Mathematical AI for Drug Discovery

Guo-Wei Wei
Mathematics
Michigan State University

http://www.math.msu.edu/~wei

Shanghai Jiao Tong University
June 16, 2020

Grant support: NIH, NSF, Pfizer, BMS, MEDC, Georgetown U, COVID-19 HPC Consortium, and MSU HPCC
Drug design and discovery

1) **Disease identification** (physiology)
2) **Target hypothesis** (biochem./mole. biol.)
3) **Virtual screening:** drug pose, binding affinity, solubility, partition coefficient, toxicity, and side-effects (biophysics/bioinformatics)
4) **Drug structural optimization in the target binding site** (biochemistry/biophysics/synthetic chem.)
5) **Preclinical in vitro and in vivo test**
6) **Clinical trials**
7) **Optimize drug’s efficacy, pharmacokinetics, and pharmacodynamics properties** (quantitative systems pharmacology)

Influenza -- flu virus

M2 channel

Amantadine

M2-A complex
Drug discovery is one of the most challenging and competitive fields

It takes about $2.6 billion and over 10 years to bring a new drug to the market.

Target selection: 1-2 years
Hit generation: 1-2 years
Lead optimization: 1-2 years
Pre-clinical development: 1-2 years
Clinical trails: 3-7 years
FDA review & approval: 1-2 years

Our concern:

Total global spending on pharmaceutical R&D was about $180 billion in 2018 and will grow to $213 billion in 2024.

NSF Budget: < $8 billion
NIH: Budget < $35 billion
The Promise of AI & Machine Learning

AlphaFold

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

---

Reductionism; Data-driven

Fundamentalism; Mechanistic

Two schools of thinking

Quantum Mechanics
\[ \mathbb{R}^{3Nn+3N} \]

Molecular Mechanics
\[ \mathbb{R}^{3N} \]

Multiscale Coarse-grain
\[ \mathbb{R}^{M} \] \( (3<M<3N) \)

Poisson-Boltzmann, PNP, etc.
\[ \mathbb{R}^{3} \]

Differentiable Manifold
\[ \mathbb{R}^{2} \]

Algebraic Topology
\[ \mathbb{R}^{1} \]

Graph Theory
\[ \mathbb{R}^{0} \]

Index Theory
\[ \mathbb{R}^{0} \]
Our Strategy

Sequence data
Structure data
Biophysics
Bioinformatics
Systems biology
Systems physiology

Drug Design & 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

Harness a century’s accomplishments in mathematics
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 Konigsberg

Augustin-Louis Cauchy, Ludwig Schläfli, 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></td>
<td><img src="image" alt="Circle" /></td>
<td><img src="image" alt="Sphere" /></td>
<td><img src="image" alt="Torus" /></td>
</tr>
<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:

\[0\text{-simplex} \quad 1\text{-simplex} \quad 2\text{-simplex} \quad 3\text{-simplex}\]

Simplicial complexes of ten points:
Persistent homology

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

\[ K = \left\{ \sum_j c_j \sigma_j^q \right\} \]

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

Boundary operator:
\[ \partial_q \sigma^q = \sum_{j=0}^q (-1)^j \{v_0, v_1, ..., \nu_j, ..., v_k\} \]

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

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

Homology group:
\[ H_q = Z_q \setminus B_q \]

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


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

(Cang & Wei, 2018)

Zixuan Cang

(Cang & Wei, 2018)

Wasserstein curves

Optimal transport
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

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

Mean curvatures of subcellular structures

(Feng, Xia, Tong and Wei, JCP, IJNMBI, 2012)
Protein (drug) binding site prediction by the product of curvature and electrostatics

(Xia, Feng, Tong & Wei, JCP 2013; Zhao, Cang, Tong & Wei, Bioinformatics, 2018)
De Rham-Hodge theory and discrete exterior calculus

Hodge decomposition:

A vector field = Harmonic + curl-free + divergent-free

(Zhao, Wang, Tong & Wei, 2018)

