Structure & Dynamics from Random Observations

Abbas Ourmazd

University of Wisconsin Milwaukee

BioXFEL
NSF Science & Technology Center
# Acknowledgments

<table>
<thead>
<tr>
<th>Institution</th>
<th>Contributors</th>
<th>Location</th>
<th>Collaborators</th>
</tr>
</thead>
<tbody>
<tr>
<td>UWM</td>
<td>A. Dashti</td>
<td>CFEL, Hamburg</td>
<td>R. Santra</td>
</tr>
<tr>
<td></td>
<td>R. Fung</td>
<td></td>
<td>O. Vendrell</td>
</tr>
<tr>
<td></td>
<td>A. Hosseinzadeh</td>
<td></td>
<td>A. Hanna</td>
</tr>
<tr>
<td></td>
<td>G. Mashayekhi</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P. Schwander</td>
<td>Columbia</td>
<td>J. Frank et al.</td>
</tr>
<tr>
<td></td>
<td>R. Sepehr</td>
<td></td>
<td>D. Giannakis et al.</td>
</tr>
<tr>
<td>BioXFEL</td>
<td>Fromme et al.</td>
<td>Northwestern</td>
<td>T. Seideman</td>
</tr>
<tr>
<td></td>
<td>Schmidt et al.</td>
<td></td>
<td>S. Ramakrishna</td>
</tr>
<tr>
<td></td>
<td>Spence et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Many others</td>
<td>SPI</td>
<td>Aquila et al.</td>
</tr>
</tbody>
</table>

Abbas Ourmazd
Concepts

- **Classic**: Best experiments ⇔ Greatest control
  - Limits extracted information

- **Data-driven**: Richest datasets ⇔ Random observations
  - Control VERY poor for new tools (e.g., XFELs)
  - Allows system to do what it wants to do

- **Search algorithms**: Information from random observations
  - TB - PB datasets
  - Extract specific information at will
  - No preconceived notions (“templates” or “priors”)

- **Tools**: Manifold Learning
  - Information from geometric structure of data
  - Plus constraints due to nature of relevant spaces
Structural Dynamics from Random Observations

- **3D Structure of objects**
  - Snapshots from unknown orientations
  - Popular with mathematicians, esp. with simulated snapshots
  - Solved problem (multiple times)
  - Examples: Viruses, ribosome, single protein molecules

- **Nanomachine work cycles**
  - Snapshots from unknown orientations & states
  - Molecular movies along work cycles
  - Map energy landscape, nanomachine thermodynamics
  - Example: Ribosome, Ca-channels

- **Ultrafast dynamics**
  - Accurate histories from low-SNR snapshots with inaccurate timing
  - Example: Ultrafast (femtosecond) bond-breaking
Caveat Emptor!

Fraction of Successful Theories

- Toy Model: 99%
- Simulation: 90%
- Sim. w/ Noise: 9%
- Exptal. Data: 0.9%

Abbas Ourmazd
Ca-Channel in Action
Cryo-EM Snapshot
Ca-Channel in Action
50-frame 3D Movie of Ca-Channel

By manifold-learning from experimental 2D snapshots
Geometric Approach

- All we have is ensemble of diffracted intensities
  - A snapshot is $I_n = (i_1, \ldots, i_p)$
  - Lives in $p$-dimensional data space
  - Intrinsic manifold dimensionality = Degrees of freedom exercised

Dimensionality reduction by Manifold Embedding
Scattering as Mapping
Random Sightings of Identical Objects

Latent Space (Orientations)

Data Space (Intensities)

\[ \Phi : \mathcal{L} \mapsto L^2(\mathbb{R}^2) \]
\[ \mathcal{L} = SO(3) \]

Unknown relation between latent & data space metrics prevents interpretation

Giannakis et al, Optics Express 12, 12799 (2012)
Embedded Manifold
What’s it Mean?

