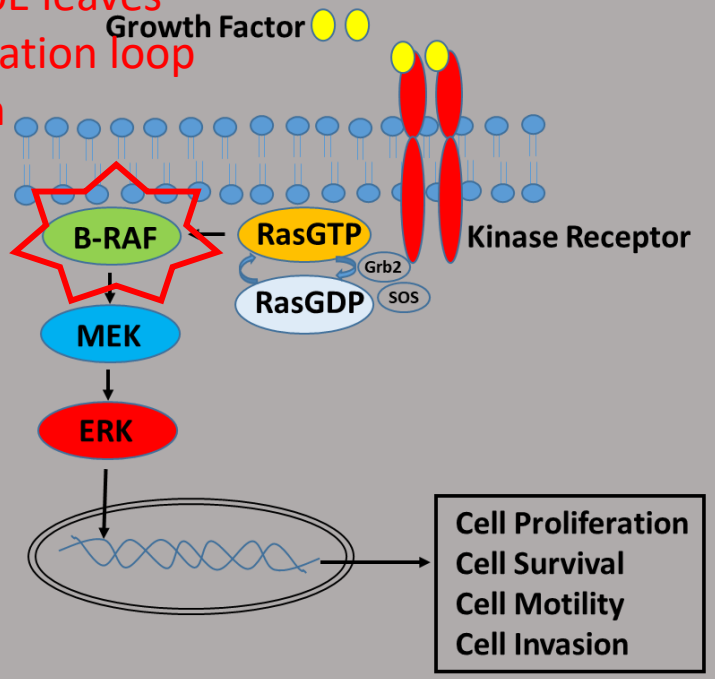




Raf inhibitor Sorafenib targeting V600E mutant BRAF

Note: sorafenib hyperactivates melanoma cancer in wild-type cells in absence of 2nd mutation

V600E leaves activation loop open



Conclusions

1. We present a series of statistical methods with integrated software pipeline for the application of high-performance computing and gaming graphics to comparative protein dynamic studies.

2. We demonstrate its potential utility for combining drug class variant screening and personalized genomic medicine

3. We emphasize the potential importance of understanding dynamics in addition to static data (i.e. sequence and structure)

