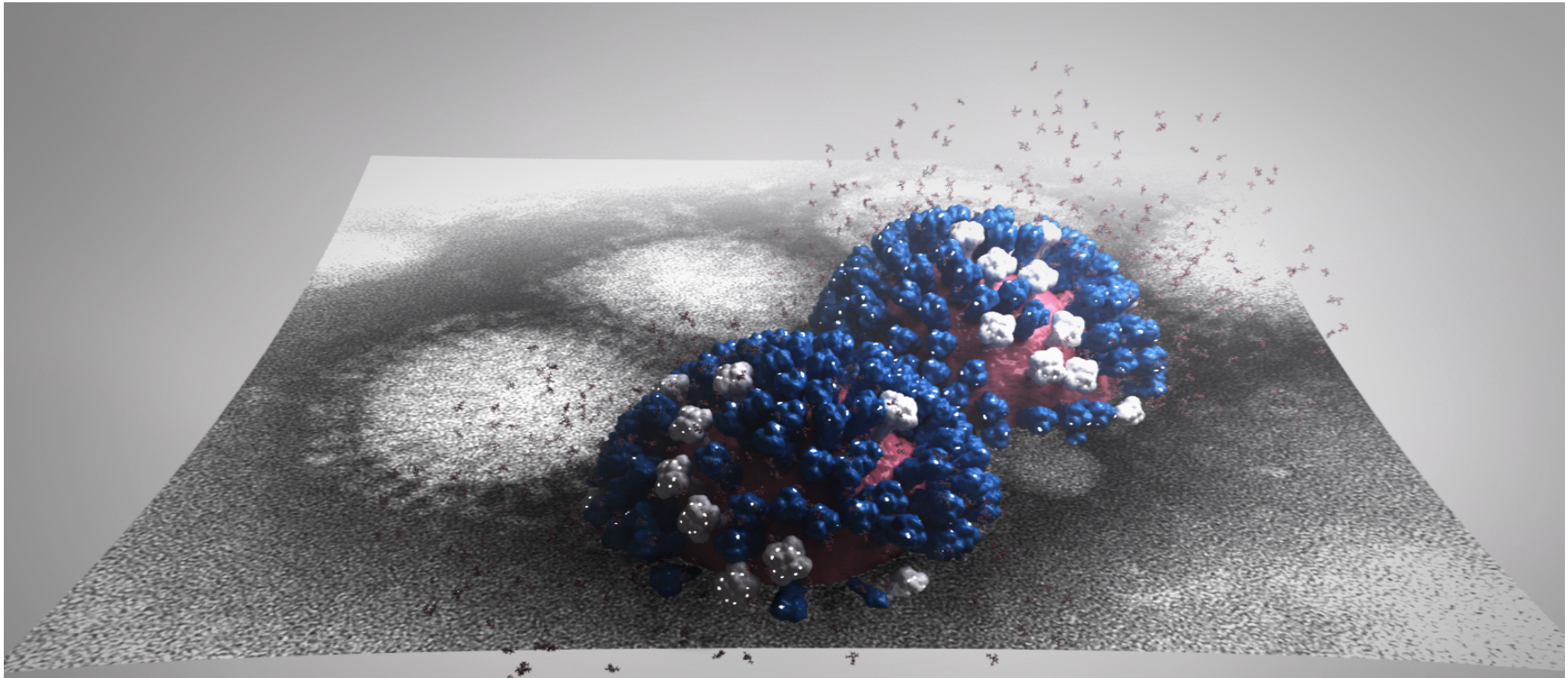
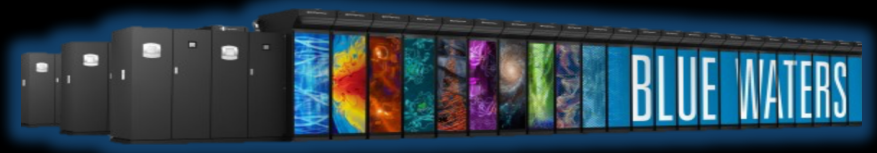


The Evolving Demands of Computational Biophysics in the Petascale Computing Era

Rommie E. Amaro . UC San Diego . OAC Webinar . Oct 2018

Convergence of HPC, data science, & data enabling transformative advances at the intersection of observational and simulation sciences



360,000 cores + GPU acceleration

National Strategic Computing Initiative (NSCI)

Exascale

MolSSI

Compute Power

HP 735
12 CPUs

1993
protein
10k atoms
100s ps

SGI Origin
128 CPUs

1997
ion channel
100k atoms
1 ns

LeMieux
3k CPUs

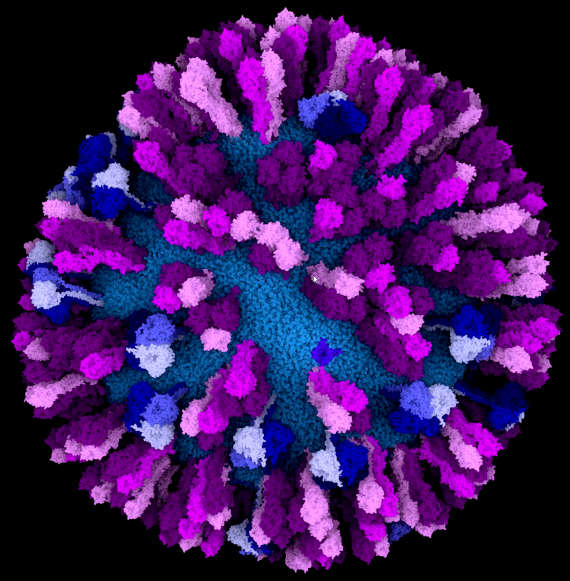
2002
ATPase
500k atoms
10s ns

Ranger
60k CPUs

2007
ribosome
2 mil atoms
100s ns

2015
Enveloped virus
160 mil+ atoms
1-100 μs

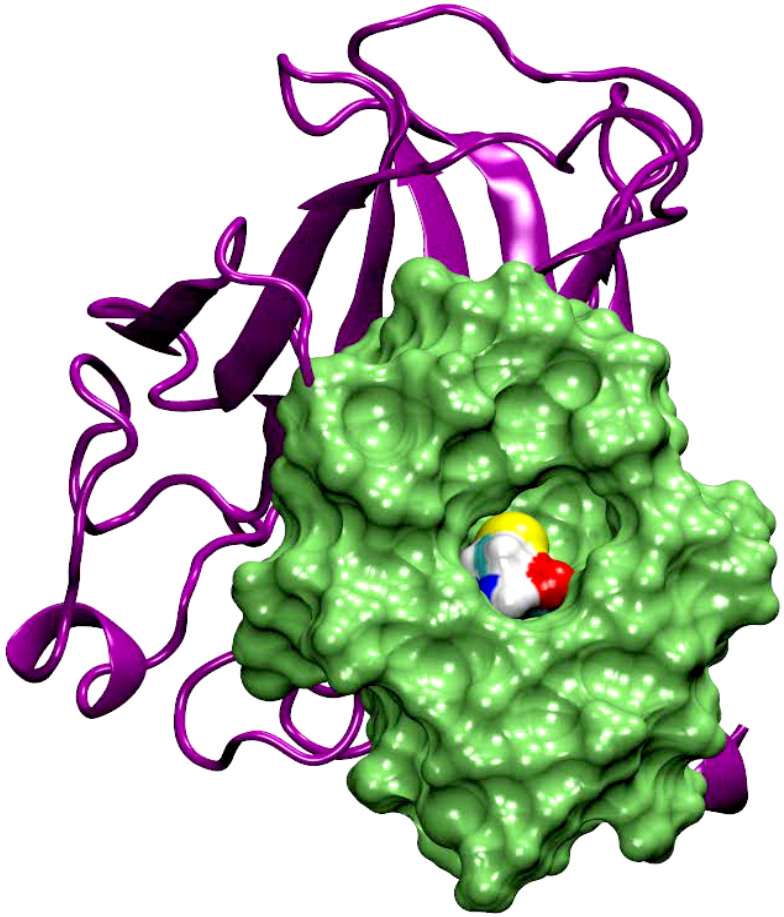
time



100 nm, 160 million atoms, μs


Influenza
Cancer
Chlamydia
Trypanosomiasis

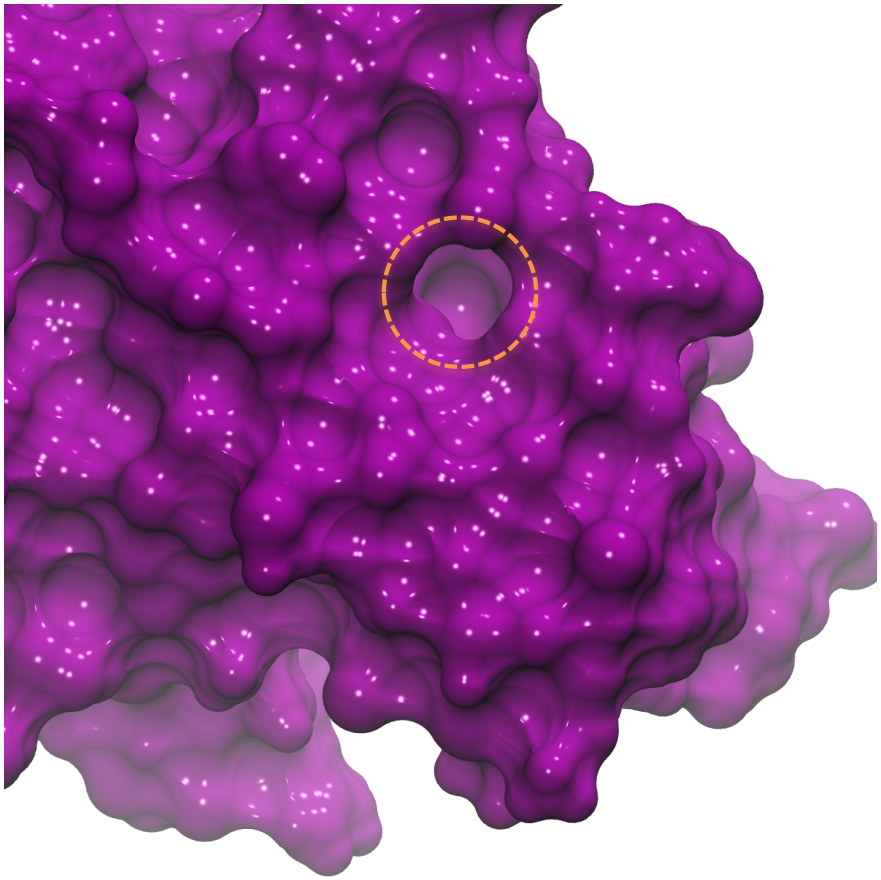
Simulations Reveal Target Flexibility



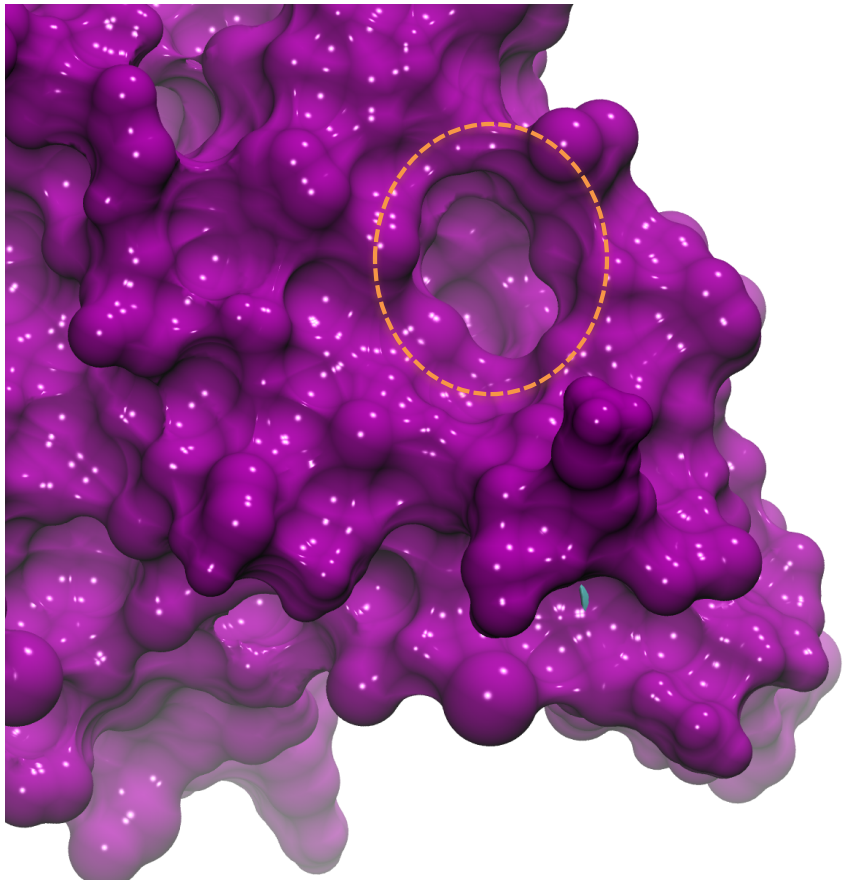
5% exposed,
matches NMR

Wassman, Baronio, Demir, et al. Nature Comm., (2013)

