DYNAMICS
FROM RANDOM OBSEVATIONS

Abbas Ourmazd
University of Wisconsin Milwaukee

NSF Science & Technology Center

ACKNOWLEDGMENTS

UWM	R. Fung
	A. Dashti
	A. Hosseinizadeh
	P. Schwander
	G. Mashayekhi
	J. Slawinska
BioXFEL	P. Fromme et al.
	M. Schmidt et al.
	J. Spence et al.
	Many others
CFEL	R. Santra
	O. Vendrell
	A. Hanna
Columbia	J. Frank et al.
NYU	D. Giannakis et al.
CUNY	A. des Georges
	S. Salah
SPI	Aquila et al.

EMERGING CONCEPTS

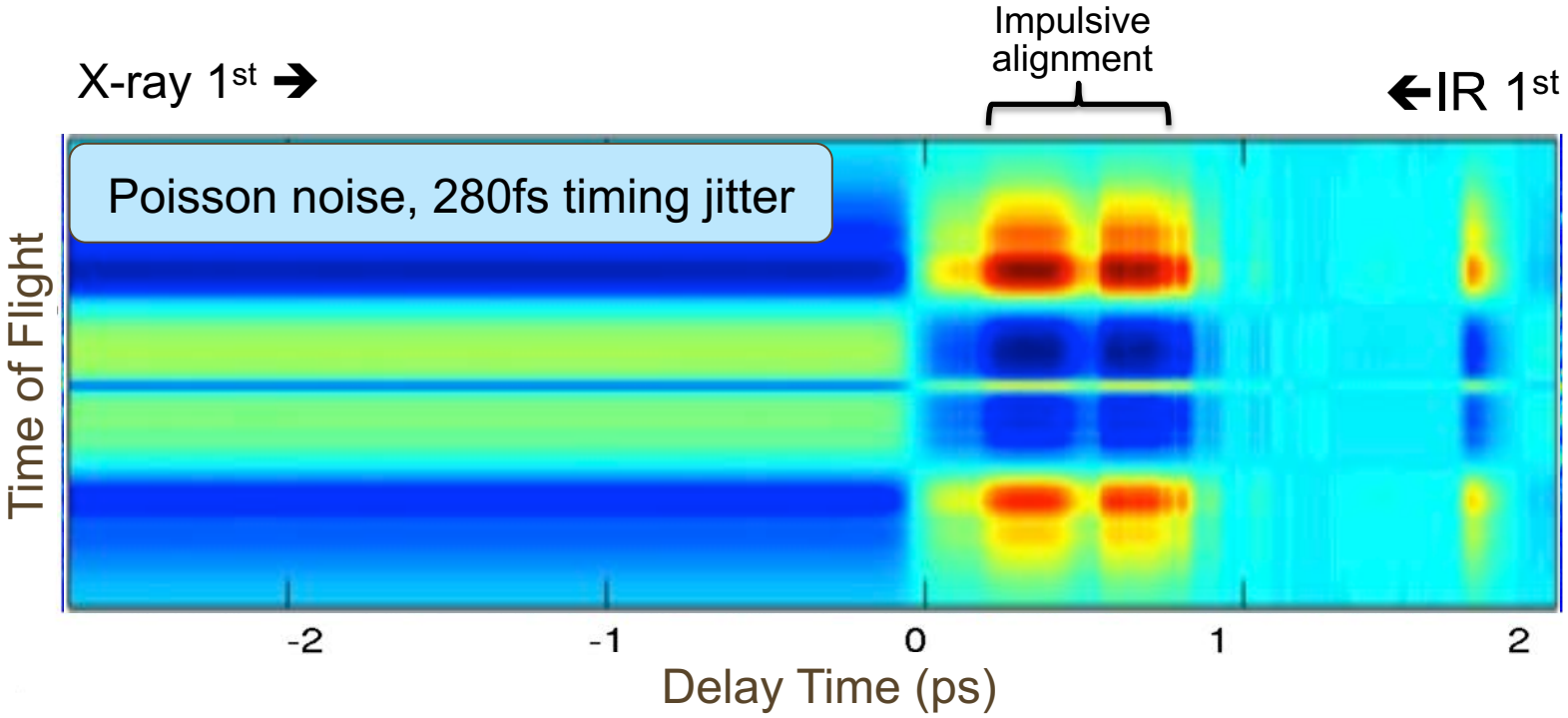
- Classic: Best experiments $\doteq$ Greatest control
  - Limits extracted information
- Data-driven: Richest datasets $\Leftrightarrow$ Random observations
  - Allows system to do what it wants to do
  - Control VERY poor for new tools (e.g., XFELs)
- Search algorithms: Information from random observations
  - TB - PB datasets
  - Extract specific information at will
  - No preconceived notions (“templates” or “priors”)
- Tools: Geometric machine learning
  - Information from intrinsic structure of data; unsupervised, mathematically rigorous

OUTLINE: FUTURE IS DATA

- Accurate dynamics from noisy data
  - Overwhelming noise, uncertain timing information
- "Dynamics" from data with no timing information
  - 3D movies of biological nanomachines in action
- Thermodynamics of function
  - Reaction coordinates, energy landscapes
- Reactions: Isomerization, ligand binding
  - Structure, conformations, dynamics, thermodynamics
  - Paths, probabilities, branching ratios
- Accessing fleeting transition states
  - "Had we but world enough and time"

BOND-BREAKING IN N_2

XFEL PUMP-PROBE (Data from Glownia et al)

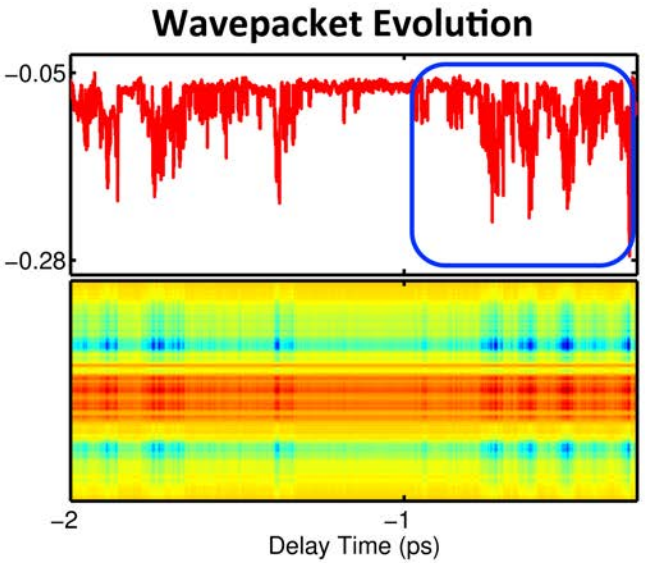


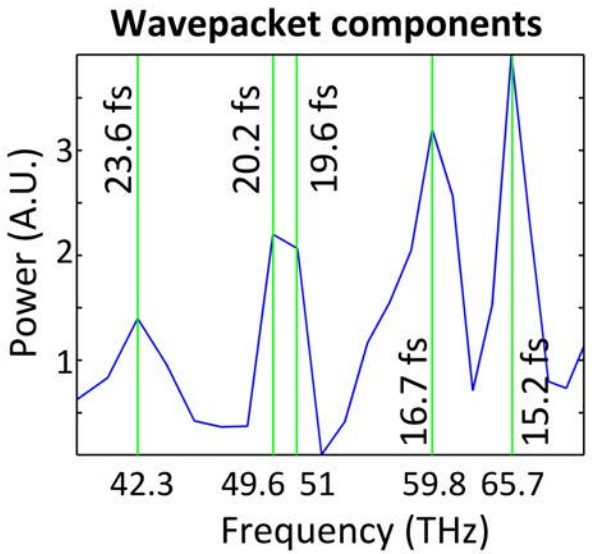
Time resolution: ~ 1fs

Original timing uncertainty: 280fs

Fung et al., *Nature* 532, 471 (2016)