- Manifold gives low-dimensional representation of data
  - Captures (nonlinear) structure of the data

- What does the embedding mean?
  - What does projection on each “new dimension” mean?
  - Fundamental problem in graph theory
  - “We don’t know!” (Coifman 2010)

- Diffusion Map EOF’s ➔ Laplace-Beltrami EF’s
  - With respect to INDUCED (aka unknown) metric

- Symmetry powerful means of injecting physical insight
  - Object symmetries
  - Symmetries of operations
Laplace-Beltrami operator
\[ \Delta_g = \left( \sqrt{|g|} \right)^{-1} \partial_i (\sqrt{|g|} g^{ij} \partial_j) \]

For nearly all scattering scenarios, metric has the form:
\[ g = g_{Top} + \text{[object-specific term]} \]

Metric of the symmetric top:
\[ g_{Top} = \begin{pmatrix} \ell_1^2 & 0 & 0 \\ 0 & \ell_1^2 & 0 \\ 0 & 0 & \ell_3^2 \end{pmatrix} \]

Taub metric in GR; classical & quantum mechanics of rigid rotors

Eigenfunctions are Wigner D-functions
- Yields snapshot orientations
- Projection on known eigenfunctions = Noise reduction
  - Giannakis et al, Optics Exp. 20, 12827 (2012)
  - Schwander et al, Optics Exp. 20, 12799 (2012)
Chaperonin Molecule
From Cryo-EM Snapshots

3D reconstruction of single molecule at 12x lower dose than currently needed
From Structure to Function

- Understanding function drives biology
  - Also (bio)chemistry & (bio)physics

- Function determines structure

- Involves *continuous changes* in structure ("conformations")
  - "Life does not cluster"

- Changes in structure involve changes in energy
  - Otherwise there would be no "definite" structure

- Conformations & energies key to function
  - "Molecular movies" and "energy landscapes"
Metric not separable into orientation & conformation parts

\[
g = \begin{pmatrix}
g(\theta,\theta) & g(c,\theta) \\
g(\theta,c) & g(c,c)
\end{pmatrix}
\]

When conformational change \(<\!\!<\) orientational effect:

- Determine orientations
- Determine conformational spectrum for each orientation
- Patch information over SO(3)
- Often requires special kernel

- Dashti et al, PNAS 111, 17492 (2014)
Biological Nanomachines

- Bombarded by “Brownian impact”
  - From the surrounding thermal bath

- Each trajectory of a single machine is unique
  - Hence irreproducible

- Only ensemble averages meaningful
  - Kinetic / thermodynamic description

- The nature of ensemble used for averaging is KEY
  - What populations can you distinguish before averaging?

- Complete description requires access to energy landscape
  - Each point gives energy of ensemble
A reaction coordinate controls a set of concerted changes.
A point in landscape represents a specific conformation & energy.
Energy Landscape of Yeast Ribosome
Molecular Movie: Minimum-energy Trajectory

Dashti et al. PNAS 2014

L1 stalk movement
Head closure
Head swivel
Ca-Channel in Action
50-frame 3D Movie of Ca-Channel

Abbas Ourmazd
Energy landscape reveals degrees of freedom
- Exercised by system during observation

Dynamics show how a “functional path” is traversed
- Building a house vs. independent movements of the limbs
- Needs timing information

Dynamical information limited by noise & timing uncertainty
- Pump-probe timing jitter (physics, chemistry, biology)
- Non-uniform reaction initiation (chemistry, biology)
- Incomplete information (climate)

Accurate dynamics despite noise & extreme timing uncertainty
- Femtosecond accuracy from data with ~300fs timing uncertainty
Dynamics of Bond-breaking in N$_2$
Experimental ToF Spectral Movie (LCLS)

Poisson noise & 280fs timing jitter
Extracted Dynamics
NLSA Mode 2

X-ray 1\(^{st}\)  IR 1\(^{st}\)

Time resolution: \(~ 1\text{fs}\)
Original timing uncertainty: 280\text{fs}
Wavepacket Dynamics

Wavepacket Evolution

Wavepacket components

Power (A.U.)