New Site Opens

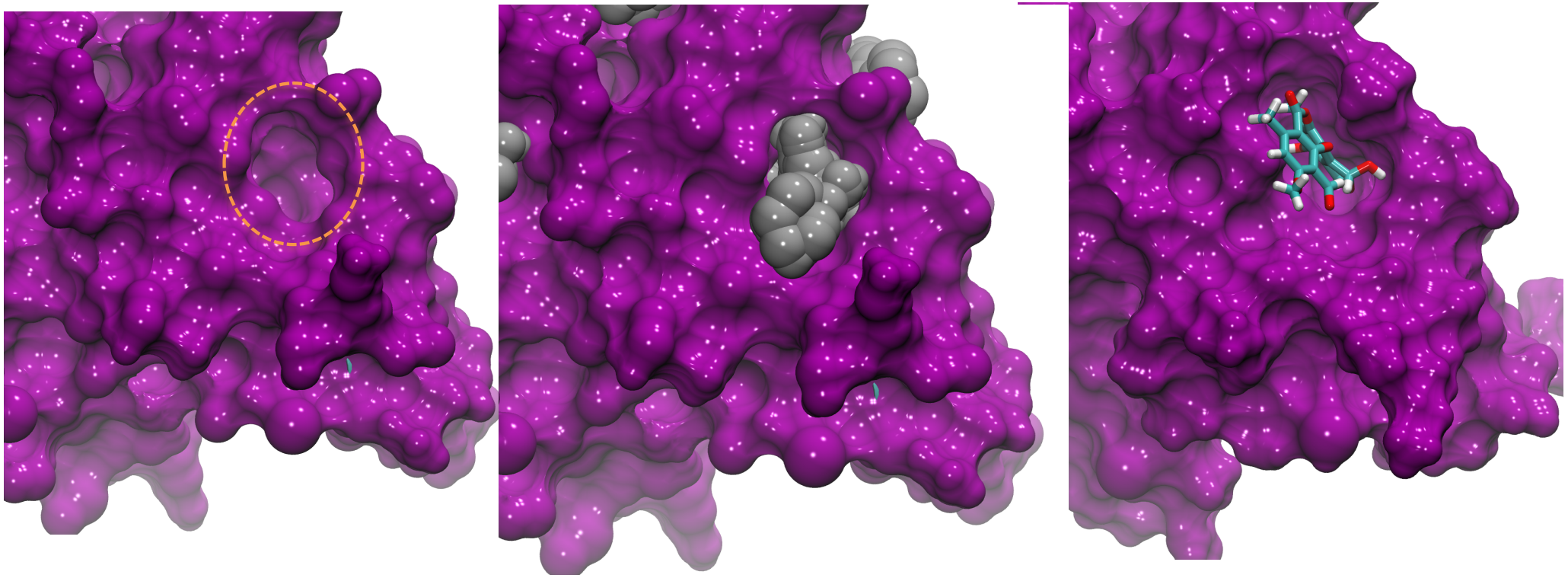


> 95 X-ray structures



“Open” MD structure

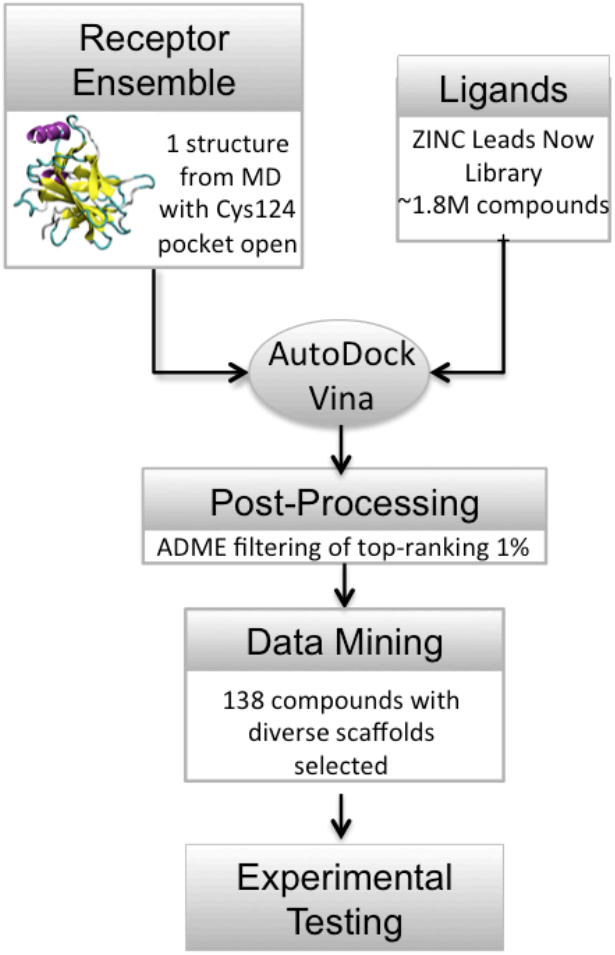
New Site is Druggable

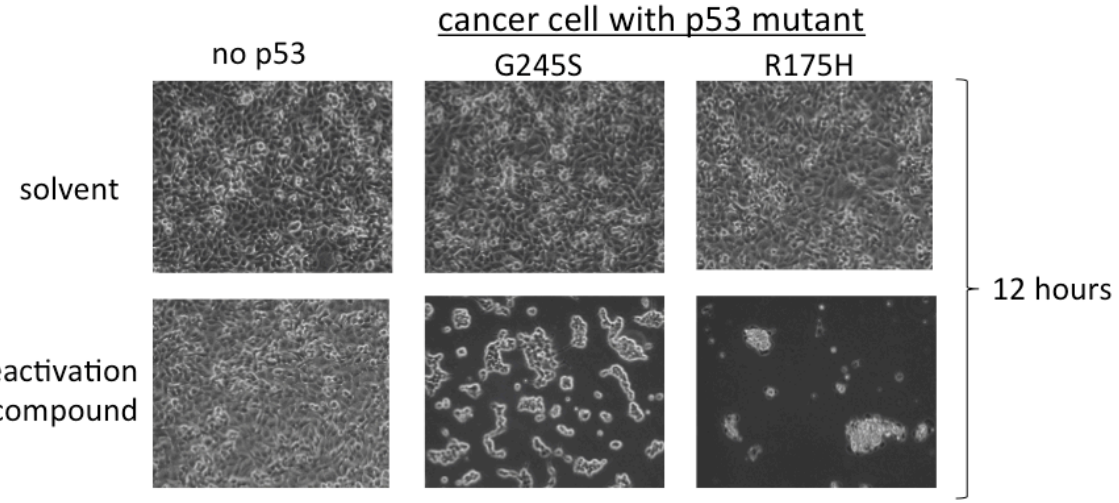


Wassman, Baronio, Demir, et al. Nature Comm., (2013)

Vajda et al., Computational Solvent Mapping: http://ftmap.bu.edu/

Our computational approach discovers more novel p53 reactivation compounds in 6 months than all the research efforts of the previous 20 years combined





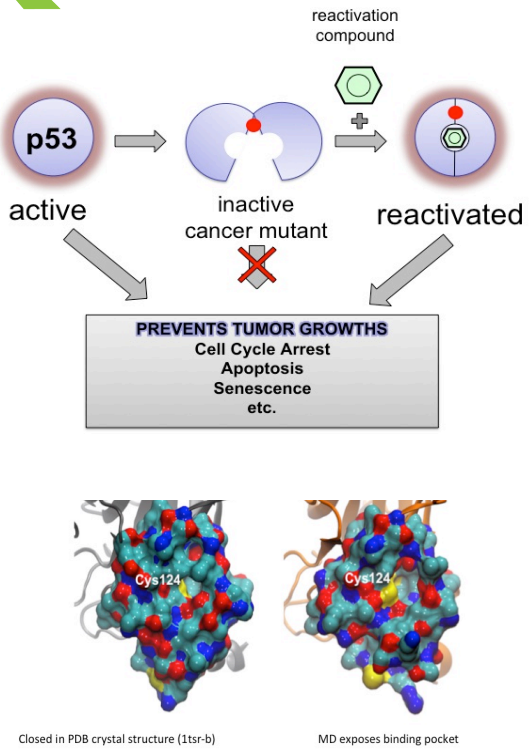
15/138 compounds tested in mammalian cancer cell lines rescue p53 activity and kill cancer cell