WAVEPACKET DYNAMICS

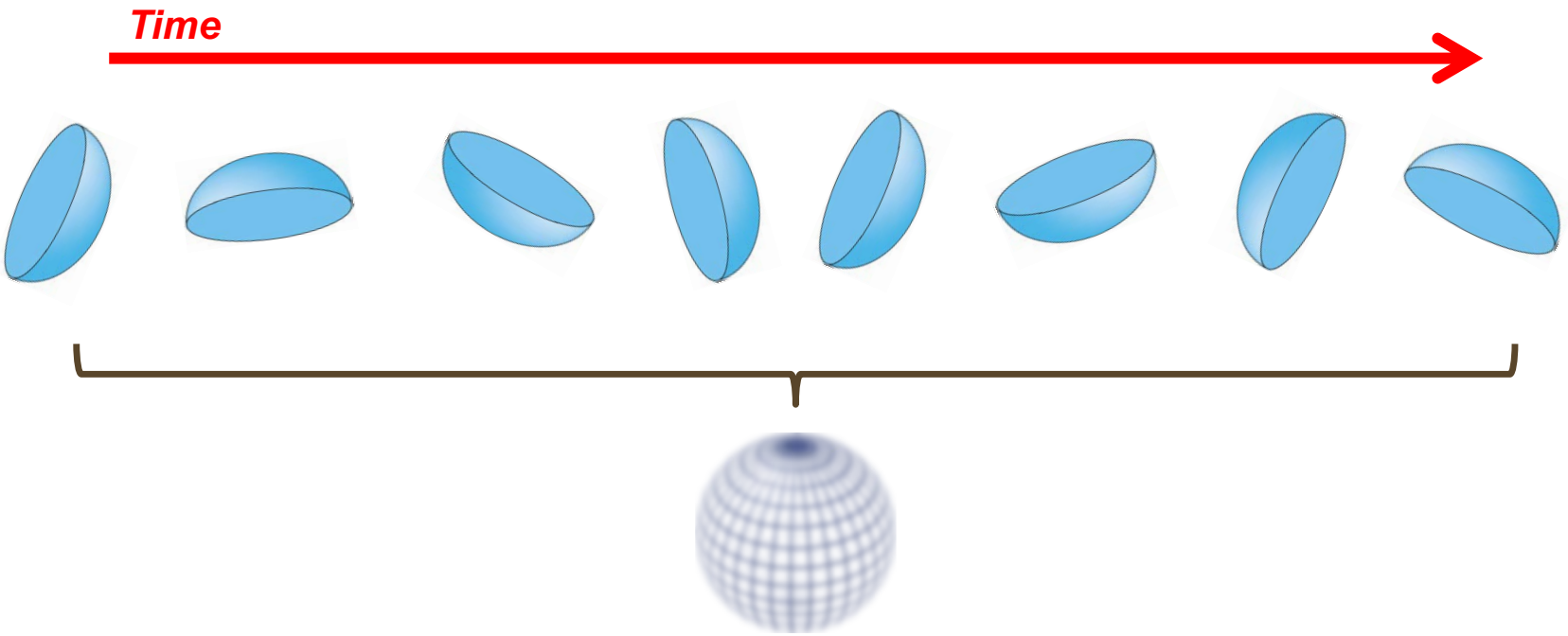




Vibrational Periods (fs)

Measured	Theory	State	Remarks
15.2	15.1	$X^2\Sigma_g^+$	Previously observed in time domain
16.7	16.7	$X^1\Sigma_g^+$	
19.6	17.7	$A^2\Pi_u$	Overlap between peaks degrades accuracy
20.2	22.4	$a^3\Pi_u$	
23.6	23.7	$A^1\Pi_u$	

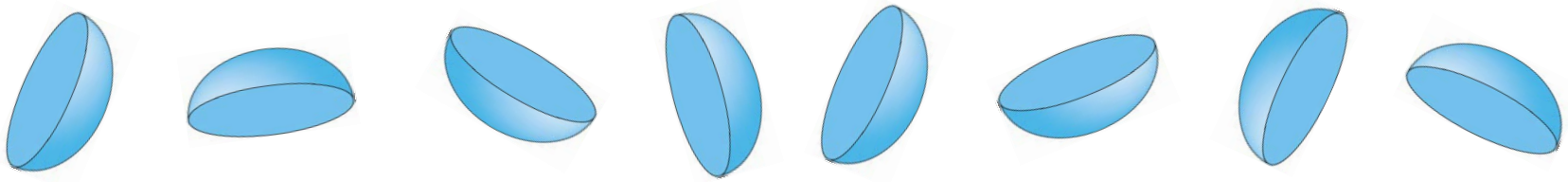
STRUCTURAL DYNAMICS



Merging snapshots from different timepoints degrades information

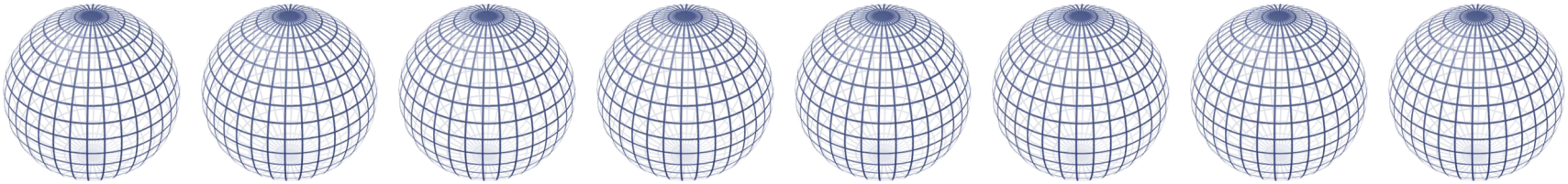
FUTURE OF STRUCTURAL DYNAMICS

Random Ewald cups of partials, inaccurate timestamps
Non-uniformly distributed in time



