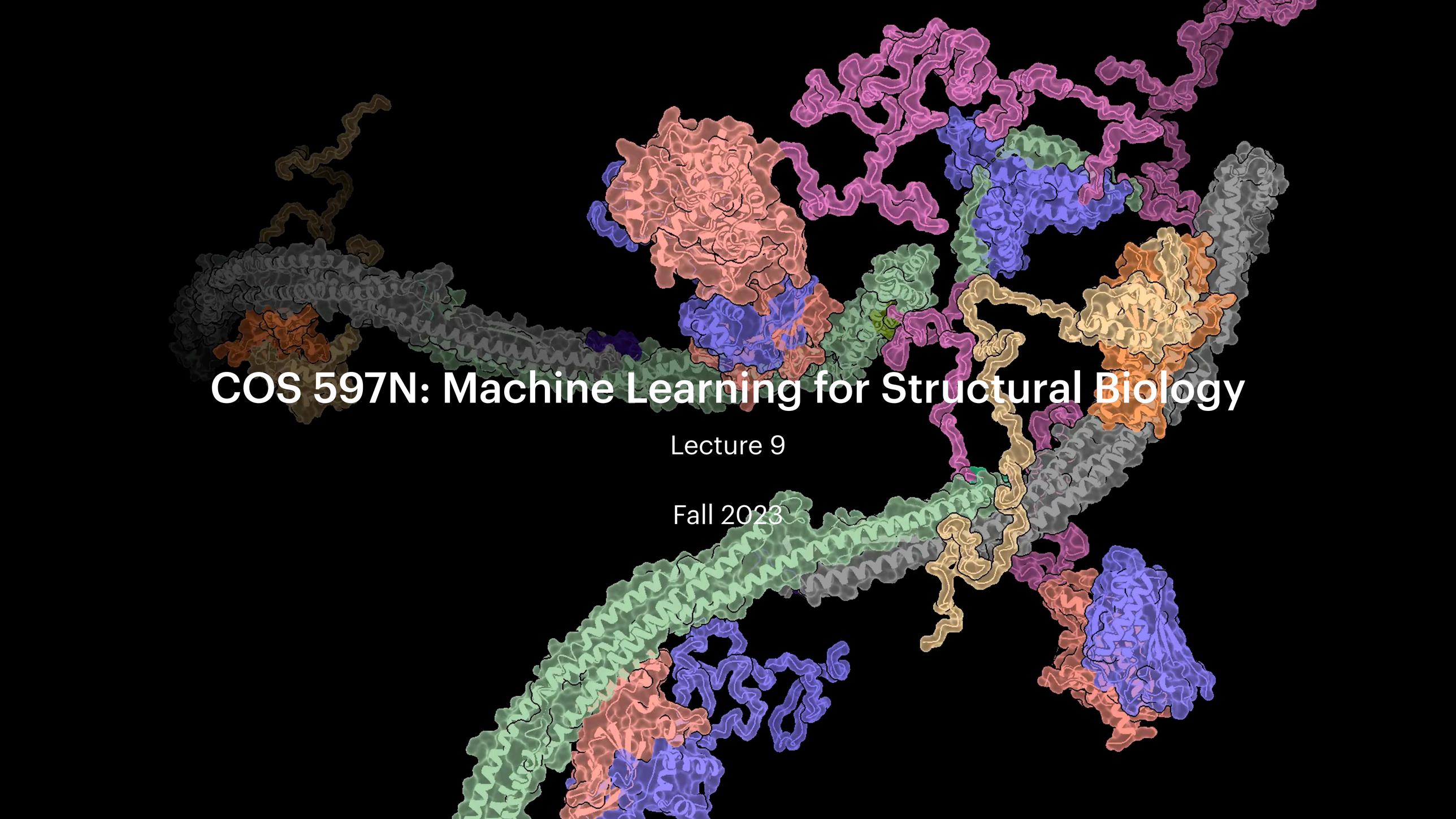


COS 597N: Machine Learning for Structural Biology

Lecture 9

Fall 2023

Week 9 Course Logistics

- Optional student-only “precept”, Tuesdays at 4:30p in CS 401.
- Today:
  - Molecular dynamics and computational drug discovery overview
  - Boltzmann Generators: Yihao and Jiahao
  - Feedback: Justin and Eugene
- Next week: Protein structure determination II: 3Dflex and ModelAngelo
- Project
  - Feedback on proposals: What is the research question? Think about scope for a semester-long class project
  - Results milestone by Monday Nov 27, 11:59p on Canvas
    - Work-in-progress report, containing an outline of the sections of the paper, and at least one figure or table containing preliminary results via Canvas
    - Graded for completion

Recap: “The Protein-Folding Problem, 50 Years on”
Dill & Maccallum, Science 2012

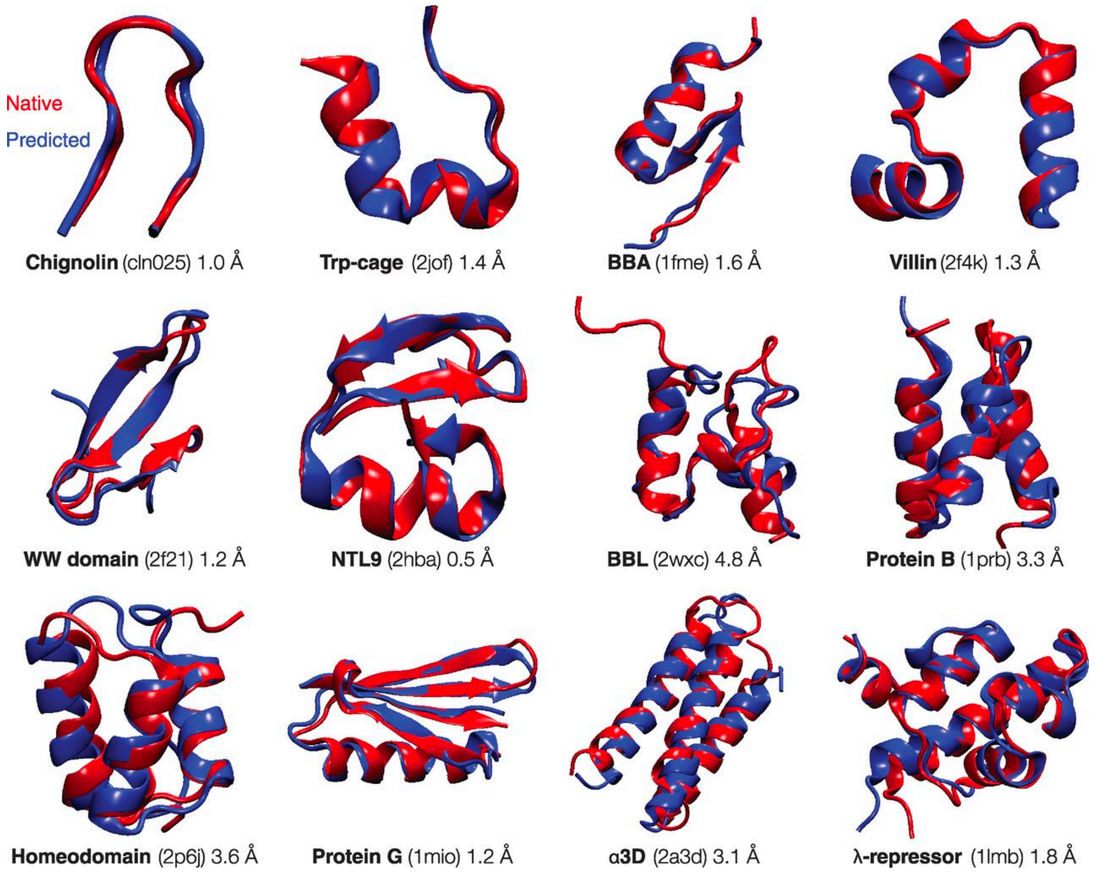
Three broad questions:

1. What is the physical code by which an amino acid sequence dictates a protein’s native structure? (Thermodynamics)
2. How can proteins fold so fast? (Kinetics)
3. Can we devise a computer algorithm to predict protein structures from their sequences?

> “Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure.”

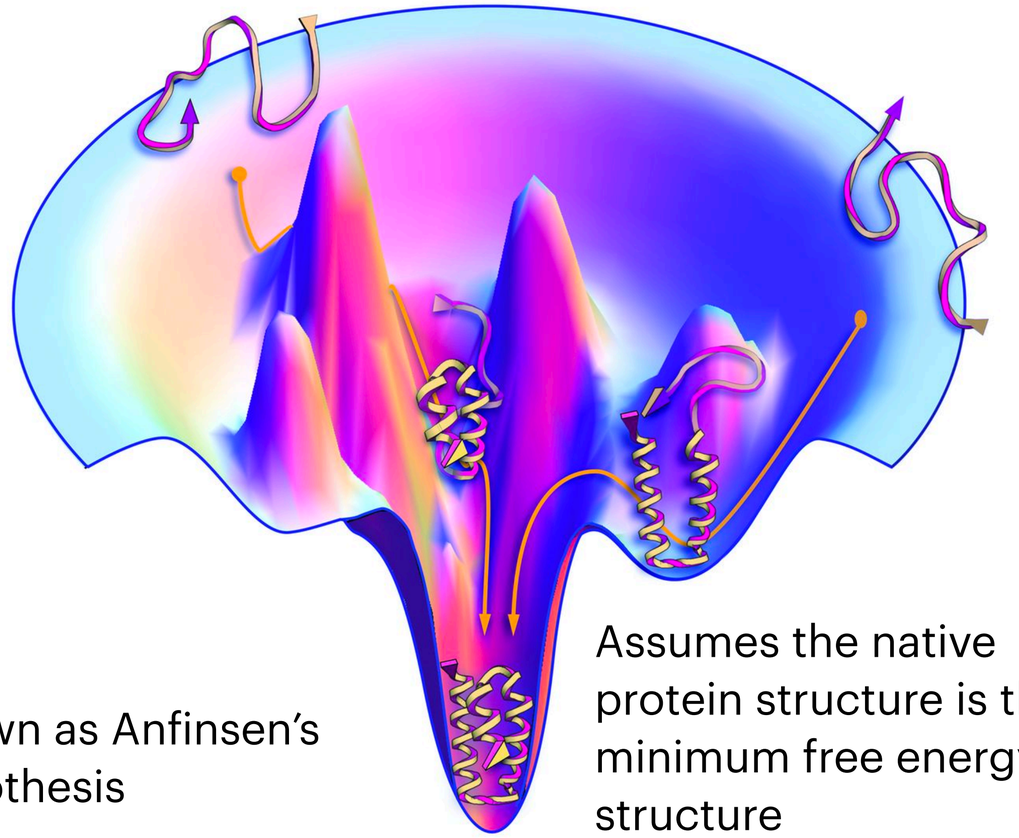
Fig 1: In 1958, Kendrew and coworkers published the first structure of a globular protein, myoglobin at 6A resolution

Protein folding is driven by physics



Lindorff-Larsen et al, Science 2011

Energy landscape theory of protein folding



Known as Anfinsen's hypothesis

Assumes the native protein structure is the minimum free energy structure

Figure from Dill & Maccallum, Science 2012

The Boltzmann distribution

- In real life, and in an MD simulation, *atoms are in constant motion*.
  - They will not simply go to an energy minimum and stay there (Where have we seen this assumption?)
- Given enough time (i.e. when the system is at equilibrium), the simulation samples the *Boltzmann distribution*
  - The probability of a particular arrangement of atoms is a function of its potential energy
    - $$p(\mathbf{x}) \propto exp(-E(\mathbf{x}))$$
  - How long do you need to simulate to reach all energetically favorable arrangements?
  - This is not the only way to explore the energy surface (i.e. sample the Boltzmann distribution) — What are other ways?

Adapted from Stanford CS279 slides

Why is sampling from the Boltzmann distribution hard?

1. How do you represent the potential energy surface of a protein?
2. Protein dynamics typically contains states separated by long timescales, i.e. folded and unfolded states

[See DESRES slides]

From silicon to medicine

Core challenges of using molecular dynamics for early-stage drug discovery

Ellen Zhong, Cory Hargus, Tom Weinreich, Caleb Jordan

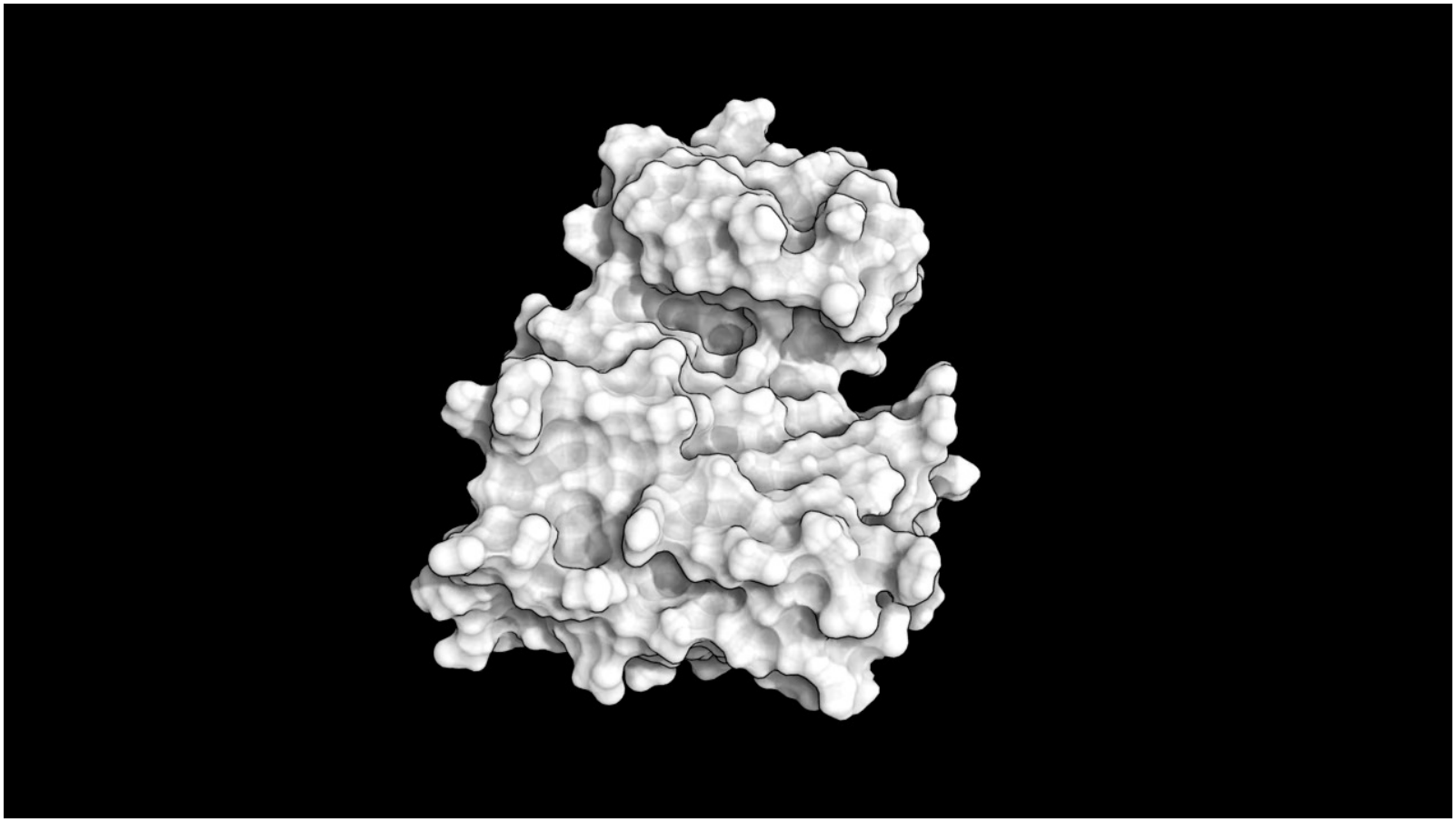
D. E. Shaw Research

November 14th, 2015

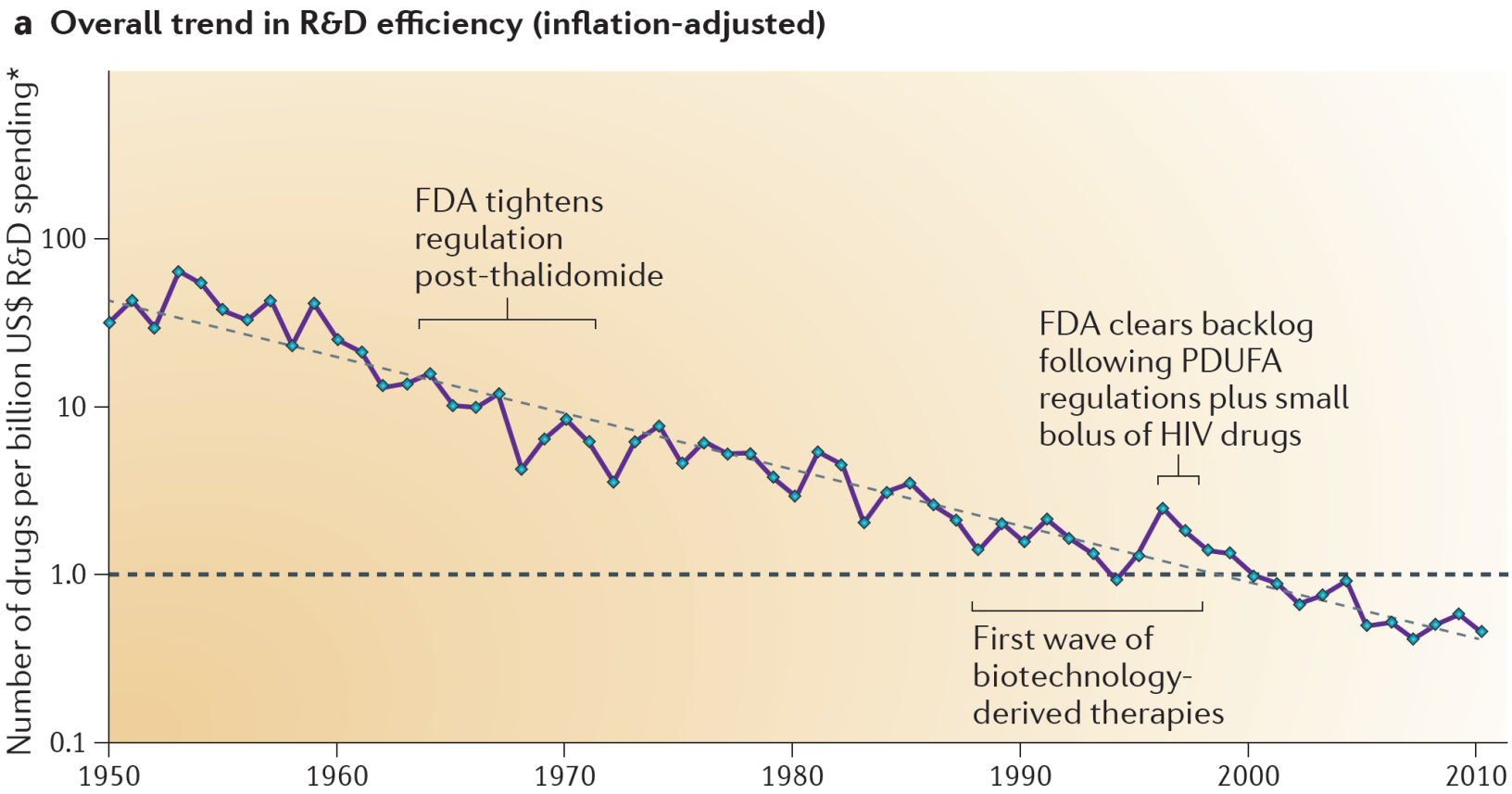
Who are we?