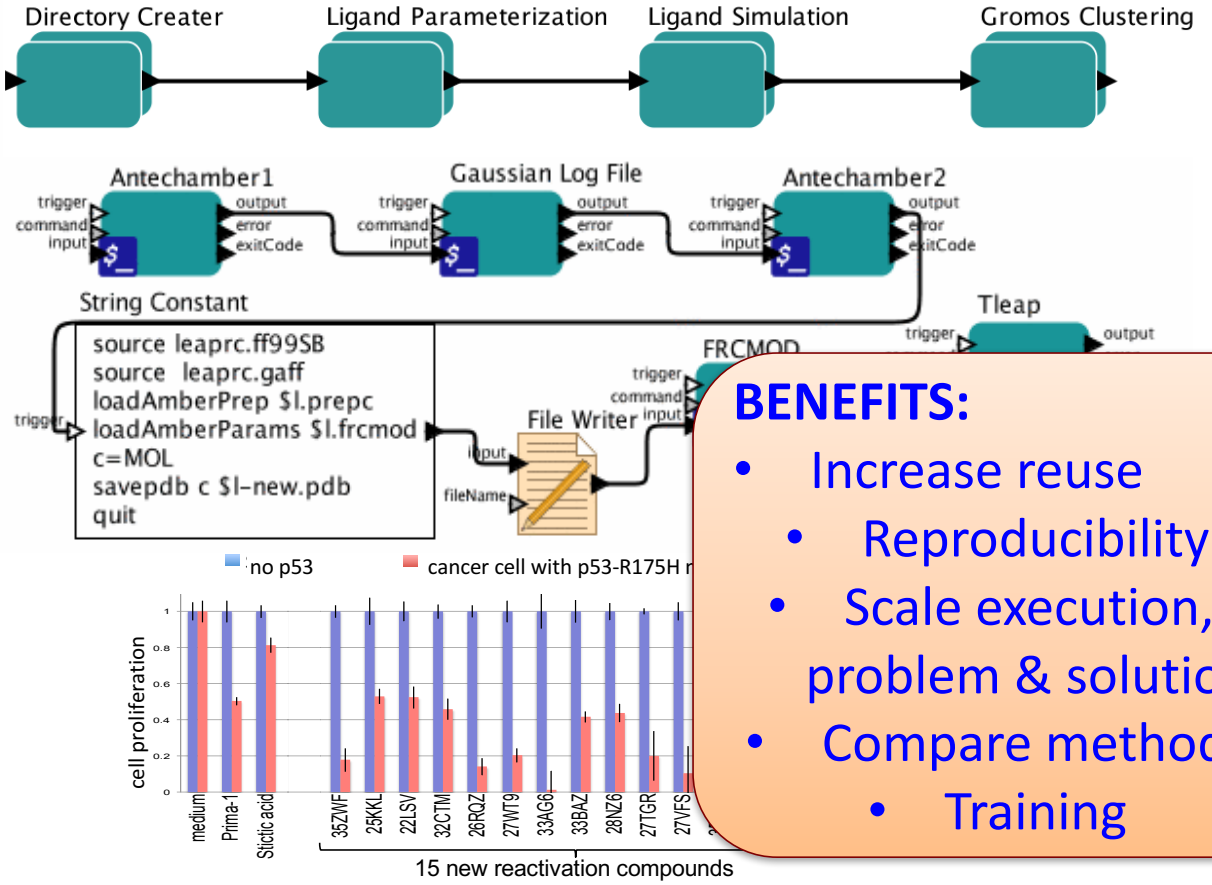
Actavalon

Challenges

- Reproducibility
- Scalability
- Interoperability
- Reliability

Scalable Drug Discovery



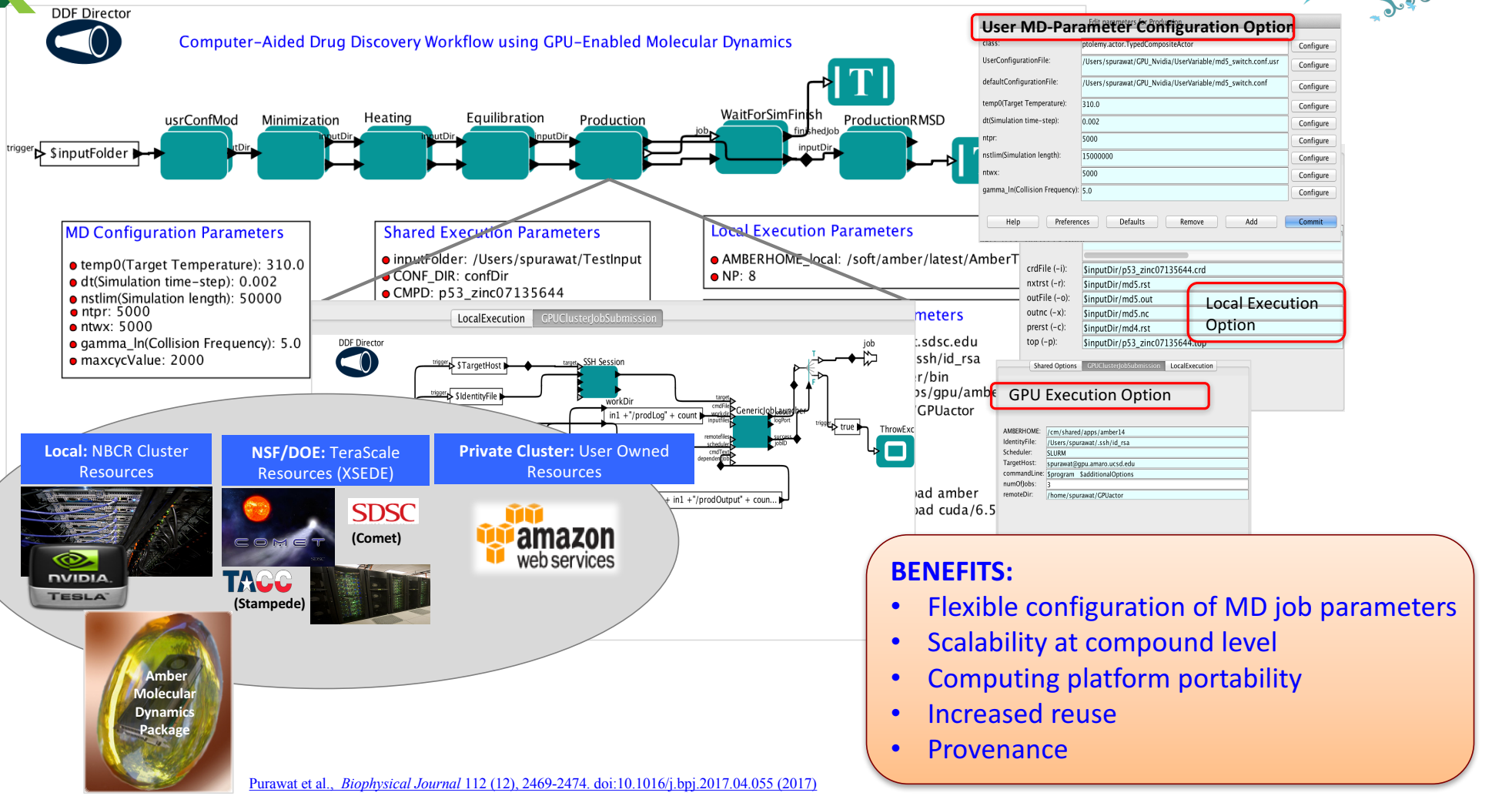


BENEFITS:

- Increase reuse
- Reproducibility
- Scale execution, problem & solution
- Compare methods
- Training

leong et al., 2014

AMBER GPU Molecular Dynamics Workbench



BENEFITS:

- Flexible configuration of MD job parameters
- Scalability at compound level
- Computing platform portability
- Increased reuse
- Provenance

Purawat et al., *Biophysical Journal* 112 (12), 2469-2474. doi:10.1016/j.bpj.2017.04.055 (2017)

Challenges

- Reproducibility
- Scalability
- Interoperability
- Reliability

Drug Design Data Resource (D3R)

blinded prediction challenges to drive advances in CADD

Central Goal: Utilize previously unpublished datasets as benchmarks for developers of protein-ligand modeling technologies

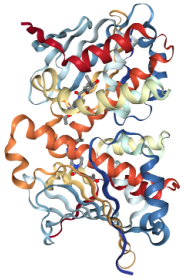
Synergy with Public Databases: Public release of more industrial crystal structures and affinity data

Broader Goals: Utilize blinded datasets to drive improvement of all CADD technologies and to foster education and dissemination of methods

More predictive CADD methods benefit everyone!

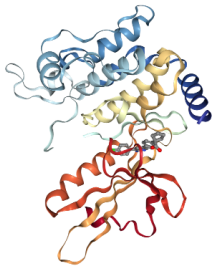
Grand Challenge 2015

35 participants, 355 submissions



HSP 90: focus on potency predictions

Data from Abbvie and Carlson's CSAR project
8 cocrystal structures (.6-2.0 Å resolution)
180 IC50s (5 nM-20 µM)
Three series: benzimidazolones, aminopyrimidines, benzophenone-like
Varied water-mediated interactions; open/closed conformations

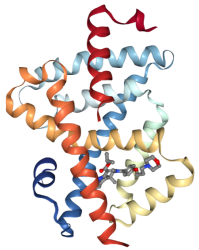


MAP4K4: focus on pose predictions

Data from Genentech
30 cocrystal structures (1.6 – 2.5 Å resolution)
18 IC50 data (3.1 nM - 10 µM)
Diverse chemotypes binding in ATP site
Open/closed P-loop structures

Grand Challenge 2

49 participants, 262 submissions



Farnesoid X Receptor (FXR): poses and potencies

Data from Roche
36 cocrystal structures (resolutions <2.6Å)
102 IC50s (0.3 nM-260 µM)
Three series + misc: sulfonamides, benzimidazoles, spiros
Helix shifts and varied water-bridges

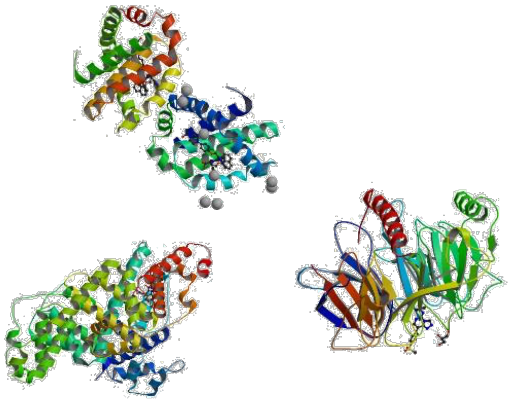
Toward Greater Statistical Power

Continuous Evaluation of Ligand Pose Predictions (CELPP)

Saturday

PDB pre-release
InChIs
Protein sequences
Forthcoming IDs

D3R scripts
Eliminate trivial ligands
Pick protein structures



Sunday

D3R releases InChIs and protein structures for docking

D3R opens for submissions

Tuesday

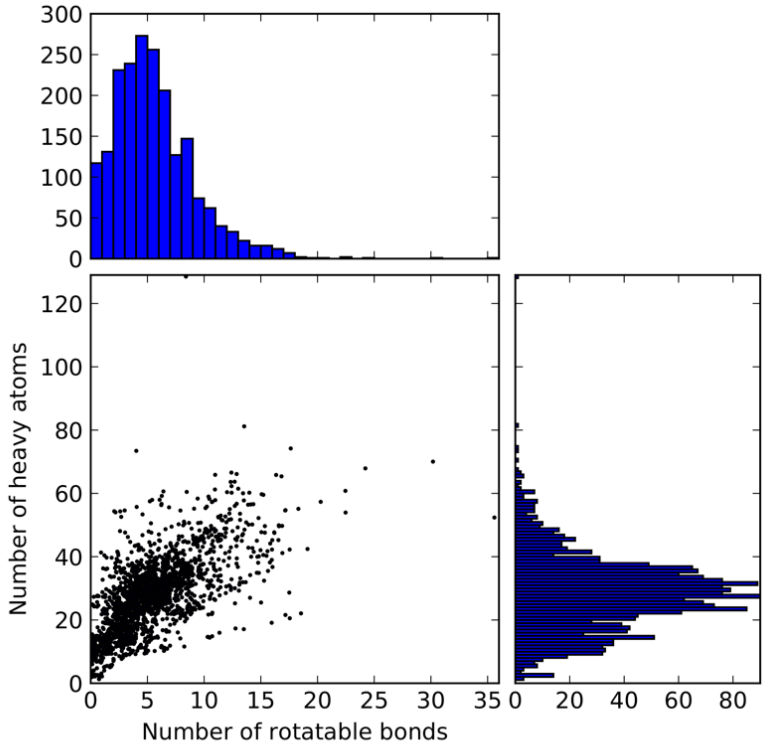
D3R submission window closes
PDB releases structures

Wednesday

D3R evaluates predictions against released structures

CELPP Challenges and Participants

>2,000 cross-docking cases over 64 weeks



Five “in-house” docking servers

- Autodock Vina
- Two GLIDE methods
- OE Fred
- rDOCK

Four anonymized external participants

In 64 weeks of running CELPP, already have order of magnitude more data / statistics than the whole 9 years of the previous efforts combined

Capturing Complex Workflows

Method 1
OMEGA, SHAFTS, Amber11

Method 2
GLIDE-CCDC-GOLD, Amber14, MMGBSa

Method 3
WaterMap, SHAPE Screening, Structural Interaction Fingerprint, DFT/B3LYP/6-31G*, GLIDE-SP-XP, Induced-fit-docking, Emodel/GlideScore-SP, Binding Pose Metadynamics