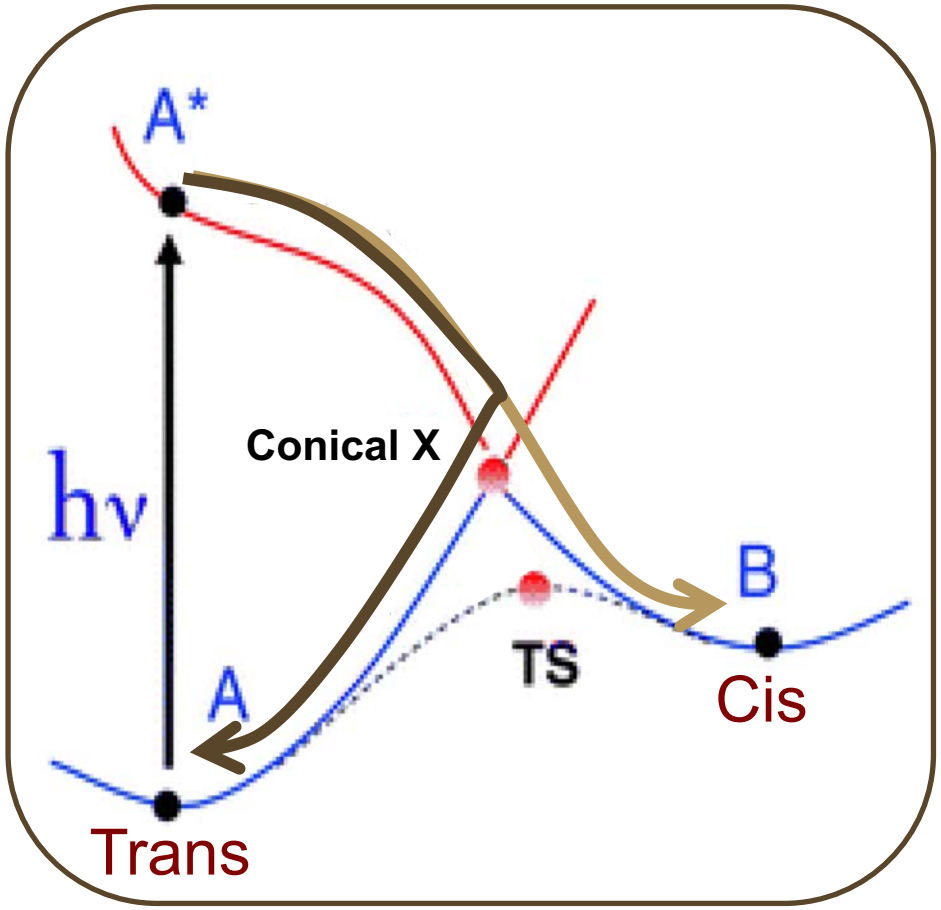
Time

Magic Algorithm

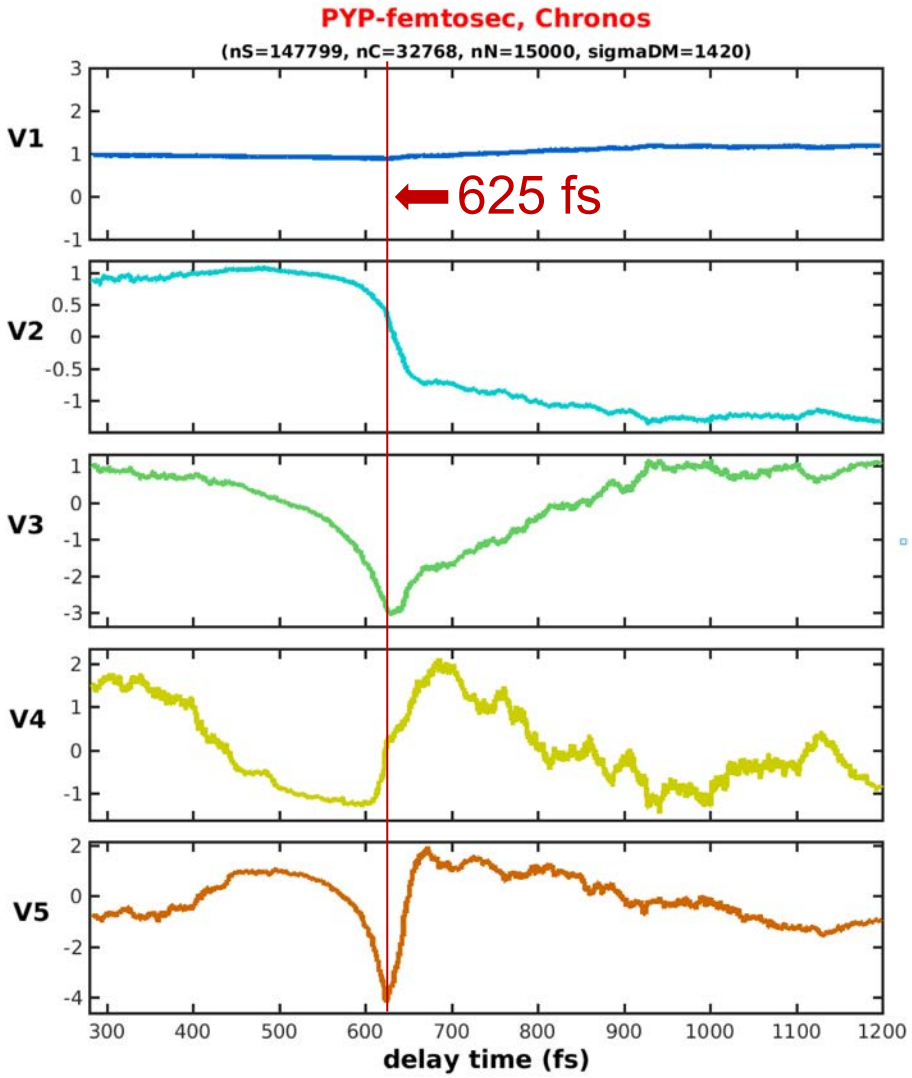


Evolution of 3D diffraction volume, accurate timestamps
uniformly distributed in time

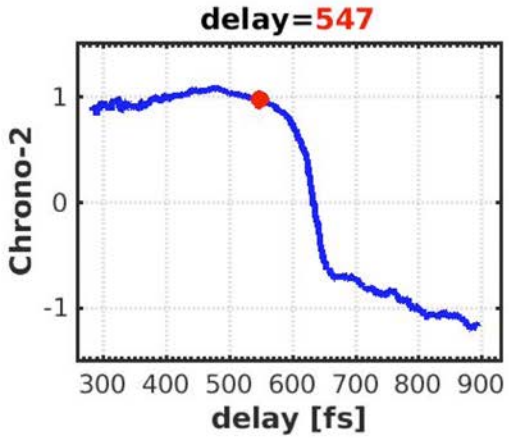
CONICAL INTERSECTIONS

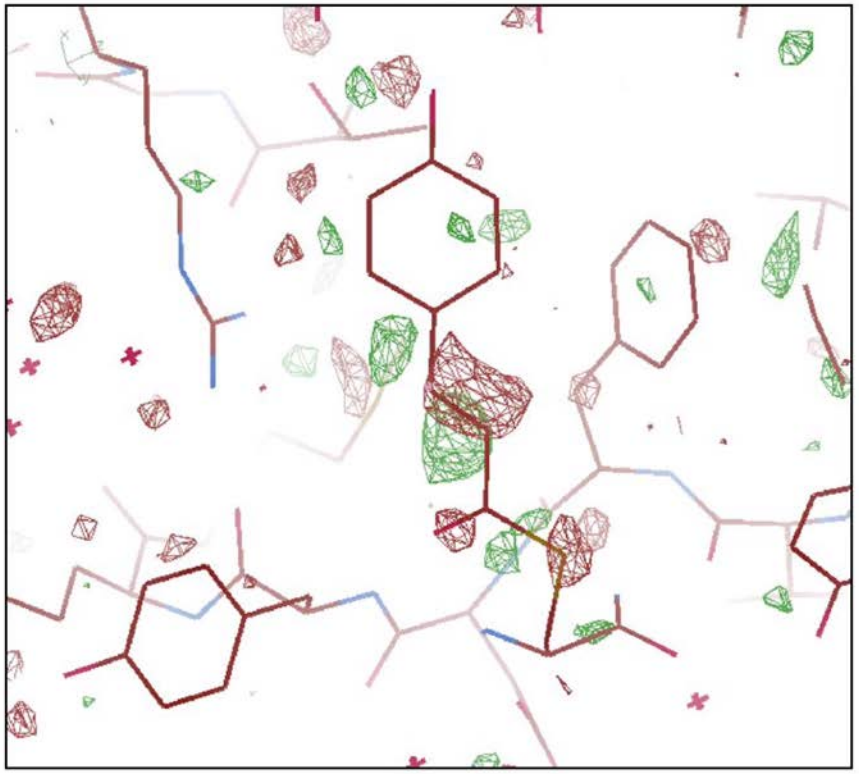


PYP: TIME EVOLUTION (CHRONOS)



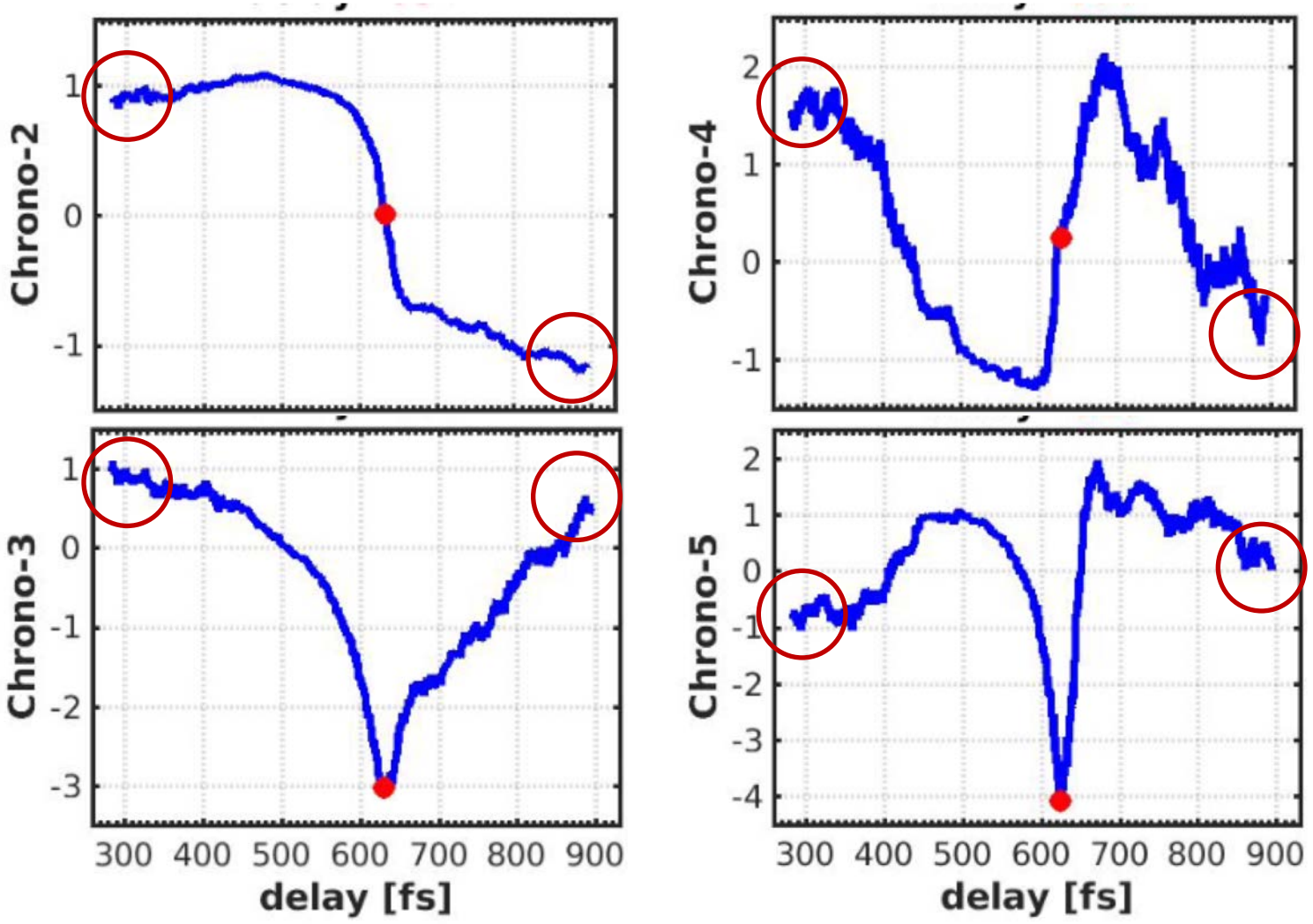
CONICAL INTERSECTION: THE MOVIE
17.5 kDa PROTEIN