- Independent research group
  - Founded in 2002 by David E. Shaw
  - ~100 chemists, computer scientists, engineers
  - Located in midtown Manhattan
- High-level goal: enable fundamental advances in human biochemistry and drug discovery
  - Main tool: molecular dynamics (MD) simulations
- Designed and build Anton
  - Special-purpose supercomputer for MD

Where Biochemistry Research Meets Simulation Meets Hardware Specialization Meets Graphics



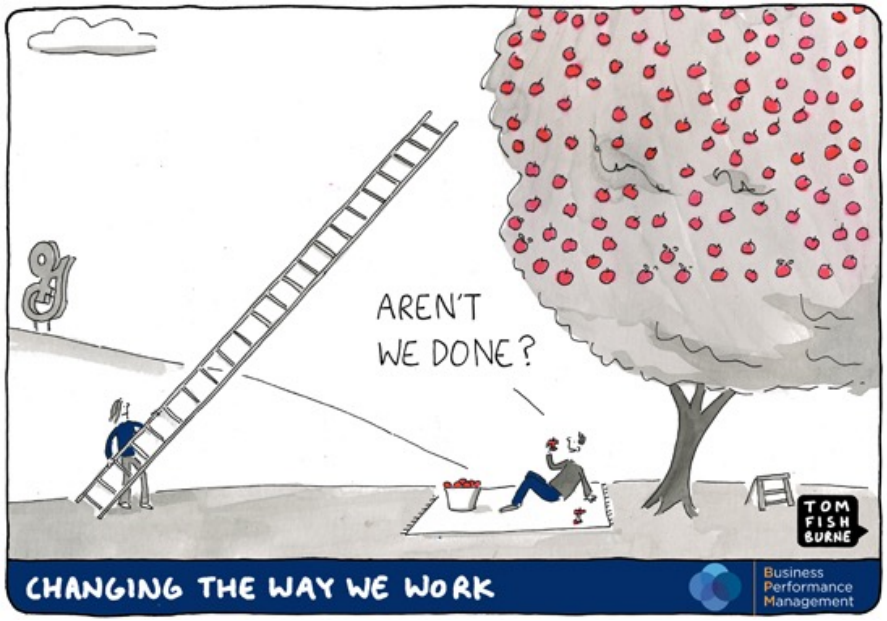
Cost of pharmaceutical R&D



Nature Reviews Drug Discovery, 2012

Why so expensive?

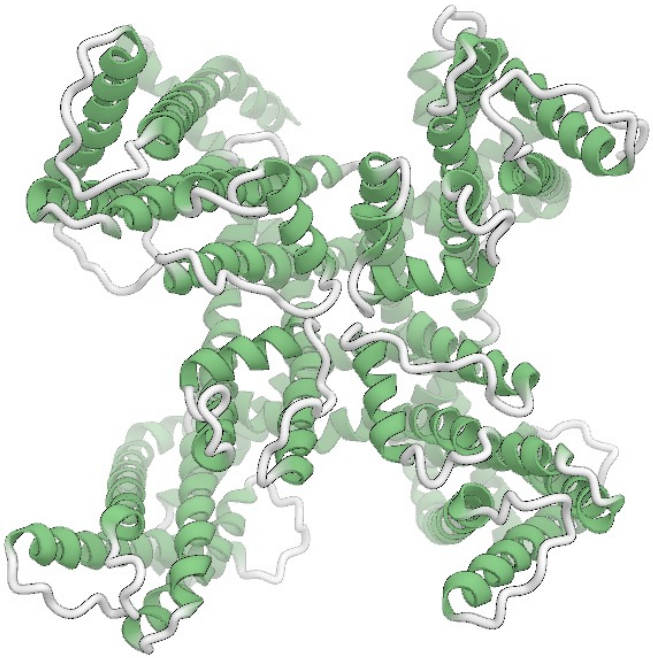
- No more low hanging fruit
- 90% of drug candidates fail in clinic
- Dozens of major diseases with well-understood biology remain undruggable

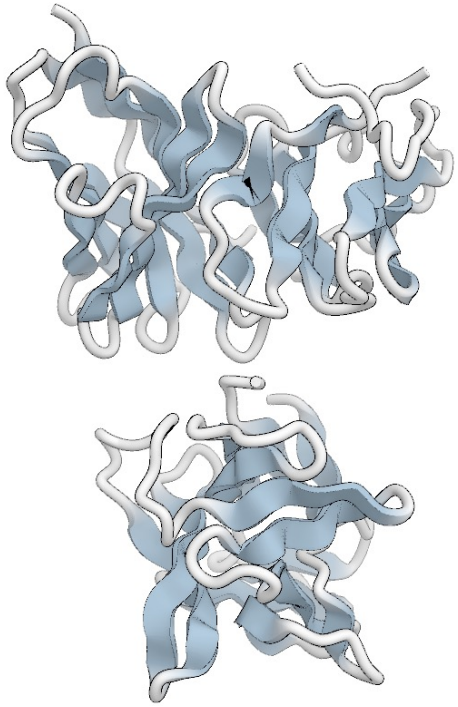


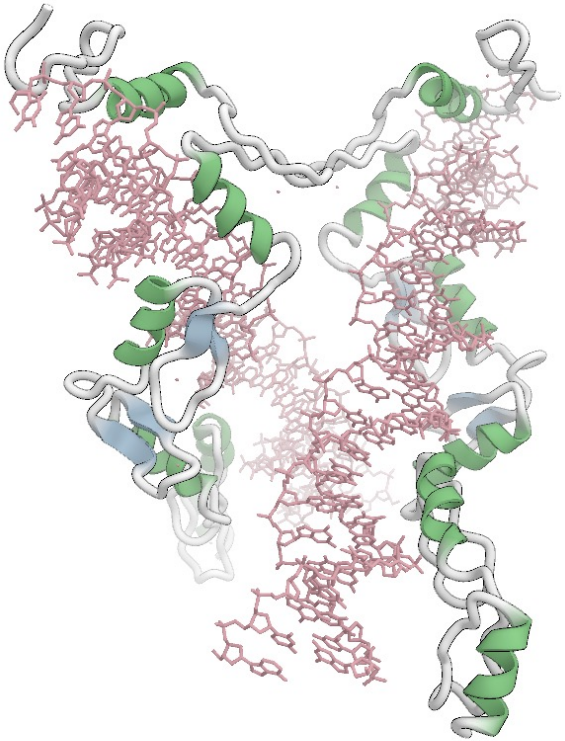
Proteins: The molecular machines of our body

Catalyzing chemical reactions
DNA replication

Inter- and intracellular signaling
And more!







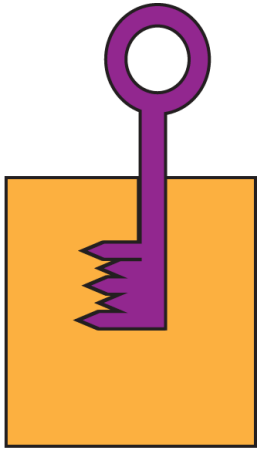
Small molecules modulate protein activity

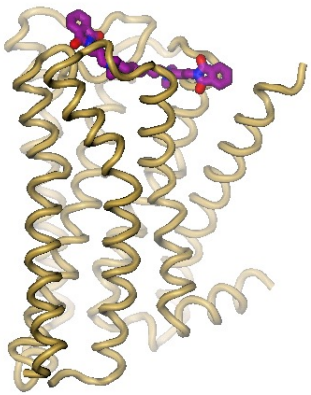
- Small molecules are like keys
  - They turn proteins on and off
  - They have a specific “key hole” that they fit into

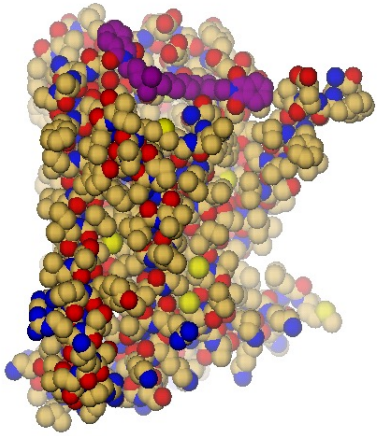
Schematic **Cartoon** **All-atoms**

Drug molecule

Protein



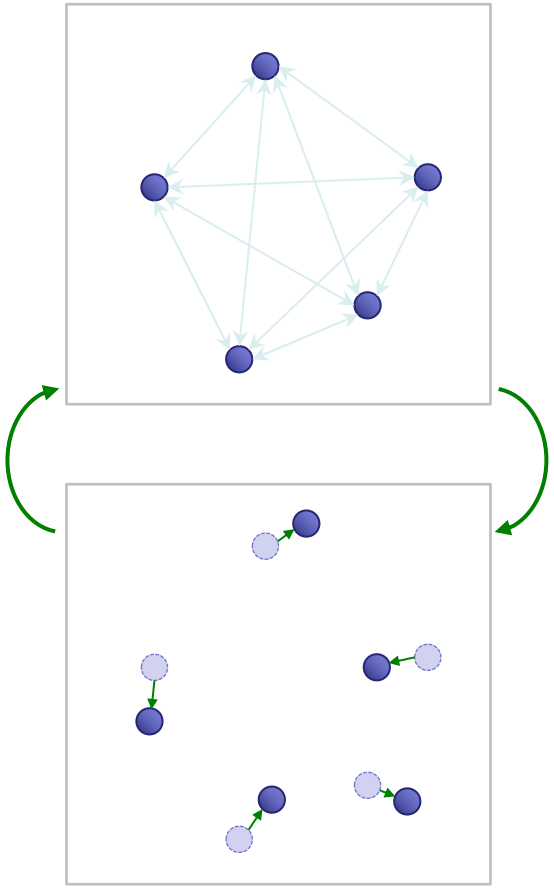




What is Molecular Dynamics?

$$F = ma$$

1. Compute the forces acting on every atom
2. Update positions and velocities based on Newton's laws of motion
3. Repeat!

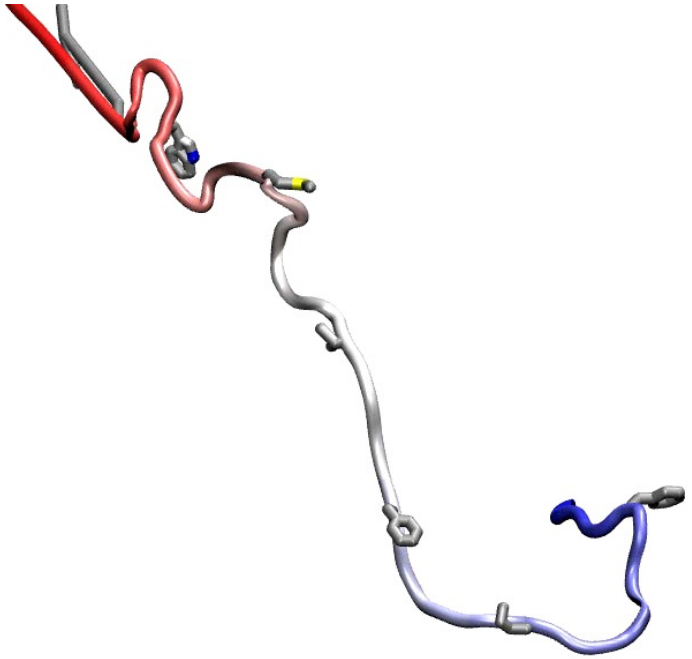


The anatomy of a simulation

Simulation provides researchers with a computational microscope

With sufficiently long and accurate simulations, you can:

- "Watch" biomolecular mechanisms
- "Watch" proteins fold
- "Watch" proteins interact with other proteins or with drug molecules



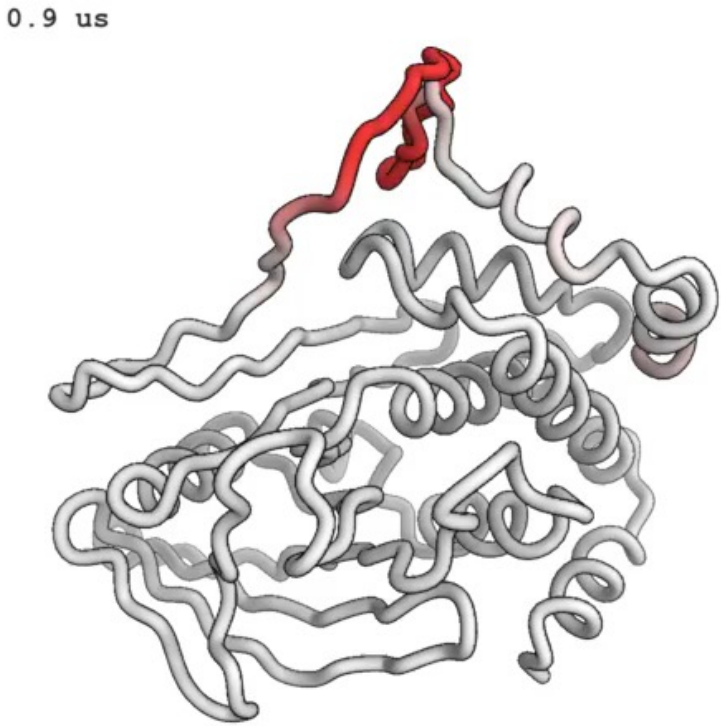
Ab initio protein folding

Lindorff-Larsen *et al*. (2011) *Science*

The Goals of Computational Drug Discovery

The road to developing a new drug begins with discovering:

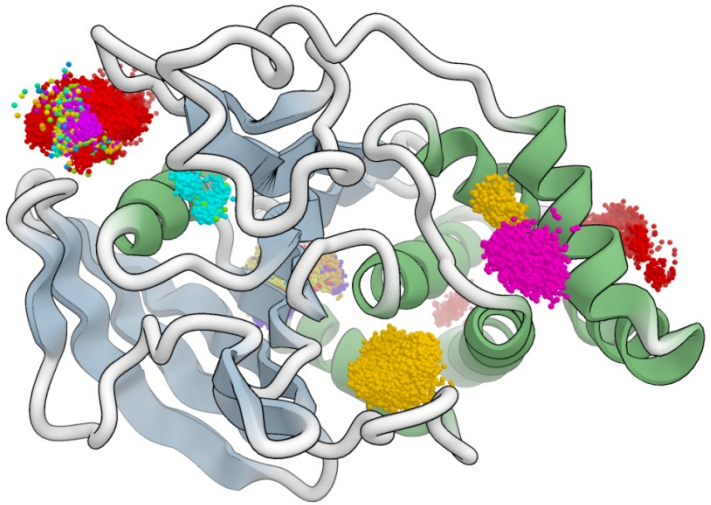
- A good binding pocket
- A good binder



The Goals of Computational Drug Discovery

The road to developing a new drug begins with discovering:

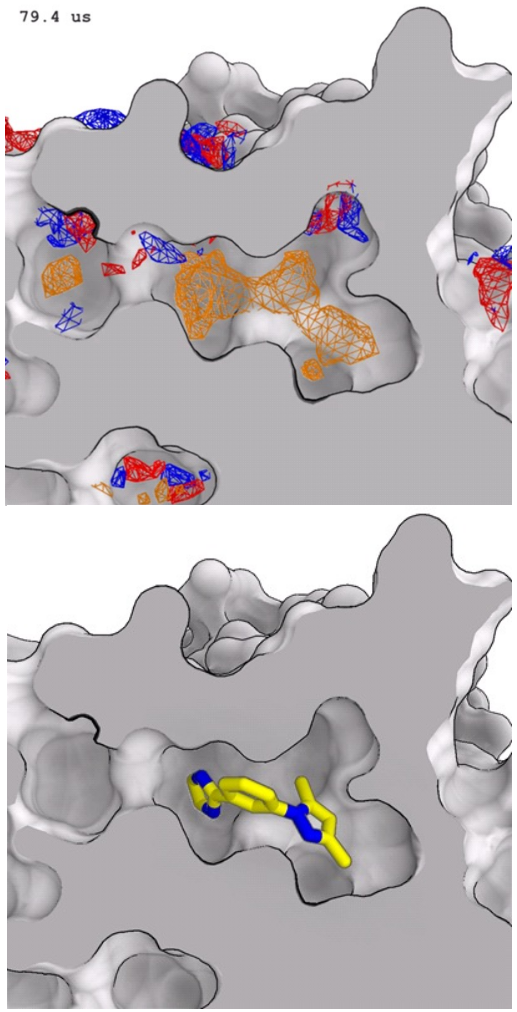
- A good binding pocket
- A good binder



The Goals of Computational Drug Discovery

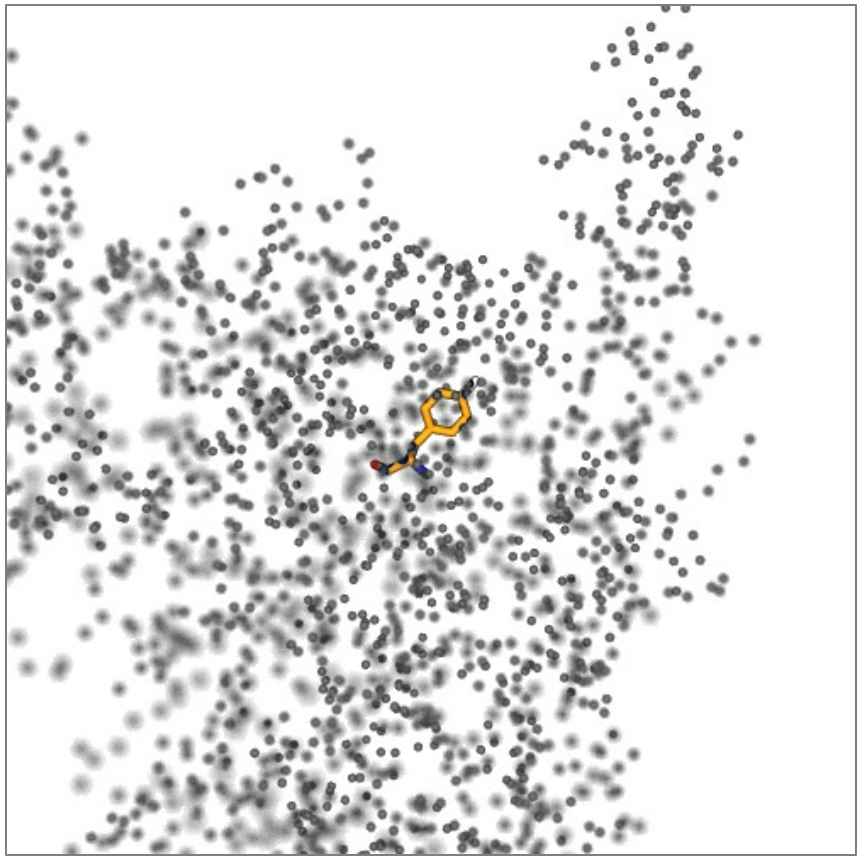
The road to developing a new drug begins with discovering:

- A good binding pocket
- A good binder



What's so hard?