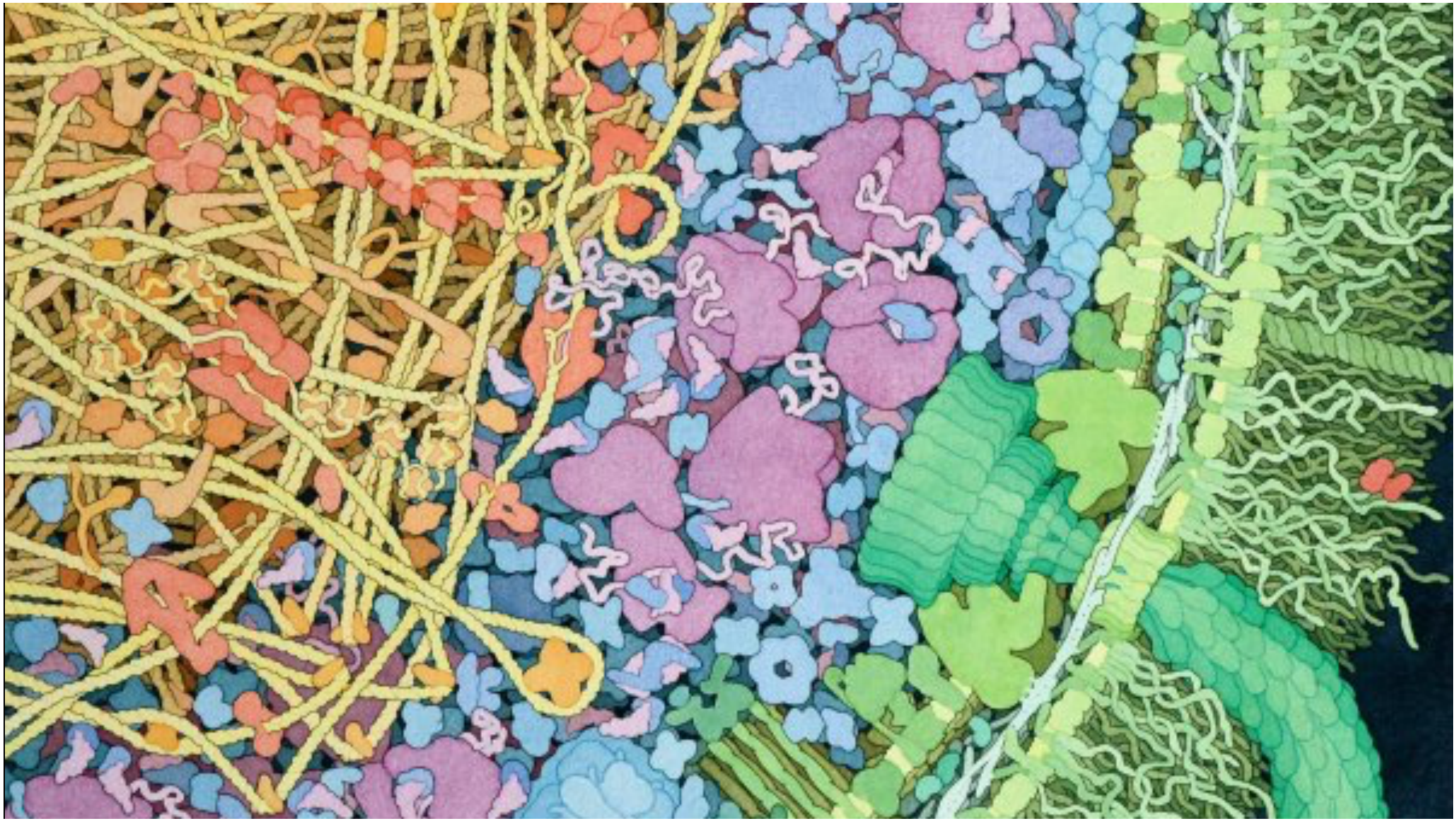
Next up:
CELPP+ for binding affinity predictions

Full description of methods

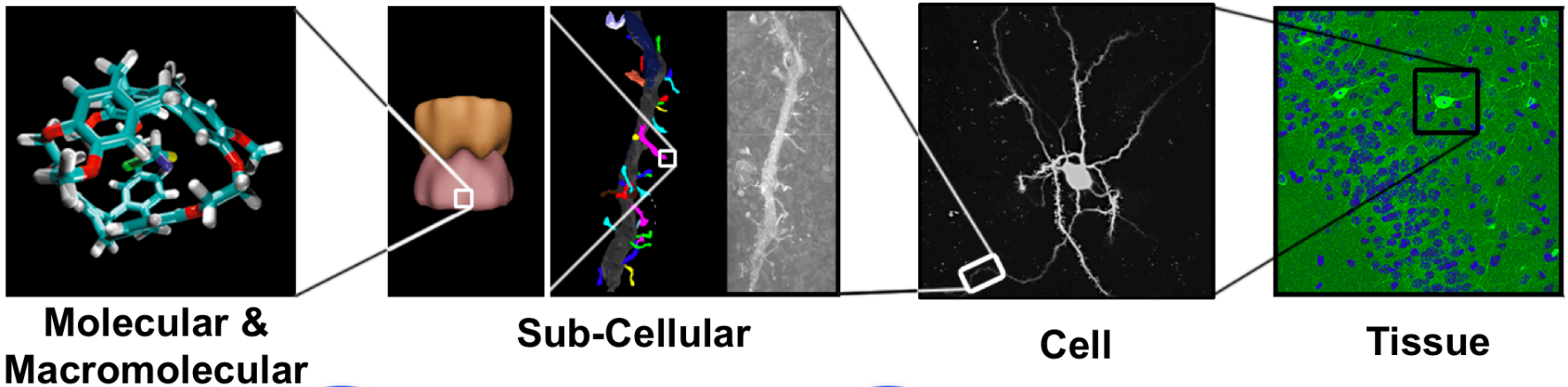
Reproducibility

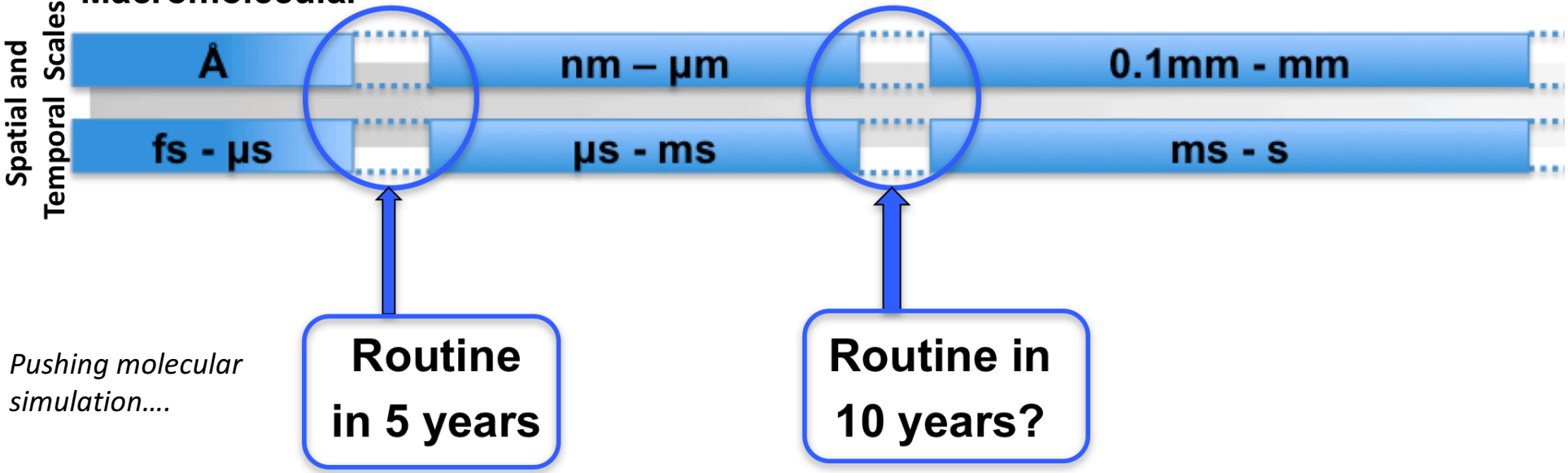
Evaluation on new datasets

Application to drug design projects



Multiscale methods bridge gaps across scales





Pushing molecular simulation….

Volume 2, Article No. 0148, 2018

Algorithmic Challenges

PERSPECTIVES

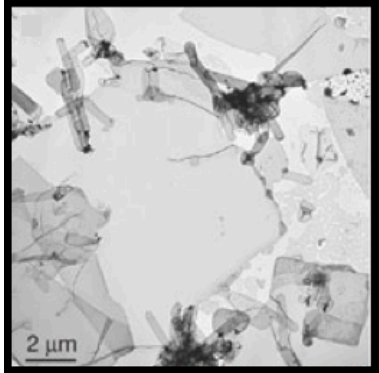
Multiscale methods in drug design bridge chemical and biological complexity in the search for cures

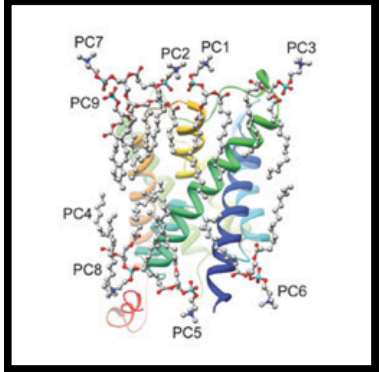
Rommie E. Amaro and Adrian J. Mulholland

Abstract | Drug action is inherently multiscale: it connects molecular interactions to emergent properties at cellular and larger scales. Simulation techniques at each of these different scales are already central to drug design and development, but methods capable of connecting across these scales will extend our understanding of complex mechanisms and our ability to predict biological effects. Improved algorithms, ever-more-powerful computing architectures and the accelerating growth of rich data sets are driving advances in multiscale modelling methods capable of bridging chemical and biological complexity from the atom to the cell.

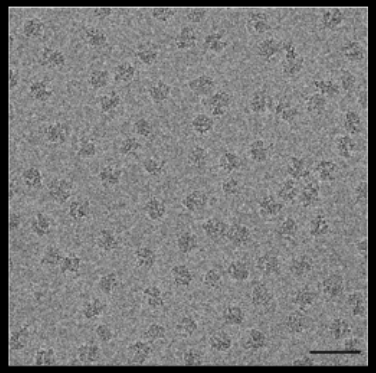
3D structural data to build visible virtual cells

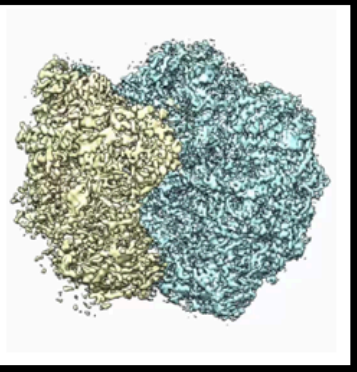
Electron crystallography
2-D crystals of membrane proteins in their native environment



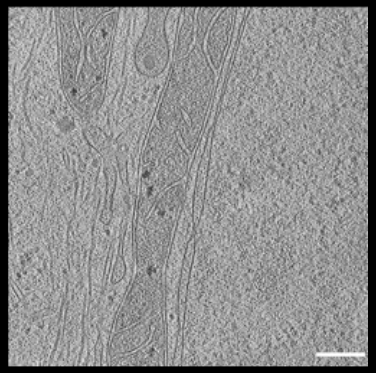


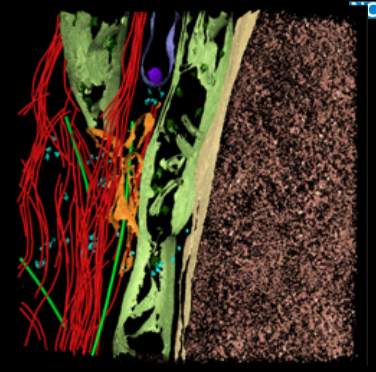
Single-particle analysis
Purified molecules in solution ~0.2-10 MDa