Maxwell theory
Quantum mechanics/Solid state theory
Quantum field theory/Yang-Mills theory

Cryo-EM data:
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

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

(Chen, Zhao, Tong & Wei, 2019)
Evolutionary de Rham-Hodge

Obtain multiscale spectral geometry & persistent topology from $k$-form Hodge Laplacians!

$$\Delta_k^{l,p} = \partial_{k+1}^l d_k^l + d_{k-1}^{l+p} \partial_k^{l+p}$$

(Chen, Zhao, Tong & Wei, 2019)
Algebraic Graph Theory for Biomolecules

Molecular graph $G(V,E)$

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

$$
\begin{pmatrix}
0 & \Phi_{12} & \Phi_{13} & 0 \\
\Phi_{12} & 0 & 0 & \Phi_{24} \\
\Phi_{13} & 0 & 0 & \Phi_{34} \\
0 & \Phi_{24} & \Phi_{34} & 0
\end{pmatrix}
$$

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

Can one hear the shape of a drum?

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

$$
\begin{pmatrix}
\Phi_{12} + \Phi_{13} & -\Phi_{12} & -\Phi_{13} & 0 \\
-\Phi_{12} & \Phi_{12} + \Phi_{24} & 0 & -\Phi_{24} \\
-\Phi_{13} & 0 & \Phi_{13} + \Phi_{34} & -\Phi_{34} \\
0 & -\Phi_{24} & -\Phi_{34} & \Phi_{24} + \Phi_{34}
\end{pmatrix}
$$

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

(Nguyen and Wei, 2018)
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!

\[
\Gamma_{ij}(\Phi) = \begin{cases} 
-\Phi(r_{ij}, \eta), & i \neq j, \\
- \sum_{j, j \neq i}^N \Gamma_{ij}, & i = j
\end{cases}
\]

\[
\Phi(r_{ij}, \eta) = \begin{cases} 
1, & r_{ij} \rightarrow 0 \\
0, & r_{ij} \rightarrow \infty \\
e^{-\frac{(r_{ij} / \eta)^\kappa}}, & \text{other}
\end{cases}
\]

\[
B_{i}^{FRI} = a(\Gamma_{ii}(\Phi))^{-1}
\]

Persistent Spectral Graph

(Wang, Nguyen, Wei, 2019)

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

- Adjoint boundary operator:
  \[ \partial_q^* : C_{q-1}(K) \rightarrow C_q(K) \]

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

- $q$-combinatorial Laplacian matrix:
  \[ L_q = B_{q+1} B_{q+1}^T + B_q^T B_q \]

- Betti numbers:
  \[ \beta_q = \dim(L_q(K)) - \text{rank} \left( L_q(K) \right) = \# \text{ of zeros eigenvalues of } L_q(K) \]

(Wang, Nguyen, Wei, 2019)

(Goldberg, Thesis, 2002; Horak, Jost, AIM, 2013; persistence!
Serrano, Gomze, Arxiv, 2019,...)
Multiscale: The Poisson-Boltzmann equation

- Discontinuous dielectric constant at the interface
- Non-smooth interface (geometric singularity)
- Singular charges (delta functions)

Chern et al, 2003; Geng, Yu, Wei, JCP, 2007; Geng, Zhao, JCP 2017

\[-\nabla \cdot (\varepsilon(r) \nabla \phi) = \sum_i q_i c_i e \frac{q_i \phi}{kT} + \sum_i \left( Q_i \delta(r - r_i) - d_i \cdot \nabla \delta(r - r_i) + \Theta_i : \nabla \nabla \delta(r - r_i) \right)\]

Point charge         Charge polarization (Amoeba)
MIBPB solution of the Poisson equation with protein interfaces

Relative solvation energy deviations over grid refinement for 947 proteins in the Amber test set (Zhao, Wei JCP, 2004; Zhou, Zhao, Wei, JCP, 2006; Yu, Geng, Wei, JCP 2007,...)

