How Math and AI are Revolutionizing Biosciences

Guo-Wei Wei
Mathematics
Biochemistry and Molecular Biology
Michigan State University
http://www.math.msu.edu/~wei

Beyond TDA - Persistent topology and its applications in data sciences
August 28, 2021

Grant support: NIH, NSF, NASA, Pfizer, BMS, MEDC, Georgetown U, COVID-19 HPC Consortium, and MSU Foundation
COVID-19 demonstrates the importance of biosciences

216,322,048 cases
4,500,679 deaths

What can mathematics do?
How will COVID-19 unfold in the future?
The Promise of AI & Machine Learning

**ALPHA FOLD**

won 25 of 43 contests and was ranked 1st among 98 competitors in CASP13, Dec. 2018.
Challenges of AI in biomolecular systems

- **Geometric dimensionality:** $\mathbb{R}^{3N}$, where $N \sim 5000$ for a protein.
- **Machine learning dimensionality:** $> 1024^3 m$, where $m$ is the number of atom types in a protein.
- **Molecules have different sizes --- non-scalable.**
- **Complexity:** intermolecular & intramolecular interactions
Given a protein with $N$ atom and an average of $n$ electrons in each atom.

Basic hypothesis:
Intrinsic physics lies on low-dimensional manifolds in a high dimensional space.

**Two schools of thinking**

- **Fundamentalism; Mechanistic**
  - Multiscale Coarse-grain: $\mathbb{R}^M$ ($3 < M < 3N$)
  - Poisson-Boltzmann, PNP, etc.: $\mathbb{R}^3$
  - Molecular Mechanics: $\mathbb{R}^{3N}$
  - Differentiable Manifold: $\mathbb{R}^2$
  - Algebraic Topology: $\mathbb{R}^1$
  - Graph Theory: $\mathbb{R}^0$

- **Quantum Mechanics**
  - QM/MM: $\mathbb{R}^K$
    - $3N < K < 3N(n+1)$
  - Index Theory: $\mathbb{R}^0$
Our Strategy

Sequence data
Structure data
Biophysics
Bioinformatics
Systems biology
Systems physiology

Biological Discovery

Algebraic topology
Differential geometry
Graph theory
Multiscale PDEs
(Harness a century’s accomplishments in mathematics)

Machine learning
Deep learning
Manifold learning
Reinforcement learning
Generative network

Euler
Lagrange
Hilbert
de Rham
Gauss
Einstein
Hodge
Chern
Cartan
Classical Topology

Möbius Strips (1858)  Klein Bottle (1882)

Torus        Double Torus

Leonhard Paul Euler (Swiss Mathematician, April 15, 1707 – Sept 18, 1783)

Seven Bridges of Königsberg

Augustin-Louis Cauchy, Ludwig Schlafli, Johann Benedict Listing, Bernhard Riemann, and Enrico Betti

Leonhard Euler (1735)
Topological invariants: Betti numbers

\( \beta_0 \) is the number of connected components.
\( \beta_1 \) is the number of tunnels or circles.
\( \beta_2 \) is the number of cavities or voids.

<table>
<thead>
<tr>
<th>Point</th>
<th>Circle</th>
<th>Sphere</th>
<th>Torus</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 = 1 )</td>
<td>( \beta_0 = 1 )</td>
<td>( \beta_0 = 1 )</td>
<td>( \beta_0 = 1 )</td>
</tr>
<tr>
<td>( \beta_1 = 0 )</td>
<td>( \beta_1 = 1 )</td>
<td>( \beta_1 = 0 )</td>
<td>( \beta_1 = 2 )</td>
</tr>
<tr>
<td>( \beta_2 = 0 )</td>
<td>( \beta_2 = 0 )</td>
<td>( \beta_2 = 1 )</td>
<td>( \beta_2 = 1 )</td>
</tr>
</tbody>
</table>
Vietoris-Rips complexes of planar point sets