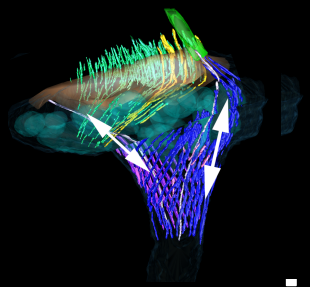


Electron tomography
Pleomorphic samples, e.g., cells and organelles

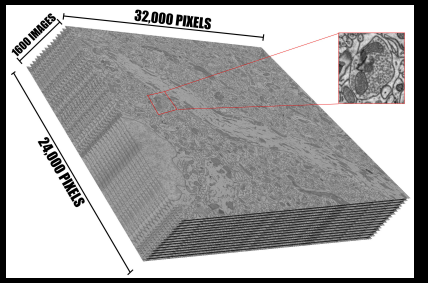


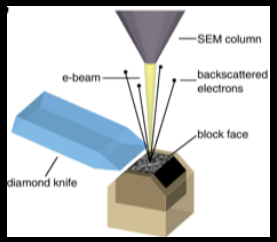


Serial Section EM
Resin-embedded samples



Serial Block EM
Resin-embedded tissues





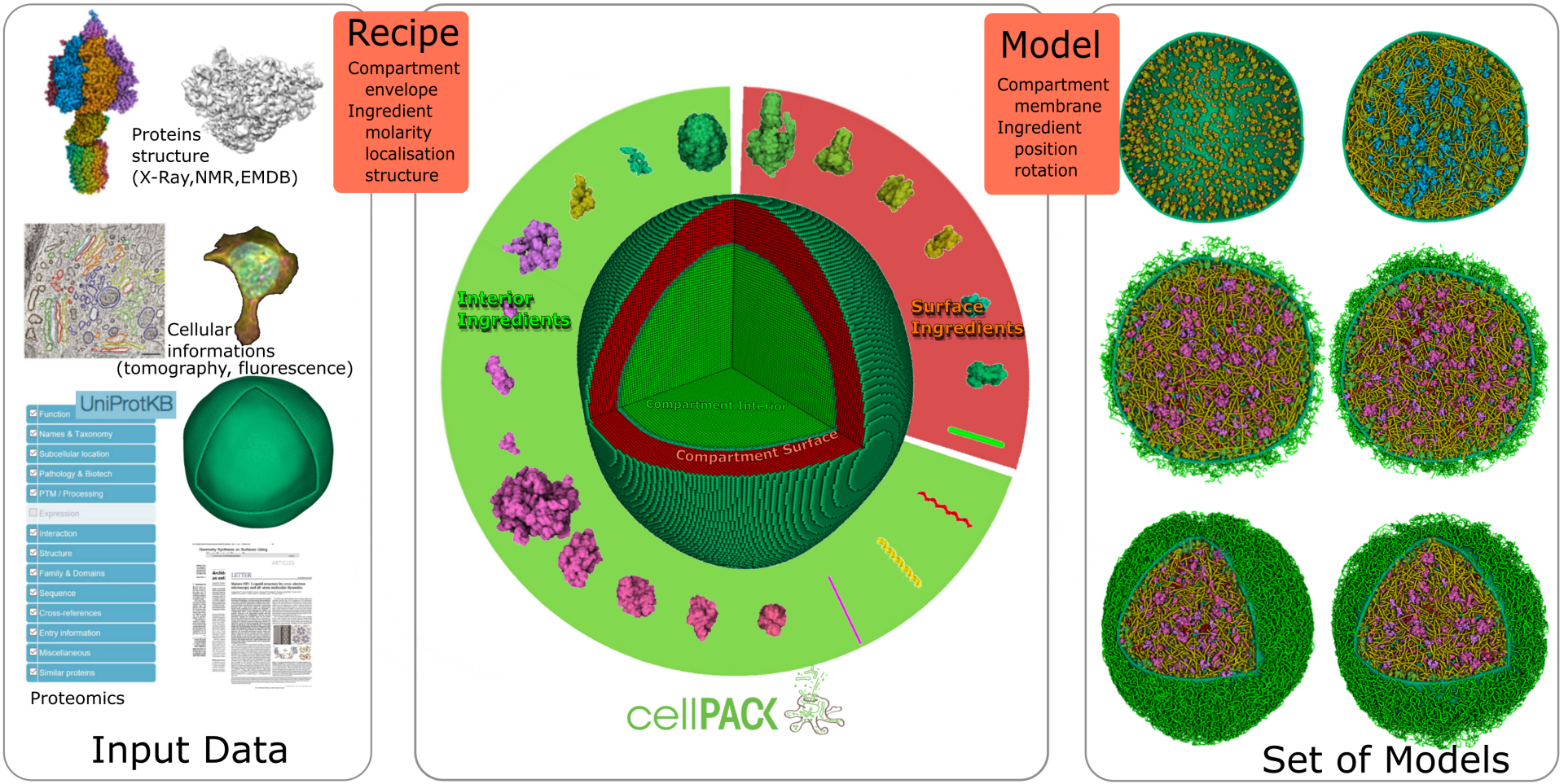
Routine dataset is 1.2 trillion pixels

- 100,000's of structures in a single dataset

Challenges & Opportunities

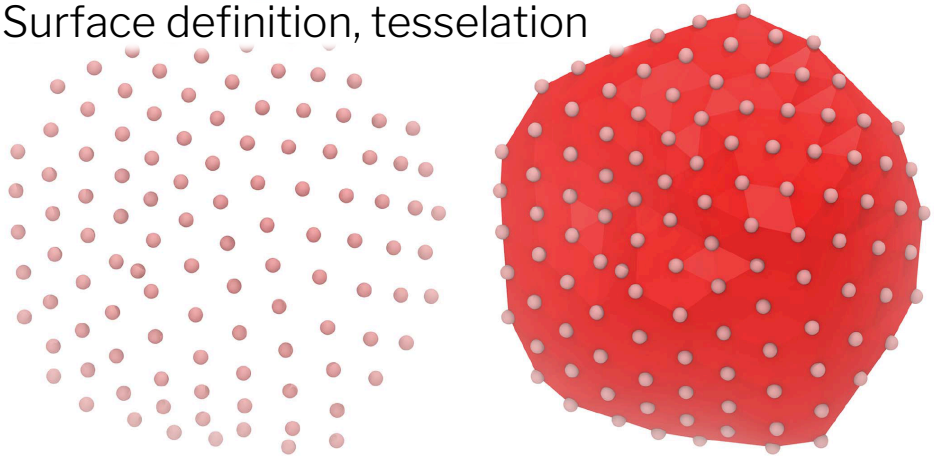
- Data Complexity
  - imaging segmentation & refinement
  - extracting signal from rich datasets: cryoEM, diffuse scatter
- Data integration
  - Bringing diverse datasets together

Cell-centered, data-centric modeling framework

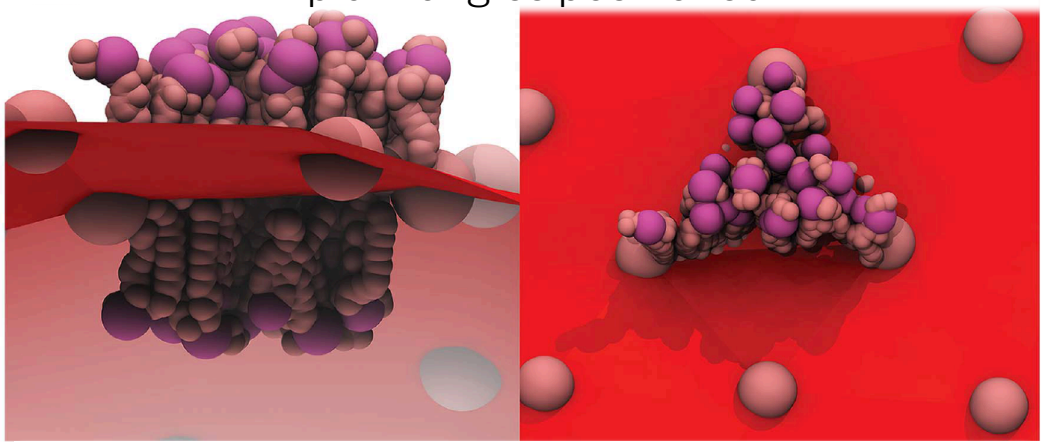


Membranes: Lipid bilayers with realistic geometries

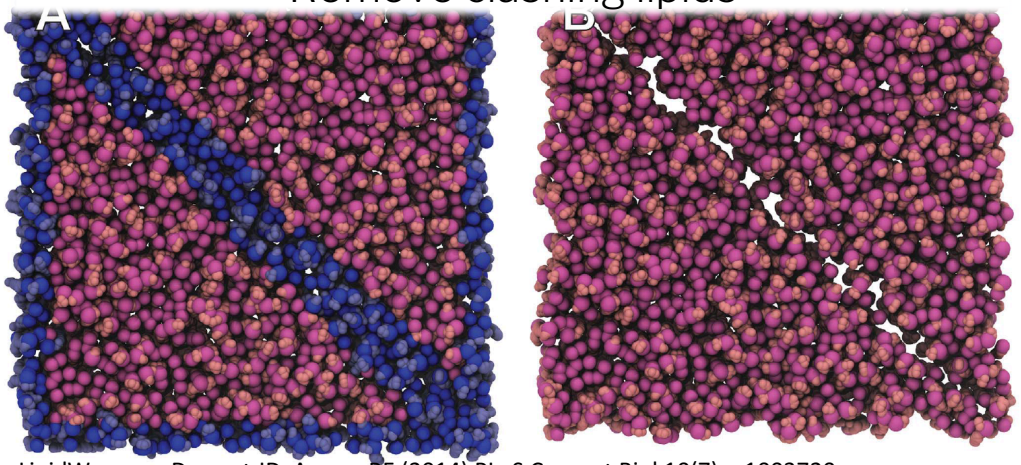
Surface definition, tesselation



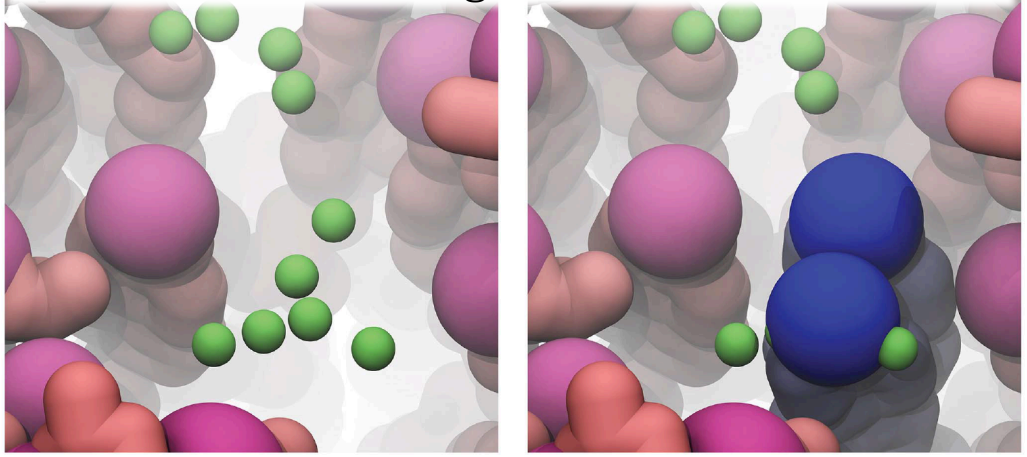
Lipid triangles positioned



Remove clashing lipids

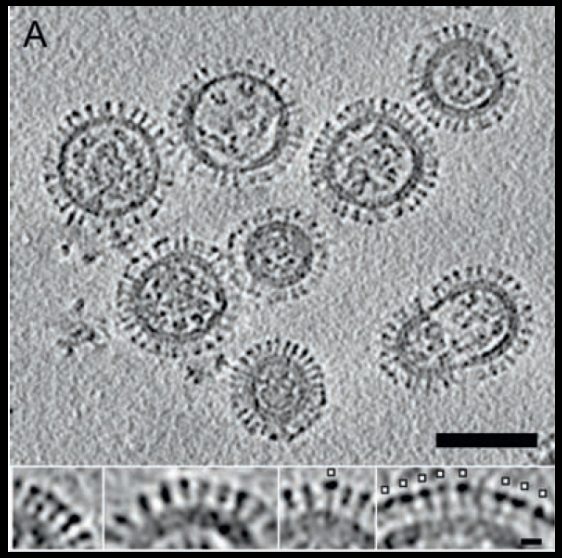


Filling in holes



LipidWrapper: Durrant JD, Amaro RE (2014) PLoS Comput Biol 10(7): e1003720.