some of our great students

Erin Coppola BME

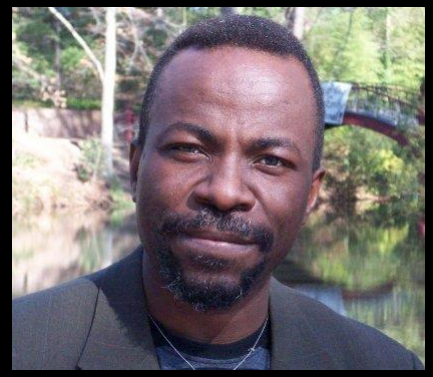
Jamie Mortensen BME

Justin Liao BME

Dr. Gregory A. Babbitt

BABBITTLAB@RIT

COMPUTATIONAL MOLECULAR EVOLUTION AND ECOLOGY



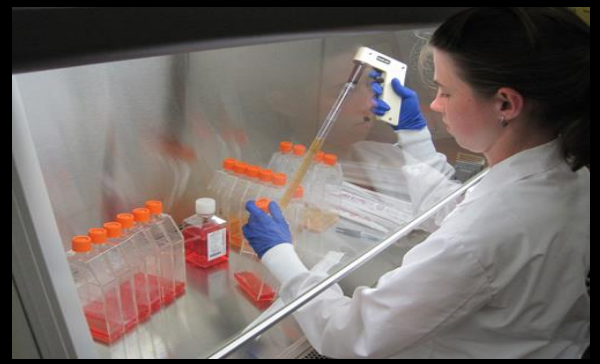
Dr. Ernest P. Fokoue

Mohammed Alawad
Bioinformatics

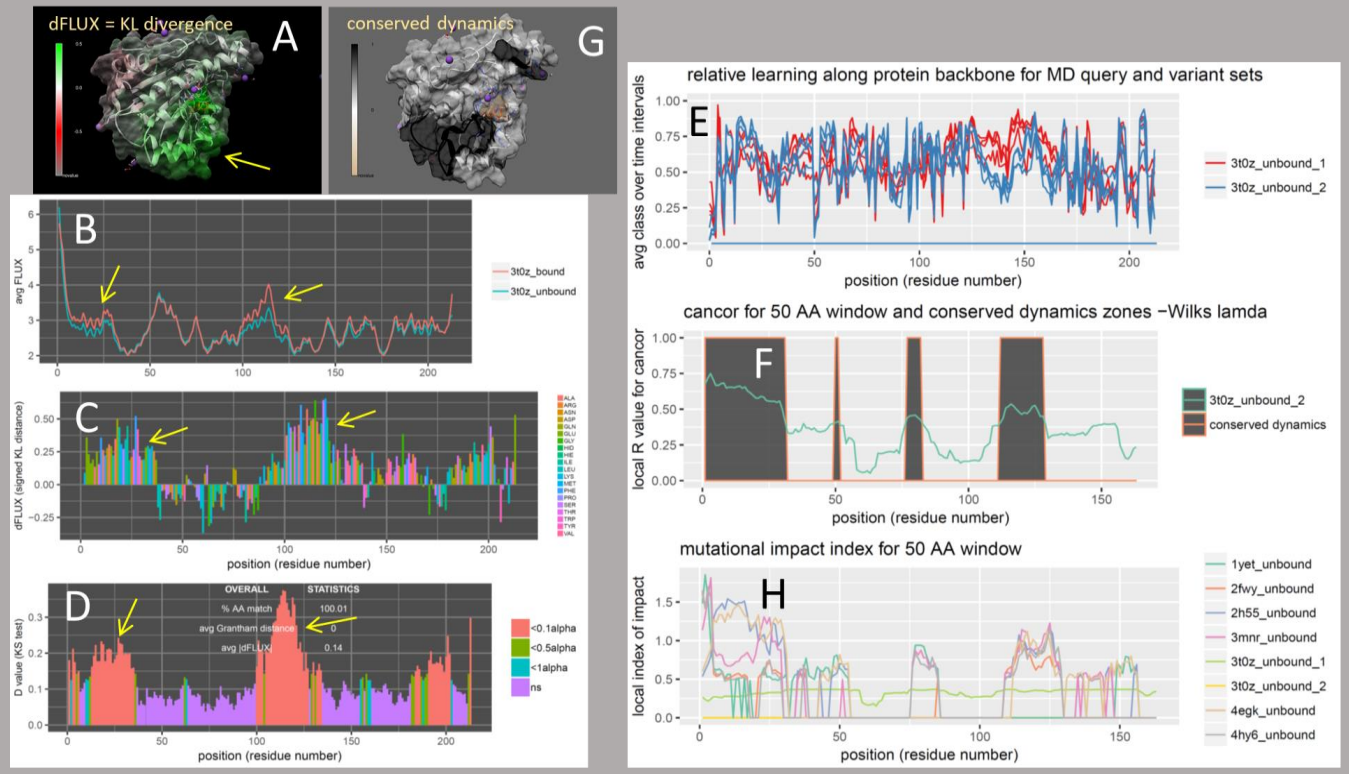
Katharina Schulze
Bioinformatics

sponsored by

NVIDIA

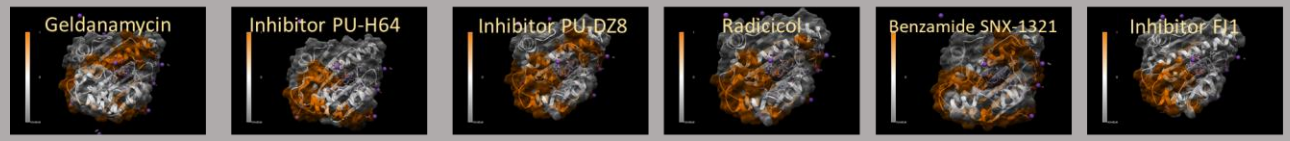


Drug class variants targeting Hsp90



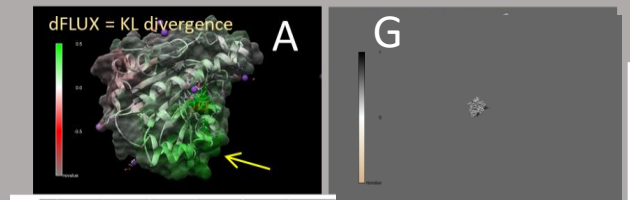
I Ligand contacts of drug binding variants on protein dynamics of Hsp90 in the ATP binding pocket