**Simplexes:**

- 0-simplex
- 1-simplex
- 2-simplex
- 3-simplex

**Simplicial complexes of ten points:**
Persistent homology

Simplexes:
- 0-simplex
- 1-simplex
- 2-simplex
- 3-simplex

**k-chain:**
\[ K = \left\{ \sum_j c_j \sigma^q_j \right\} \]

**Chain group:**
\[ C_q(K, \mathbb{Z}_2) \]

**Boundary operator:**
\[ \partial_q \sigma^q = \sum_{j=0} \left\{ (-1)^j \{v_0, v_1, ..., \hat{v}_j, ..., v_k \} \right\} \]

**Cycle group:**
\[ Z_q = \text{Ker} \partial_q \]

**Boundary group:**
\[ B_q = \text{Im} \partial_{q+1} \]

**Homology group:**
\[ H_q = Z_q / B_q \]

**Betti number:**
\[ \beta_q = \text{Rank}(H_q) \]

Xia, Wei, IJNMBE, 2014;
Xia, Feng, Tong, Wei, JCC, 2015
Algebraic Topology

Vietoris-Rips complexes, **persistent homology** and **topological fingerprint**

(Xia, Wei, 2014)
Topological fingerprints of an alpha helix

Beta barrel

Microtubule

(Xia & Wei, IJNMBE, 2014, 2015)
Algebraic Topology

2D persistent homology of protein unfolding (1UBQ)

(Xia & Wei, JCC, 2015)
Persistent cohomology
(incorporating non-geometric information in topology)

Wasserstein curves
Optimal transport

Zixuan Cang
SIAM JMDS
2020
Minimal Surfaces
A way to minimize energy and maximize stability

Viral morphology

Leonhard P. Euler
(Swiss Mathematician, April 15, 1707 – Sept 18 1783)

Joseph L. Lagrange
(Italian Mathematician, January 25 1736 – April 10, 1813)

Man-made life, Mycoplasma mycoides
Differential geometry based minimal surface model

\[ G = \int \gamma [\text{area}] \, dr \quad \text{area} = |\nabla S| \]

where \( G \) is the surface energy, \( \gamma \) is the surface tension, and \( S \) is a surface characteristic function:

Generalized Laplace-Beltrami flow:

\[ \frac{\partial S}{\partial t} = |\nabla S| \left[ \nabla \cdot \frac{\gamma \nabla S}{|\nabla S|} \right] \]

Mean curvature

(Bates, Wei, Zhao, 2006; JCC, 2008; Zhao, Cang, Tong & Wei, Bioinformatics 2018)
Differential Geometry (Connections & curvature forms)

Mean curvatures of subcellular structures

Kelin Xia

Gauss

Minimum

Maximum

Shape index

Curvedness

(HIV receptor)

(Feng, Xia, Tong and Wei, JCP, IJNMBI, 2012)
De Rham-Hodge theory and discrete exterior calculus

Hodge decomposition:

\[ \text{A vector field} = \text{Harmonic} + \text{curl-free} + \text{divergent-free} \]

(Zhao, Wang, Chen, Tong & Wei, BMB, 2020)

Cryo-EM data:

\[ \text{Input} = \text{Normal Gradient} \oplus \text{Tangential Curl} \oplus \text{Tangential Harmonic} \oplus \text{Central Harmonic} \]
Algebraic topology

Multiscale analysis

Differential geometry

Evolutionary de Rham-Hodge theory

Persistent homology

Manifold convergence

Multiscale analysis
Evolutionary de Rham-Hodge

Filtration of a manifold

\[ M_0 \xrightarrow{\mathcal{I}_{0,1}} M_1 \xrightarrow{\mathcal{I}_{1,2}} M_2 \xrightarrow{\mathcal{I}_{2,3}} \ldots \xrightarrow{\mathcal{I}_{n-1,n}} M_n \xrightarrow{\mathcal{I}_{n,n+1}} M \]