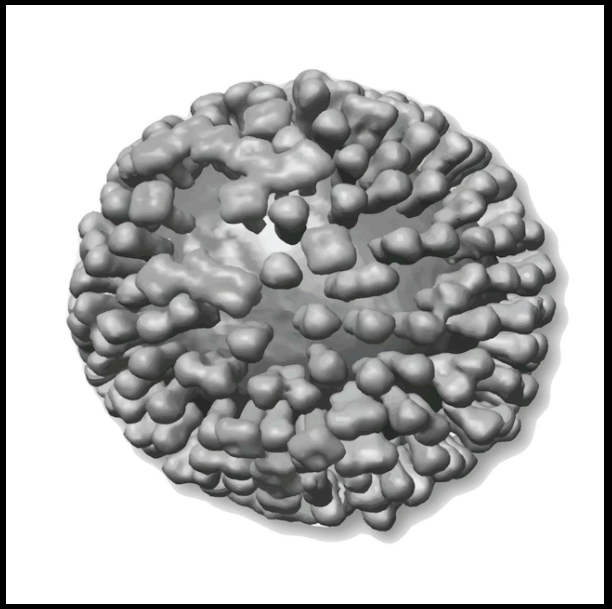
Moving from single protein to whole virus



Alasdair Steven, NIH

Fully Atomic Reconstructions

PyMolecule
LipidWrapper
CellPACK



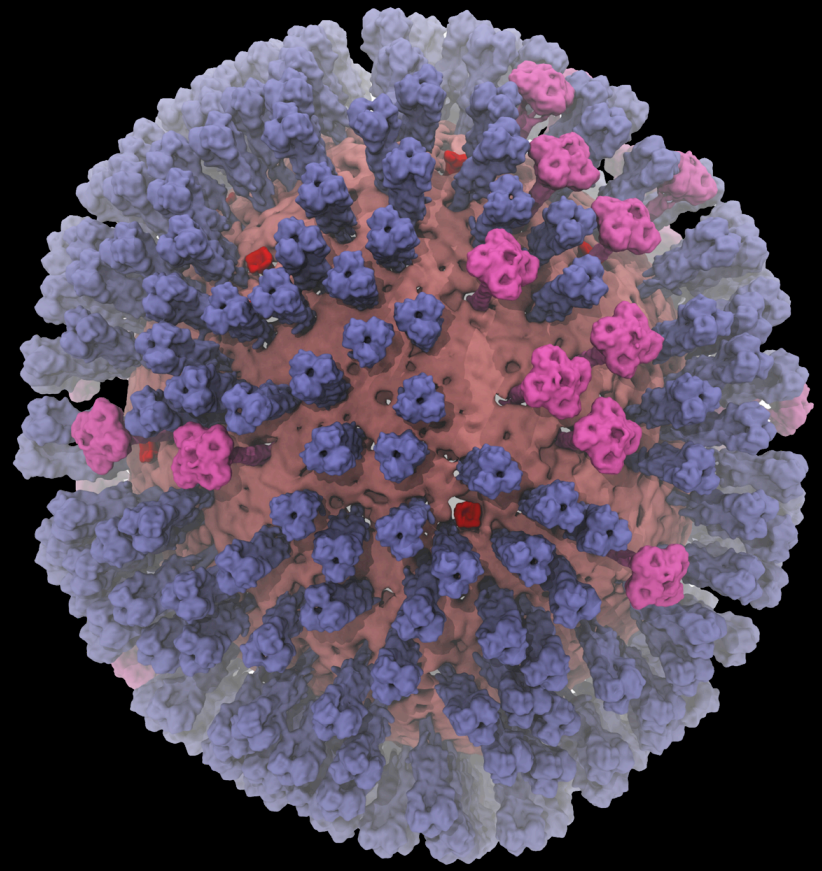
- Improved sense of the physical arrangement of biological entities in complex biological milieu
- Enables simultaneous study of multiple components
- Mesoscale molecular models as a platform for other simulation approaches (e.g., Brownian dynamics, Mcell, lattice boltzmann MD)

... leads us to new avenues of investigation, not possible on the single protein scale

Johnson et al, Nature Methods (2014),; Durrant & Amaro, PLOS Comp Bio (2014).

Petascale Molecular Dynamics Simulation of Fully Lipid Enveloped Virus

- Largest biological system ever simulated at atomic level (~160 million atoms)
- 4.5 ns/day using 114,688 CPUs
- 158 ns total simulation
- Saving every 20 ps ➔ ~25 TB of data
- Collaboration with TCBG P41



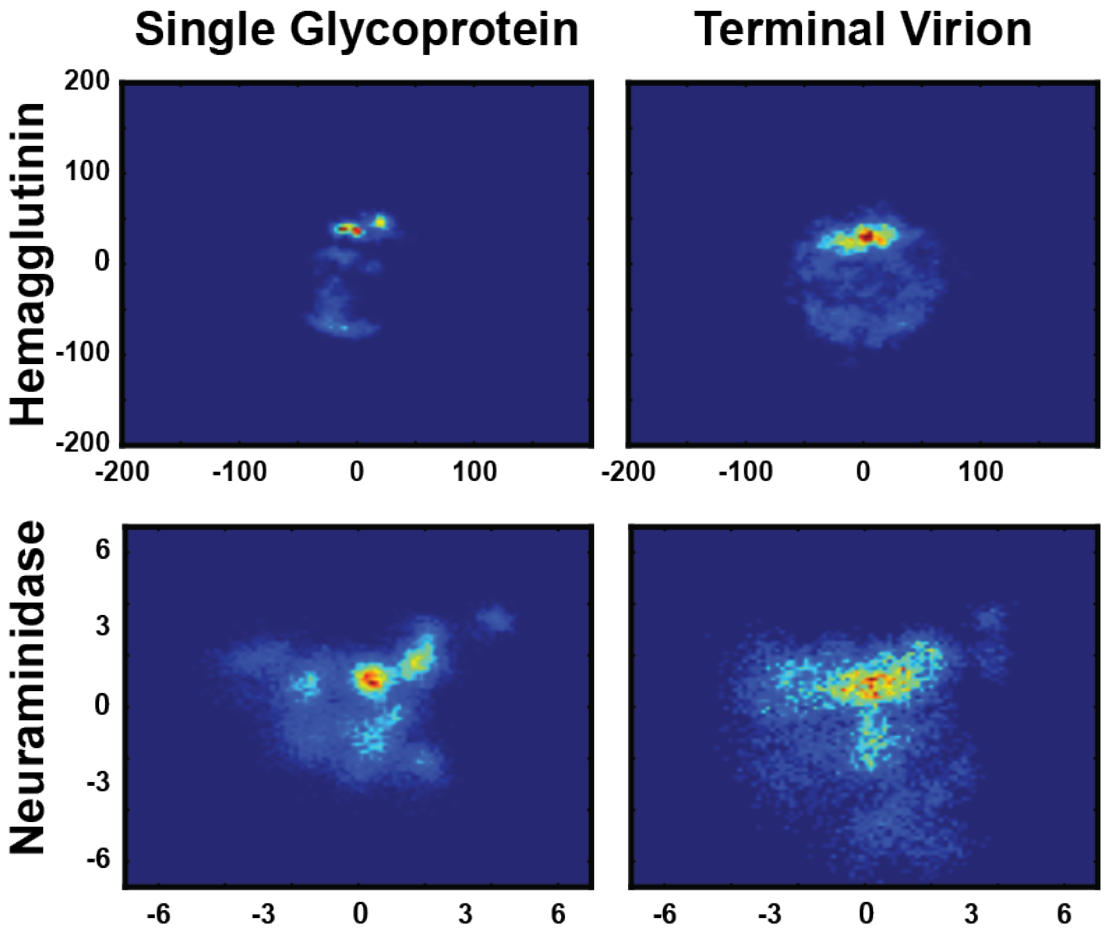
~ 160 million all-atom MD simulation with NAMD on Blue Waters

Challenges

- Accessibility
  - To increasingly large datasets
  - To the Big Machines (eg Blue Waters, Titan)
  - To farms of GPUs

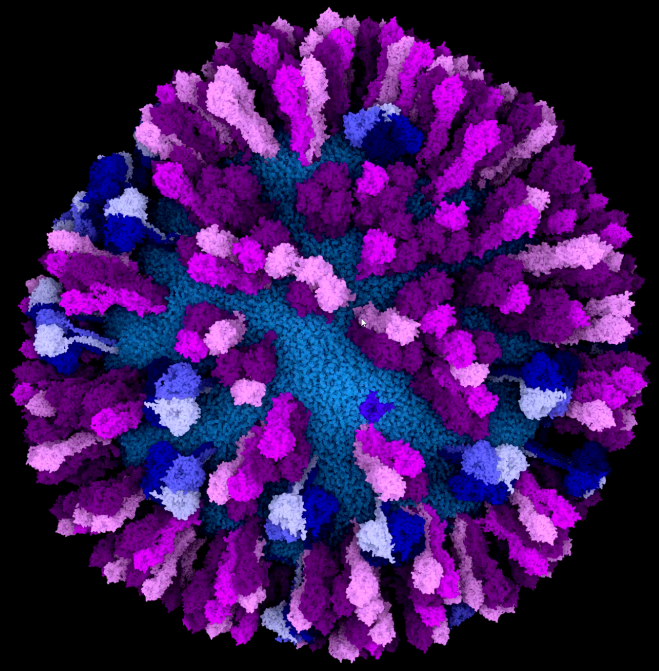
- Visualization, analysis methods, etc
- Data interaction at scale

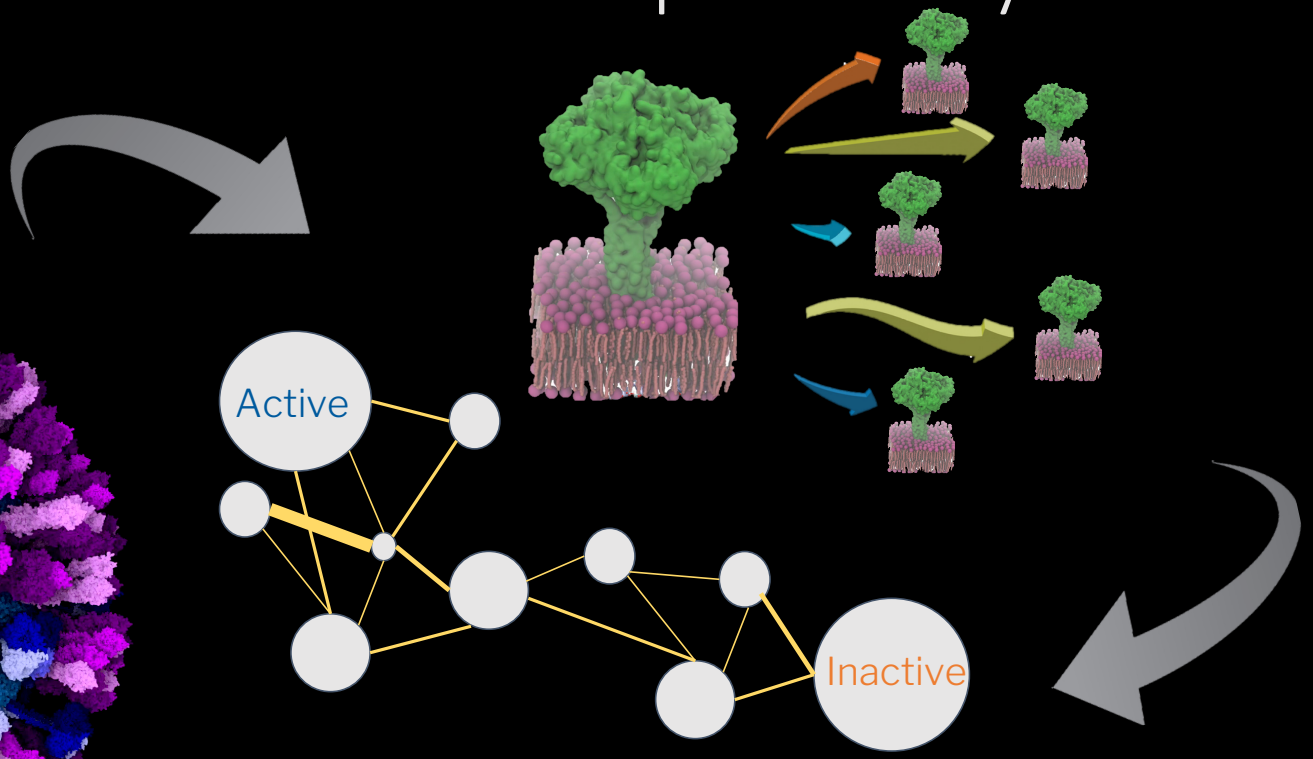
Dynamics in the packed, crowded virus different