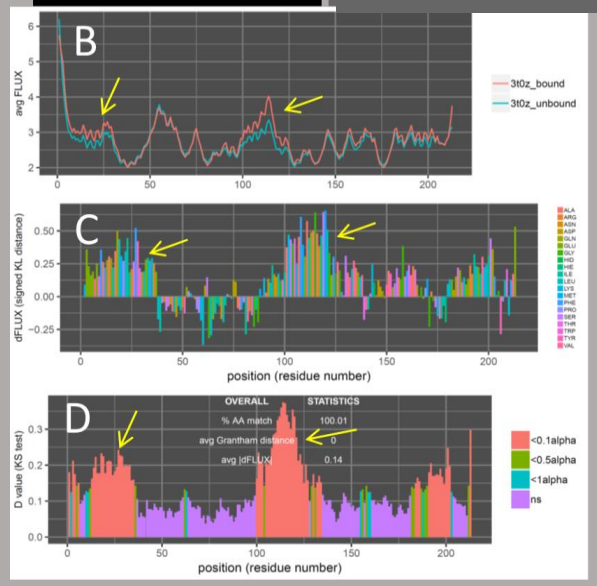
ligand	PDB	% sim	Val 186	Thr 184	Trp 162	Tyr 139	Phe 138	Lys 112	Alal 111	Leu 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 93	Lys 58	Ala 55	Asp 54	Asn 51
ATP	3t0z	100		x			x				x		x	x		x		x		x
Geldanamycin	1yet	87.5		x			x	x			x			x		x		x		x
Inhibitor PU-H64	2fwy	75		x	x		x			x			x			x		x		x
Inhibitor PU-DZ8	2h55	75		x	x		x			x			x		x	x	x	x		x
Radicicol	4egk	62.5	x	x			x			x						x	x	x	x	x
Benzamide SNX-1321	3mnr	37.5		x	x	x				x		x			x	x	x	x		
Inhibitor FJ1	4hy6	0			x	x			x											

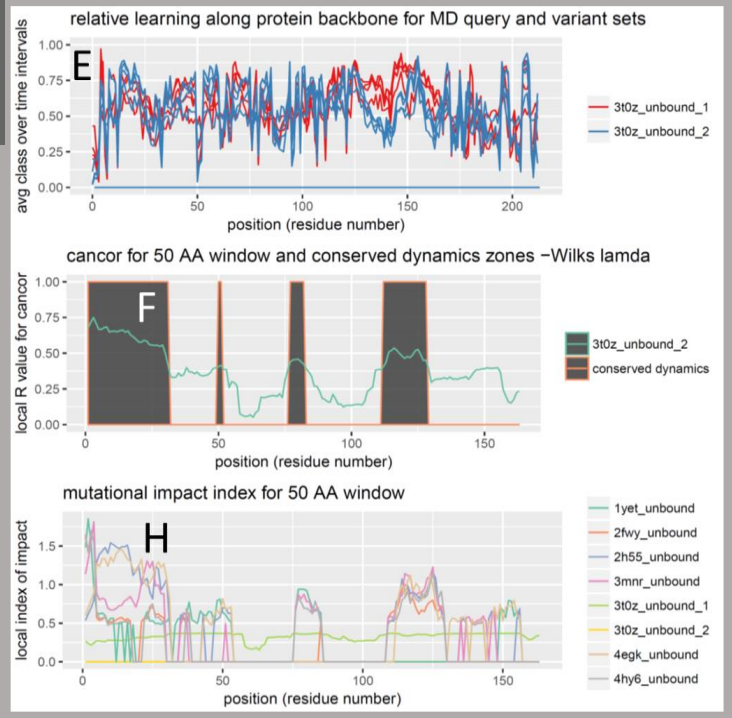


J Impacts of drug binding variants on protein dynamics of Hsp90 that are significantly different than binding to ATP

