# Mathematics based graph neural network for drug design

Duc Nguyen

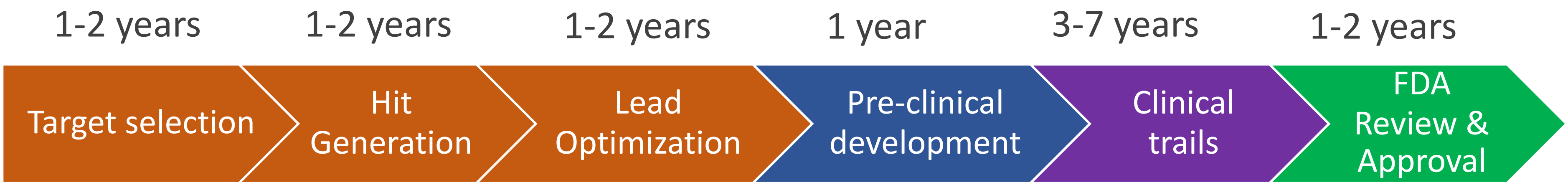
Department of Mathematics

University of  Kentucky

The Fourth TSIMF Conference on Computational and Mathematical Bioinformatics and Biophysics

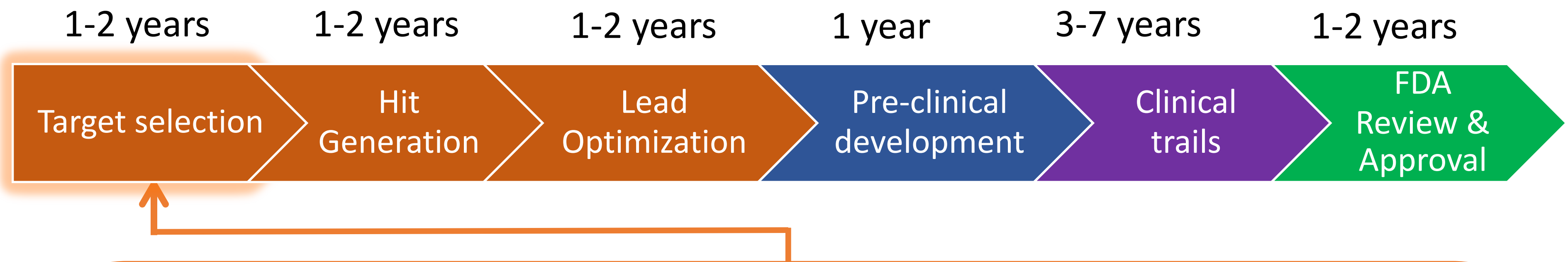
December 12-15, 2021

# Rational Drug Discovery



| | | | | | |
|---|---|---|---|---|---|
| 1-2 years | 1-2 years | 1-2 years | 1 year | 3-7 years | 1-2 years |
| Target selection | Hit Generation | Lead Optimization | Pre-clinical development | Clinical trails | FDA Review & Approval |