Electrostatic binding energies of 14 RNA-protein complexes over grid refinement (Nguyen, Wang, Wei, JCC, 2015)

(Wang, Wei, 2015)
Mathematical deep learning

Protein-ligand complex

Element specific groups

Mathematical representations

Machine learning prediction

Algebraic topology

Differential geometry

Graph theory

And/or

PDE
Mathematical learning architecture in the Wei lab

1D convolutional neural networks for mixed features

Multitask convolutional neural networks

Artificial neural networks for multi-task solubility and partition coefficient predictions

Artificial neural networks for single- and multi-task toxicity predictions

Topological Multi-Task Deep Learning

Convolutional neural network assisted multitask gradient boosting trees

(Cang and Wei, PLOS CB, 2017)

(Cang and Wei, Bioinformatics, 2017)

(Wu and Wei, JCC, 2018)

(Wu and Wei, JCIM, 2018)

(Cang and Wei, PLOS CB, 2017)

(Wang et al., Nature MI, 2020)
Topological learning based predictions

Classification of ligands & decoys
DUD database 128,374 protein-ligand/decoy pairs

Predicting mutations on 2648 globular proteins (Cang and Wei, Bioinformatics, 2017)

Predicting mutations on 223 membrane proteins
(Cang and Wei, PLOS CB, 2018)

Prediction RMSD of LogP (Star set)

Binding affinity prediction of PDBBind v2013 core set of 195 protein-ligand complexes

Cang, Mu and Wei, PLOS CB, 2018

Wu and Wei, JCC, 2018

Cang and Wei, PLOS CB, 2017
Performance of algebraic graph learning (AGL-Score) (Nguyen & Wei, JCIM, 2019)

a) Scoring Power (CASF-2013)

- AGL-Score: 0.792
- EIC-Score: 0.774
- PLEC-nn: 0.770
- RF::VinaElem: 0.752
- ΔvinaRF20: 0.686
- X-ScoreHM: 0.614
- ΔSAS: 0.606
- ChemScore@SYBYL: 0.592
- ChemPLP@GOLD: 0.579
- PLP1@DS: 0.568
- AutoDock Vina: 0.564
- G-Score@SYBYL: 0.558
- ASP@GOLD: 0.556
- ASE@MOE: 0.544
- ChemScore@GOLD: 0.536
- D-Score@SYBYL: 0.526
- Alpha-HB@MOE: 0.511
- LUDI3@DS: 0.487
- GoldScore@GOLD: 0.483
- Affinity-dG@MOE: 0.482
- NHA: 0.478
- MWT: 0.473
- LigScore2@DS: 0.456
- GlideScore-SP: 0.452
- Jain@DS: 0.408
- PMF@DS: 0.364
- GlideScore-XP: 0.277
- London-dG@MOE: 0.242
- PMF@SYBYL: 0.221

b) Ranking Power (CASF-2013)

- AGL-Score: 60%
- X-ScoreHM: 59%
- ChemPLP@GOLD: 59%
- ΔvinaRF20: 55%
- PLP2@DS: 55%
- GoldScore@GOLD: 55%
- ChemScore@SYBYL: 54%
- Affinity-dG@MOE: 54%
- LigScore1@DS: 52%
- Alpha-HB@MOE: 52%
- G-Score@SYBYL: 52%
- LUDI1@DS: 52%
- AutoDock Vina: 49%
- D-Score@SYBYL: 49%
- ΔSAS: 49%
- PMF@DS: 49%
- ASP@GOLD: 49%
- ChemScore@GOLD: 46%
- London-dG@MOE: 43%
- PMF@SYBYL: 43%
- GlideScore-SP: 43%
- Jain@DS: 42%
- ASE@MOE: 40%
- GlideScore-XP: 35%

Docking Power (CASF-2013)