Drug class variants targeting Hsp90

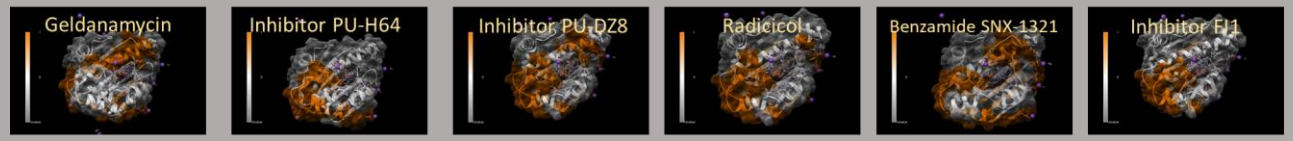






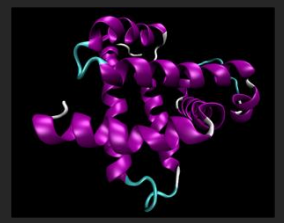
I Ligand contacts of drug binding variants on protein dynamics of Hsp90 in the ATP binding pocket

ligand	PDB	% sim	Val 186	Thr 184	Trp 162	Tyr 139	Phe 138	Lys 112	Alal 111	Leu 107	Asn 106	Asp 102	Met 98	Gly 97	Ile 96	Asp 93	Lys 58	Ala 55	Asp 54	Asn 51
ATP	3t0z	100		x			x				x		x	x		x		x		x
Geldanamycin	1yet	87.5		x			x	x			x			x		x		x		x
Inhibitor PU-H64	2fwy	75		x	x		x			x			x			x		x		x
Inhibitor PU-DZ8	2h55	75		x	x		x			x			x		x	x	x	x		x
Radicicol	4egk	62.5	x	x			x			x						x	x	x	x	x
Benzamide SNX-1321	3mnr	37.5		x	x	x				x		x			x	x	x	x		
Inhibitor FJ1	4hy6	0			x	x			x											



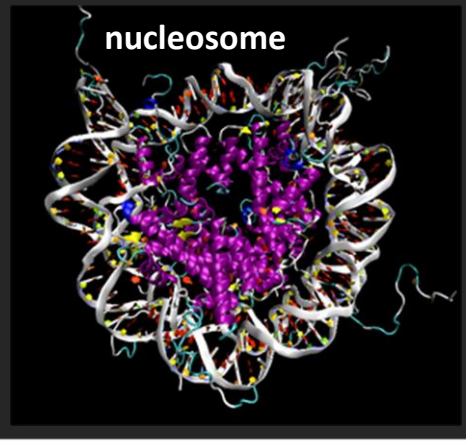
J Impacts of drug binding variants on protein dynamics of Hsp90 that are significantly different than binding to ATP