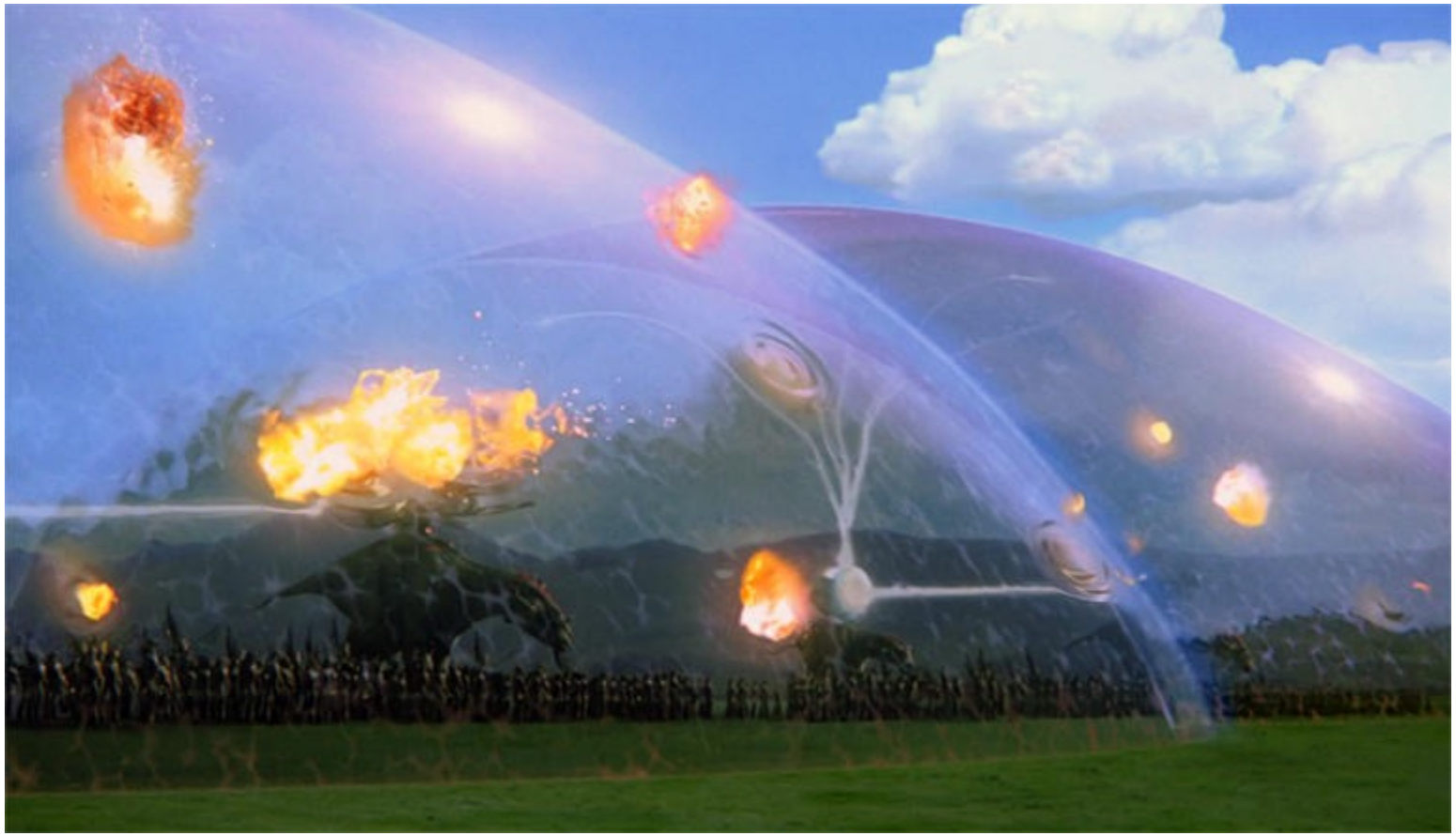
- Sampling
  - Many body problem (10^4 – 10^6 atoms)
  - Limiting frequency is X-H stretch: ~10 fs (1 fs = 10^{-15} s)
- Force field design
  - Model development
  - Model parameterization



Force fields in popular culture

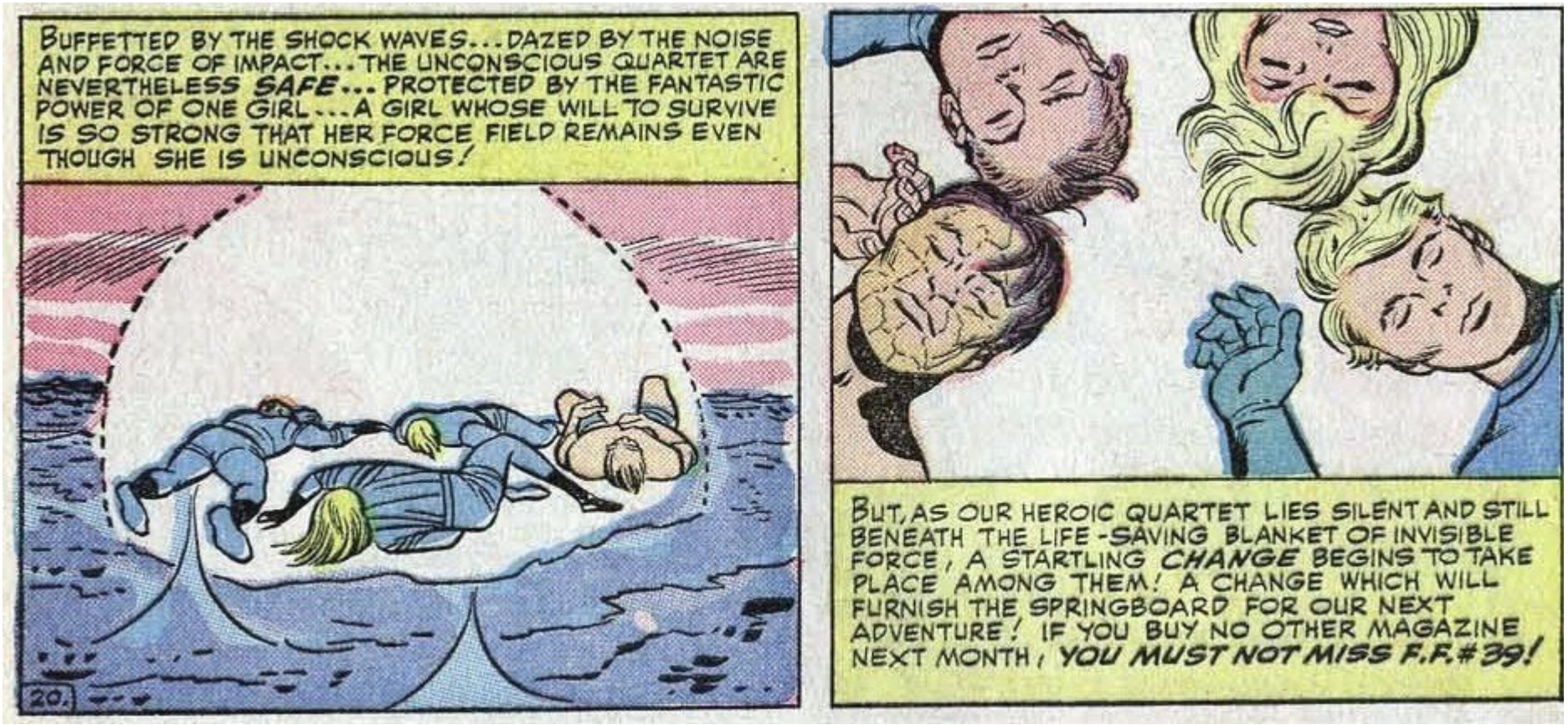
Violet Parr (of The Incredibles)

Force fields in popular culture



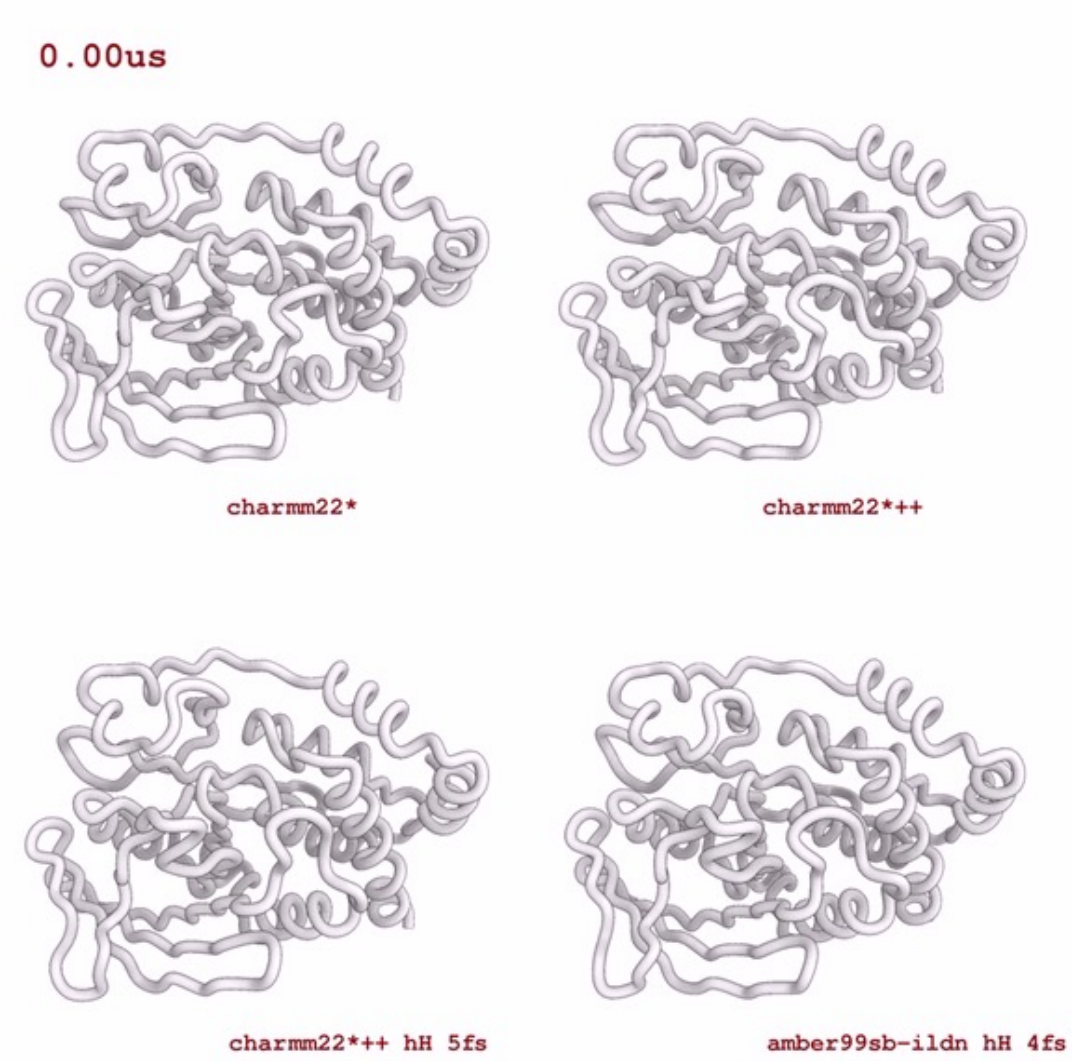
Battle of Naboo from "Star Wars Episode I: The Phantom Menace."

Force fields in popular culture

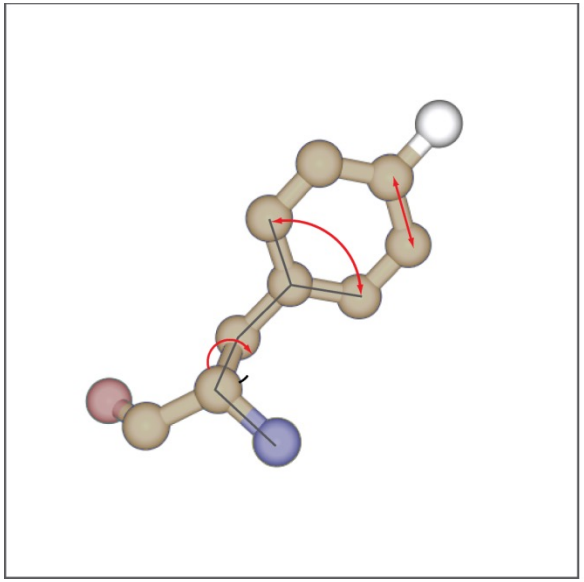


The Invisible Woman (of the Fantastic Four)

Loopy simulations!



Molecular mechanics force field

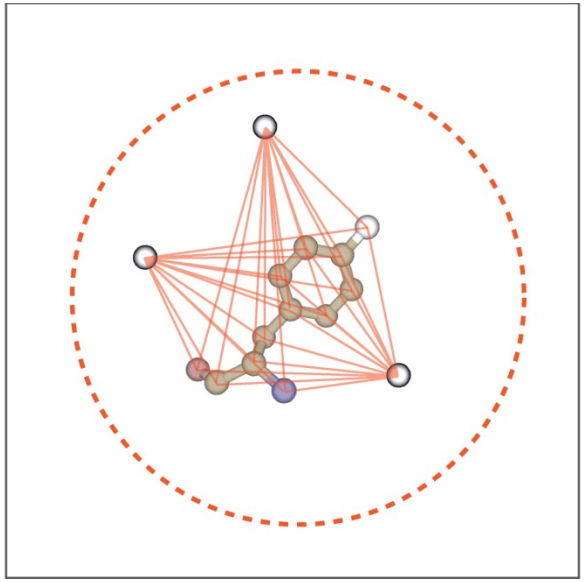


bonded terms

$$\sum_{\text{bonds}} k_b (r - r_0)^2$$

$$+ \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2$$

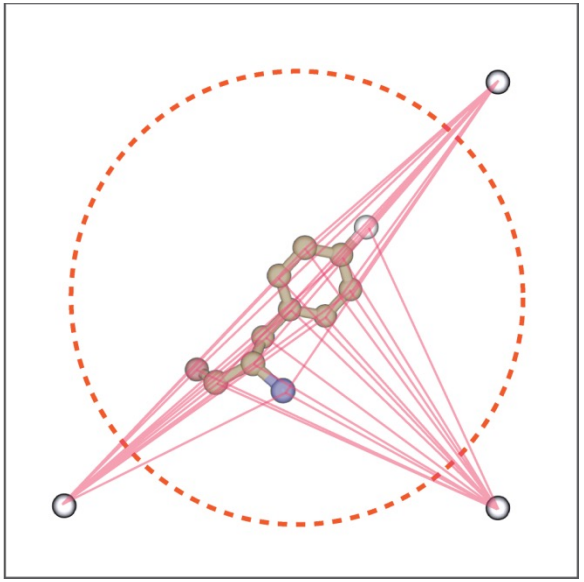
$$+ \sum_{\text{torsions}} A[1 - \cos(n\tau - \varphi)]$$



near / local non-bonded:
VdW & electrostatic

$$\sum_{i,j<i} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

$$+ \sum_{i,j<i} \frac{q_i q_j}{r_{ij}}$$

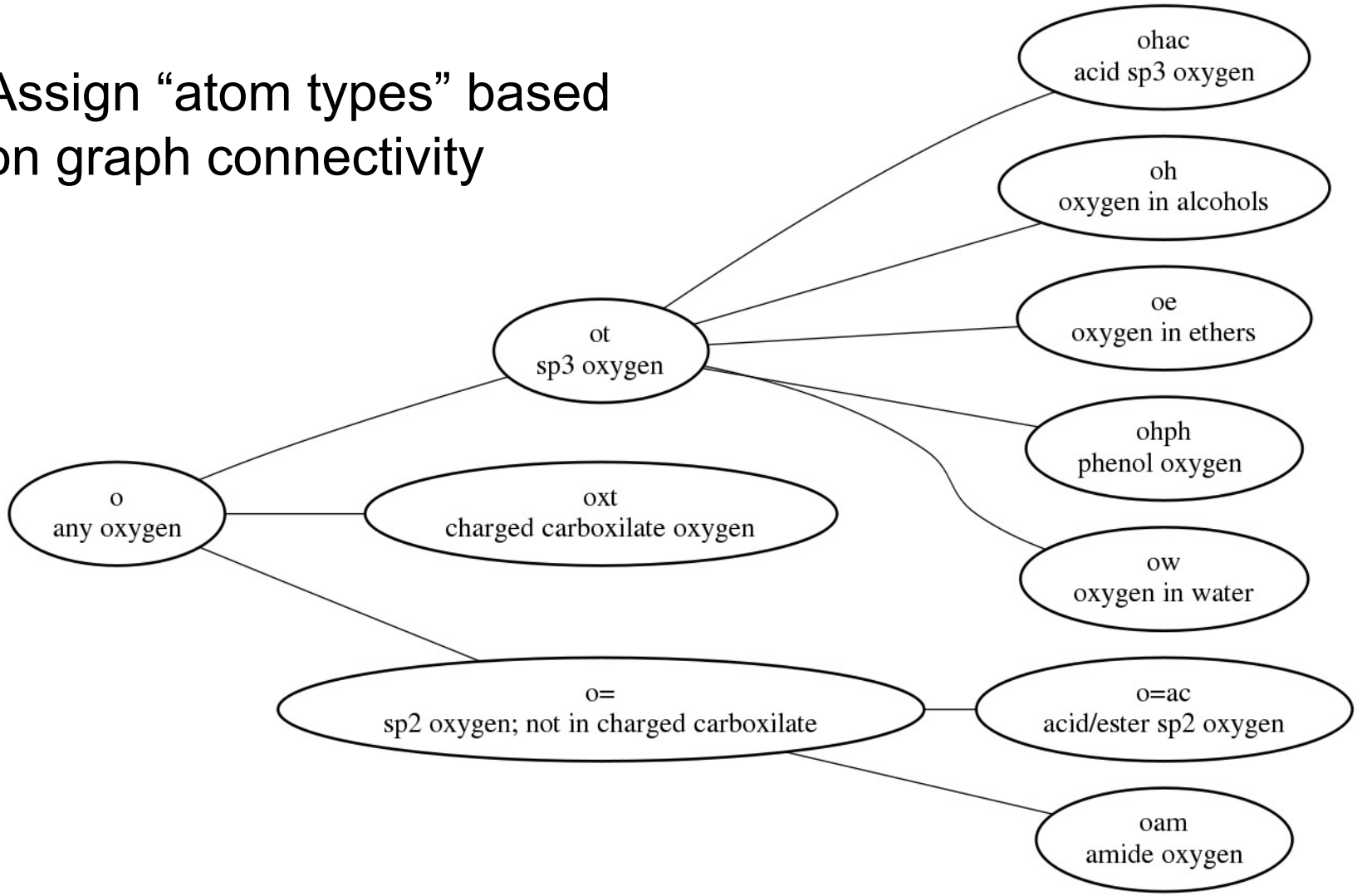


far / global non-bonded:
electrostatic

$$\text{FT}\left[\sum_{i,j<i} \frac{q_i q_j}{r_{ij}}\right]$$

Ab initio force field development

Assign “atom types” based on graph connectivity



Ab initio force field development

- Search through parameter space for the (global) optimum

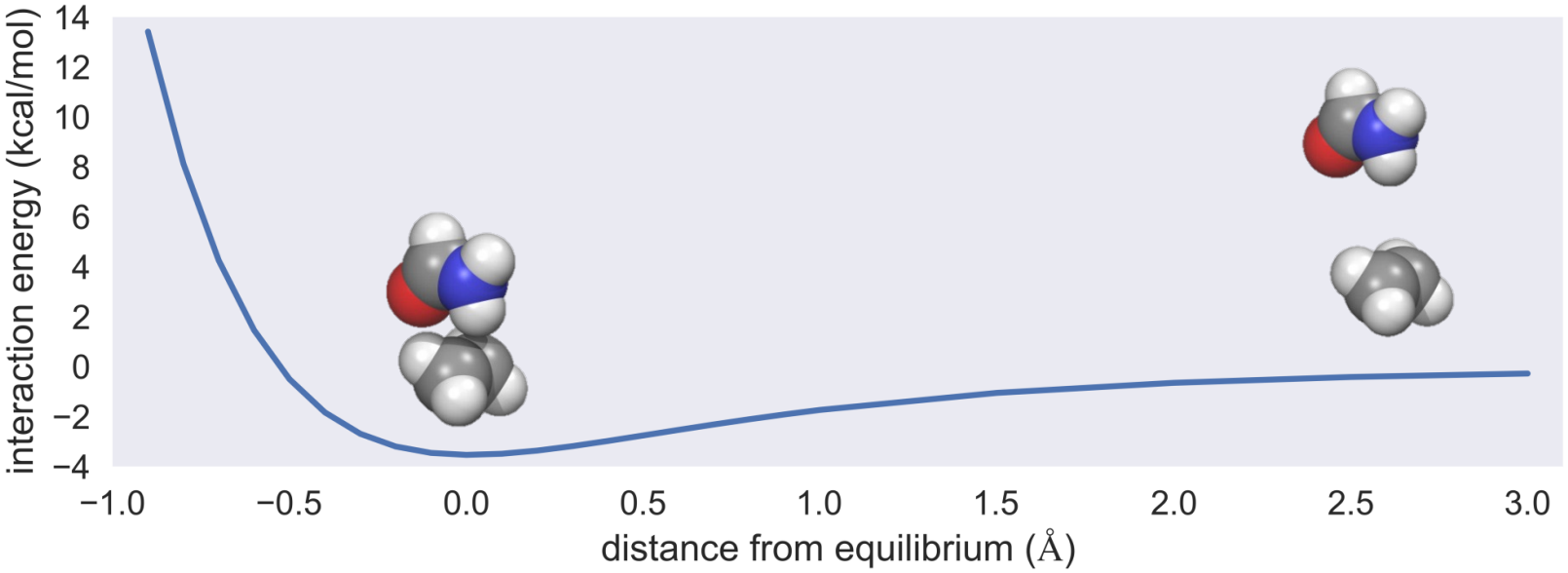
$$\text{obj} = \sum_{i=1}^{N} w_i(\hat{x}_i - x_i)^2 \text{ where } \hat{x}_i\text{: MM, } x_i\text{: QM}$$

- Energies, geometries, monomer properties, etc.
- Weights are often Boltzmann: $w_i = \exp(-\frac{E_i}{k_B T})$
- Validate with experimental observables
  - Density
  - Enthalpy of vaporization
  - Solvation free energy

Quantum Mechanics (QM) Data

- Quantum Chemistry
  - Møller–Plesset perturbation theory
  - Coupled cluster
- Example: Interaction energy