- AGL-Score: 90%
- ΔvinaRF20: 87%
- AutoDock Vina: 85%
- ChemPLP@GOLD: 81%
- GlideScore-SP: 78%
- ChemScore@GOLD: 78%
- LigScore2@DS: 77%
- PLP1@DS: 77%
- Alpha-HB@MOE: 75%
- GlideScore-XP: 74%
- ASP@GOLD: 72%
- GoldScore@GOLD: 71%
- Affinity-dG@MOE: 63%
- X-ScoreHM: 61%
- ChemScore@SYBYL: 60%
- London-dG@MOE: 60%
- LUDI1@DS: 59%
- PMF@SYBYL: 52%
- PMF04@DS: 52%
- ASE@MOE: 51%
- Jain@DS: 48%
- G-Score@SYBYL: 45%
- D-Score@SYBYL: 18%
Our competitive edge

Binding affinity ranking of 362 Pfizer compounds (fully blind)

-ML-K-CompH
-ML-D-Lig
-ML-D-Lig,PF
-ML-K-Lig,PF
-ML-K-Comp
-ML-D-Comp
-ML-K-CompH
-Glide-D
-Vina-D

Our predictions

Pfizer pays $1 million annually for this software license

Free software
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

Math based GAN

Predicted complex

Final predictions to be compared with experiments

Drug pose

Experimental

Predicted

Predicted binding affinity (kcal/mol)

Experimental binding affinity (kcal/mol)

(Nguyen et al, JCAMD, 2018)

**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)
- Free Energy Set 2 (partials)

**Stage 2**
- Scoring (partials)
- Free Energy Set 1 (partials)
- Free Energy Set 2 (partials)

*(Nguyen et al, JCAMD, 2018)*

Dr D Nguyen
### D3R Grand Challenge 3 (2017-2018)

(Nguyen et al, JCAMD, 2018)

<table>
<thead>
<tr>
<th>Pose Prediction</th>
<th>Cathepsin Stage 1A</th>
<th>Cathepsin Stage 1B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pose Predictions</strong> (partials)</td>
<td>Pose Prediction</td>
<td>Pose Prediction</td>
</tr>
<tr>
<td><strong>Affinity Rankings excluding Kds &gt; 10 µM</strong></td>
<td><strong>Affinity Rankings excluding Kds &gt; 10 µM</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Cathepsin Stage 1</strong></td>
<td><strong>Cathepsin Stage 2</strong></td>
<td></td>
</tr>
<tr>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
<td></td>
</tr>
<tr>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
<td></td>
</tr>
<tr>
<td><strong>VEGFR2</strong></td>
<td><strong>JAK2 SC2</strong></td>
<td></td>
</tr>
<tr>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
<td></td>
</tr>
<tr>
<td>JAK2 SC3</td>
<td>TIE2</td>
<td></td>
</tr>
<tr>
<td>Scoring</td>
<td>Scoring</td>
<td></td>
</tr>
<tr>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
<td></td>
</tr>
<tr>
<td>4/4</td>
<td>3/3</td>
<td></td>
</tr>
<tr>
<td>4/4</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>4/4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Active / Inactive Classification</th>
<th>VEGFR2</th>
<th>JAK2 SC2</th>
<th>p38-α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
<td></td>
</tr>
<tr>
<td>JAK2 SC3</td>
<td>TIE2</td>
<td>ABL1</td>
<td></td>
</tr>
<tr>
<td>Scoring</td>
<td>Scoring</td>
<td>Scoring</td>
<td></td>
</tr>
<tr>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
<td></td>
</tr>
<tr>
<td>1/5</td>
<td>1/4</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>1/4</td>
<td>1/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affinity Rankings for Cocrystallized Ligands</th>
<th>Cathepsin Stage 1</th>
<th>Cathepsin Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
<td>Scoring (partials)</td>
</tr>
<tr>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
<td>Free Energy Set</td>
</tr>
<tr>
<td>3/17</td>
<td>19/44</td>
<td></td>
</tr>
<tr>
<td>9/17</td>
<td>3/20</td>
<td></td>
</tr>
<tr>
<td>1/4</td>
<td>1/4</td>
<td></td>
</tr>
</tbody>
</table>