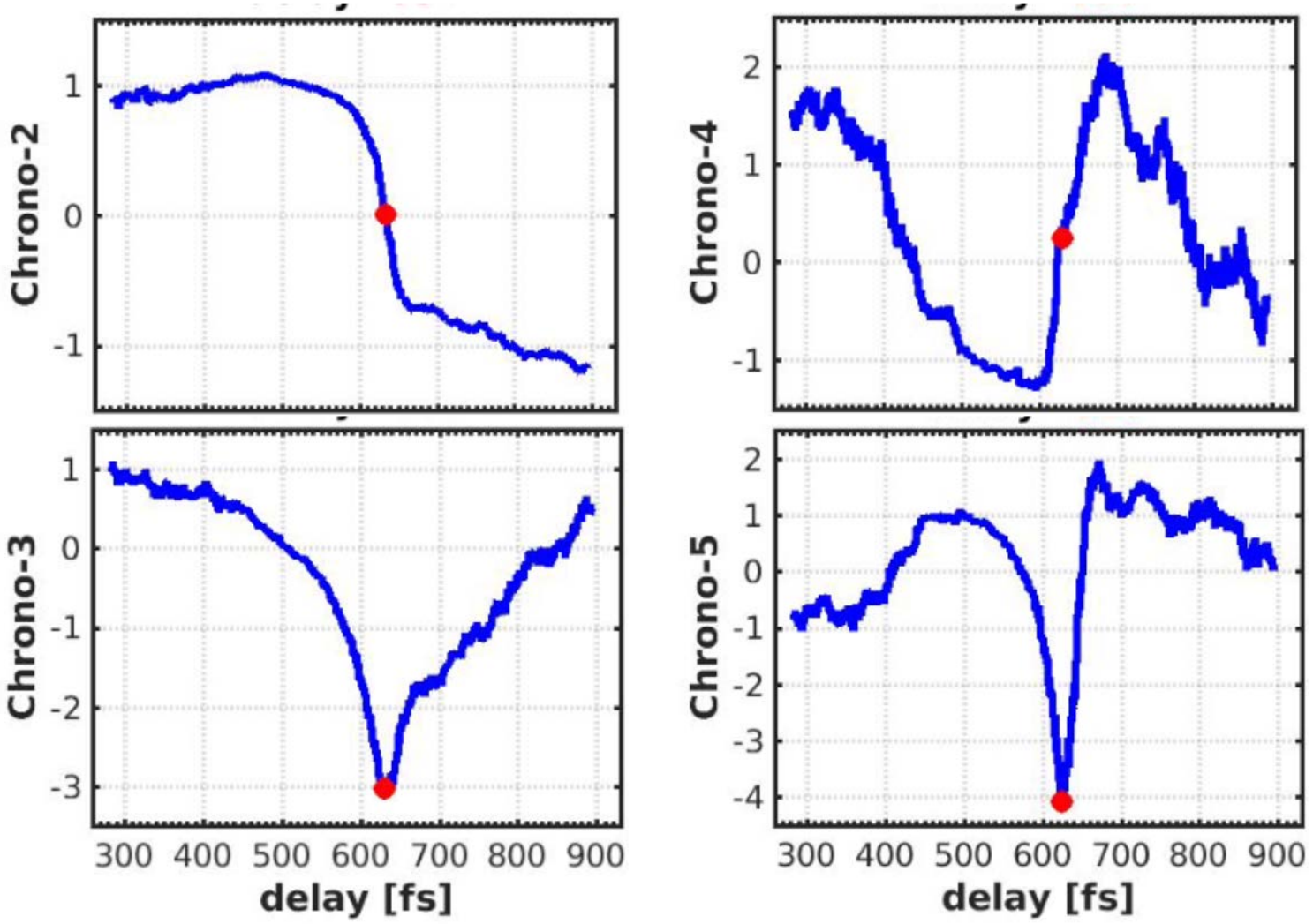


Time resolution: ~1 fs
Spatial resolution: ~ 3 Å

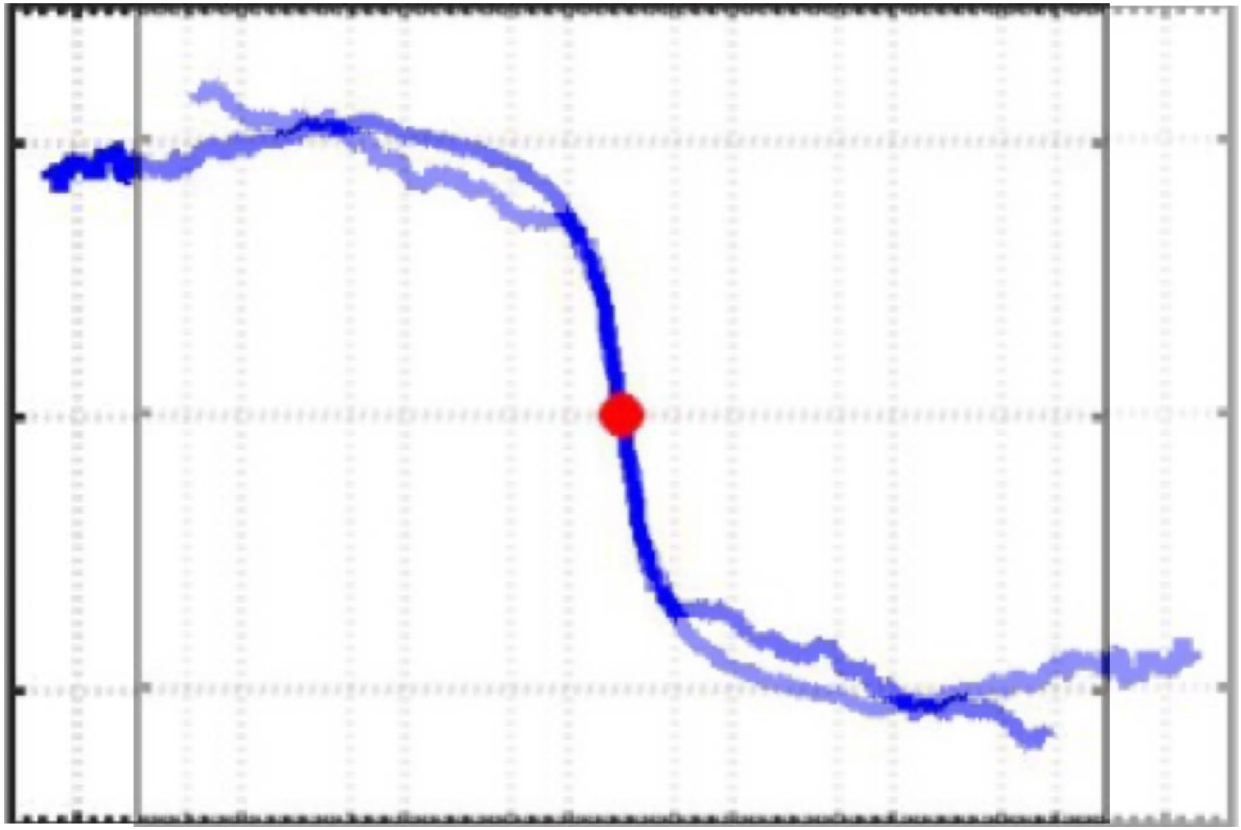
CHRONO TYPES
SUCCESS & FAILURE: TWICE



CHRONO SYMMETRIES
INVERSION & NO-INVERSION: TWICE



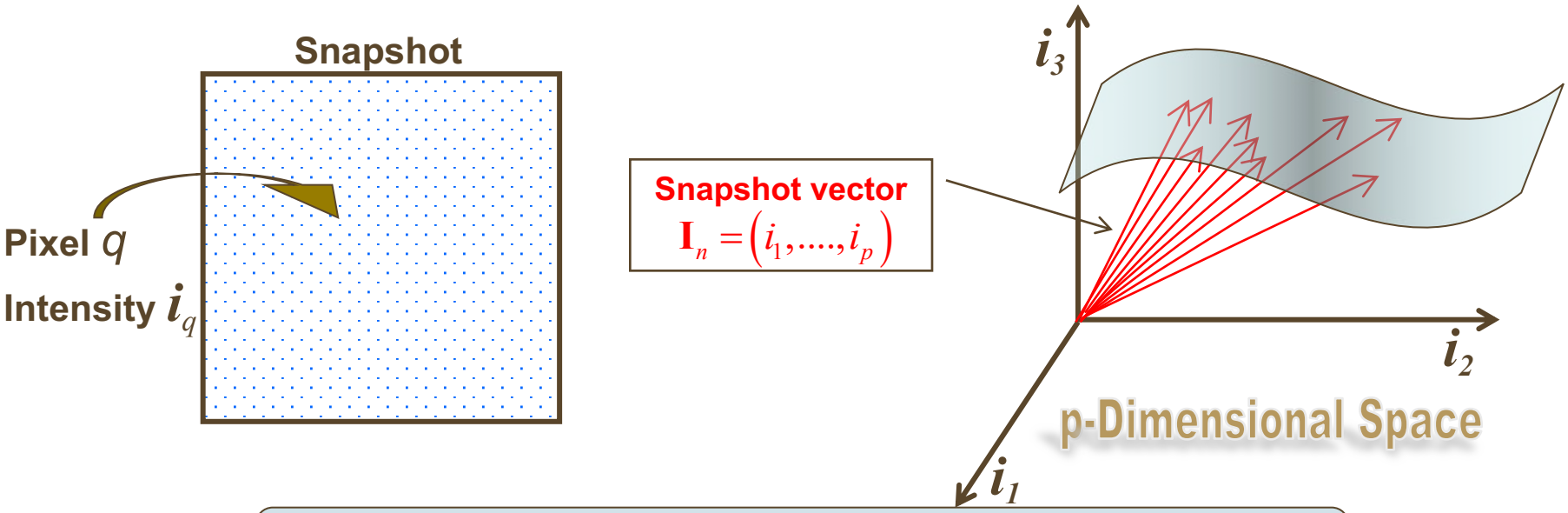
“INVERSION SYMMETRY”
CHRONO-2



Time-reversal symmetry?

LEARNING FROM DATA GEOMETRY

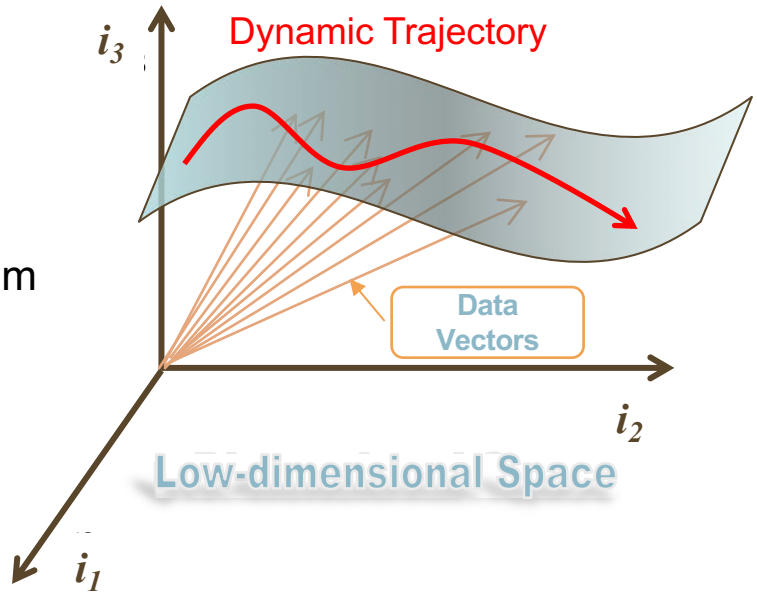
- All we have is ensemble of diffracted intensities
  - A snapshot is: $\mathbf{I}_n = \left(i_1, ..., i_p\right)$; lives in p-dimensional data space
  - Intrinsic manifold dimensionality = No. of degrees of freedom exercised