De Rham complexes induced by filtration

\[
\begin{align*}
\Omega^0_n(M_0) &\xrightarrow{d^0} \Omega^1_n(M_0) &\xrightarrow{d^1} \Omega^2_n(M_0) &\xrightarrow{d^2} \Omega^3_n(M_0) \\
\Omega^0_n(M_1) &\xrightarrow{d^0} \Omega^1_n(M_1) &\xrightarrow{d^1} \Omega^2_n(M_1) &\xrightarrow{d^2} \Omega^3_n(M_1) \\
\Omega^0_n(M_2) &\xrightarrow{d^0} \Omega^1_n(M_2) &\xrightarrow{d^1} \Omega^2_n(M_2) &\xrightarrow{d^2} \Omega^3_n(M_2) \\
&\ldots &\ldots &\ldots \\
\end{align*}
\]

(Chen, Zhao, Tong & Wei, DCDS-B 2020)
Obtain multiscale spectral geometry & persistent topology from $k$-form Hodge Laplacians! $\Delta^{l,p}_k = \partial_k^{l+1} d_k^l + d_{k-1}^{l+p} \partial_k^{l+p}$

(Chen, Zhao, Tong & Wei, DCDS-B 2020)
Algebraic Graph Theory for Biomolecules

Molecular graph $G(V,E)$

Adjacency matrix of $G(V_{ON},E)$

Can one hear the shape of a drum?

Laplacian matrix of $G(V_{ON},E)$

Eigenvalues: $\lambda_1^A, \lambda_2^A, ...$

(Nguyen and Wei, JCIM, 2019)
Geometric Graph Theory

- **Multiscale weighted colored graphs (MWCG)**
- MWCG is about 40% more accurate than Gaussian network model (GNM) in B-factor prediction, based on 364 proteins.

HIV capsid (313,236 residues) would take GNM 120 years to compute!

Persistent Spectral Graph (Persistent Laplacian) (Wang, Nguyen, Wei, IJNMBE, 2020)

- Simplexes ($\sigma^q$):
  - 0-simplex
  - 1-simplex
  - 2-simplex
  - 3-simplex

- $K$-chain: $K = \left\{ \sum_j w_j \sigma_j^q \right\}$

- Chain group: $C_q(K, \mathbb{Z}_2)$

- Boundary operator: $\partial_q: C_q(K) \to C_{q-1}(K)$
  \[ \partial_q \sigma^q = \sum_{j=0}^{q} (-1)^j \{ v_0, v_1, ..., \hat{v}_j, ..., v_q \} \]

- Adjoint boundary operator: $\partial_q^*: C_{q-1}(K) \to C_q(K)$

- $q$-combinatorial Laplacian operator: $\Delta_q = \partial_{q+1} \partial_{q+1}^* + \partial_q^* \partial_q$

- $q$-combinatorial Laplacian matrix: $\mathcal{L}_q = B_{q+1} B_{q+1}^T + B_q^T B_q$

- Betti numbers: $\beta_q = \text{dim}(\mathcal{L}_q(K)) - \text{rank}(\mathcal{L}_q(K))$ = # of zeros eigenvalues of $\mathcal{L}_q(K)$

(Wang, Nguyen, Wei, IJNMBE, 2020) (Goldberg, Thesis, 2002; Horak, Jost, AIM, 2013; Serrano, Gomze, Arxiv, 2019, ...)

Software: HERMES
Mathematical deep learning

Protein-ligand complex
Element specific groups
Mathematical representations
Machine learning prediction

Algebraic topology
Differential geometry
Graph theory
And/or
PDE
Drug Design Data Resource (D3R) Grand Challenges

- Funded in part by National Institute of General Medical Sciences
- Hosted at the University of California, San Diego
- Annually since 2015

GC4 (2018 -2019): 55 research groups
Drug Design Data Resource (D3R) Grand Challenge

**Given data**
- Primary structure: Amino acid sequence
- Math based GAN

**Math based GAN**
- Generative Adversarial Networks
- Input math feature vector
- Training set
- Discriminator
- Generator

**Predicted complex**

**Final predictions to be compared with experiments**

**Experimental** vs **Predicted** binding affinity (kcal/mol)

**Drug pose**

(Nguyen et al, JCAMD, 2018)

