Integration of deep learning and mathematics for drug discovery

Duc Nguyen[1], Zixuan Cang[1], Kaifu Gao[1], and Guo-Wei Wei[1,2,3]

[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*

Introduction

- Drug discovery is the process of finding new medications based on the knowledge of the identified target. Drug discovery is one of the most challenge tasks in the biological sciences since it takes at least 10 years and cost more than $2.6 billion for a novel medicine to travel from its initial discovery to the marketplace, as illustrated in Figure 1.

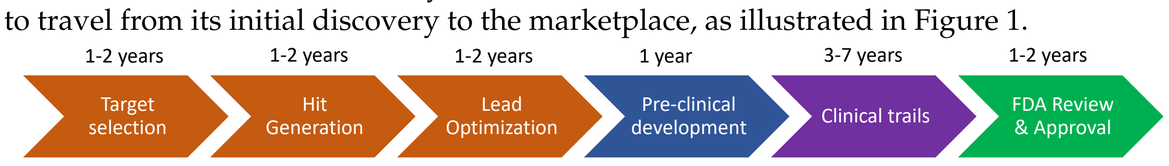
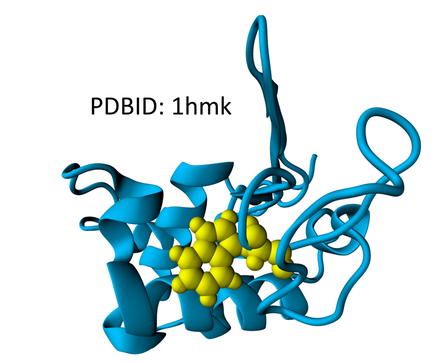


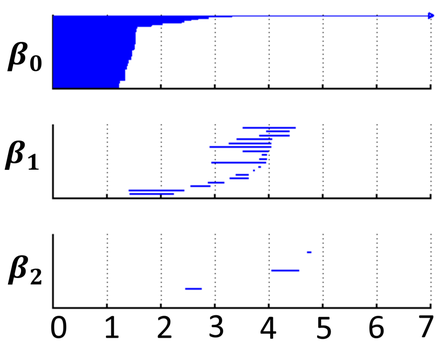
Figure 1: Illustration of the drug discovery cycle.

- Computer-aided drug design (CADD) technology plays a crucial part in drug discovery. Specifically, CADD models are utilized to identify hit-finding activities, optimize, and predict the molecular properties but these models are still massively complex, costly and time-consuming. Therefore, enhancing the CADD process is an urgent need.
- We have developed sophisticated mathematical representations integrated with advanced deep learning architectures to predict the 3D shapes of small molecules as well as their bioactivities at high accuracy with unprecedented speed.

Persistent topology representations

Cang and Wei (2017)





k-chain: $\sum_i c_i \sigma_i^k$

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

Boundary operator:

$$\partial_k \sigma^k = \sum_{i=0}^{k} [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k]$$

Filtration process:

$$\emptyset = X_0 \subseteq X_1 \subseteq \cdots \subseteq X_{m-1} \subseteq X_m = X$$

Persistence

$$\mathcal{B}_k(X) = \mathrm{Im}(\partial_{k+1})$$
$$\mathcal{Z}_k(X) = \mathrm{Ker}(\partial_k)$$
$$\mathcal{H}_k(X) = \mathcal{Z}_k(X)/\mathcal{B}_k(X)$$
$$\beta_k = \mathrm{rank}(\mathcal{H}_k)$$

Differential geometry representations

- **Element interactive densities**

Set of element types: $\mathcal{C} = \{H, C, N, O, S, P, F, Cl, \cdots\}$

Element interactive density

$$\rho_{kk'}(\mathbf{r}, \eta_{kk'}) = \sum_j w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{kk'}), \quad \mathbf{r} \in B(\mathbf{r}_i, r_i), \alpha_j = \mathcal{C}_{k'}; \|\mathbf{r}_i - \mathbf{r}_j\| > r_i + r_j + \sigma, \forall \alpha_i \in \mathcal{C}_k; k \neq k',$$