Zixuan Cang

Dr. D. Nguyen

Dr. D. Nguyen

#### Pose Predictions

<table>
<thead>
<tr>
<th>BACE Stage 1A</th>
<th>BACE Stage 1B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pose Predictions</strong> (Partials)</td>
<td><strong>Pose Prediction</strong> (Partials)</td>
</tr>
<tr>
<td>1/2</td>
<td>2/2</td>
</tr>
<tr>
<td>3/3</td>
<td>1/2</td>
</tr>
</tbody>
</table>

#### Affinity Predictions

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

<table>
<thead>
<tr>
<th>BACE Stage 1</th>
<th>BACE Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combined Ligand and Structure</strong> (No participation)</td>
<td><strong>Combined Ligand and Structure</strong></td>
</tr>
<tr>
<td><strong>Ligand Based Scoring</strong> (Partials) (No participation)</td>
<td><strong>Ligand Based Scoring</strong> (No participation)</td>
</tr>
<tr>
<td><strong>Structure Based Scoring</strong> (Partials) (No participation)</td>
<td><strong>Structure Based Scoring</strong> (Partials)</td>
</tr>
<tr>
<td><strong>Free Energy Set</strong> (No participation)</td>
<td><strong>Free Energy Set</strong></td>
</tr>
</tbody>
</table>

**Dr. Kaifu Gao Dr. D Nguyen**

---

(RMSD: ~0.5 Å)
### Target Selection: SARS-CoV-2

**Viral Protease**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGFRKMAFSGKVE</td>
<td>SARS-CoV: 96.1%</td>
</tr>
<tr>
<td>GCMVQVTCGTTTTLN</td>
<td></td>
</tr>
<tr>
<td>GLWLDDVYVCPRHV</td>
<td></td>
</tr>
<tr>
<td>ICTSEDMLNPNYED</td>
<td></td>
</tr>
<tr>
<td>LLIR ...</td>
<td></td>
</tr>
<tr>
<td>DEFTPFDVVRQC</td>
<td></td>
</tr>
<tr>
<td>SGVTFQ</td>
<td></td>
</tr>
</tbody>
</table>

**Our homology structures**

Conservative structures!

- Our prediction
- Experimental structure

\[ \text{RMSD (C}_{\alpha} \text{)} = 0.94 \text{ A} \]

*(Nguyen, Gao, Chen, Wang, Wei, 2020)*

**Polypeptide chain**

- Protease
- Drug
- Viral proteins
- Replicase

*GOOD!*
Potentially highly potent drugs for COVID-19

Over 314 SARS 3CL protease inhibitors
Near 20,000 X-ray protein-ligand complex structures

Screening of 8,565 drugs in DrugBank for COVID-19

1,553 FDA approved drugs
7,012 experimental drugs

Algebraic topology
differential geometry
Graph theory

(Gao, Nguyen, Chen, Wang, Wei, JPCL, 2020)
Table 1: A summary of the top 20 potential anti-SARS-CoV-2 drugs from 1553 FDA-approved drugs with their predicted binding affinities (unit: kcal/mol), IC$_{50}$ (μM), and corresponding brand names.