**Pose Predictions**
- BACE Stage 1A
  - Pose Predictions (Partials)

**Affinity Predictions**
- Cathepsin Stage 1
  - Combined Ligand and Structure Based Scoring
  - Ligand Based Scoring (No participation)
  - Structure Based Scoring
  - Free Energy Set

- BACE Stage 1
  - Combined Ligand and Structure (No participation)
  - Ligand Based Scoring (Partials) (No participation)
  - Structure Based Scoring (Partials) (No participation)
  - Free Energy Set (No participation)

- BACE Stage 2
  - Combined Ligand and Structure
  - Ligand Based Scoring (Partials)
  - Structure Based Scoring (Partials)
  - Free Energy Set

### D3R Grand Challenge 3 (2017-2018)

**Pose Prediction**
- Cathepsin Stage 1A
  - Pose Predictions (Partials)

**Affinity Rankings excluding Kds > 10 μM**
- Cathepsin Stage 1
  - Scoring (Partials)
  - Free Energy Set

- VEGFR2
  - Scoring (Partials)

- JAK2 SC2
  - Scoring (Partials)

- JAK2 SC3
  - Scoring
  - Free Energy Set

**Active / Inactive Classification**
- VEGFR2
  - Scoring (Partials)

- JAK2 SC2
  - Scoring (Partials)

- TIE2
  - Scoring (Partials)
  - Free Energy Set 2

**Affinity Rankings for Cocrytalized Ligands**
- Cathepsin Stage 1
  - Scoring (Partials)
  - Free Energy Set

### D3R Grand Challenge 2 (2016-2017)

**Given:** Farnesoid X receptor (FXR) and 102 ligands

**Tasks:** Dock 102 ligands to FXR, and predict their poses, binding free energies and energy ranking

**Stage 1**
- Pose Predictions (Partials)
  - Scoring (Partials)
  - Free Energy Set 1 (Partials)

**Stage 2**
- Scoring (Partials)
  - Free Energy Set 1 (Partials)
  - Free Energy Set 2 (Partials)
Life cycle of SARS-CoV-2 in host cells

1. ACE2 binds to SARS-CoV-2.
2. TMPRSS2 cleaves the S protein.
3. Replicase complex forms 5’ 3’ RNA genome.
4. Translation of N nucleocapsid protein.
5. N nucleocapsid enters the cytoplasm.
6. Nucleocapsid and other proteins assemble into NPs.
7. Release of viral genome from NPs.
8. Nucleocapsid (N) is translated.
9. Spike (S), Membrane (M), Envelope (E) are translated.
10. Virus release from host cells.
What governs SARS-CoV-2 transmission and evolution (new variants)?
Mutations Strengthened SARS-CoV-2 Infectivity

We discovered the mechanism of viral transmission and evolution

Jiahui Chen¹, Rui Wang¹, Menglun Wang¹ and Guo-Wei Wei¹,²,³

1 - Department of Mathematics, Michigan State University, MI 48824, USA
2 - Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA
3 - Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Correspondence to Guo-Wei Wei: wei@math.msu.edu
https://doi.org/10.1016/j.jmb.2020.07.009
Edited by Anna Panchenko

Received 4 June 2020;
Received in revised form 9 July 2020;
Accepted 17 July 2020
Available online 23 July 2020

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infectivity is a major concern in coronavirus disease 2019 (COVID-19) prevention and economic reopening. However, rigorous determination of SARS-CoV-2 infectivity is very difficult owing to its continuous evolution with over 10,000 single nucleotide polymorphisms (SNP) variants in many subtypes. We employ an algebraic topology-based machine learning model to quantitatively evaluate the binding free energy changes of SARS-CoV-2 spike glycoprotein (S protein) and host angiotensin-converting enzyme 2 receptor following mutations. We reveal that the SARS-CoV-2 virus becomes more infectious. Three out of six SARS-CoV-2 subtypes have become slightly more infectious, while the other three subtypes have significantly strengthened their infectivity. We also find that SARS-CoV-2 is slightly more infectious than SARS-CoV according to computed S protein-angiotensin-converting enzyme 2 binding free energy changes. Based on a systematic evaluation of all possible 3686 future mutations on the S protein receptor-binding domain, we show that most likely future mutations will make SARS-CoV-2 more infectious. Combining sequence alignment, probability analysis, and binding free energy calculation, we predict that a few residues on the receptor-binding motif, i.e., 452, 489, 500, 501, and 505, have high chances to mutate into significantly more infectious COVID-19 strains.