Interactions with neighboring molecules matter

Cell-scale Markov state models of protein dynamics



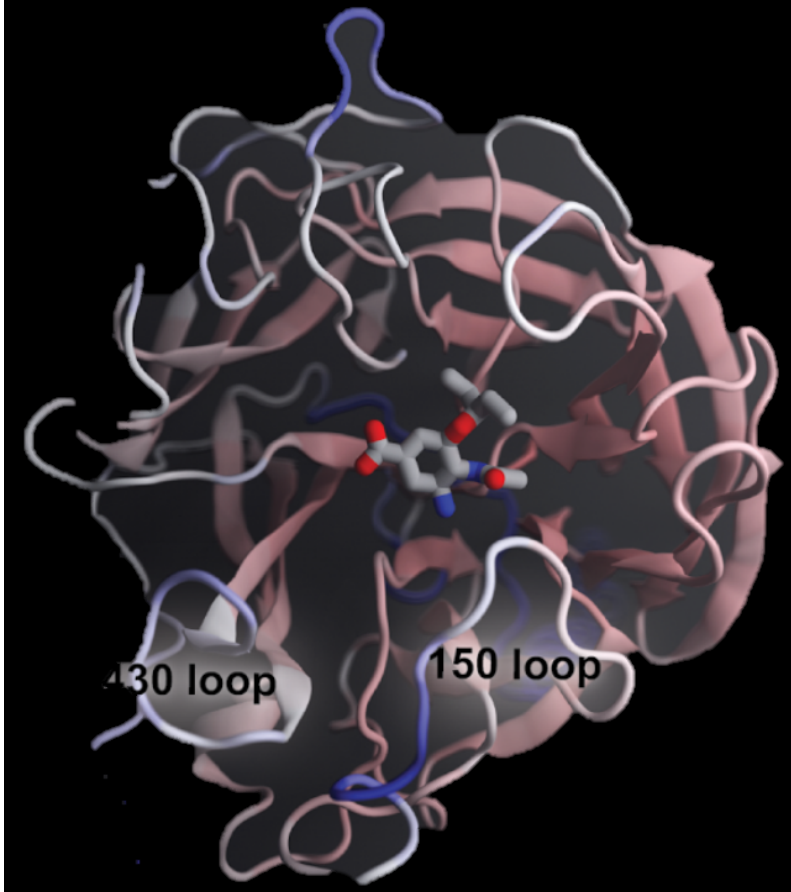


Markov state models define metastable states and transitions between states

Allows one to extract long timescale dynamics from many short timescale simulations

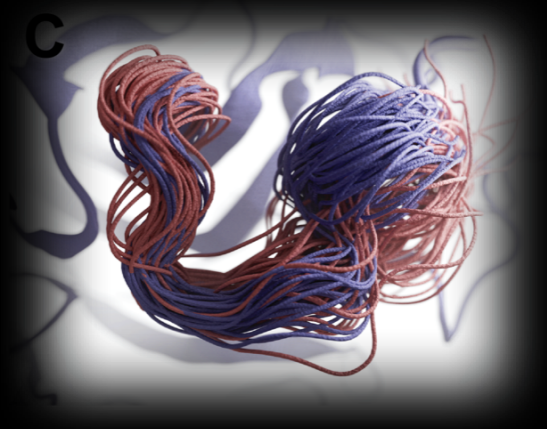
Swope, Pande, Schutte, Noe…

MSMs characterize loop dynamics & druggable pockets



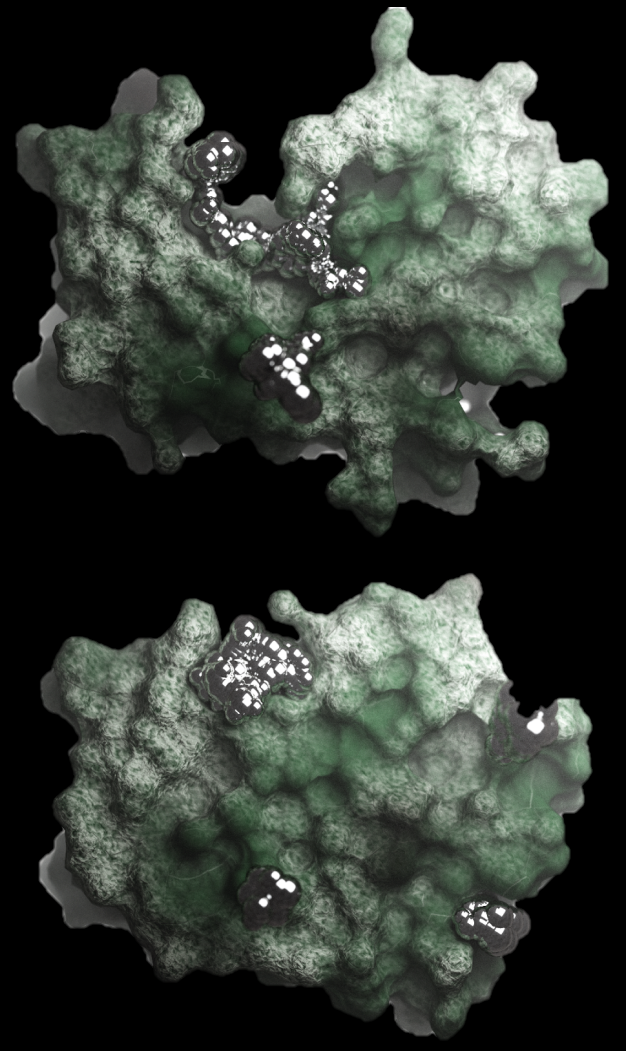
Virion has 30 NAs, 236 HAs
Enough sampling to make a Markov state model (MSM) of NA loop dynamics

2-state Macrostate model
open/closed



MFPT for the 150-loop:

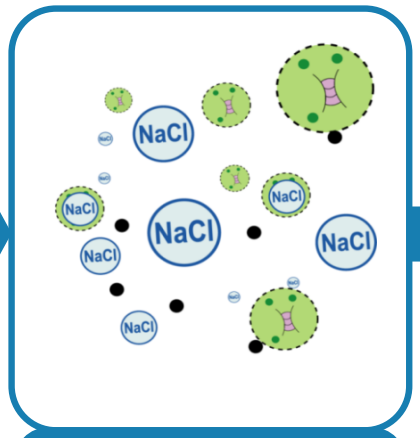
- *open to closed* 52.9ns
- *closed to open* 198.4 ns



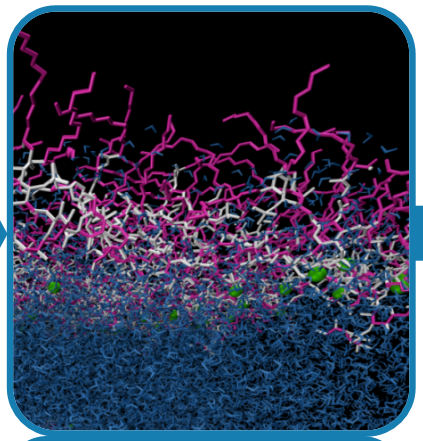
Impact of Sea Spray Aerosols on Chemistry of the Environment