2.36 fs
20.2 fs
19.6 fs
16.7 fs
15.2 fs

Frequency (THz)

42.3
49.6
51
59.8
65.7

Vibrational Periods (fs)

<table>
<thead>
<tr>
<th>Measured</th>
<th>Theory</th>
<th>State</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.2</td>
<td>15.1</td>
<td>$X^2\Sigma^+_g$</td>
<td>Previously observed in time domain</td>
</tr>
<tr>
<td>16.7</td>
<td>16.7</td>
<td>$X^1\Sigma^+_g$</td>
<td></td>
</tr>
<tr>
<td>19.6</td>
<td>17.7</td>
<td>$A^2\Pi_u$</td>
<td>Overlap between peaks degrades accuracy</td>
</tr>
<tr>
<td>20.2</td>
<td>22.4</td>
<td>$a^3\Pi_u$</td>
<td></td>
</tr>
<tr>
<td>23.6</td>
<td>23.7</td>
<td>$A^1\Pi_u$</td>
<td></td>
</tr>
</tbody>
</table>
“Super-resolution” Dynamics

- A series of concatenated frames contains weak time-arrow
  - The longer the series, the stronger the arrow
  - NO AVERAGING!

- High-frequency signal damped by timing uncertainty
  - Exponentially, but that does not mean to zero!

- Use Singular Value Decomposition (SVD) to extract signal
  - Linear algebra
  - BUT data structures rarely linear

- Perform SVD on curved hyperplane (plus some tricks)
  - Intrinsically curved data manifolds
Extracting Dynamics

- Snapshots do not capture all system variables
  - Some variables not measured (“projection”)

- Manifold of snapshots \( x_t \) not true representation of system
  - \( x_t \): Measured data vectors

- Form concatenated vectors \( X_t = (x_t, x_{t-\delta t}, \cdots, x_{t-(c-1)\delta t}) \)
  - \( c \) snapshots in each concatenated vector \( X_t \)
  - Successive concatenated vectors displaced by \( \delta t \)

- Manifold defined by \( X_t \) reveals “true” system dynamics
  - As though all relevant system variables had been measured

- System dynamics from “time-lagged embedding”
  - Packard (1980); Takens (1981); Sauer et al. (1991); Deyle & Sughara (2011)
Decomposes data into spatial & temporal eigenmodes
- Equivalent to two PCA’s
- Finds eigenmodes of $XX^T$ and $X^T X$
- LINEAR analysis

Spatial eigenmodes: *topos*
- Correlated spatial variations

Temporal eigenmodes: *chronos*
- Dynamic development of each topo

Each chrono associated with a topo
- Temporal evolution of each spatial eigenmode
Manifold Embedding

- Finds intrinsic (nonlinear) structure of data
  - “Curved” hypersurface; manifold

- Embeds manifold in *Euclidean* space
  - Describes in terms of orthogonal coordinates
  - “Unfurls” curved manifold
SVD on Manifold
Nonlinear Laplacian Spectral Analysis

- Project data on to manifold
  - Describe each point in terms of manifold coordinates

\[ A = X \mu \Psi \]

- Perform SVD on manifold
  - In essence performs SVD of matrix with \(10^{12}\) elements!

- Project back into time domain

Abbas Ourmazd

28
Super-resolution Dynamics
Does it really work?

Abbas Ourmazd
Conclusions

- Random observations contain a wealth of information
  - Ultralow SNR
  - Unknown viewing angle, state of machine
  - Uncertain timing

- Geometric machine learning key to this information
  - With a few (simple) tricks

- General constraints help decrypt the information
  - Incorporated in geometrical approach as manifold isometries

- Access to “super-resolution” dynamics
  - Precise timestamp not same as accurate timestamp
  - Dynamics need not be a function of time!

- Applicable to all vector data
Summary

- **3D structure recovery**
  - No orientational information, low SNR (down to ~1/100)
  - Demonstrated with experimental cryo-EM snapshots

- **Conformational movies and energy landscapes**
  - No orientational or conformational information, low SNR
  - Thermodynamic properties of nanomachines
  - Demonstrated with experimental cryo-EM snapshots

- **Accurate dynamics**
  - Factor of 300 beyond timing uncertainty
  - Demonstrated with spectral snapshots
Extracted Dynamics
NLSA Mode 2

Detailed information on bond-breaking modes