<table>
<thead>
<tr>
<th>DrugID</th>
<th>Name</th>
<th>Brand name</th>
<th>Predicted binding affinity</th>
<th>IC$_{50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB01123</td>
<td>Proflavine</td>
<td>Bayer Pessaries, Molca, Septicide</td>
<td>-8.37</td>
<td>0.72</td>
</tr>
<tr>
<td>DB01243</td>
<td>Chloroxine</td>
<td>Capitrol</td>
<td>-8.24</td>
<td>0.89</td>
</tr>
<tr>
<td>DB08998</td>
<td>Demexiptiline</td>
<td>Deparon, Tinoran</td>
<td>-8.14</td>
<td>1.06</td>
</tr>
<tr>
<td>DB00544</td>
<td>Fluorouracil</td>
<td>Adrucil</td>
<td>-8.11</td>
<td>1.11</td>
</tr>
<tr>
<td>DB03209</td>
<td>Oteracil</td>
<td>Teysuno</td>
<td>-8.09</td>
<td>1.16</td>
</tr>
<tr>
<td>DB13222</td>
<td>Tilbroquinol</td>
<td>Intetrix</td>
<td>-8.08</td>
<td>1.18</td>
</tr>
<tr>
<td>DB01136</td>
<td>Carvedilol</td>
<td>Coreg</td>
<td>-8.06</td>
<td>1.22</td>
</tr>
<tr>
<td>DB01033</td>
<td>Mercaptopurine</td>
<td>Purinethol</td>
<td>-8.04</td>
<td>1.26</td>
</tr>
<tr>
<td>DB08903</td>
<td>Bedaquiline</td>
<td>Sirturo</td>
<td>-8.02</td>
<td>1.29</td>
</tr>
<tr>
<td>DB00257</td>
<td>Clotrimazole</td>
<td>Canesten</td>
<td>-8.00</td>
<td>1.35</td>
</tr>
<tr>
<td>DB00878</td>
<td>Chlorhexidine</td>
<td>Betasept, Biopatch</td>
<td>-8.00</td>
<td>1.35</td>
</tr>
<tr>
<td>DB00666</td>
<td>Nafarelin</td>
<td>Synarel</td>
<td>-8.00</td>
<td>1.35</td>
</tr>
<tr>
<td>DB01213</td>
<td>Fomepizole</td>
<td>Antizol</td>
<td>-7.98</td>
<td>1.39</td>
</tr>
<tr>
<td>DB01656</td>
<td>Roflumilast</td>
<td>Daxas, Daliresp</td>
<td>-7.97</td>
<td>1.41</td>
</tr>
<tr>
<td>DB00676</td>
<td>Benzyl benzoate</td>
<td>Ascabin, Ascabiol, Ascarbin, Tenutex</td>
<td>-7.96</td>
<td>1.45</td>
</tr>
<tr>
<td>DB06663</td>
<td>Pasireotide</td>
<td>Signifor</td>
<td>-7.95</td>
<td>1.47</td>
</tr>
<tr>
<td>DB08983</td>
<td>Etofibrate</td>
<td>Lipo Merz Retard, Liposec</td>
<td>-7.94</td>
<td>1.48</td>
</tr>
<tr>
<td>DB06791</td>
<td>Lanreotide</td>
<td>Somatuline</td>
<td>-7.94</td>
<td>1.48</td>
</tr>
<tr>
<td>DB00027</td>
<td>Gramicidin D</td>
<td>Neosporin Ophthalmic</td>
<td>-7.94</td>
<td>1.48</td>
</tr>
<tr>
<td>DB00730</td>
<td>Thiabendazole</td>
<td>Mintezol, Tresaderm, and Arbotect</td>
<td>-7.93</td>
<td>1.51</td>
</tr>
</tbody>
</table>
Top inhibitors in the FDA approved drugs

(a) Proflavine, -8.37 kcal/mol

(b) SARS-CoV-2 protease and Proflavine complex

(c) Chloroxine, -8.24 kcal/mol

(d) SARS-CoV-2 protease and Chloroxine complex

(e) Demexiptiline, -8.14 kcal/mol

(f) SARS-CoV-2 protease and Demexiptiline complex
SARS-CoV-2 antibody therapies
Binding analysis of SARS-CoV-2 antibody therapies

(Chen, Gao, Wang, Nguyen, Wei, 2020)

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Predicted BA (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR3022</td>
<td>-15.4</td>
</tr>
<tr>
<td>B38</td>
<td>-14.7</td>
</tr>
<tr>
<td>CB6</td>
<td>-13.5</td>
</tr>
<tr>
<td>ACE2</td>
<td>-11.9</td>
</tr>
<tr>
<td>BD23</td>
<td>-10.8</td>
</tr>
<tr>
<td>H11-D4</td>
<td>-10.3</td>
</tr>
<tr>
<td>S309</td>
<td>-9.9</td>
</tr>
<tr>
<td>P2B-2F6</td>
<td>-9.6</td>
</tr>
</tbody>
</table>