© 2020 Elsevier Ltd. All rights reserved.
We predicted key mutation sites in prevailing variants
Mutations at 501 and 452 in prevailing SARS-CoV-2 variants
We discovered the mechanism of viral transmission and evolution of all mutations on the RBD, which potentially increases the complexity of antiviral drug and vaccine development. This global analysis indicates that mutations on the RBD strengthen the binding of S protein and ACE2, leading to more infectious SARS-CoV-2.

We hypothesize that natural selection favors those mutations that enhance the viral transmission and if our predictions are correct, the predicted infectivity strengthening mutations will outpace predicted infectivity weakening mutations over time. Figure 3 illustrates the increase in the frequency of each strengthening mutations occurred. It is interesting to note that overall, infectivity-strengthening mutations grow faster than infectivity-weakening mutations, which also reveals that SARS-CoV-2 subtypes having infectivity-strengthening mutations are able to infect more people. Specifically, frequencies of S477N, N439K, V483A, and V367F are higher than those of other mutations, indicating these mutations have a stronger transmission capacity.

The SARS-CoV-2 genotypes are clustered into six clusters or subtypes based on their single nucleotide

All experiments, if done correctly, confirm our hypothesis

Figure 3. The time evolution of 89 SARS-CoV-2 S protein RBD mutations. The red lines represent the mutations that strengthen the infectivity of SARS-CoV-2 (i.e., $\Delta \Delta G$ is positive), and the blue lines represent the mutations that weaken the infectivity of SARS-CoV-2 (i.e., $\Delta \Delta G$ is negative). Many mutations overlap their trajectories. Here, the collection date of each genome sequence that deposited in GISAID is applied.
Mutation-induced binding free energy changes for spike protein-ACE-2 complex (more infectious)

The most observed 100 out of 651 RBD mutations from 506,768 patient genome sequences

The odd for these 100 most observed mutations to be here accidentally is smaller than one chance in 1.2 nonillion!

\((1.2 \times 10^{30})\)

Wang, Chen, Gao, Wei, Genomics, 2021
Genome-Math-AI modeling of protein-protein binding affinity changes following mutations

1,489,884 patients ➔ Viruses ➔ Sequencing ➔ Genotyping

GISAID

24,478 single mutations

Convolution ➔ Topological barcodes + Other inputs

ACE2

683 RBD mutations

130 antibodies

Pooling & Dropout ➔ Convolution ➔ Flattening ➔ BFE changes
Comparison between experimental data (Top panel, Science, 370(6521): 1208, 2020) and machine learning predicted RBD-mutation-induced BFE changes (Bottom panel) for the SARS-CoV-2 S protein and CTC-445.2 complex. The high similarity between these heatmaps demonstrates the reliability of our machine learning predictions. Our model was extensive validated and trained with tens of thousands of experimental data points.
Prediction of mutational impacts to Eli Lilly antibody therapy

The UK variant (N501Y) makes this antibody slightly more effective.

The South Africa variant (E484K) will reduce Eli Lilly antibody efficacy by about 25 times.

The California variant (L452R) will reduce Eli Lilly antibody efficacy by about 2.7 times.

Atlas of emerging variants

2 co-mutations:
- Delta
- Lambda

3 co-mutations:
- Xi
- Omicron
- Mu

4 co-mutations:
- Beta
- Beta plus
- Gamma
- Nu

Delta plus
The last frontier of science is biology.

The last frontier of biology is mathematics.