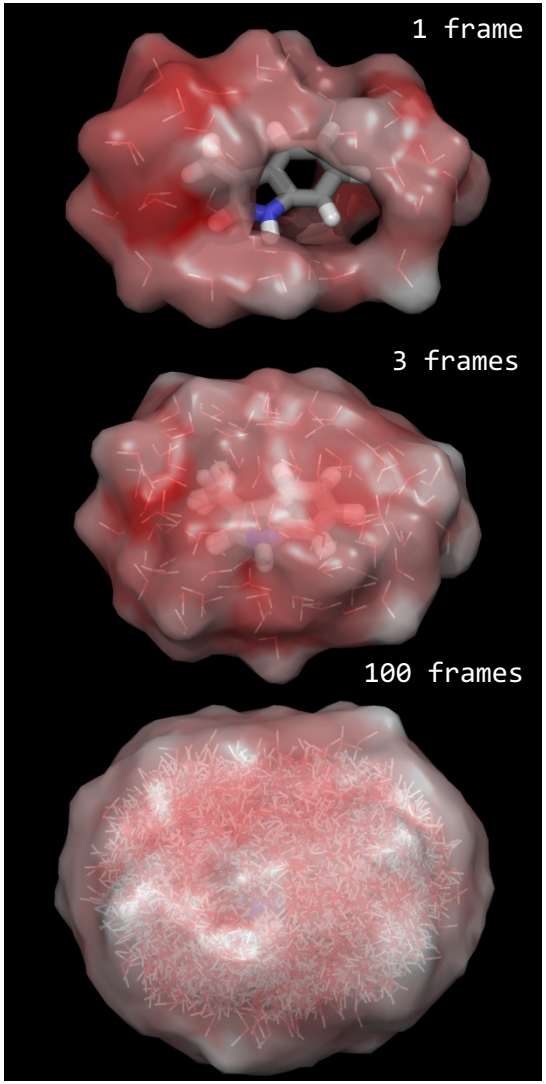
$$\Delta E = E_{\mathrm{AB}} - E_{\mathrm{A}} - E_{\mathrm{B}}$$



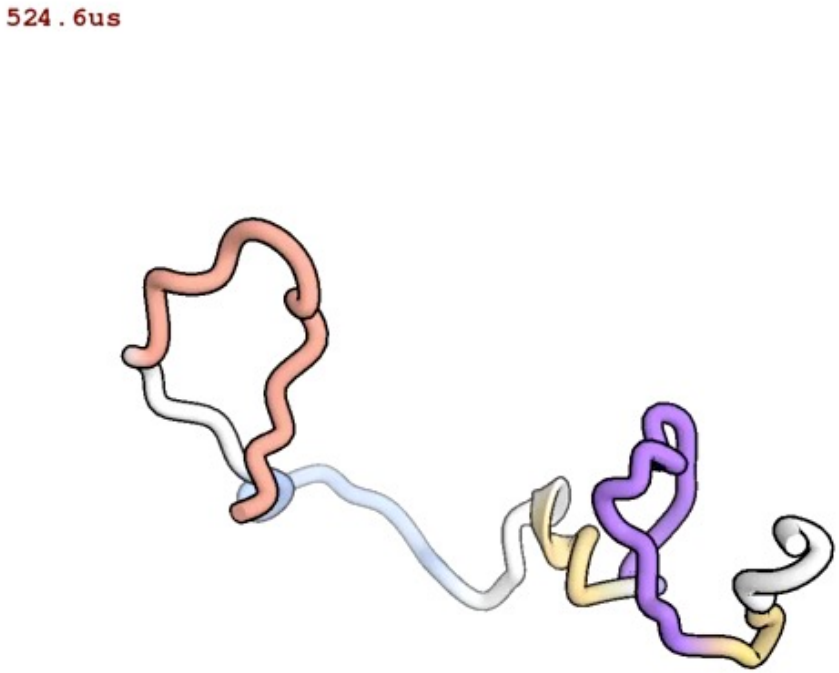
Quantum Mechanics (QM) Data

- Quantum Chemistry
  - Møller–Plesset perturbation theory
  - Coupled cluster
- Example: Micrsolvation energy

$$E_{\text{solvation}} = E_{\text{all}} - E_{\text{solvent}} - E_{\text{solute}}$$

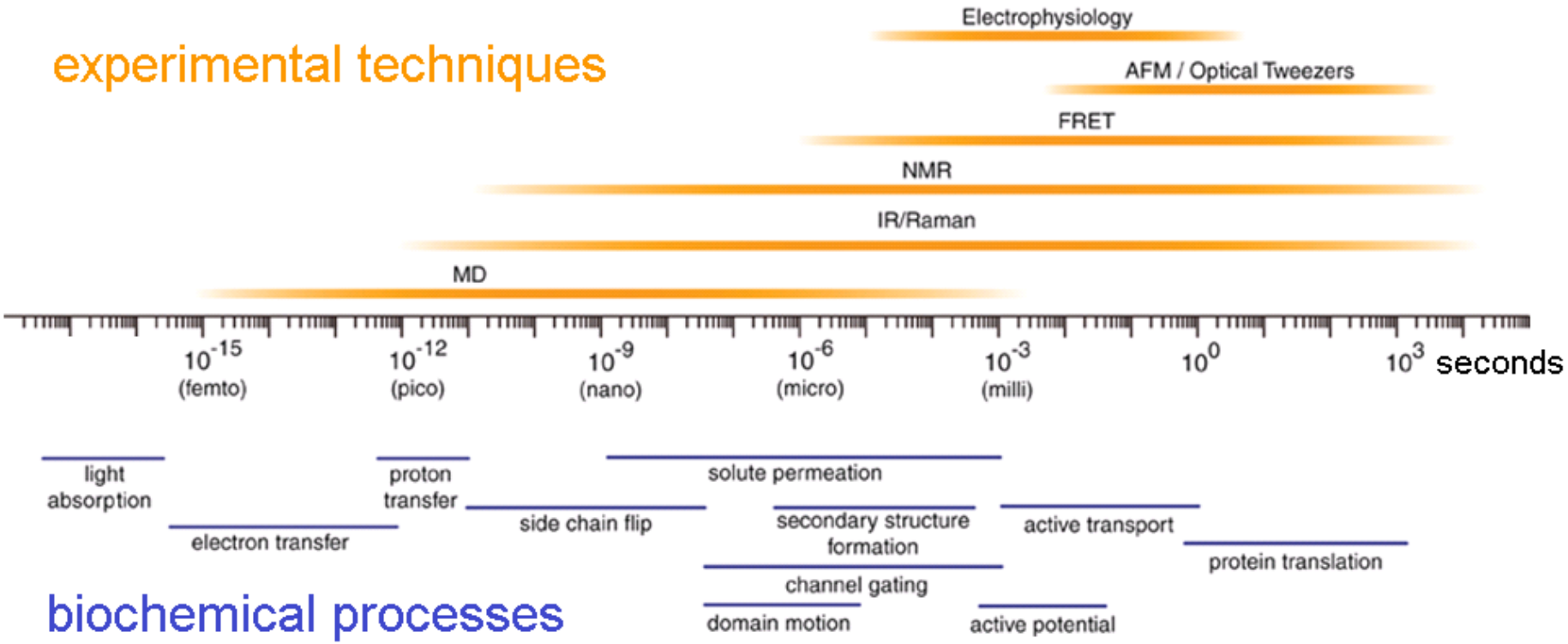


But can we correctly fold proteins?




ANTON2

How long does biological stuff take?



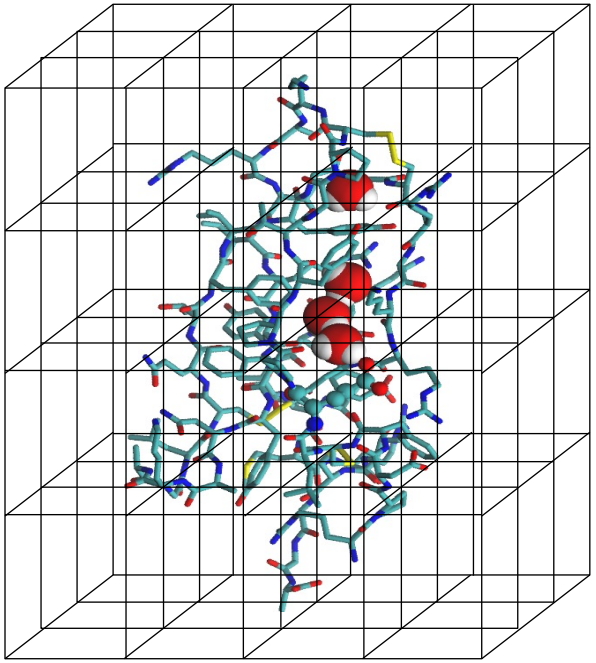
Sources of Speedup on Anton

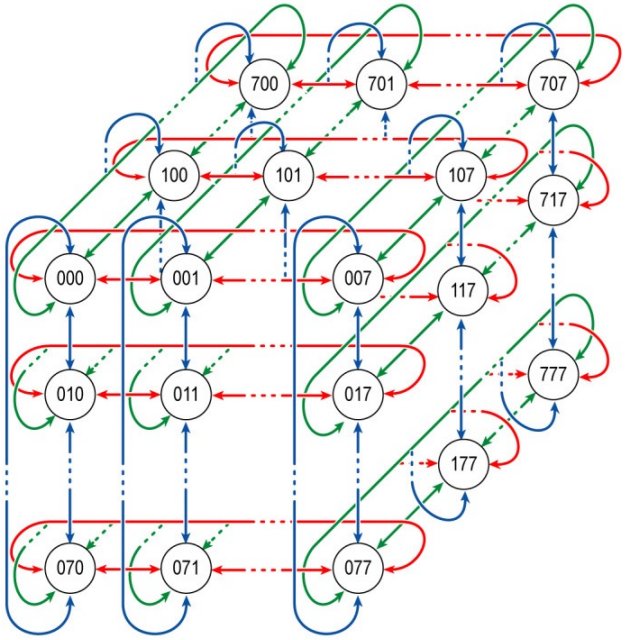
- Parallelization in the space domain
  - Up to 2K separate computation nodes working all at once
  - Huge networking and systems task
- Judicious use of arithmetic specialization
  - Tailored for speed: Custom ASIC designed in-house
  - Flexibility, programmability only where needed
- Fast, Carefully choreographed communication
  - Data flows to just where it's needed
  - Almost never need to access off-chip memory

Sources of Speedup on Anton

- Parallelization in the space domain
  - Up to 2K separate computation nodes working all at once
  - Huge networking and systems task
- Judicious use of arithmetic specialization
  - Tailored for speed: Custom ASIC designed in-house
  - Flexibility, programmability only where needed
- Fast, Carefully choreographed communication
  - Data flows to just where it's needed
  - Almost never need to access off-chip memory

Spatial Decomposition





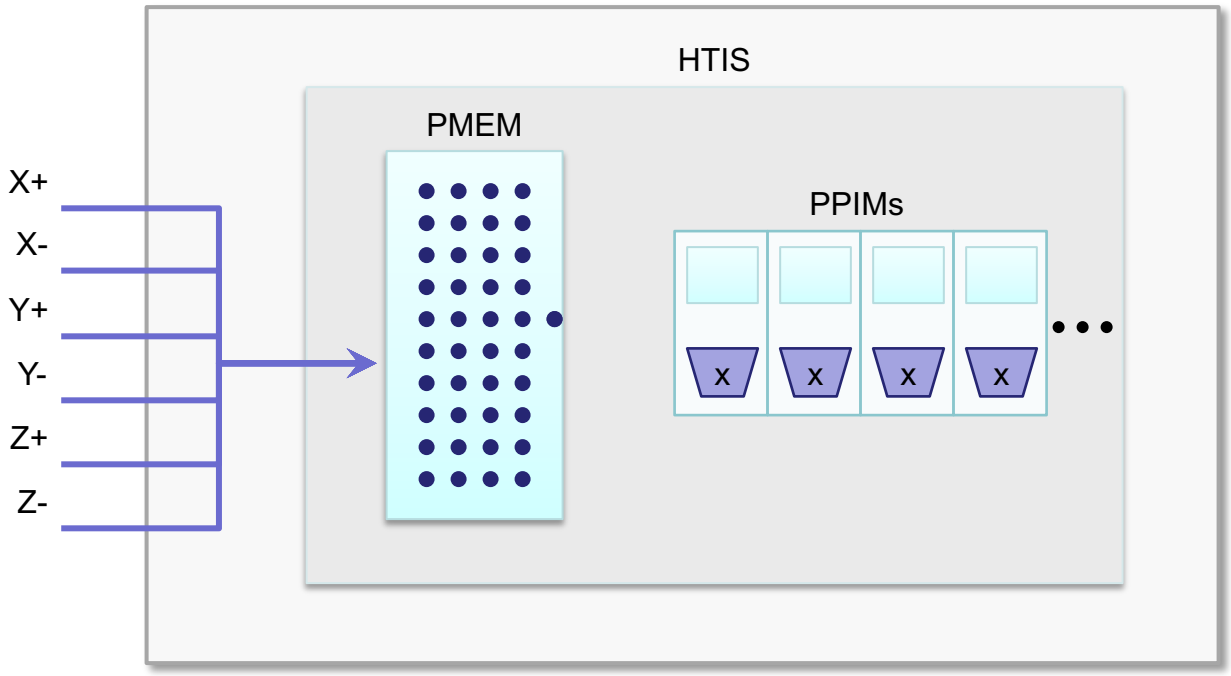
D. E. Shaw, “Exploiting 162-Nanosecond End-to-End Communication Latency on Anton”

Sources of Speedup on Anton

- Parallelization in the space domain
  - Up to 2K separate computation nodes working all at once
  - Huge networking and systems task
- Judicious use of arithmetic specialization
  - Tailored for speed: Custom ASIC designed in-house
  - Flexibility, programmability only where needed
- Fast, Carefully choreographed communication
  - Data flows to just where it's needed
  - Almost never need to access off-chip memory

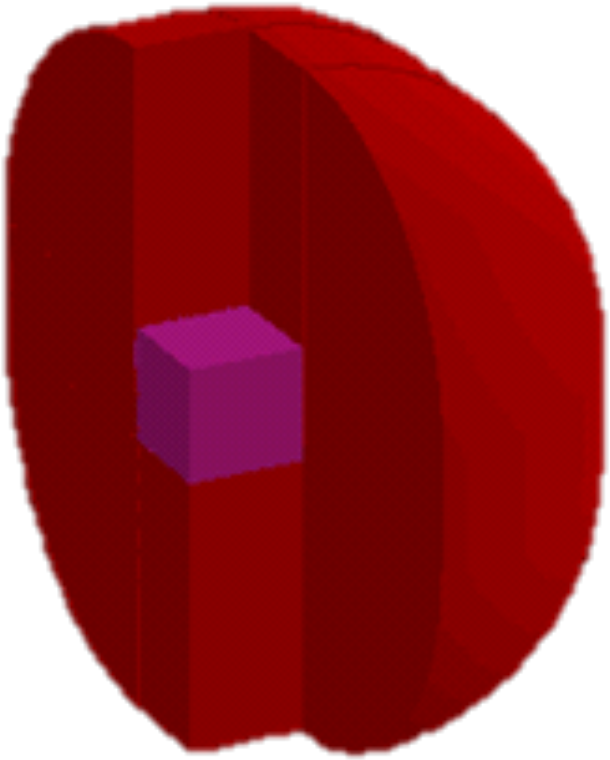
Particle-Particle Interaction in Hardware

1) Place the atoms in particle memory (PMEM)
2) Store one set of atoms in the PPIMs
3) Stream another set through the PPIMs

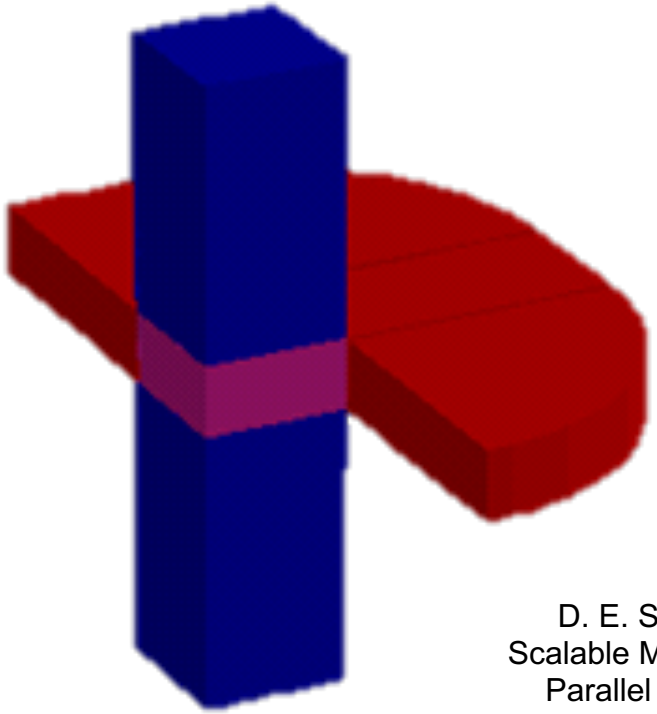


D. E. Shaw, "High-Throughput Pairwise Point Interactions in Anton, a Specialized Machine for Molecular Dynamics Simulation"

Compute Interactions on Neutral Territory



Traditional Method

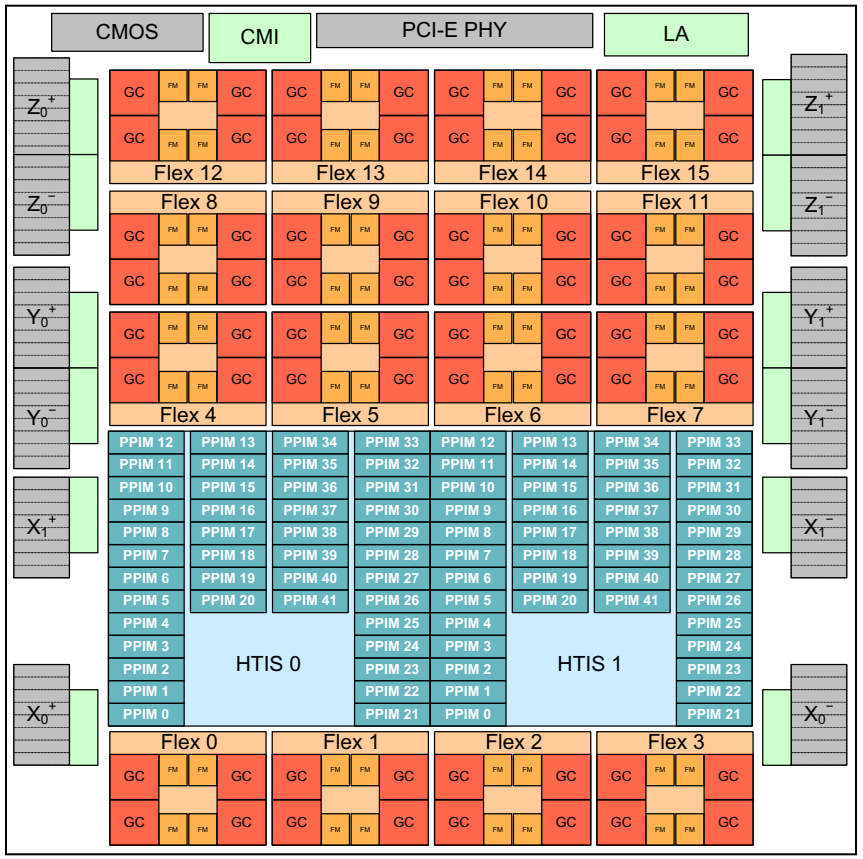


NT Method

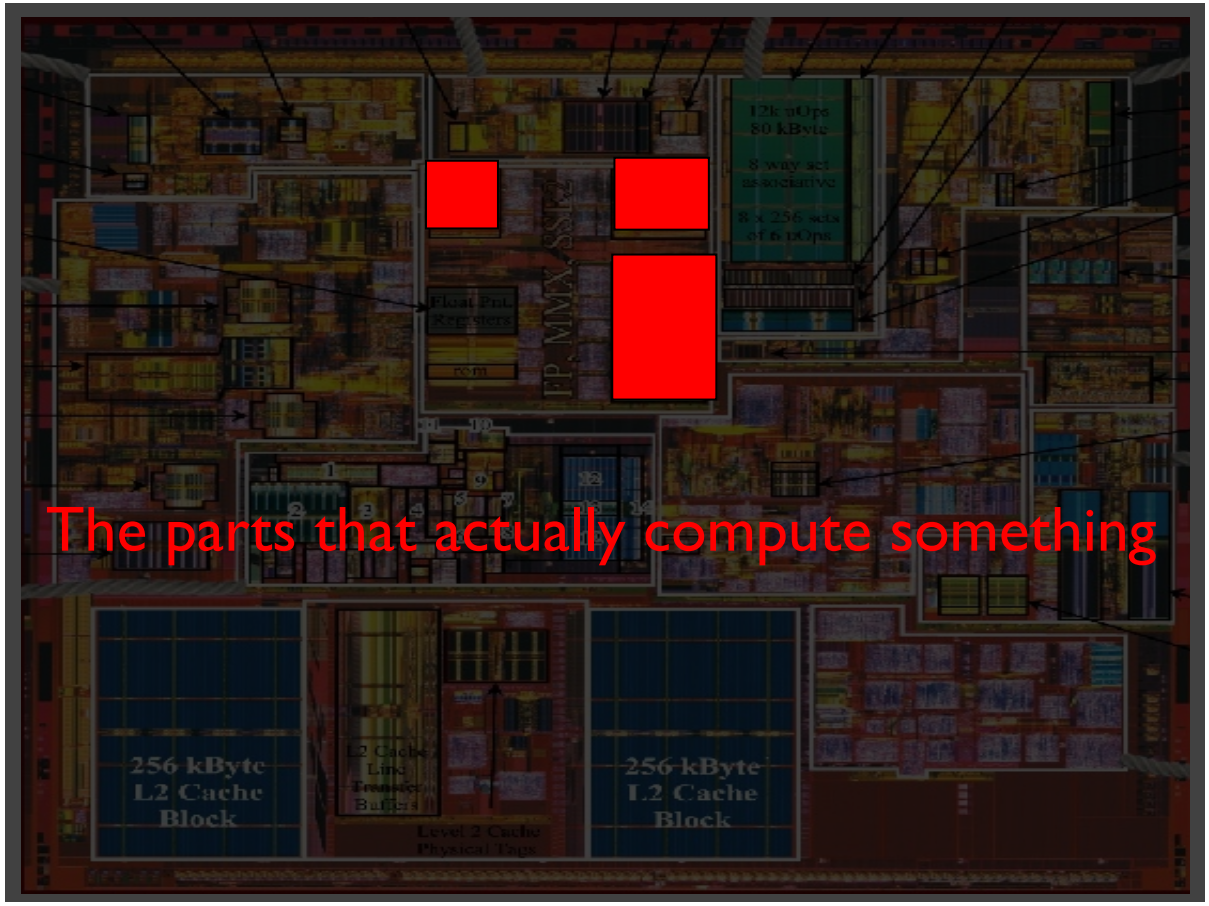
D. E. Shaw, “A Fast, Scalable Method for the Parallel Evaluation of Distance-Limited Pairwise Particle Interactions”, *J. Comput. Chem.,* 2005

What Does it Actually Look Like?