(Eli Lilly)

X. Sunney Xie
# Binding analysis of SARS-CoV antibody therapies

(Chen, Gao, Wang, Nguyen, Wei, 2020)

<table>
<thead>
<tr>
<th>MOLECULE</th>
<th>Predicted BA (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80R</td>
<td>-17.3</td>
</tr>
<tr>
<td>ACE2</td>
<td>-11.5</td>
</tr>
<tr>
<td>VHH-72</td>
<td>-9.7</td>
</tr>
<tr>
<td>m396</td>
<td>-9.4</td>
</tr>
<tr>
<td>S230</td>
<td>-7.7</td>
</tr>
<tr>
<td>F26G19</td>
<td>-6.7</td>
</tr>
</tbody>
</table>
SARS-CoV-2 Genotyping

9212 Single Mutations in 17496 SARS-CoV-2 Genomes

3CL and RdRp: drug targets; S: vaccine target; N, RdRp, and E: diagnostic targets

Date=20200602

https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html
SARS-CoV-2 transmission and evolution

Ms. Rui Wang

(Wang, Hozumi, Yin, Wei, JCIM, 2020)
Mutations on SARS-CoV-2

Wang, Hozumi, Yin, Wei, 2020

Cluster A
Cluster B
Cluster C
Cluster D
Mutations on SARS-CoV-2 proteins

Nucleocapsid protein
503 mutations

Spike protein
1004 mutations
Receptor binding domain

Papain-like protease
187 mutations

RNA polymerase
607 mutations

Endoribonuclease
256 mutations

Main protease
203 mutations

Main protease
Binding domain

Envelope protein
52 mutations
### Mutations on SARS-CoV-2 diagnostic targets

<table>
<thead>
<tr>
<th>Primer</th>
<th>MC</th>
<th>SC</th>
<th>Cluster I</th>
<th>Cluster II</th>
<th>Cluster III</th>
<th>Cluster IV</th>
<th>Cluster V</th>
<th>Cluster VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RX7038-N1 primer (Fw)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>10</td>
<td>22</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>RX7038-N1 primer (Rv)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>20</td>
<td>68</td>
<td>0</td>
<td>36</td>
<td>25</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>RX7038-N2 primer (Fw)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3</td>
<td>16</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>RX7038-N2 primer (Rv)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4</td>
<td>15</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>RX7038-N3 primer (Fw)&lt;sup&gt;6&lt;/sup&gt;</td>
<td>7</td>
<td>168</td>
<td>7</td>
<td>149</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>RX7038-N3 primer (Rv)&lt;sup&gt;6&lt;/sup&gt;</td>
<td>8</td>
<td>30</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>N1-U.S.-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>9</td>
<td>298</td>
<td>4</td>
<td>281</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>N2-U.S.-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>7</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>N3-U.S.-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>11</td>
<td>29</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>N-Sarbeco-F&lt;sup&gt;b&lt;/sup&gt;[3]</td>
<td>8</td>
<td>26</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>N-Sarbeco-P&lt;sup&gt;b&lt;/sup&gt;[3]</td>
<td>8</td>
<td>21</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>N-Sarbeco-R&lt;sup&gt;b&lt;/sup&gt;[3]</td>
<td>9</td>
<td>57</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>28</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>N-China-F&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>15</td>
<td>8507</td>
<td>4</td>
<td>84</td>
<td>8186</td>
<td>24</td>
<td>9</td>
<td>199</td>
</tr>
<tr>
<td>N-China-R&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>11</td>
<td>24</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>N-China-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>N-HK-F&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>3</td>
<td>44</td>
<td>0</td>
<td>2</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>N-HK-R&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>5</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>N-JP-F&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>4</td>
<td>11</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>N-JP-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>6</td>
<td>13</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>N-TL-F&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>12</td>
<td>61</td>
<td>6</td>
<td>41</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>N-TL-R&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>8</td>
<td>18</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>N-TL-P&lt;sup&gt;[10]&lt;/sup&gt;</td>
<td>5</td>
<td>28</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>E-Sarbeco-F1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>E-Sarbeco-R2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
## Mutations on SARS-CoV-2 diagnostic targets (continue)