Manifold represents information content of the data

GEOMETRIC MACHINE LEARNING

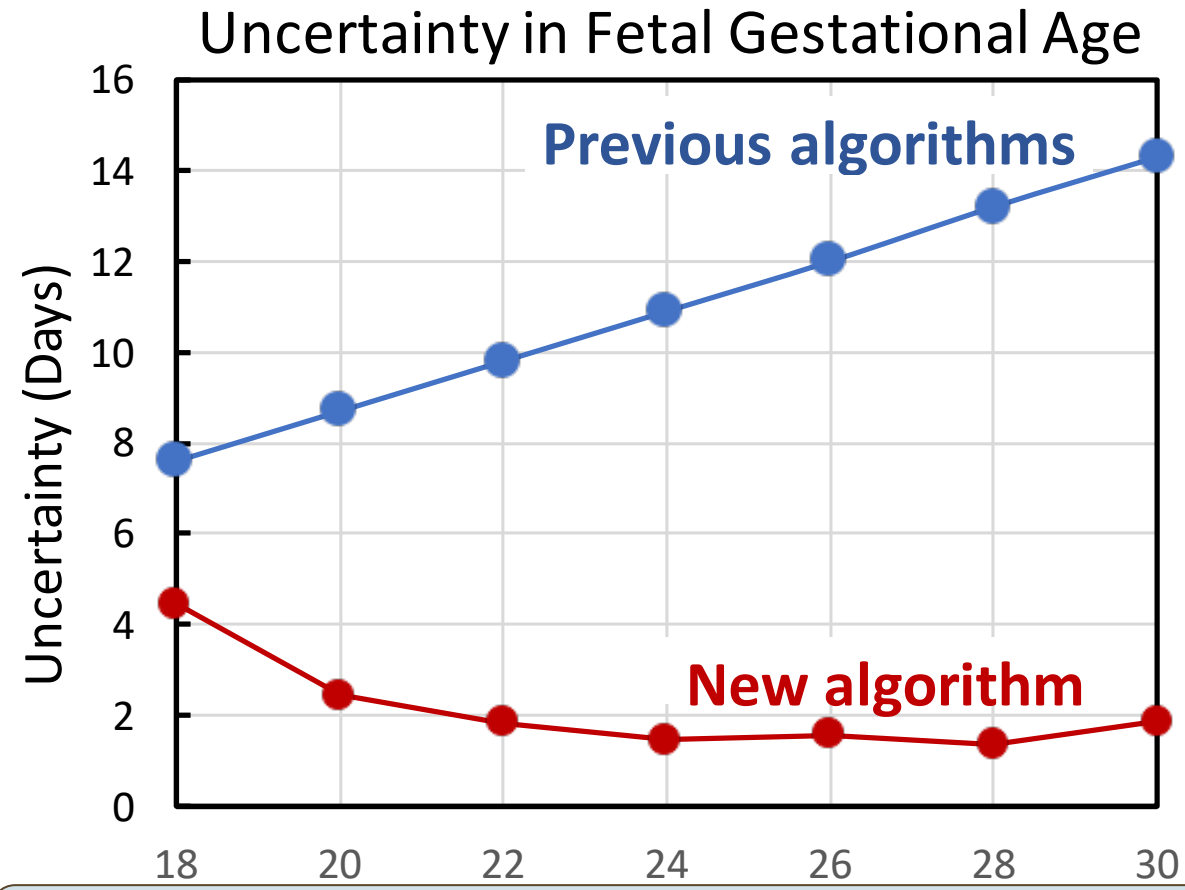
- Learn curved manifold from data-cloud
  - Eigenfunctions of Laplace-Beltrami operator
- Yields reaction coordinates
  - Manifold dimensionality = No. of degree of freedom
- Perform all operations in curved manifold
  - Riemannian geometry a la GR
- Extract information by navigating on manifold
  - Manifold dimensionality = No. of degree of freedom
- Dynamics: trajectory on manifold
  - Defined by time-stamped ensemble of snapshots