Multiscale Paradigm of Chemical & Biological Complexity

Production



Composition



Structure & Reactivity

Atmospheric Aerosol Processes

- Water Uptake
- Cloud Droplet Formation
- Ice Nucleation
- Optical Properties

Mm - µm	~10 µm - 0.5µm	µm-nm	nm-Å
Days-s	ms – µs	µs - ns	ns - fs

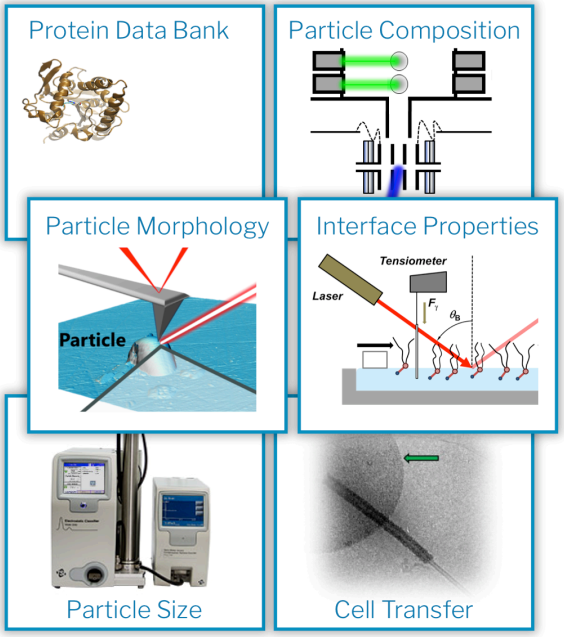
++ Complexity

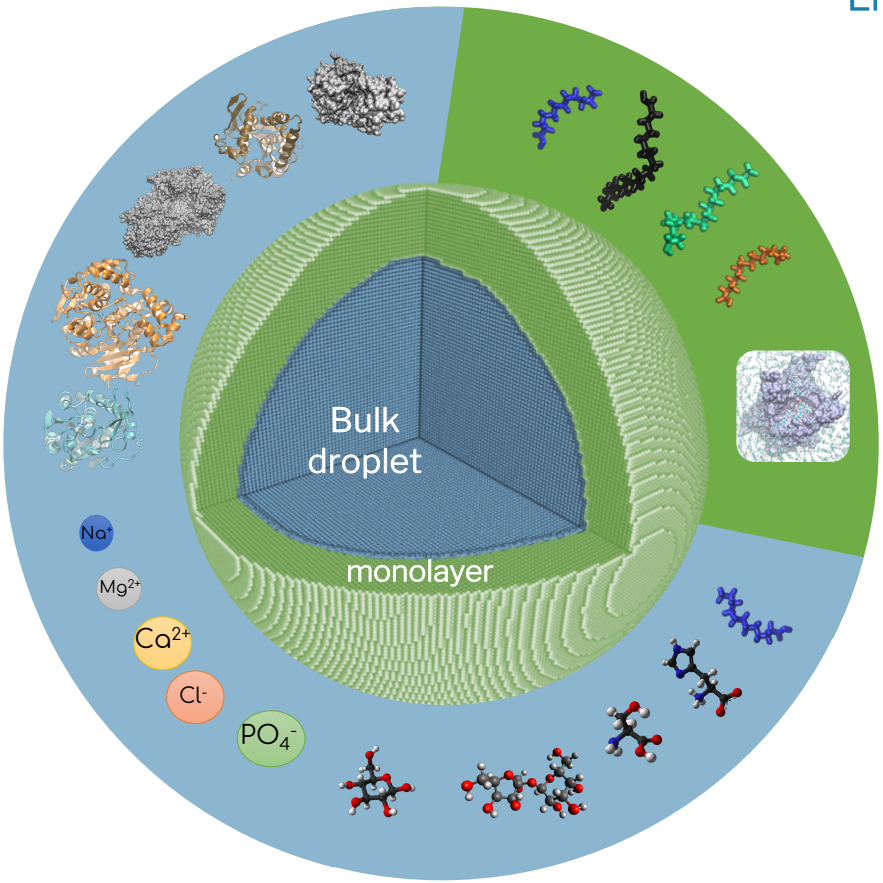
CAICE theory

Data Driven Simulations Bridge Biological & Chemical Complexity

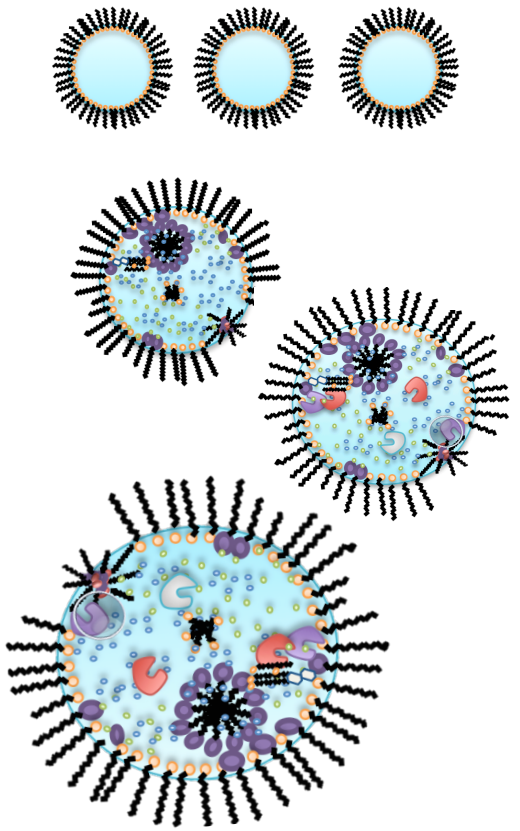
CAICE center for aerosol impacts on chemistry of the environment

Input Data Sets



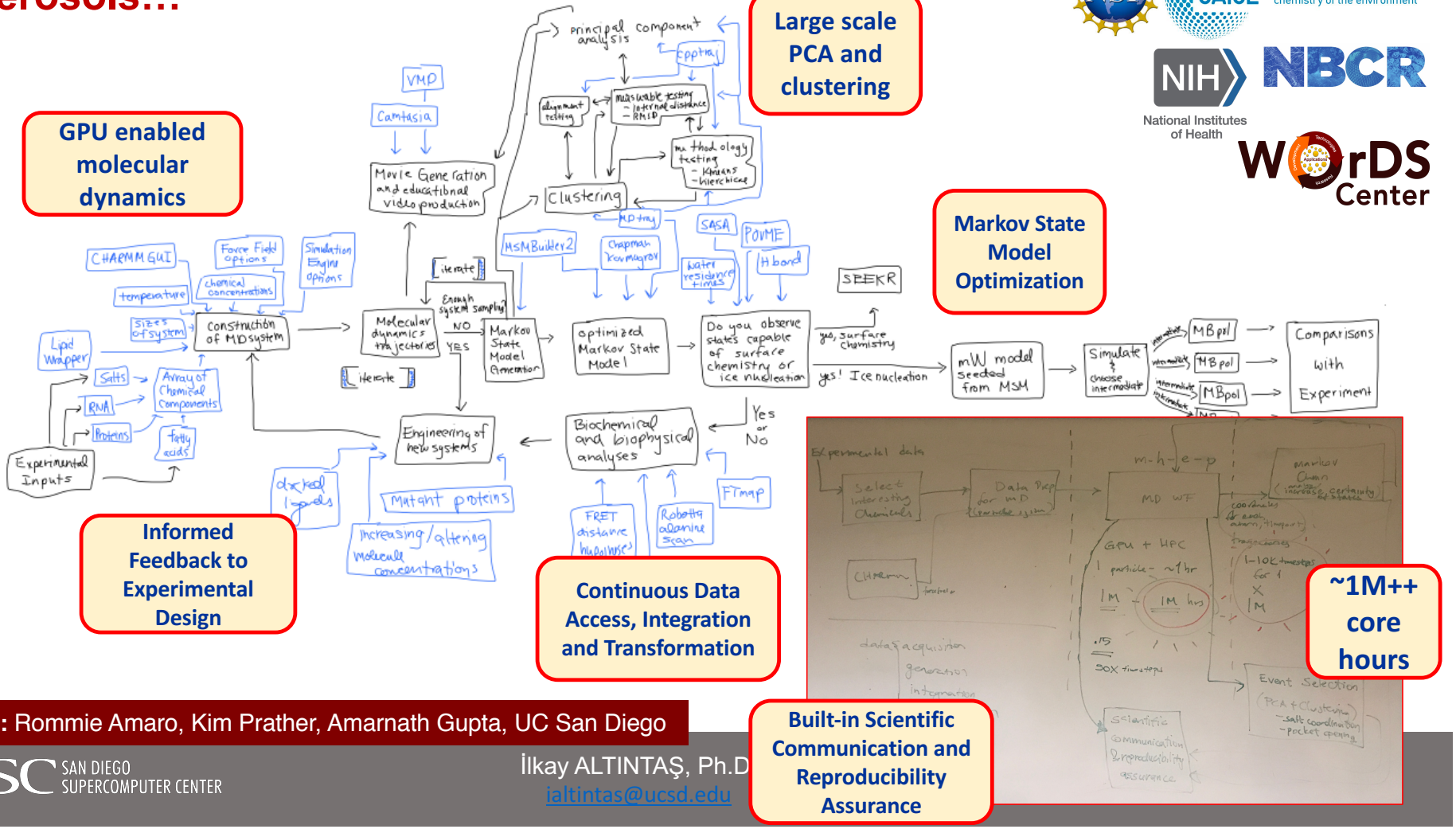


Ensembles of Molecular Models
Micron diameter



cellPACK

When computational biophysics meets sea spray aerosols…



Collaborators: Rommie Amaro, Kim Prather, Amarnath Gupta, UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

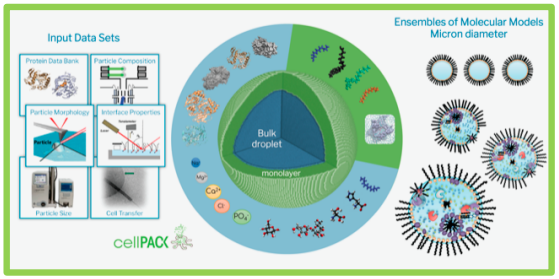
İlkay ALTINTAŞ, Ph.D
ialtintas@ucsd.edu

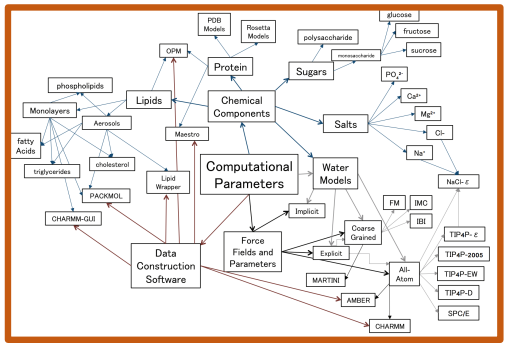
NSF CAICE center for aerosol impacts on chemistry of the environment

NIH National Institutes of Health NBCR

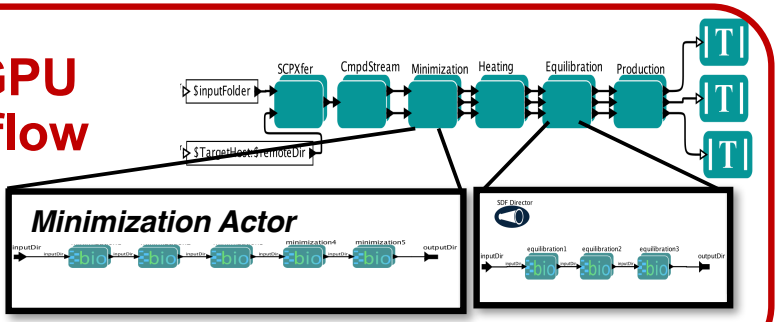
WorDS Center

Problem solving happens at the application integration level...

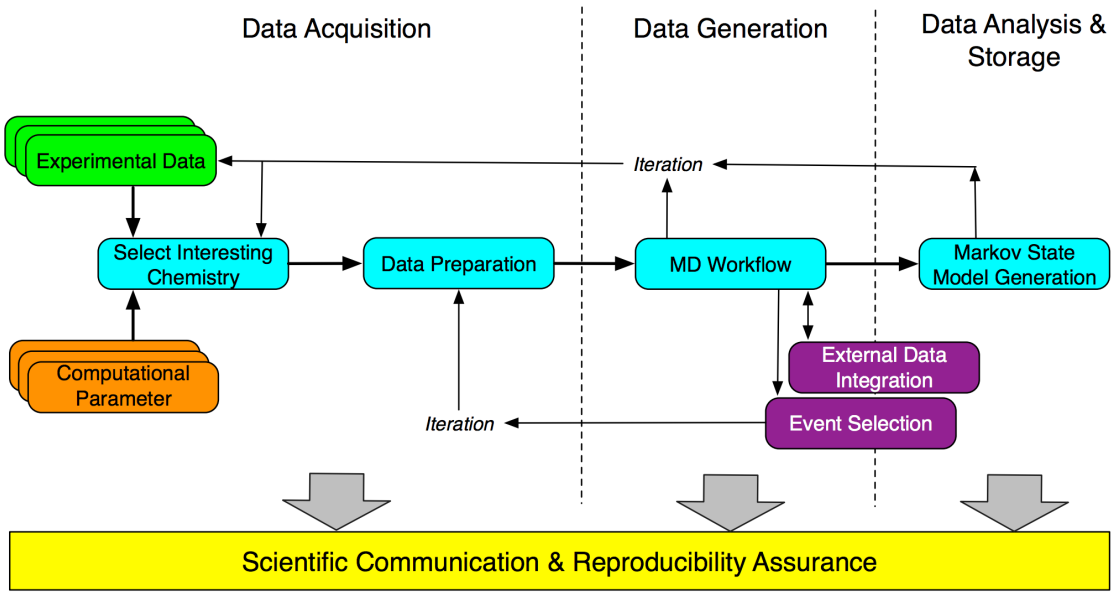




AMBER GPU MD Workflow



A Kepler Workflow Tool for Reproducible AMBER GPU Molecular Dynamics, Purawat, Ieong, Malmstrom, Chan, Yeung, Walker, Altintas, Amaro.. DOI: 10.1016/j.bpj.2017.04.055



İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

Challenges

Student / postdoc / scientist training!

Acknowledgements

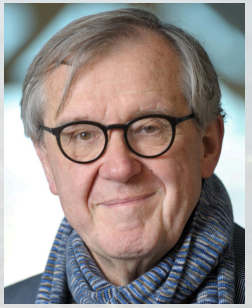
http://amarolab.ucsd.edu

http://nbcr.ucsd.edu

Lane Votapka

Ben Jagger

Cameron Abrams, Drexel
Rob Elber, UT Austin
Frank Noe, Freie Univ Berlin
Gary Huber, UCSD
Adam Van Wynsberghe, Hamilton College



COMP ACS DIVISION OF COMPUTERS IN CHEMISTRY

UC San Diego

NIH National Institute of General Medical Sciences
Basic Discoveries for Better Health

NBCR

NSF

NIH DIRECTOR'S NEW INNOVATOR AWARD

THEORETICAL and COMPUTATIONAL BIOPHYSICS GROUP
NIH RESOURCE FOR MACROMOLECULAR MODELING AND BIOINFORMATICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Hellman Fellows

COMPUTE THE CURE

NIH National Institute of Allergy and Infectious Diseases
Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases.

Acknowledgement

DISCLAIMER: This work was supported by NSF through the NSF Center for Aerosol Impacts on Chemistry of the Environment (CAICE), CHE-1801971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).