<table>
<thead>
<tr>
<th>Diagnostic Code</th>
<th>Mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-Sarbeco-F1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2 8 1 1 2 0 0 0 4</td>
</tr>
<tr>
<td>E-Sarbeco-R2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2 4 0 2 0 0 1 1 1</td>
</tr>
<tr>
<td>E-Sarbeco-P1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4 21 2 18 1 0 0 0 0</td>
</tr>
<tr>
<td>nCoV-IP2-12669Fw&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2 2 0 1 1 0 0 0 0</td>
</tr>
<tr>
<td><strong>nCoV-IP2-12759Rv&lt;sup&gt;c&lt;/sup&gt;</strong></td>
<td><strong>6 52 16 14 0 13 2 7</strong></td>
</tr>
<tr>
<td>nCoV-IP2-12696bProbe(+)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>3 8 0 0 0 2 0 0 6</td>
</tr>
<tr>
<td>nCoV-IP4-14059Fw&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1 1 0 0 0 0 0 0 1</td>
</tr>
<tr>
<td>nCoV-IP4-14146Rv&lt;sup&gt;c&lt;/sup&gt;</td>
<td>5 19 0 7 1 5 0 0 6</td>
</tr>
<tr>
<td>nCoV-IP4-14084Probe(+)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>7 19 0 6 4 3 0 0 6</td>
</tr>
<tr>
<td>RdRP-SARSr-F2&lt;sup&gt;d&lt;/sup&gt;</td>
<td>5 31 9 0 0 2 0 0 20</td>
</tr>
<tr>
<td>RdRP-SARSr-R1 [3]&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>RdRP-SARSr-P2 [3]&lt;sup&gt;d&lt;/sup&gt;</td>
<td>3 6 0 6 0 0 0 0 1</td>
</tr>
<tr>
<td>ORF1ab-China-F [10]</td>
<td>4 9 3 2 3 0 0 0 1</td>
</tr>
<tr>
<td>ORF1ab-China-R [10]</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>ORF1ab-China-P [10]</td>
<td>4 8 0 3 5 0 0 0 0</td>
</tr>
<tr>
<td>ORF1b-nsp14-HK-F [10]</td>
<td>3 3 0 0 0 2 0 0 1</td>
</tr>
<tr>
<td>ORF1b-nsp14-HK-R [10]</td>
<td>5 13 1 5 2 0 0 0 5</td>
</tr>
<tr>
<td>ORF1b-nsp14-HK-P [10]</td>
<td>3 11 0 0 1 1 1 1 8</td>
</tr>
</tbody>
</table>
Math-AI model for protein-protein binding affinity changes following mutations

Chen, Gao, Wang, Nguyen, Wei, 2020

Topological data analysis

Convolution

Pooling & Dropout

Flattening

Chen, Gao, Wang, Nguyen, Wei, 2020

Other features

\[ H_0 \]

Input
Binding affinity changes for spike protein-ACE-2 complex from SRAS-CoV to SARS-CoV-2 (more infectious)

Chen, Gao, Wang, Nguyen, Wei, 2020

Dr. Jiahui Chen

Chen, Gao, Wang, Nguyen, Wei, 2020)
Mutation induced binding affinity changes for spike protein-ACE-2 complex (getting more infectious!)

Figure 2: Overall binding affinity changes $\Delta \Delta G$ on the receptor-binding domain (RBD). The blue color region marks the binding affinity changes on the receptor-binding motif (RBM). The height of each bar indicates the predicted $\Delta \Delta G$. The color indicates the occurrence frequency in the GISAID genome dataset. (Chen, Gao, Wang, Nguyen, Wei, 2020)
The last frontier of science is biology.
The last frontier of biology is mathematics.