PREGNANCY & NEWBORN OUTCOME PREDICTION

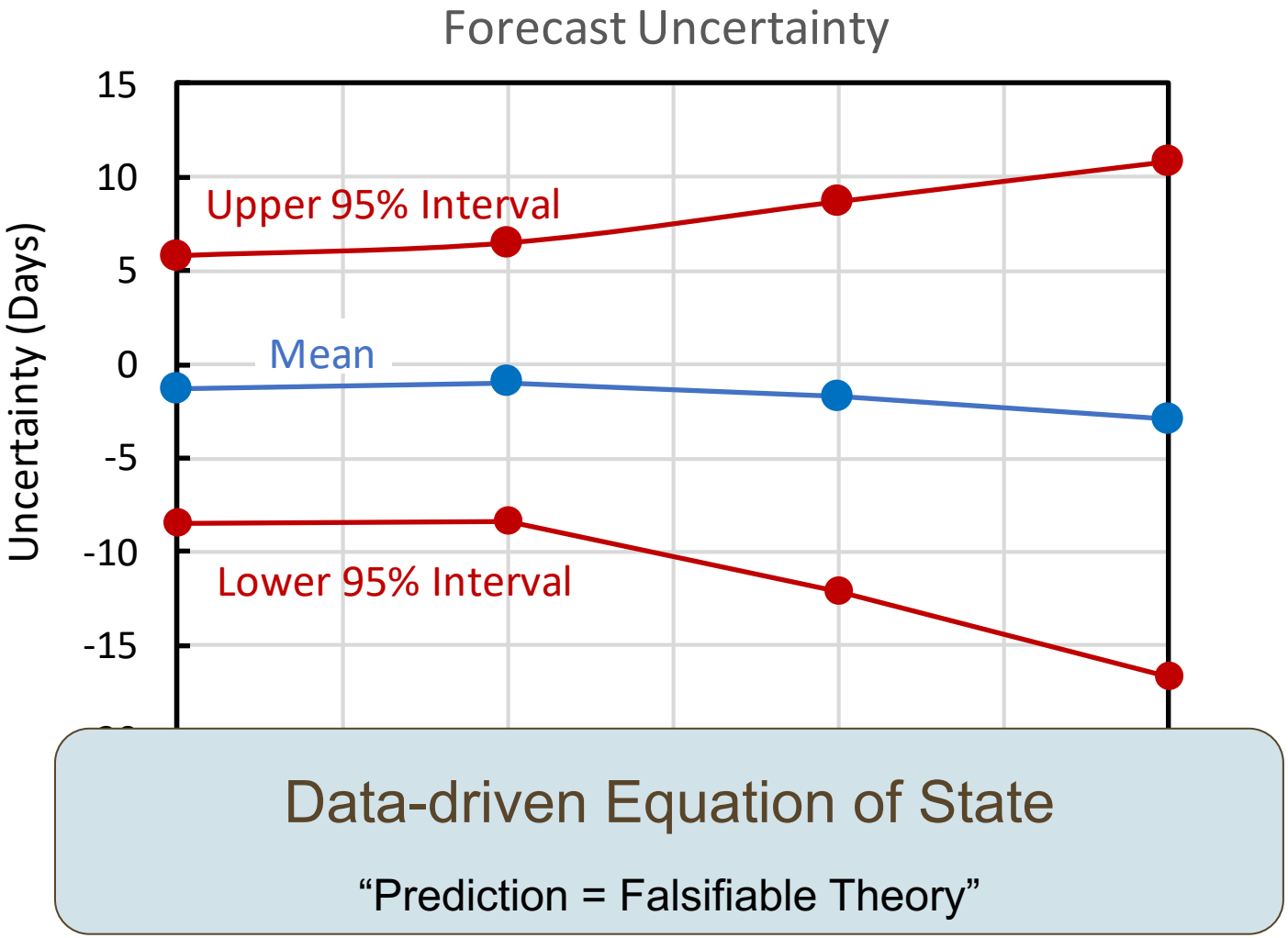
Each year ~3M babies die within the first month of life. This represents ~40% of all under 5yr deaths. Preterm birth is the single largest contributor to mortality and morbidity burden. Decreasing this burden in LMICs is constrained by a lack of tools to identify women at risk of PTB where existing interventions could be applied (progesterone, antenatal steroids, facility based deliveries, antibiotics). Furthermore, accurate assessment of gestational age both prenatally and at birth is critical to capacity to deliver antenatal interventions early and safely and to determining which infants are preterm SGA or full-term SGA, and which are only preterm. Ultrasonography combined with predictive pregnancy and newborn outcome algorithms could fill the current gap of field-ready gestational age assessment tools that are essential for both epidemiologic studies (i.e., is a low birthweight infant preterm or term and undergrown), and early identification of at risk pregnancies in order to widen the treatment and referral window.

FETAL AGE ESTIMATES FROM SONOGRAMS



First improvement since 1812 (no typo)

PREDICTING INDIVIDUAL FETAL DEVELOPMENT



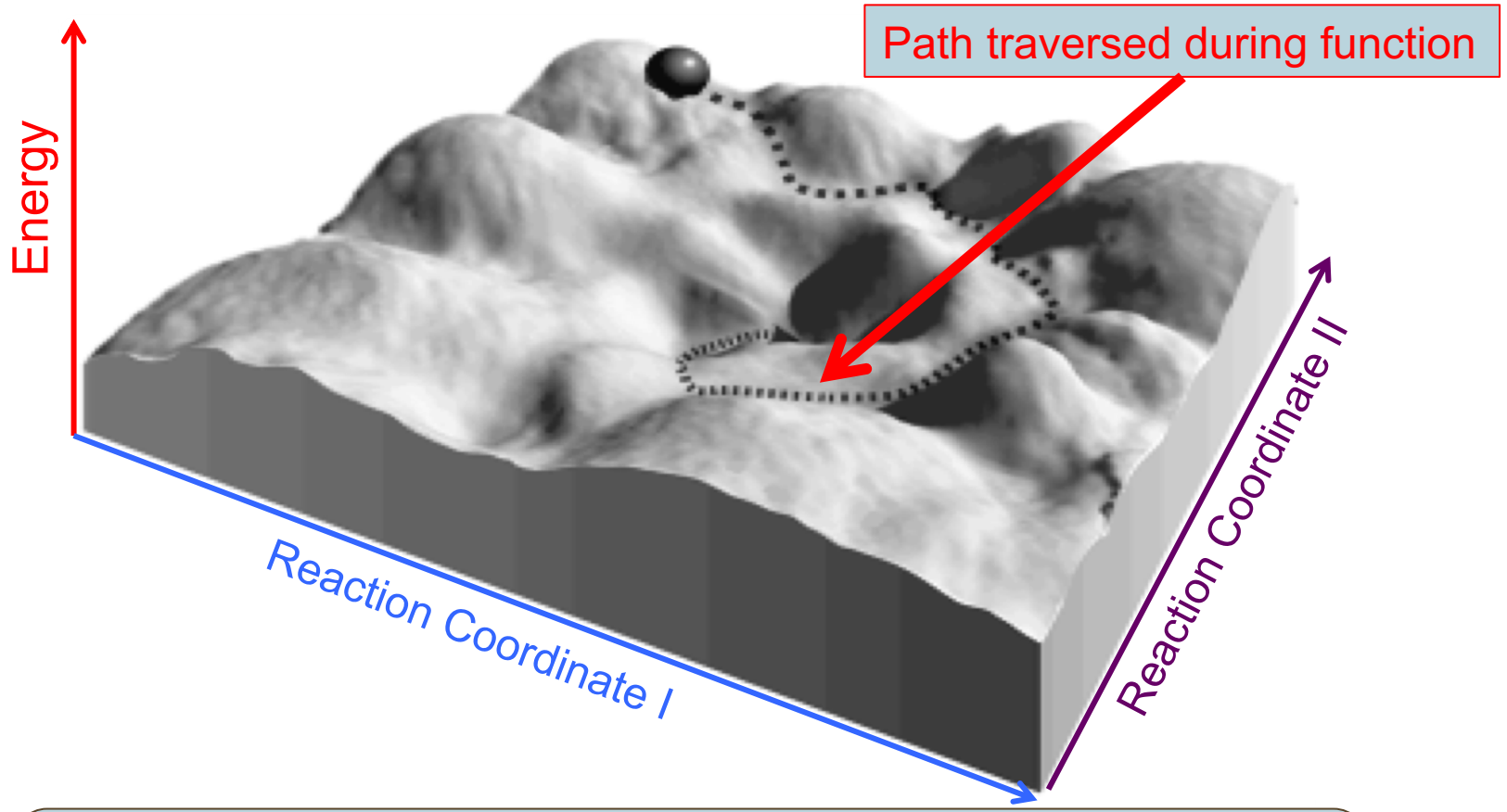
Data-driven Equation of State

"Prediction = Falsifiable Theory"