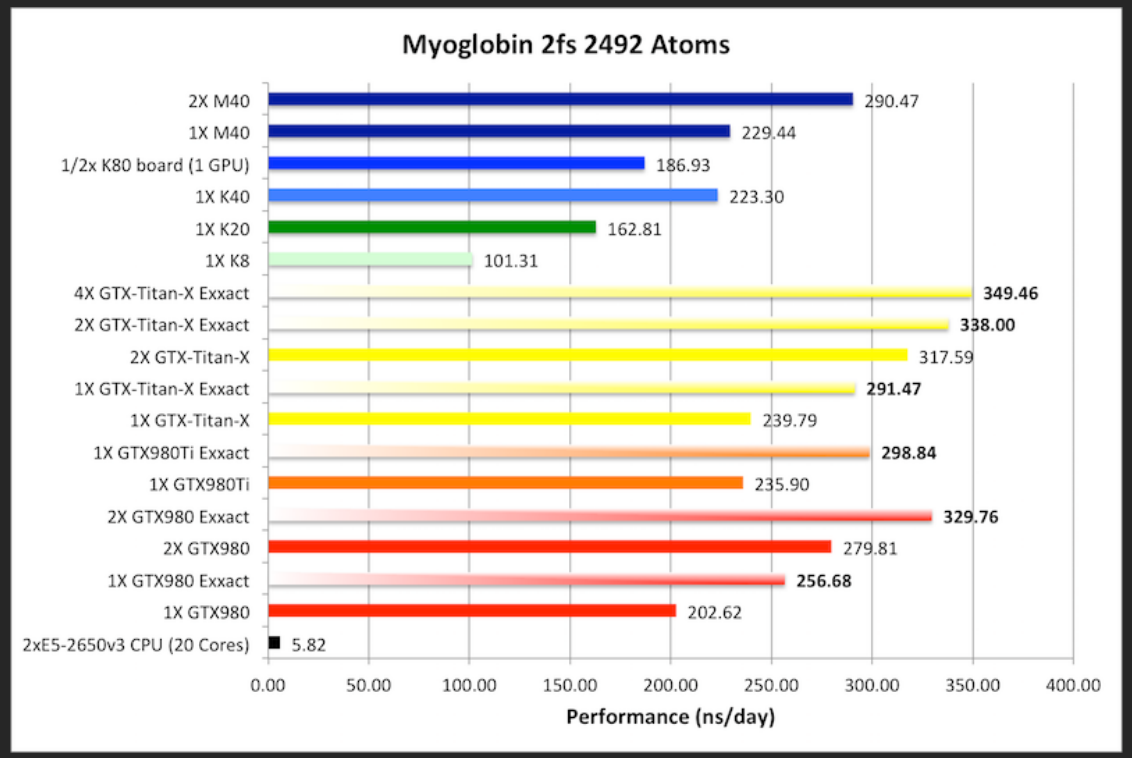
myoglobin

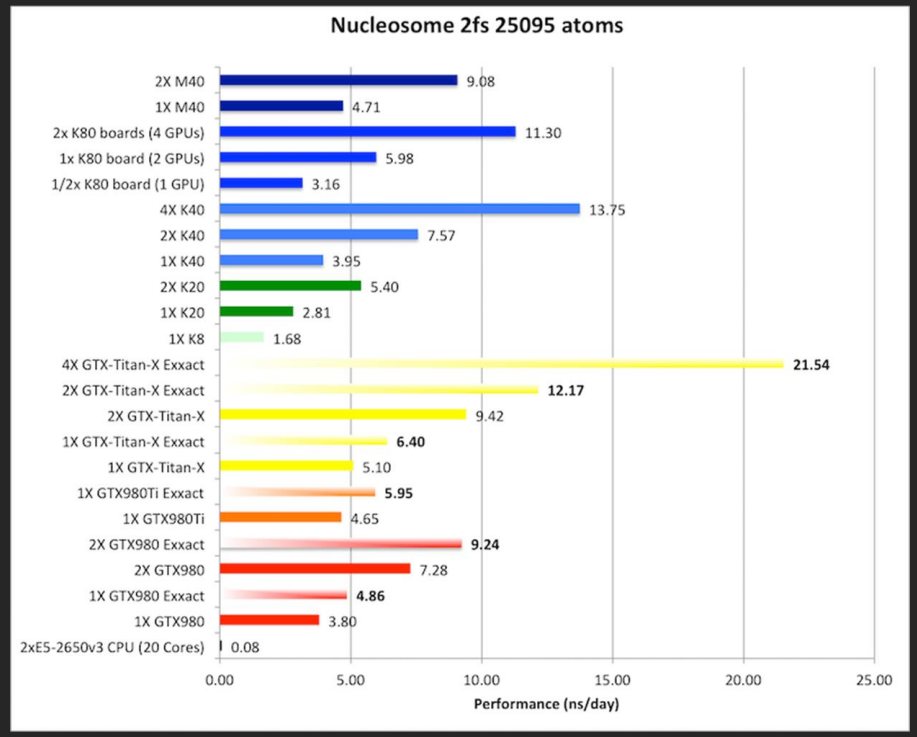


COMPUTATIONAL BIOCHEMISTRY

= molecular dynamic (MD) simulation
= 'live action' molecular biology
= upper system limit 3-5 million atoms
= numerical approximation time step of 2 femtosecs







http://ambermd.org/gpus/benchmarks.htm

The world is a dynamic mess of jiggling things

Richard P Feynman

Richard Feynman