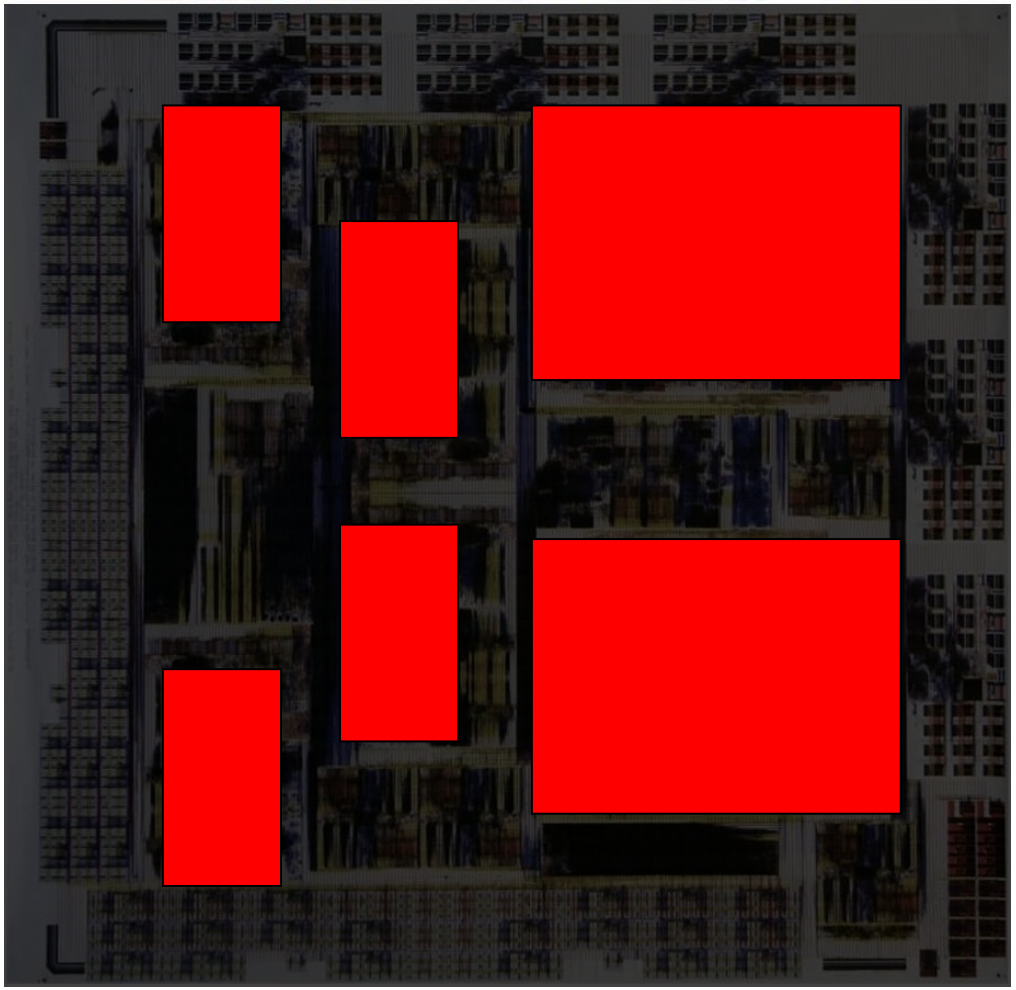
- Floorplan of entire chip
- Broken up into blocks with different functions
- Many PPIMs for interacting particles



Chip Real Estate: Commodity CPU



Chip Real Estate: Anton ASIC

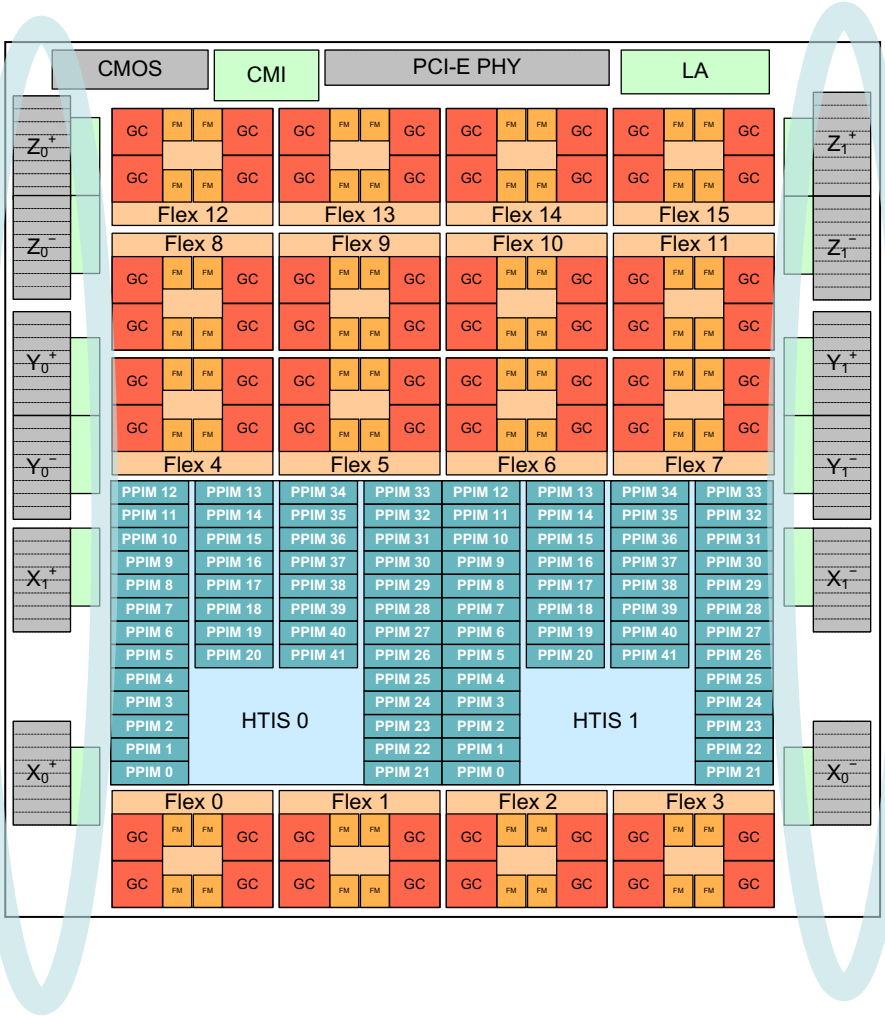


Sources of Speedup on Anton

- Parallelization in the space domain
  - Up to 2K separate computation nodes working all at once
  - Huge networking and systems task
- Judicious use of arithmetic specialization
  - Tailored for speed: Custom ASIC designed in-house
  - Flexibility, programmability only where needed
- Fast, Carefully choreographed communication
  - Data flows to just where it's needed
  - Almost never need to access off-chip memory

On-chip High-Speed Network

- Quickly serialize and deserialize communications
- Dedicated hardware for each network dimension
- Very uncommon in computer architecture



Very Low Communication Latency

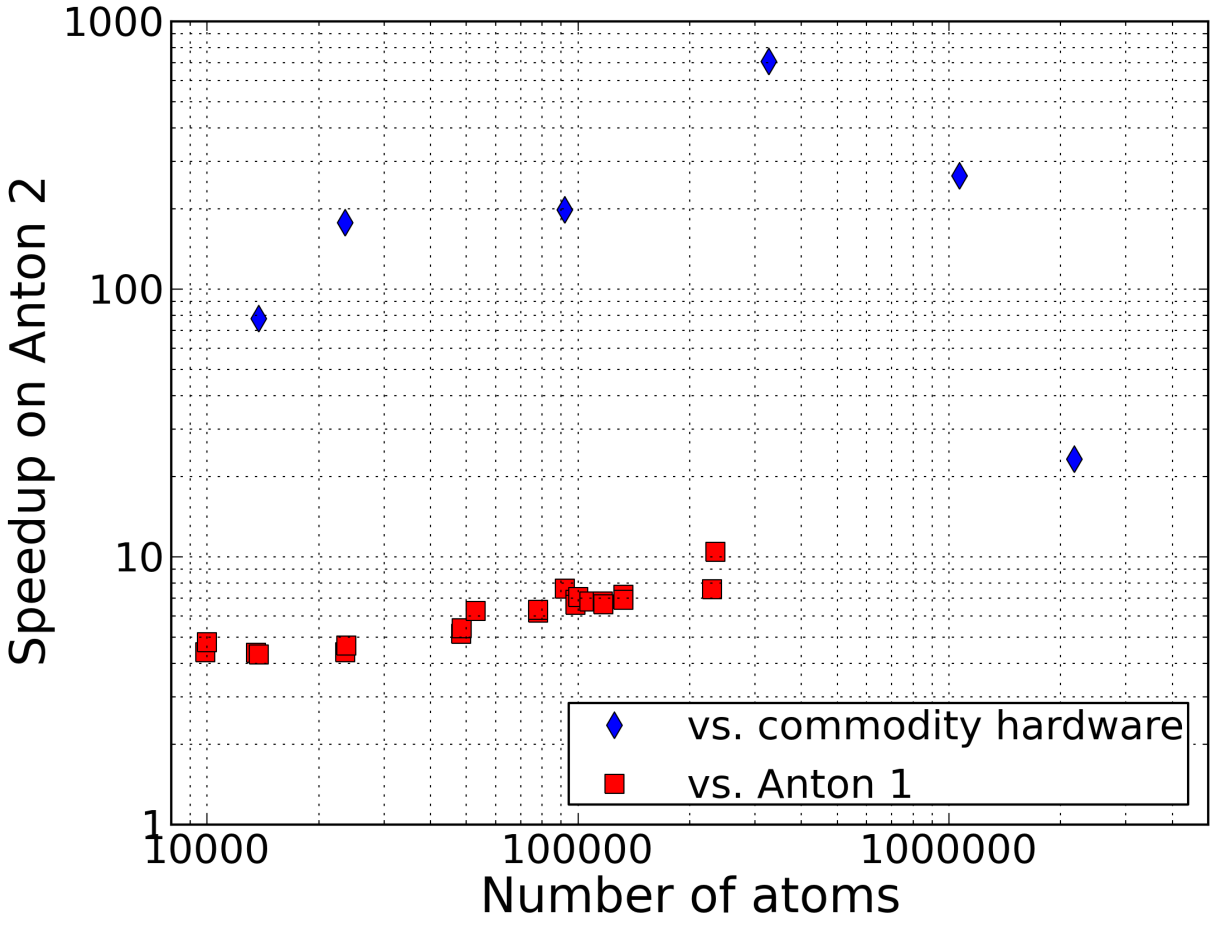
Machine name	Latency (μs)	Date
Anton	0.16	2009
Altix 3700 BX2	1.25	2006
QsNet^{II}	1.28	2005
Columbia	1.6	2005
Sun Fire	1.7	2002
EV7	1.7	2002
J-Machine	1.8	1993
QsNET	1.9	2001
Roadrunner (InfiniBand)	2.16	2008
Cray T3E	2.75	1996
Blue Gene/P	2.75	2008
Blue Gene/L	2.8	2005
ASC Purple	4.4	2005
Cray XT4	4.5	2007
Red Storm	6.9	2005
SR8000	9.9	2001

Survey of published internode software-to-software latency measurements

- Custom on-chip and off-chip network
- Network interface directly on the ASIC
- Optimized for low latency, small messages
- Cut out the software stack

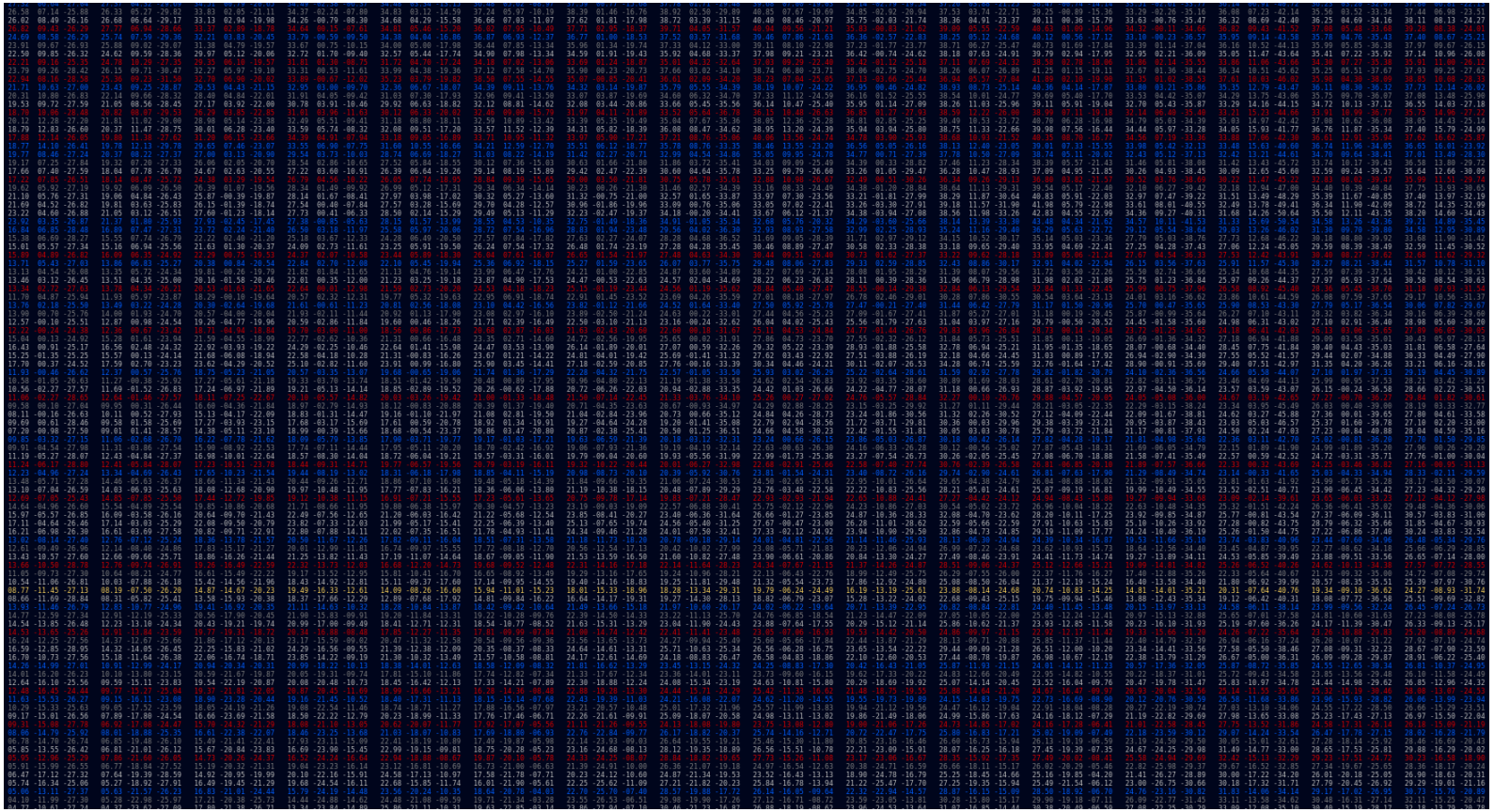
D. E. Shaw, "Exploiting 162-Nanosecond End-to-End Communication Latency on Anton"