DYNAMICS WITHOUT TIMING INFORMATION
THERMODYNAMICS OF FUNCTION

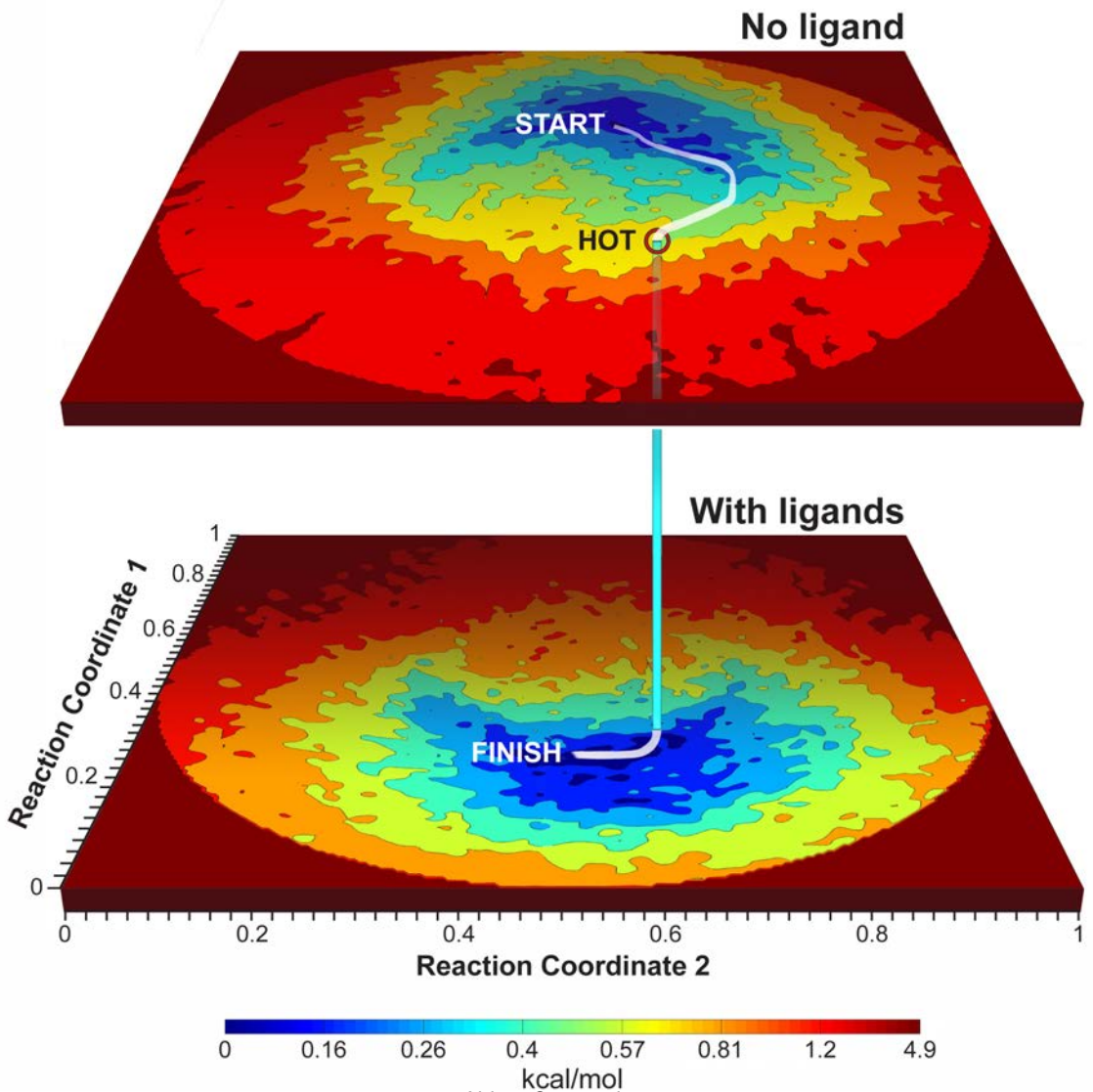
- Molecular machines undergo complex conformational changes
  - A “reaction coordinate” characterizes a set of concerted changes
- There can be multiple reaction coordinates
  - Must be orthogonal to be meaningful
  - Reaction coordinates control conformation
- Changes in structure involve changes in energy
  - Otherwise there would be no “definite” structure
- Function equivalent to traversing a path on energy landscape
  - Dimensionality of landscape = No. of reaction coordinates

FUNCTION & ENERGY LANDSCAPES

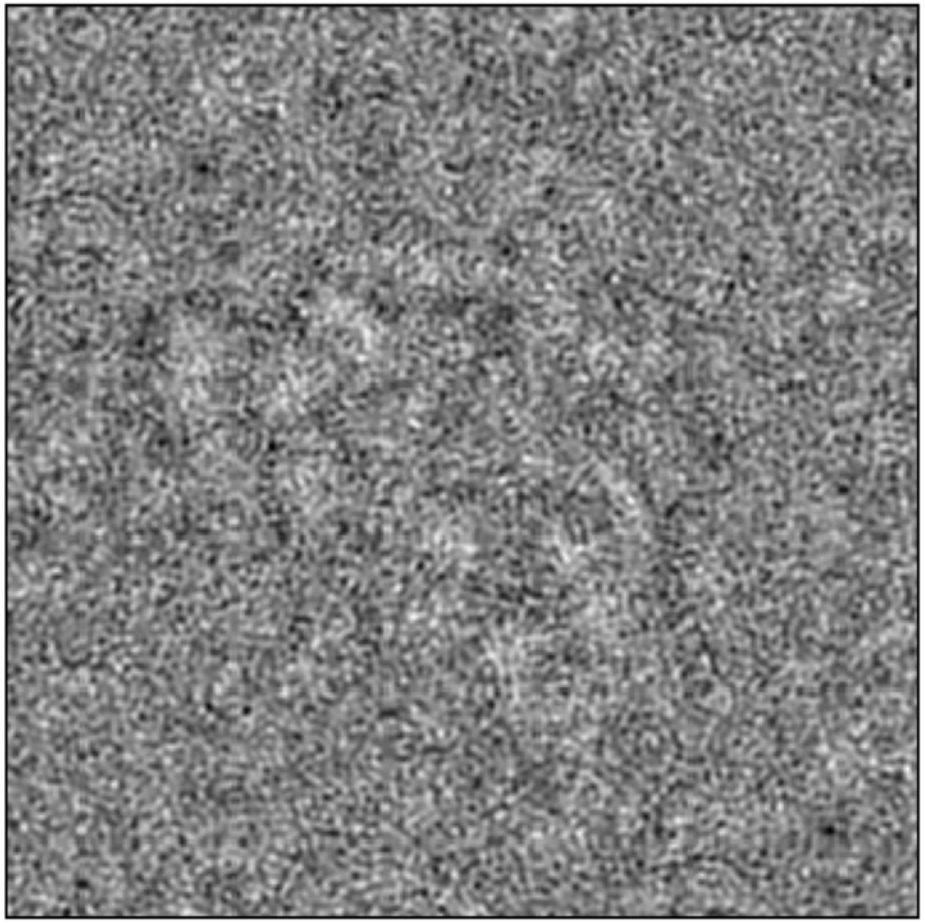


A reaction coordinate controls a set of concerted structural changes
A point in landscape represents a specific conformation & energy
Highest accessible energy ~ $k_B T \, ln(N_{snapshots})$