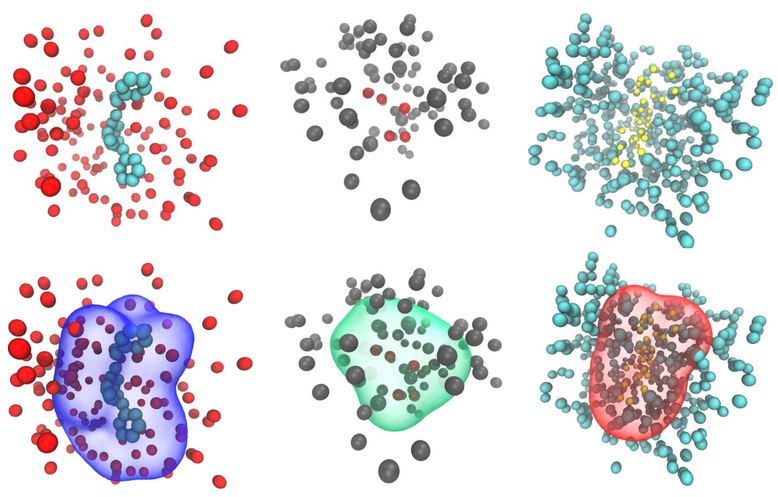
$$\Phi\left(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j\|\right) = 1, \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to 0,$$
$$\Phi\left(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j\|\right) = 0, \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to \infty.$$

- **Element interactive curvatures**

Element interactive manifolds (EIMs)

$$\rho_{kk'}(\mathbf{r}, \eta_{kk'}) = c\rho_{\max}, \quad 0 \leqslant c \leqslant 1 \quad \text{and} \quad \rho_{\max} = \max\{\rho_{kk'}(\mathbf{r}, \eta_{kk'})\}.$$

Element interactive curvatures (EICs): Gaussian, mean, minimum, maximum curvatures are defined on EIMs



Graph theory representations

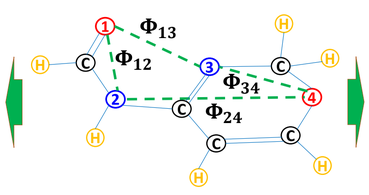
- **Multiscale weighted colored subgraphs**: $G(\mathcal{V}, \mathcal{E}_{kk'})$

$$\mathcal{V} = \{(\mathbf{r}_j, \alpha_j) | \mathbf{r}_j \in \mathbb{R}^3; \alpha_j \in \mathcal{C}; j = 1, 2, \ldots, N\}$$
$$\mathcal{E}_{kk'} = \{\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{kk'}) | \alpha_i = \mathcal{C}_k, \alpha_j = \mathcal{C}_{k'}; i, j = 1, 2, \ldots, N; \|\mathbf{r}_i - \mathbf{r}_j\| > r_i + r_j + \sigma\}$$

- **Weighted colored matrices**:

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

Laplace matrix



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

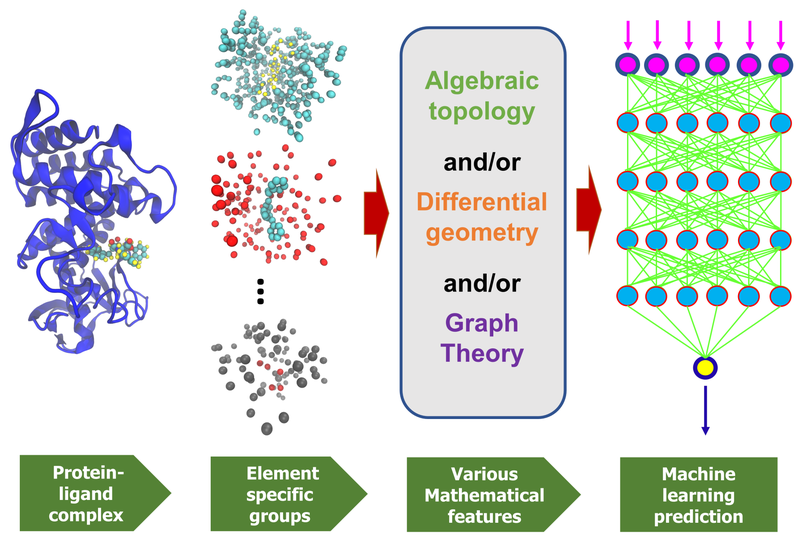
Adjacency matrix

- **Graph invariants**:

Geometric invariant: $\mathrm{Tr}L(\eta_{kk'})$

Algebraic invariant: eigenvalues of adjacency and Laplacian matrices

Math Deep Learning Models (MathDL)



D3R Grand Challenge Results

Nguyen et al. (2019)

D3R Grand Challenge 4 (2018-2019)

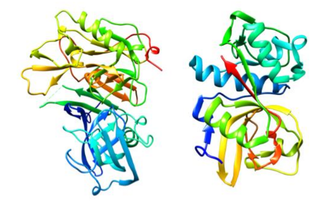
Pose Predictions

BACE Stage 1A
Pose Predictions (Partials) 1/2 3/3

BACE Stage 1B
Pose Prediction (Partials) 2/2 1/2

Affinity Predictions

Cathepsin Stage 1
Combined Ligand and Structure Based Scoring 2/5 2/3 2/4
Ligand Based Scoring (No participation)
Structure Based Scoring 2/4 3/3 3/3
Free Energy Set 1/7 1/7 2/5



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)

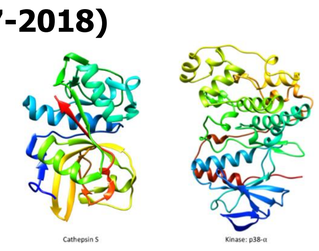
BACE Stage 2
Combined Ligand and Structure
Ligand Based Scoring (No participation)
Structure Based Scoring (Partials)
Free Energy Set 3/4 1/4

D3R Grand Challenge 3 (2017-2018)

Pose Prediction

Cathepsin Stage 1A
Pose Predictions (partials)

Cathepsin Stage 1B
Pose Prediction



Affinity Rankings excluding Kds > 10 μM

Cathepsin Stage 1
Scoring (partials)
Free Energy Set

Cathepsin Stage 2
Scoring (partials)
Free Energy Set

VEGFR2
Scoring (partials)

JAK2 SC2
Scoring (partials)

p38-α
Scoring

JAK2 SC3
Scoring
Free Energy Set

TIE2
Scoring
Free Energy Set 2

ABL1
Scoring (partials)

Active / Inactive Classification

VEGFR2
Scoring (partials)

JAK2 SC2
Scoring (partials)

p38-α
Scoring (partials)

JAK2 SC3
Scoring
Free Energy Set

TIE2
Scoring (partials)
Free Energy Set 1

ABL1
Scoring (partials)

Affinity Rankings for Cocrystalized Ligands

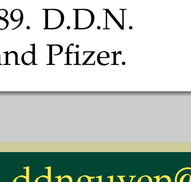
Cathepsin Stage 1
Scoring (partials)
Free Energy Set

Cathepsin Stage 2
Scoring (partials)
Free Energy Set

D3R Grand Challenge 2 (2016-2017)

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)



Conclusion

- Chemical and biological data have been successfully encoded in low-dimensional representations by our advanced mathematical approaches.
- Integration of deep learning and mathematics yields a winning pose and affinity prediction model in D3R Grand Challenges 2, 3 and 4.

References

Cang, Z. X. and Wei, G. W. (2017). TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLOS Computational Biology*, 13(7):e1005690, https://doi.org/10.1371/journal.pcbi.1005690.

Nguyen, D. D., Gao, K., Wang, M., and Wei, G.-W. (2019). Mathdl: Mathematical deep learning for d3r grand challenge 4. *arXiv preprint arXiv:1909.07784*.

Acknowledgment

This work was supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473 and NIH grant GM126189. D.D.N. and G.W.W. are also funded by Bristol-Myers Squibb and Pfizer.

http://www.math.msu.edu/~wei/ http://ddnguyen.org wei@msu.edu ddnguyen@msu.edu