Anton 2 Performance

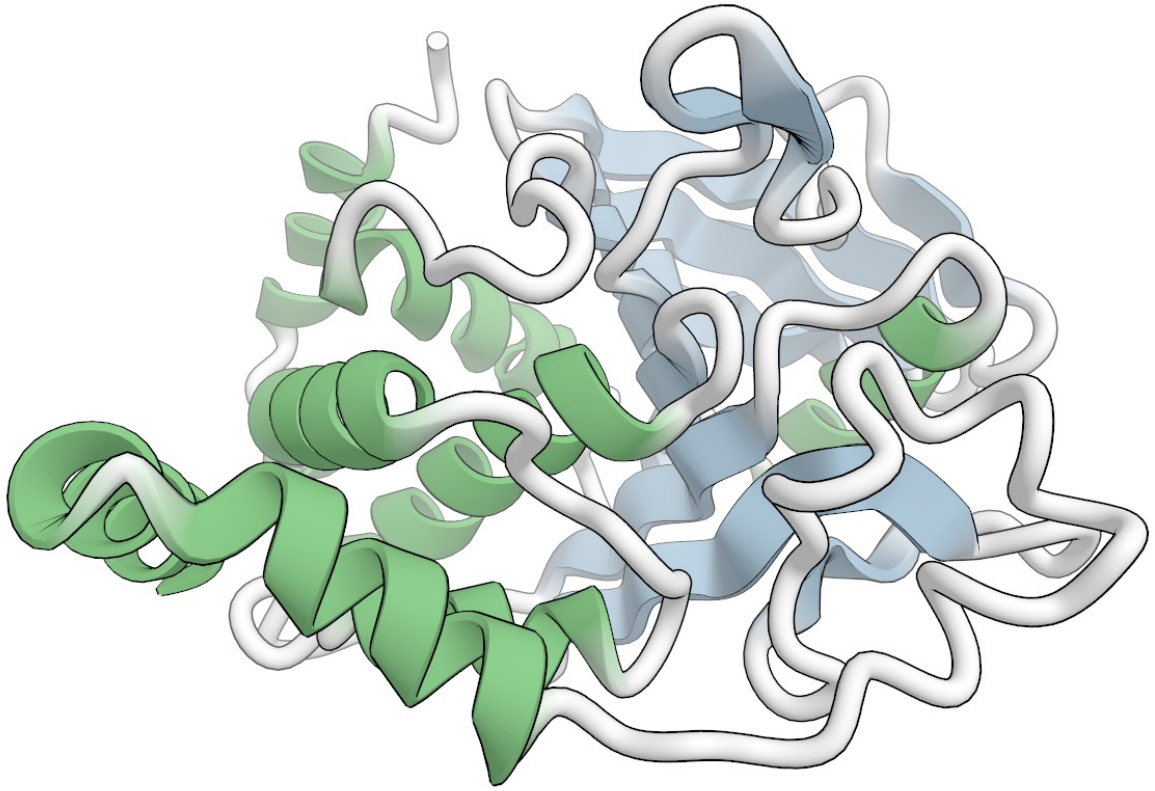


Visualization

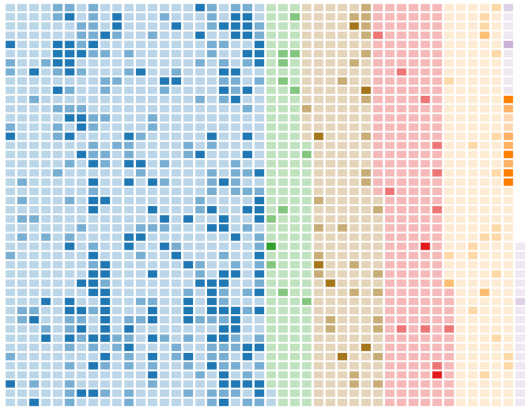
This is Molecular Dynamics

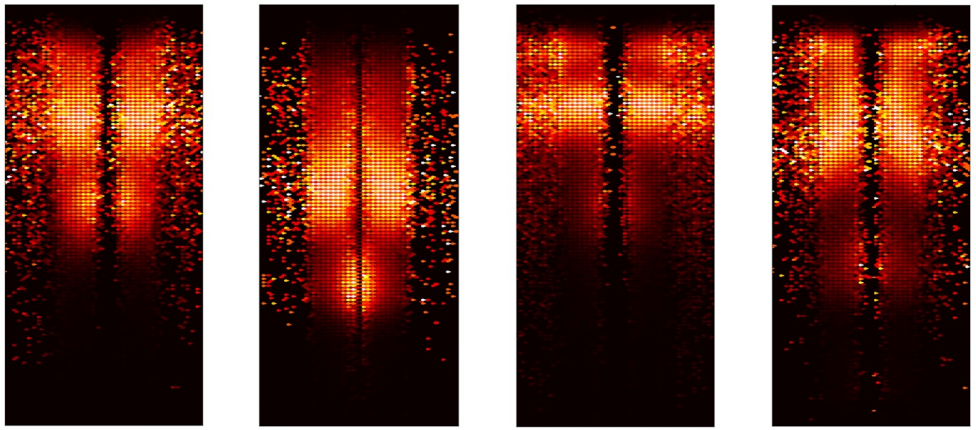


This is How a Scientist Sees a Protein



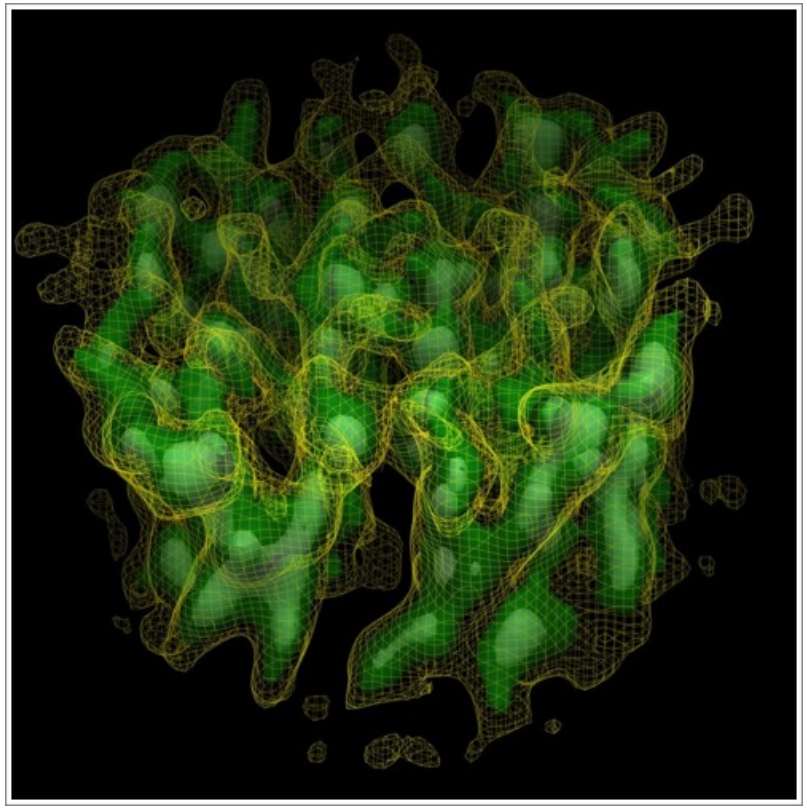
What is scientific visualization?



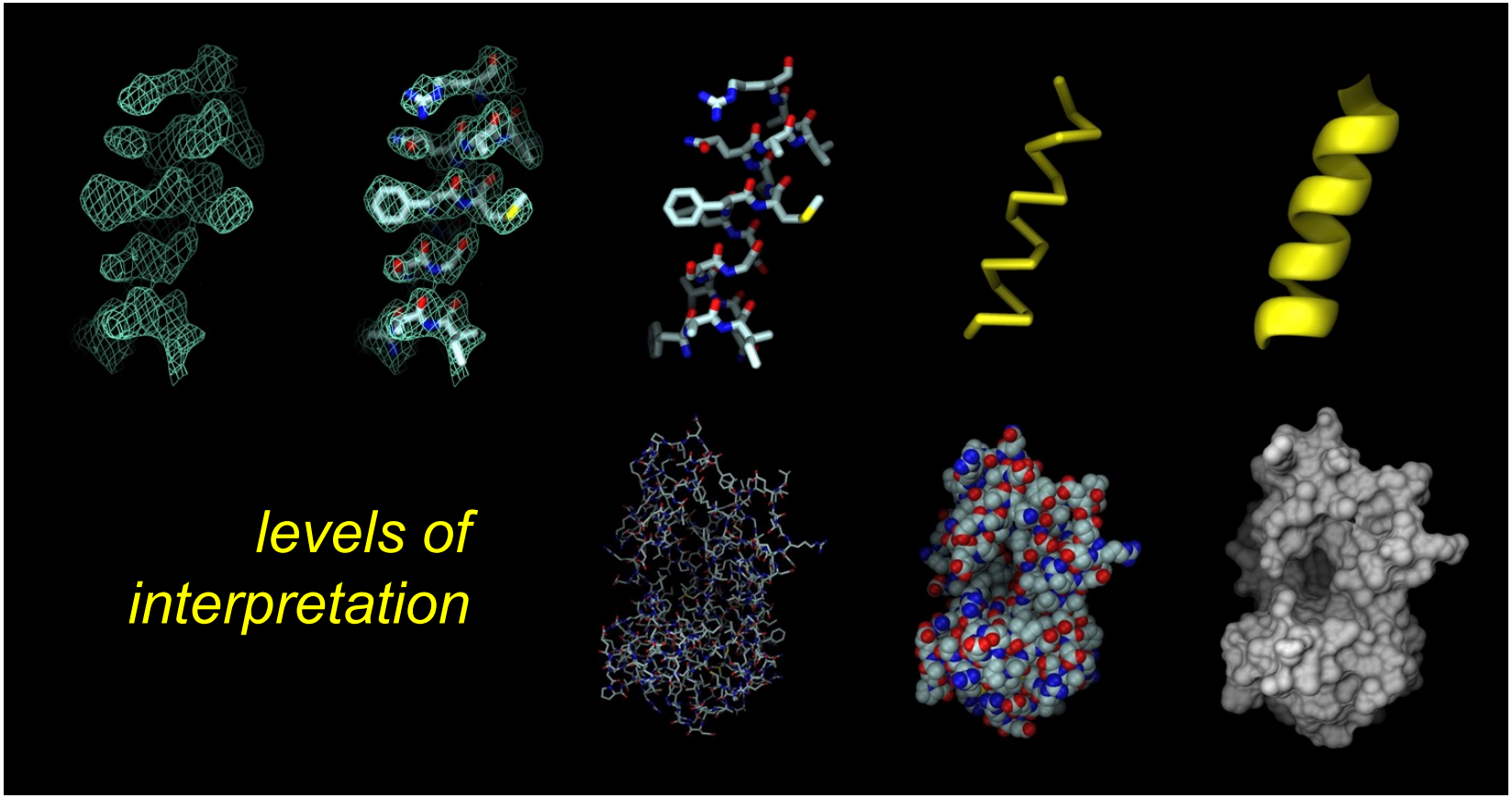


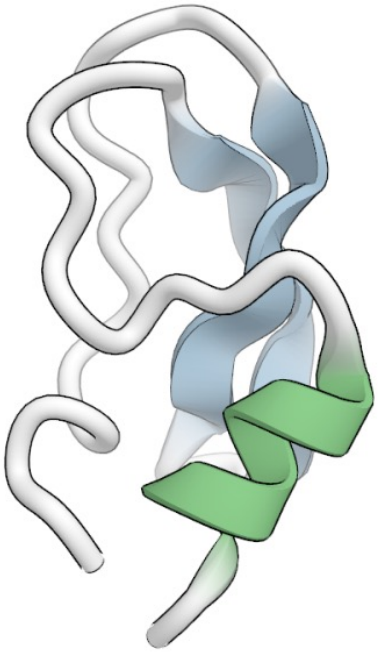
mapping abstract data to a visual representation

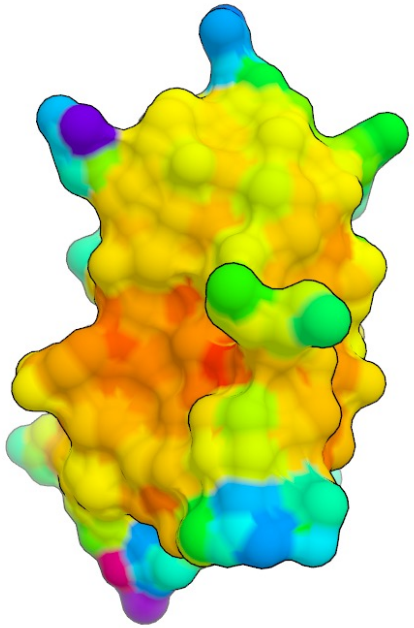
So, what do proteins look like?



X-Ray crystallography, NMR

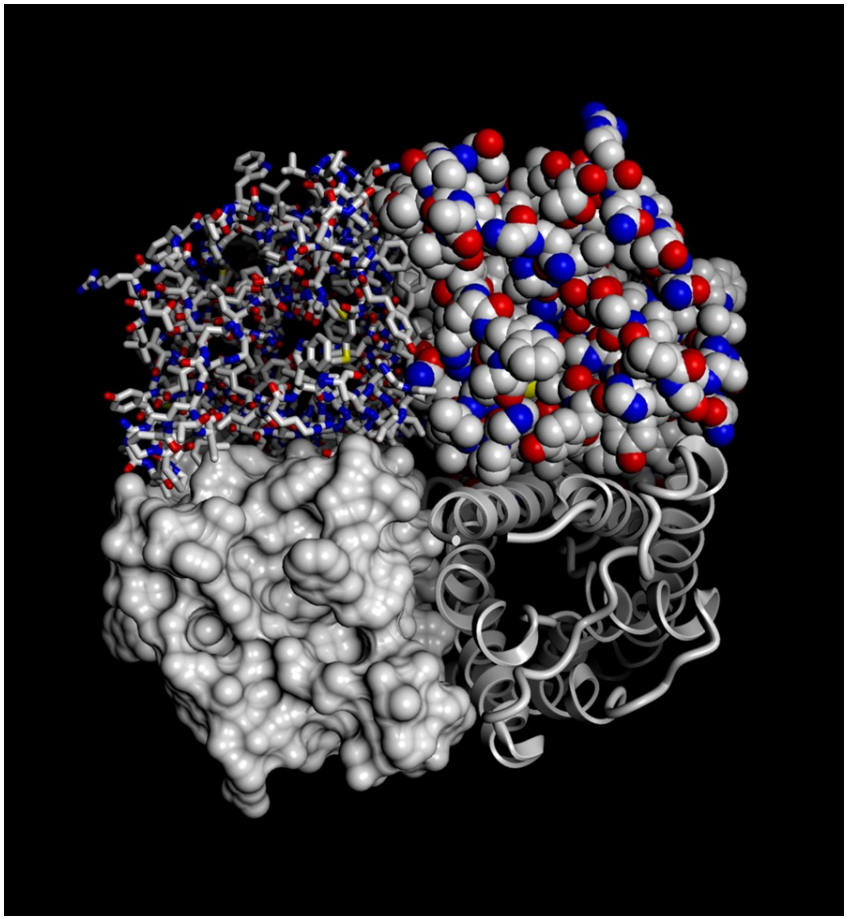

levels of
interpretation

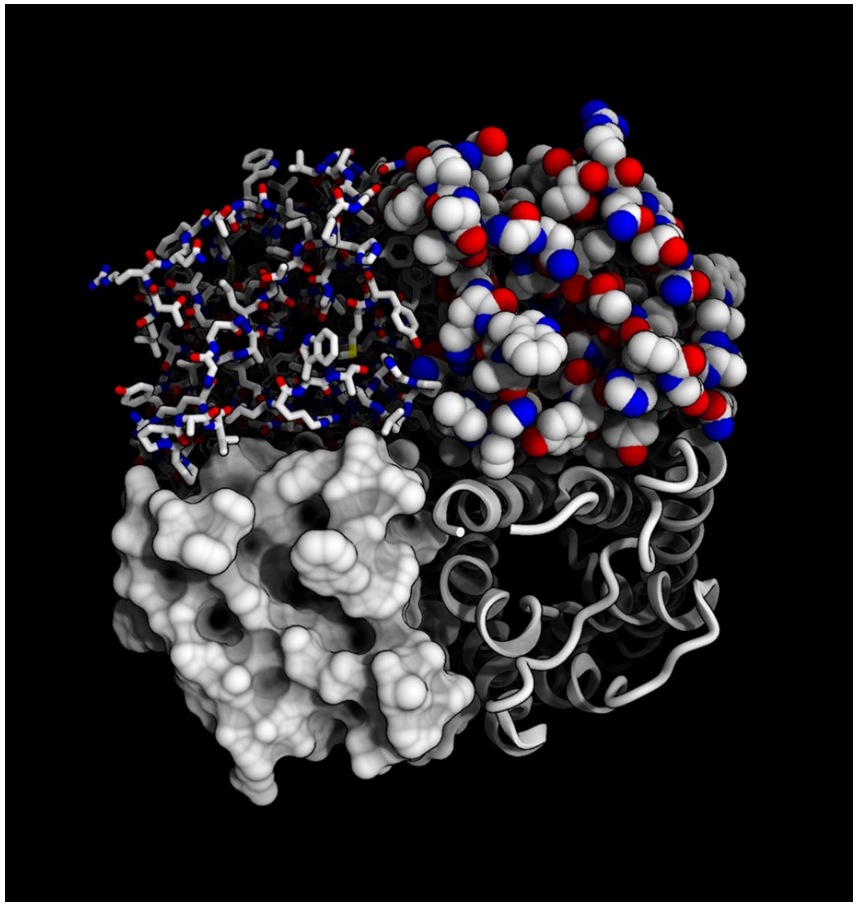




shape and color

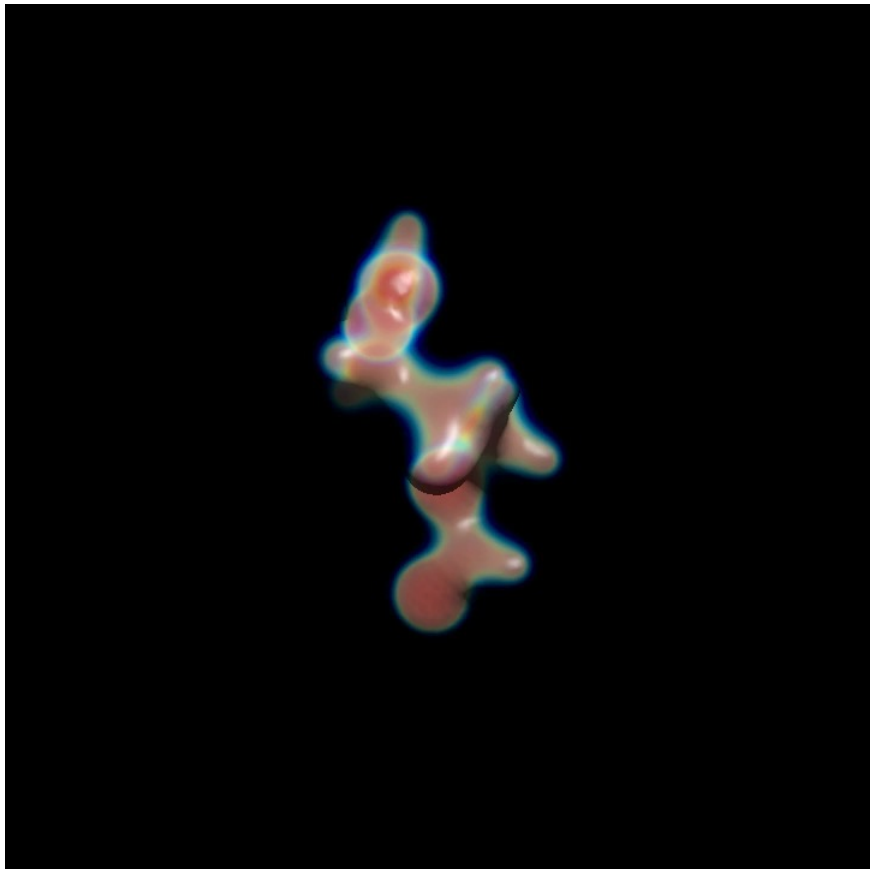
Physical Realism vs Scientific Visualization