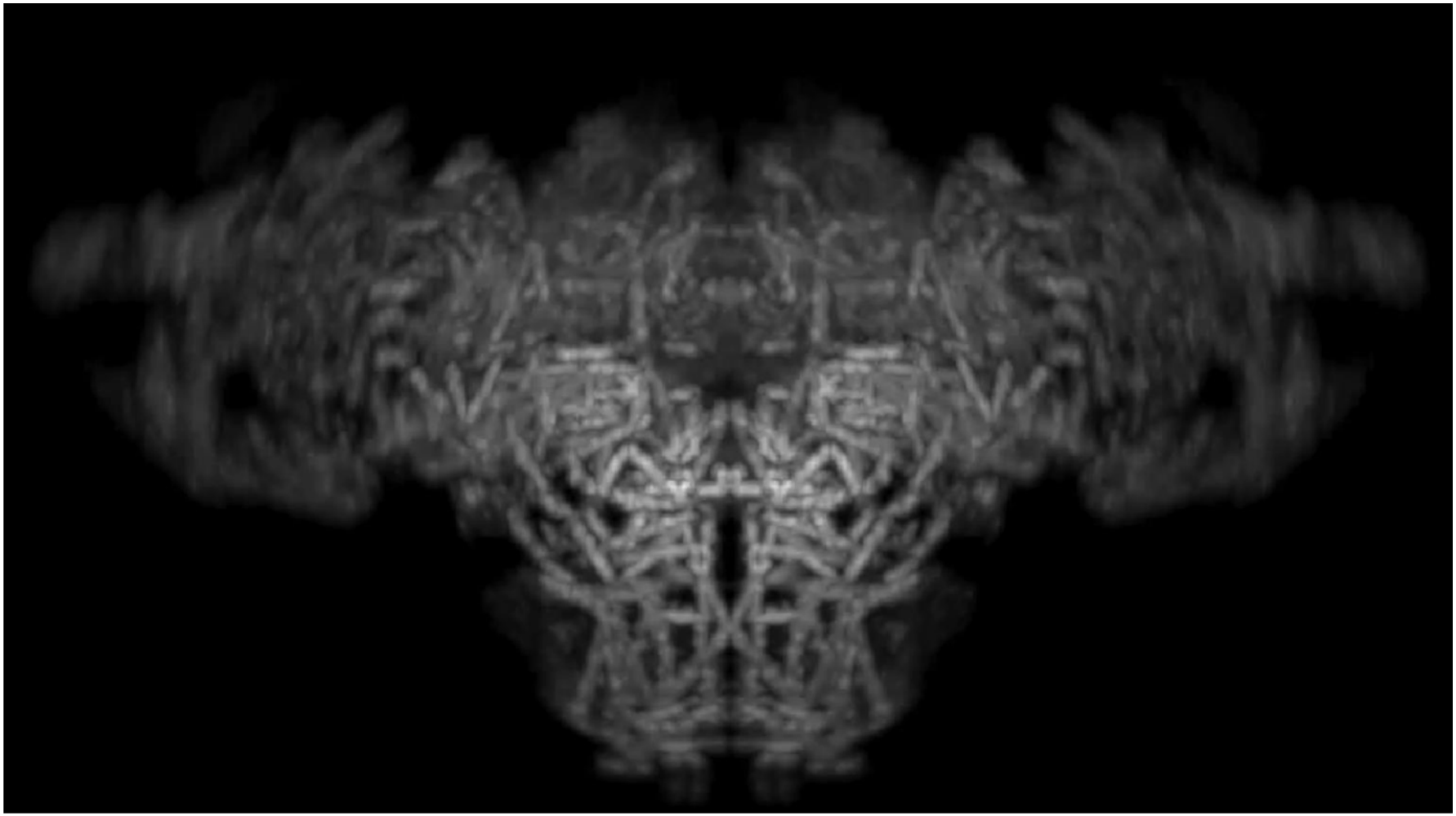
ENERGY LANDSCAPES OF Ca-CHANNEL
PATH TO LIGAND BINDING



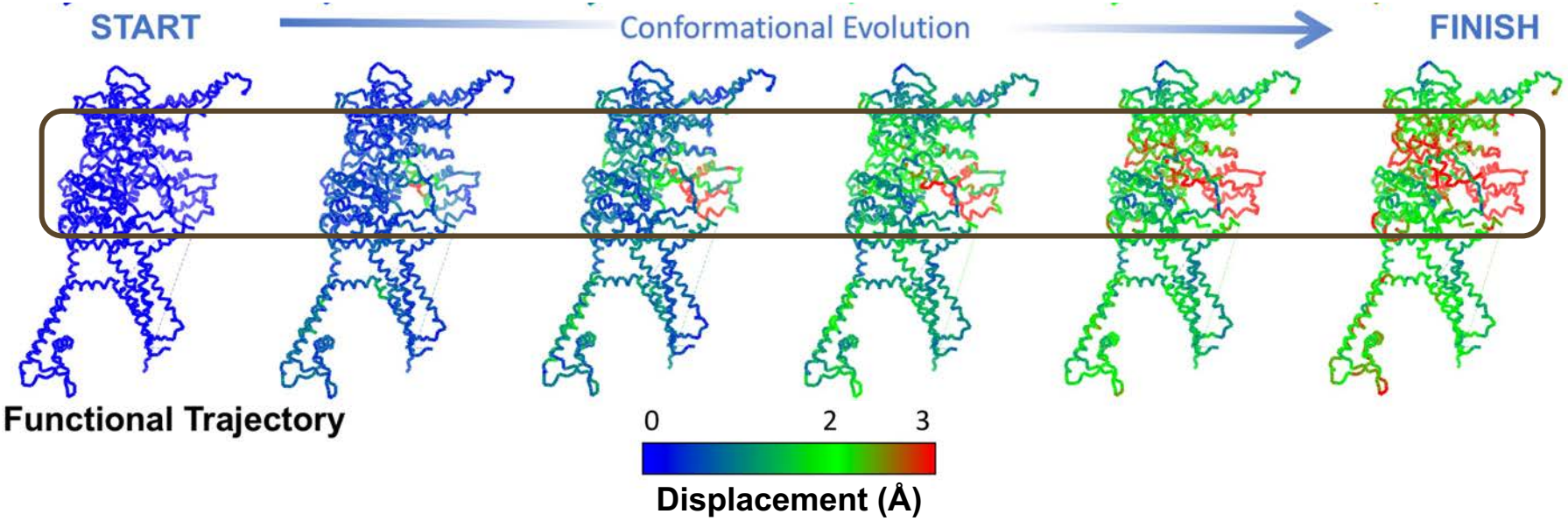
CRYO-EM SNAPSHOT OF Ca-CHANNEL
NO TIMING INFORMATION



Ca-CHANNEL IN ACTION
50-FRAME 3D MOVIE, NO TIMING INFORMATION



NEW INSIGHT: ALLOSTERIC SIGNAL TRANSDUCTION



Structural dynamics of allosteric signal transduction

Extracted from an equilibrium ensemble

Dashti et al.; Unpublished

"DYNAMICS FROM FLUCTUATIONS"

- In equilibrium, number of occupied states at energy E:

$$N(E) = N_{total} \exp - \left(\frac{E}{k_B T} \right)$$

- All states of system sighted with diminishing probability

- Highest energy state in dataset corresponds to state sighted only once

$$E_{highest} = k_B T \ln \left(\frac{N_{total}}{1} \right) = k_B T \ln \left(N_{total} \right)$$

- Total number of snapshots determines highest accessible energy

- Emerging datasets offer access to states with energies of up to $18 k_B T$
  - XFELs promise ~10^8 snapshots per experiment

FUTURE DIRECTIONS

- Function accessible with & without timing
  - Optically and chemically induced reactions
  - Femtosecond movies with atomic resolution
- Energy landscapes reveal functional paths
  - Rigorous, general framework for thermodynamics of function
  - Degrees of freedom, reaction coordinates
  - Detailed mechanisms (cf. ligand binding)
- Emerging large datasets promise access to transition states
  - Using ensembles in equilibrium, without any timing information
- Data-driven methods key to the future
  - Data-driven prediction tantamount to falsifiable theory

DATA DELUGE & MACHINE LEARNING

- No single ML technique a panacea
- Deep Learning excellent at providing answers
  - Chess, poker, facial recognition, instrument control, etc.
  - “Give me a lock, I’ll give you the key”
- Science seeks understanding
  - “Give me a lock, I’ll tell you how it works”
- “The deeper the learning, the shallower the understanding”
  - Difficult to understand nonlinear multilayer systems
- Machine-learning route to understanding
  - In geometric terms (as in GR Field Equations)

FINANCIAL SUPPORT

U.S. DEPARTMENT OF ENERGY Office of Science

BES DE-SC0002164: Underlying dynamical techniques

NSF BioXFEL

STC 1231306 : Data-analytical techniques & models

BILL & MELINDA GATES foundation

HEALTHY BIRTH, GROWTH & DEVELOPMENT

knowledge integration

OPP1159524: Fetal GA estimation