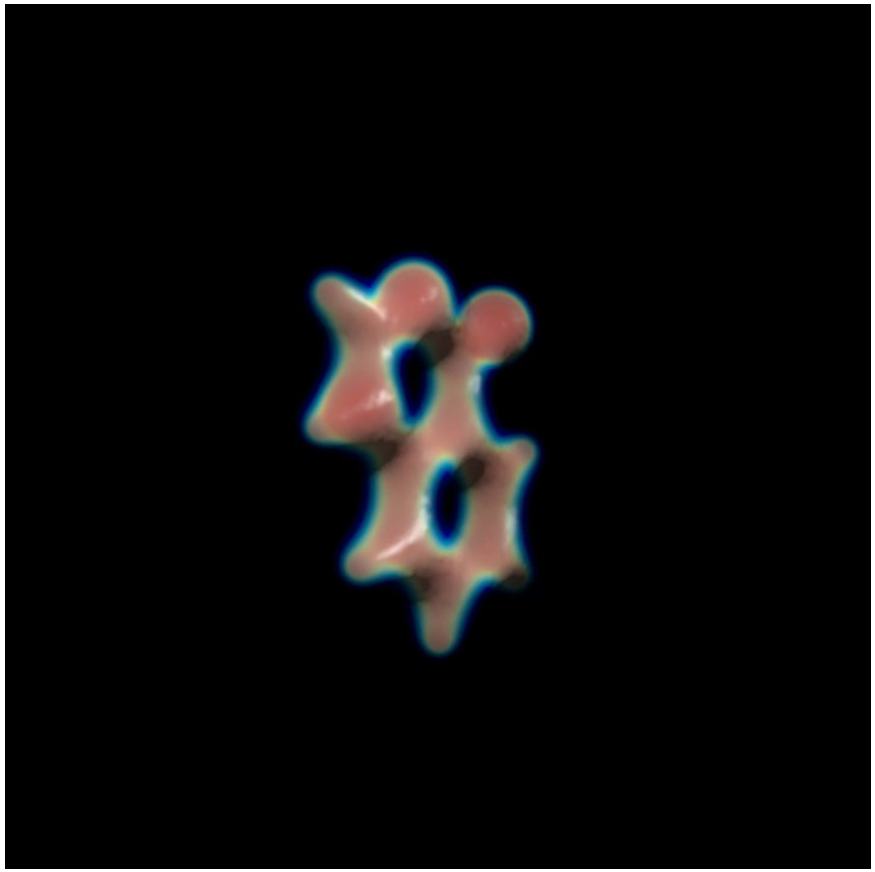




And a few more options…

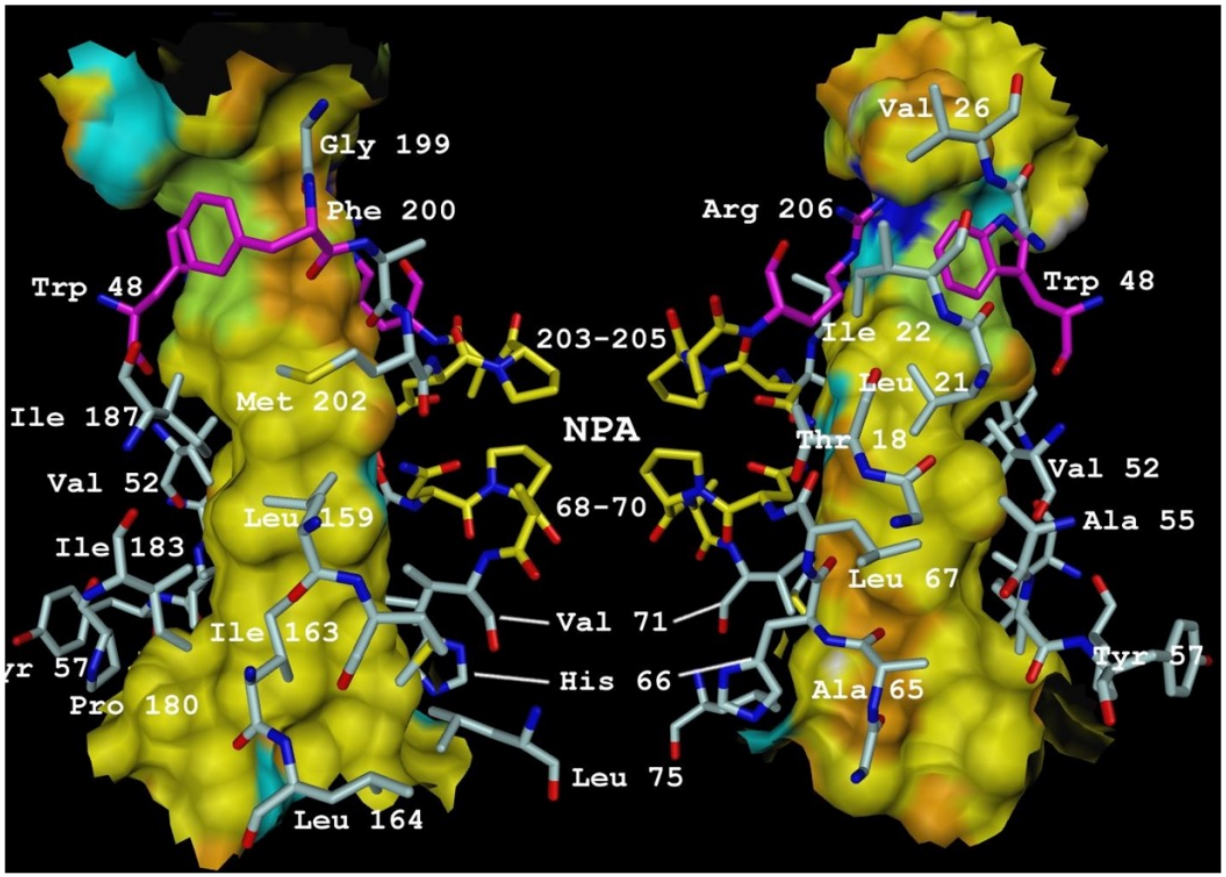
- complexity: show just enough to get point across
- use a common orientation – if possible - throughout several images or movies
- sparsely emphasize with colors
- emphasize depth with depth cueing (fog, DOF, actual 3D)
- simple lighting (hemisphere as opposed to more complex phong or BRDF)
- outlines for emphasized contours
- ambient occlusion where appropriate
- subtle shadows where appropriate
- not aiming for photorealism, but clarity (it still has to look nice, though)

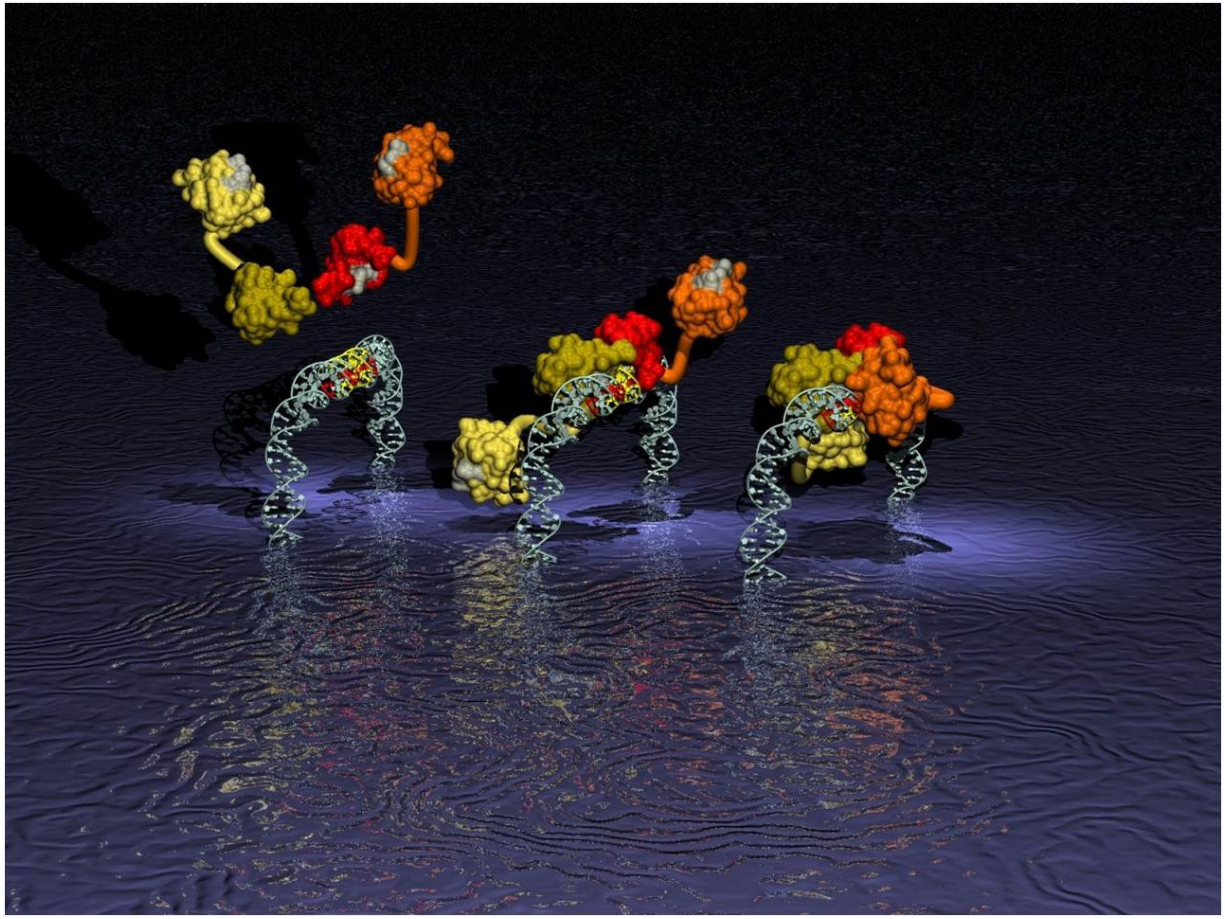




electron density fields

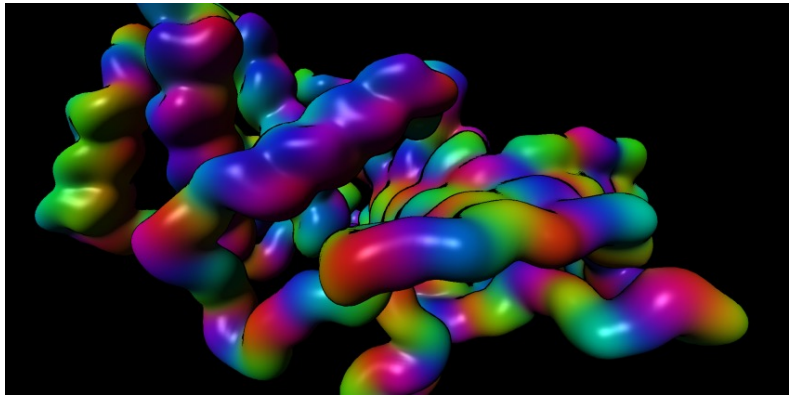
And beyond…

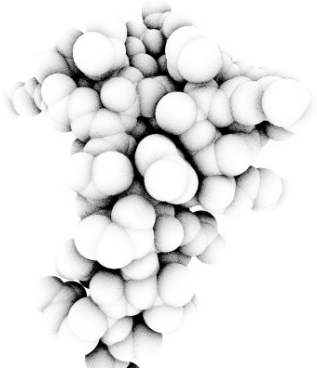

Gly 199
Phe 200
Trp 48
203-205
Met 202
Ile 187
NPA
Val 52
Leu 159
68-70
Ile 183
Ile 163
Val 71
His 66
Pro 180
Leu 75
Leu 164
Val 26
Arg 206
Trp 48
Ile 22
Leu 21
Thr 18
Val 52
Ala 55
Leu 67
Tyr 57
Ala 65

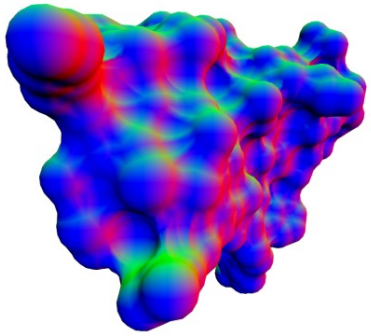


Bro, do you even code?

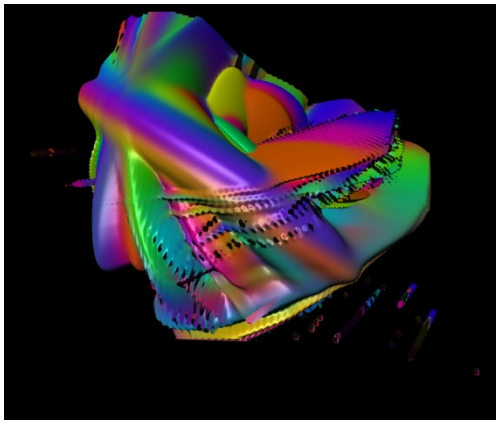
- trajectory server and architecture (C, C++)
- chemical structures, analysis, and databases
- pre-processing like smoothing, alignment
- old and new rendering styles (OpenGL, GLSL)
- animation scripting (bash, python)
- post-processing, annotation, movie encoding
- UI development

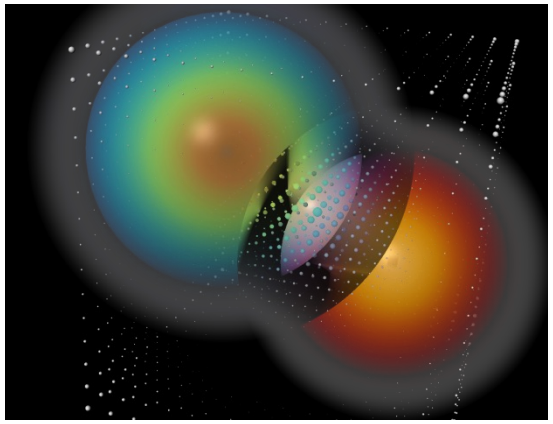


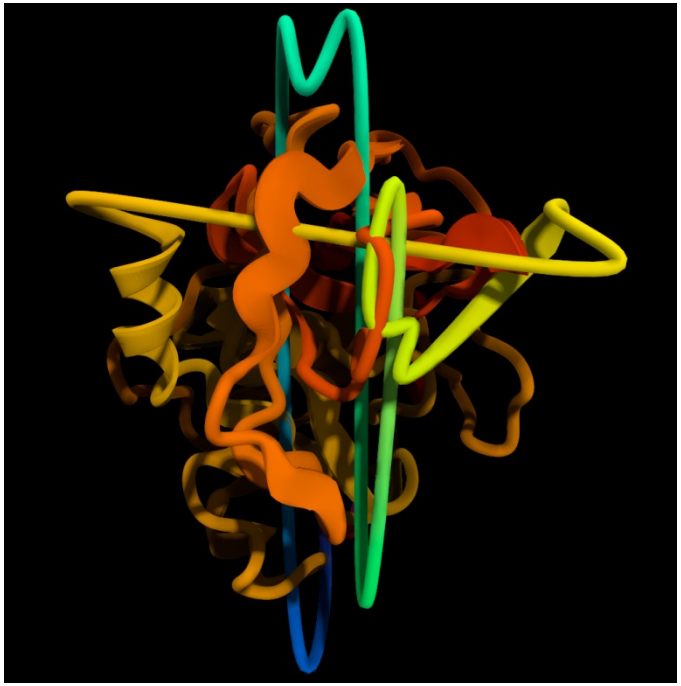




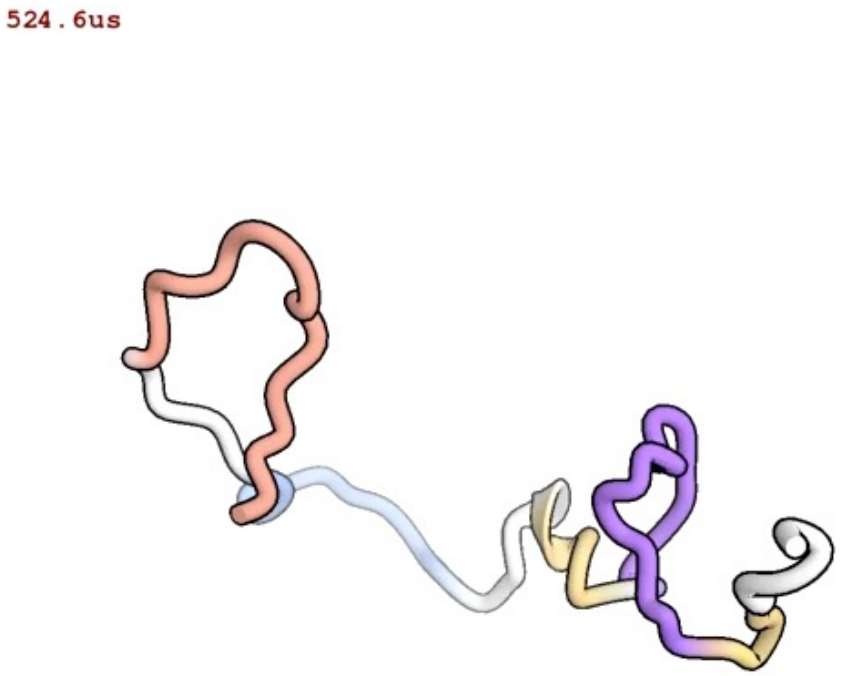
you use printf, I use…



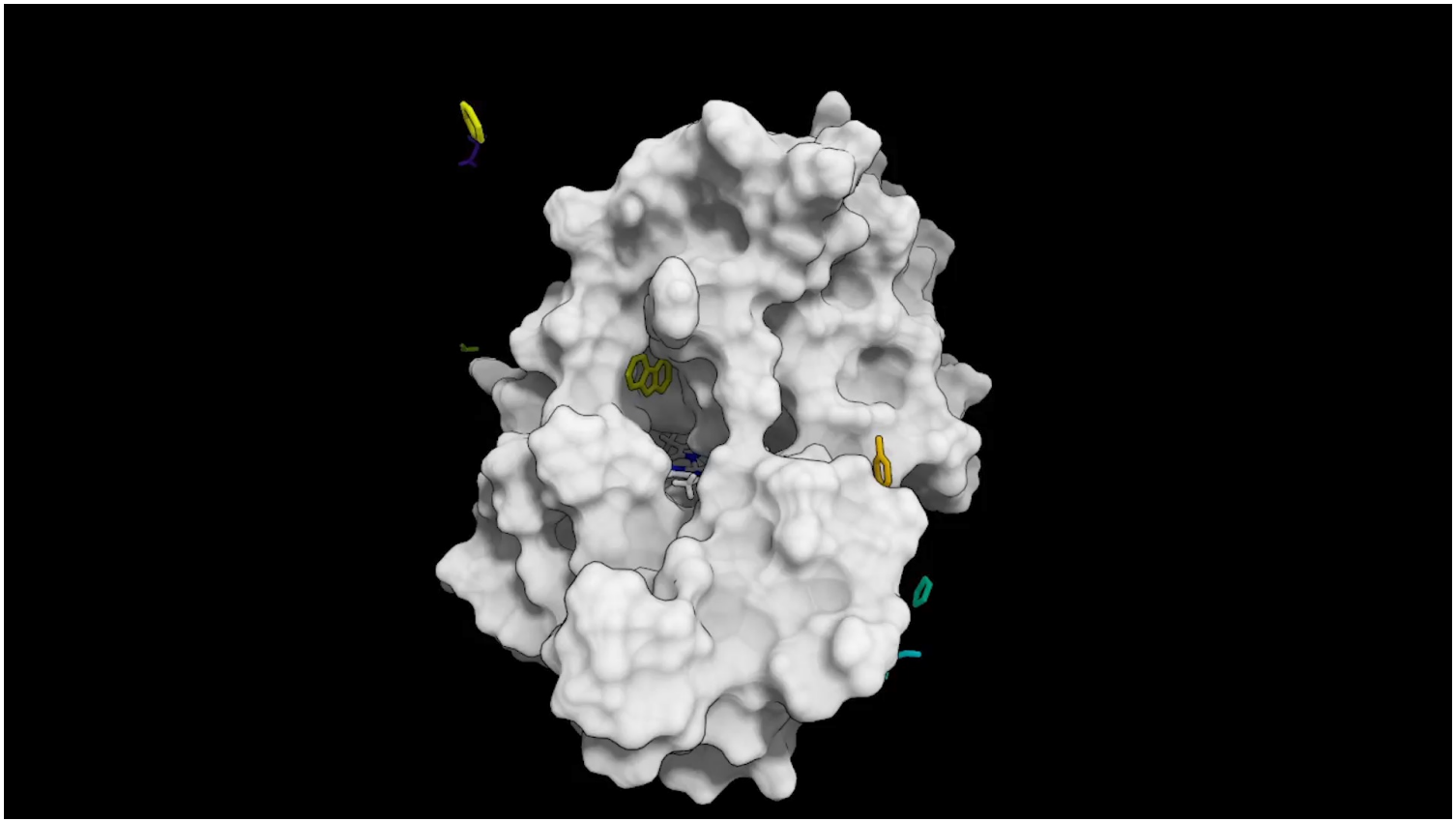




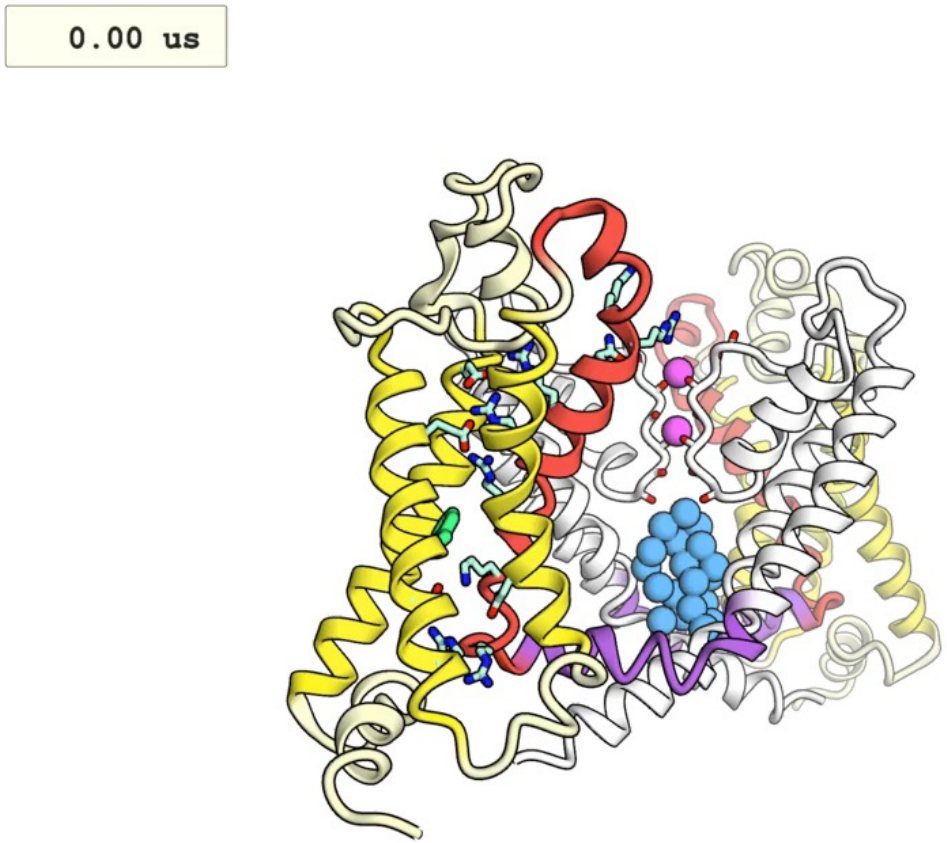
you get segfaults, I get…



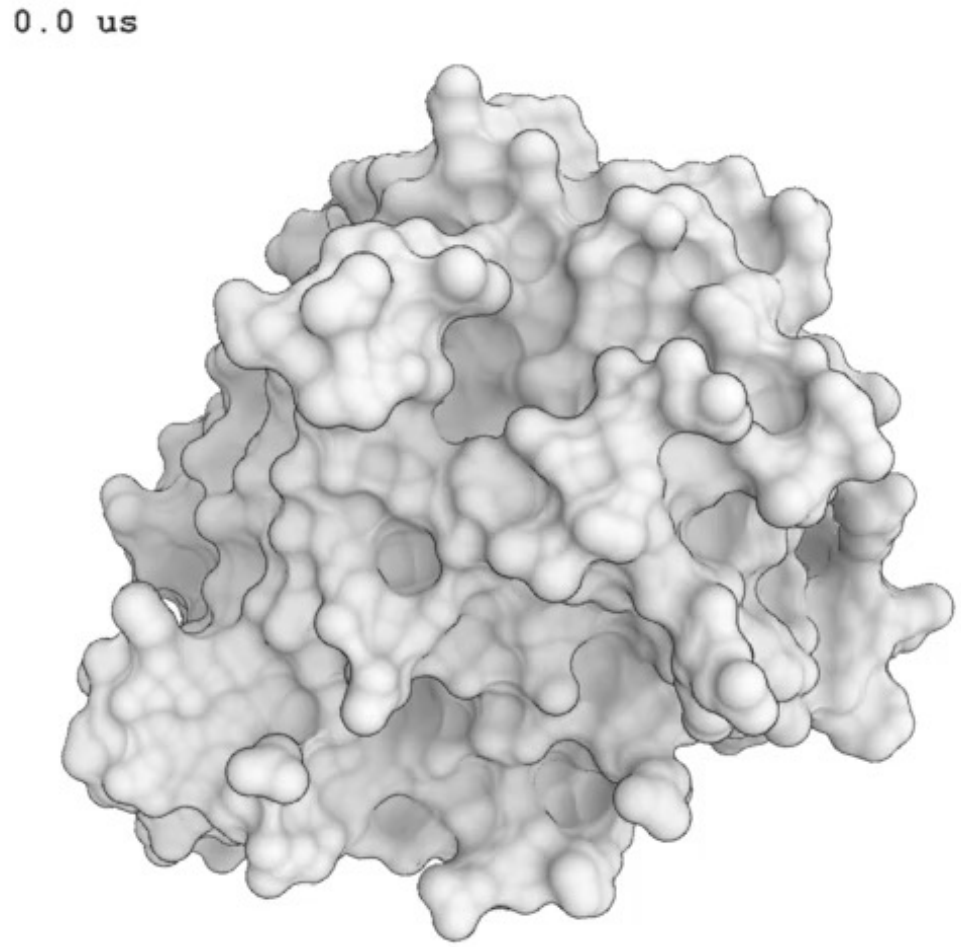
folding simulations



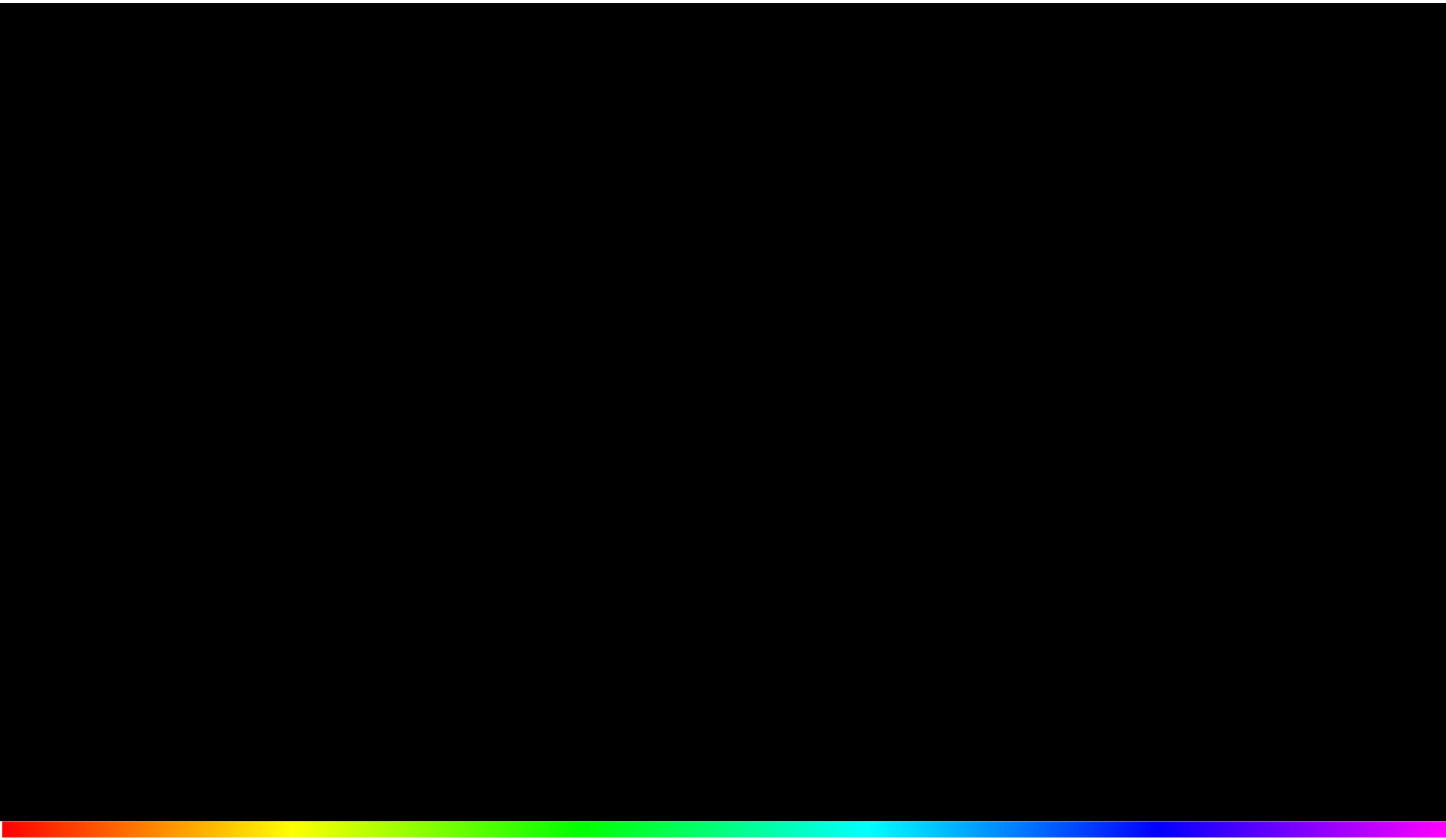
fragment screens



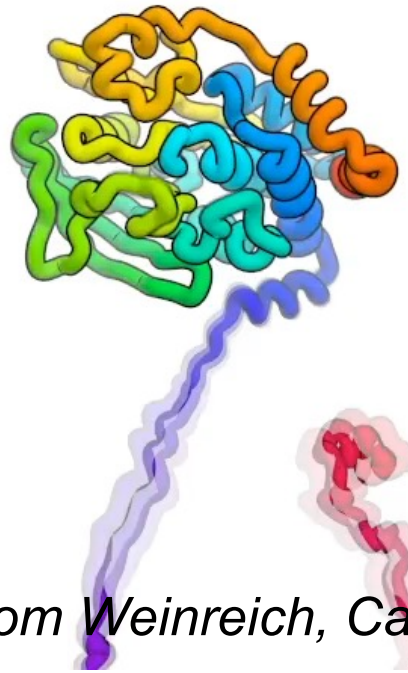
ion channels



drug discovery



Thank you!



Ellen Zhong, Cory Hargus, Tom Weinreich, Caleb Jordan

Contact us: careers@deshawresearch.com

How Fast-Folding Proteins Fold

Kresten Lindorff-Larsen, Stefano Piana, Ron Dror, David Shaw. Science, 2011.

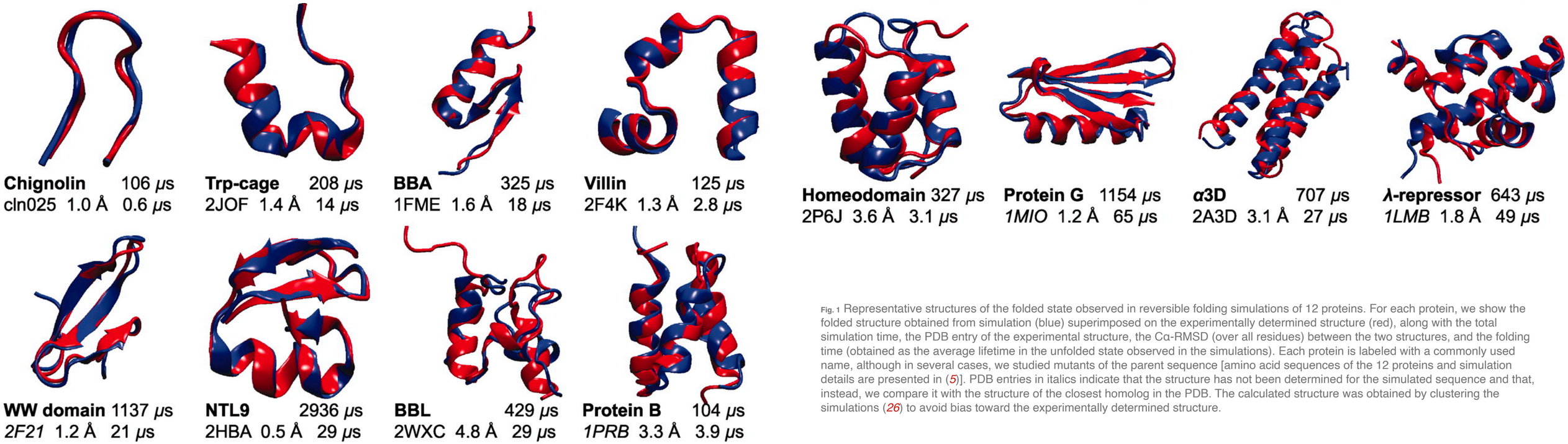


Fig. 1 Representative structures of the folded state observed in reversible folding simulations of 12 proteins. For each protein, we show the folded structure obtained from simulation (blue) superimposed on the experimentally determined structure (red), along with the total simulation time, the PDB entry of the experimental structure, the Cα-RMSD (over all residues) between the two structures, and the folding time (obtained as the average lifetime in the unfolded state observed in the simulations). Each protein is labeled with a commonly used name, although in several cases, we studied mutants of the parent sequence [amino acid sequences of the 12 proteins and simulation details are presented in (*5*)]. PDB entries in italics indicate that the structure has not been determined for the simulated sequence and that, instead, we compare it with the structure of the closest homolog in the PDB. The calculated structure was obtained by clustering the simulations (*26*) to avoid bias toward the experimentally determined structure.

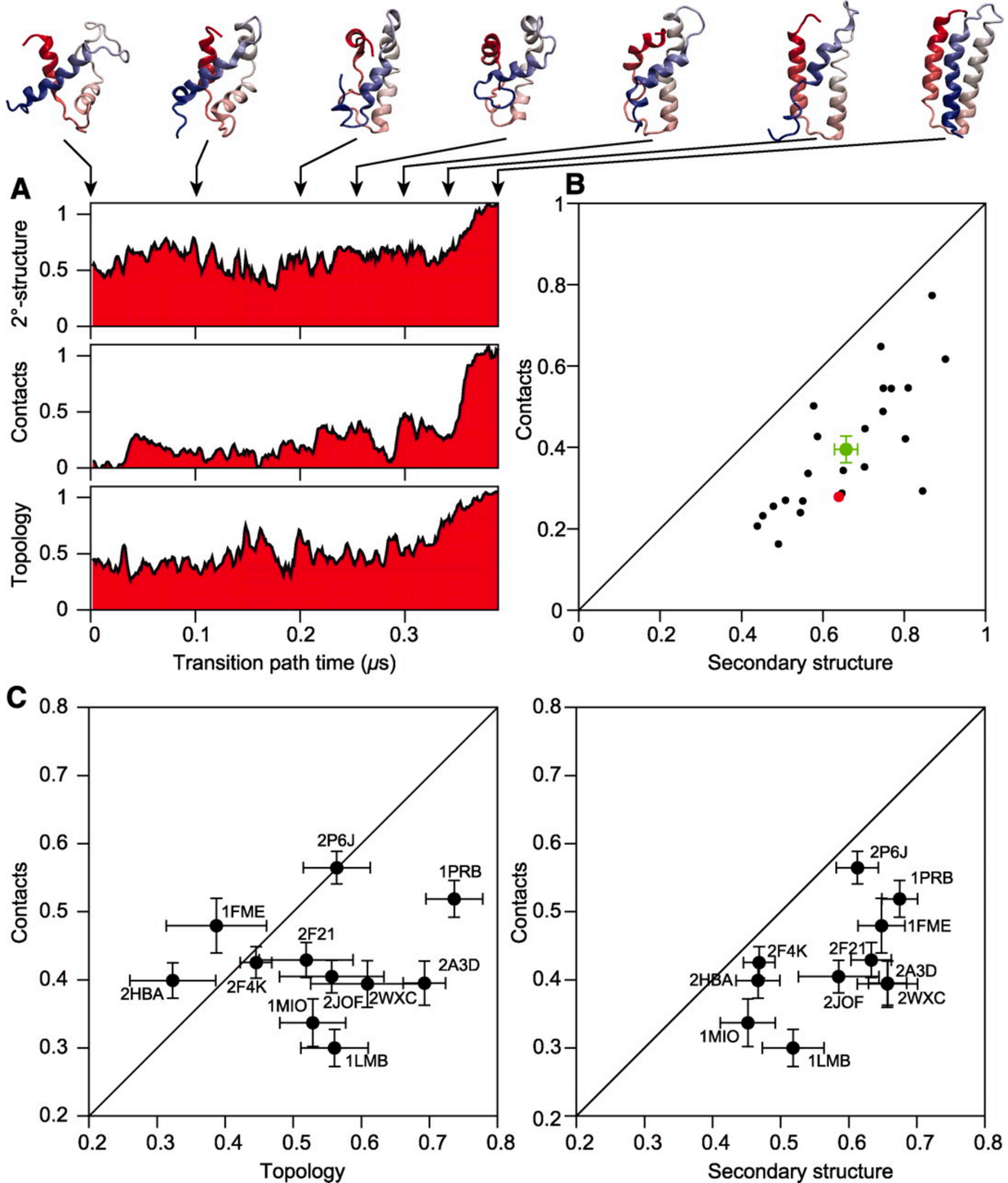


Fig. 2 Formation of topology, native contacts, and secondary structure during protein folding. (**A**) The three panels show the accumulation of native secondary structure, nonlocal native contacts, and native topology during a single folding event for α3D. Each of the three quantities was normalized such that the average value in the unfolded state was zero, and the average value in the folded state was one. Above the three panels we show seven representative structures from this transition path, with the corresponding time points shown with arrows. This analysis was repeated for each of the 24 folding and unfolding events observed for this protein, and for each of these transitions, the relative orders of formation of secondary structure, contacts, and topology were quantified by integration of these time series (with the resulting integrals, corresponding to the area under the curves, here represented by the area of the red shading). High values of this integral thus correspond to early formation of the corresponding quantity during a folding event. (**B**) The 24 transitions of α3D in a scatter plot are represented, with each of the black points corresponding to the time series integral for a single folding event (unfolding events were analyzed in reverse). The red point corresponds to the folding event shown in (A), and the green point represents the average of the time series integrals over all 24 transitions (error bars represent SEM). (**C**) We repeated this analysis for 11 of the 12 proteins (chignolin was omitted because of its small size). Each point shows the average value over all folding and unfolding events observed for one protein [as described above for the green point in (B)]. Each point is labeled with the PDB code of the relevant protein (see also Fig. 1). Most proteins fall below the diagonal in these plots, showing that topology and secondary structure develop earlier than the full set of native contacts.

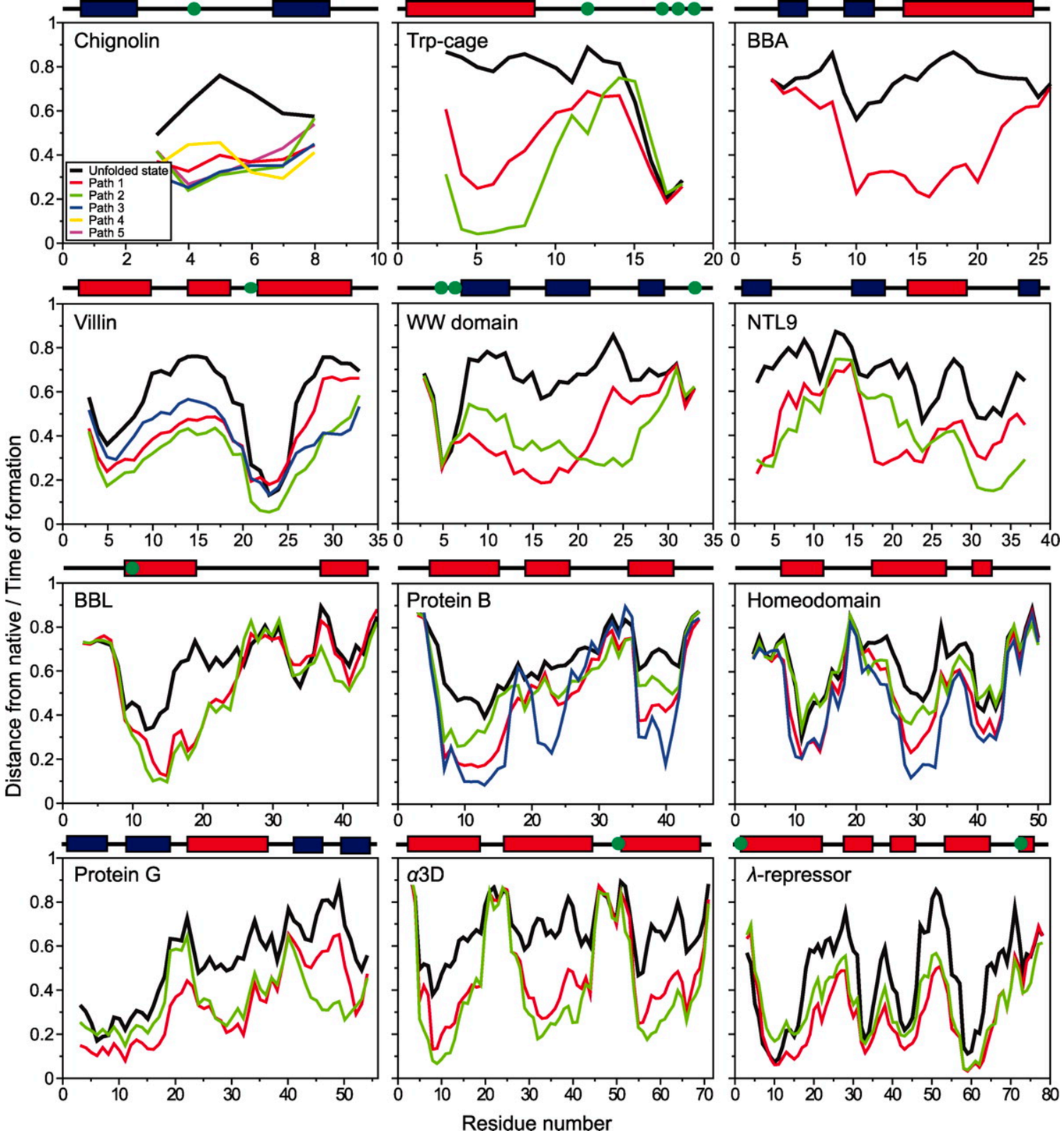


Fig. 3 Order of native structure formation along the transition pathway and the average distance from the native conformation in the unfolded state. The colored lines represent a quantity that measures when an amino acid residue adopts a nativelike structure (with a small value indicating early formation); the different colors represent the results for the different folding pathways that we obtained, as described in the main text. The average fraction of native structure in the unfolded state is shown by the black lines. The positions of helices (red) and sheets (blue) in the native state are shown above each graph together with the location of proline residues (green circles). Note that proline residues are often located at initiation sites; we speculate that this observation can be explained by the fact that proline has a restricted conformational space available and thus facilitates the local ordering of the polypeptide backbone.