# Rational Drug Discovery



1-2 years — Target selection

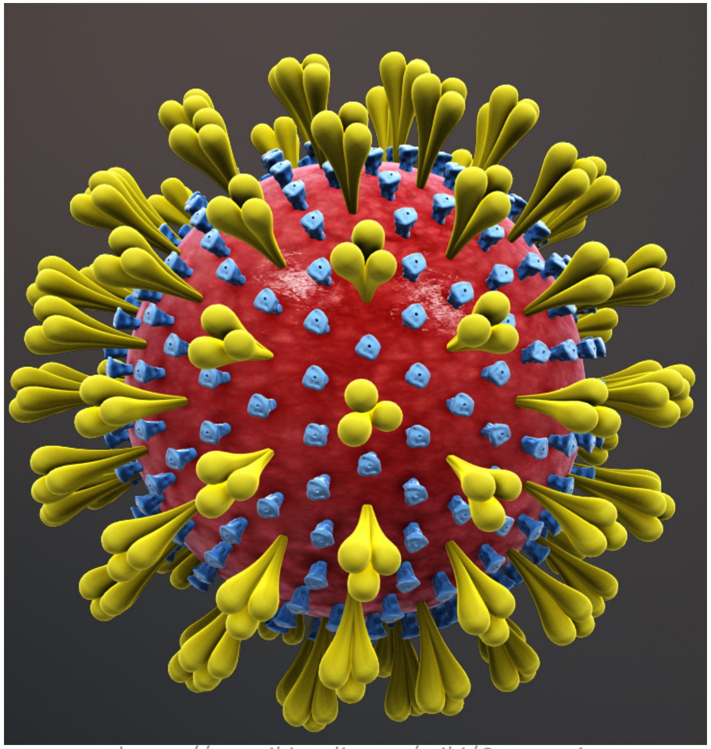
1-2 years — Hit Generation

1-2 years — Lead Optimization

1 year — Pre-clinical development

3-7 years — Clinical trails

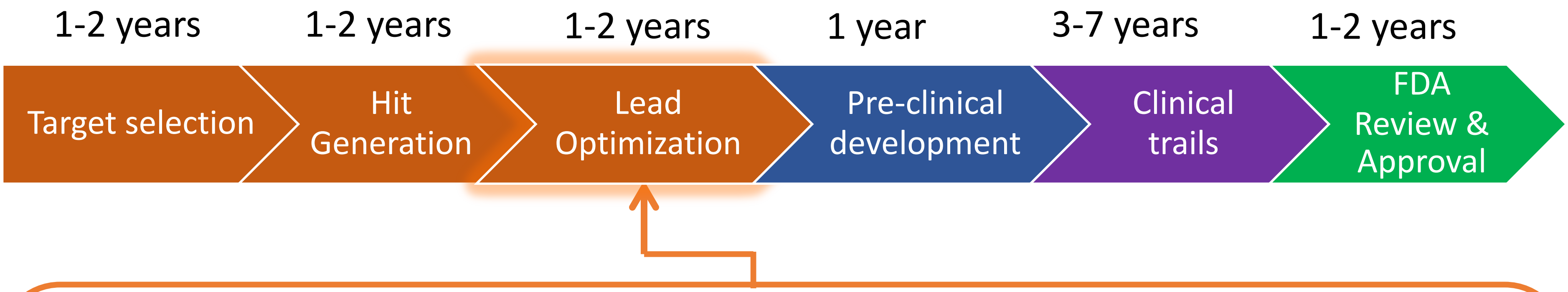
1-2 years — FDA Review & Approval



https://en.wikipedia.org/wiki/Coronavirus

- **COVID-19 (SARS-CoV-2)**
- First reported on Dec 30, 2019
- Global health emergency by WHO on 01-29-20
- As of 12-14-21: 5.31M dead cases, > 271M infected cases

# Rational Drug Discovery



1-2 years
Target selection

1-2 years
Hit Generation

1-2 years
Lead Optimization

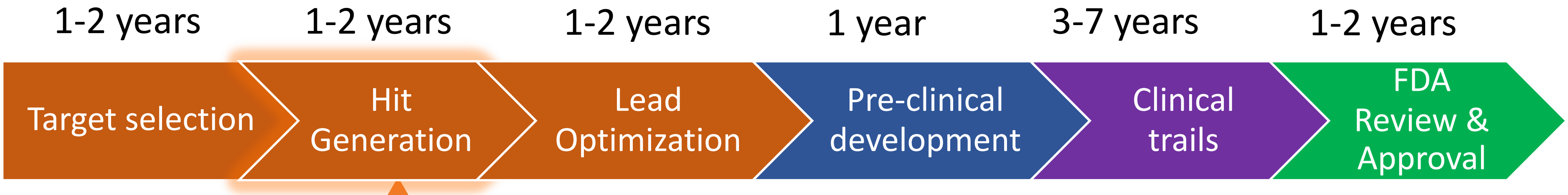
1 year
Pre-clinical development

3-7 years
Clinical trails

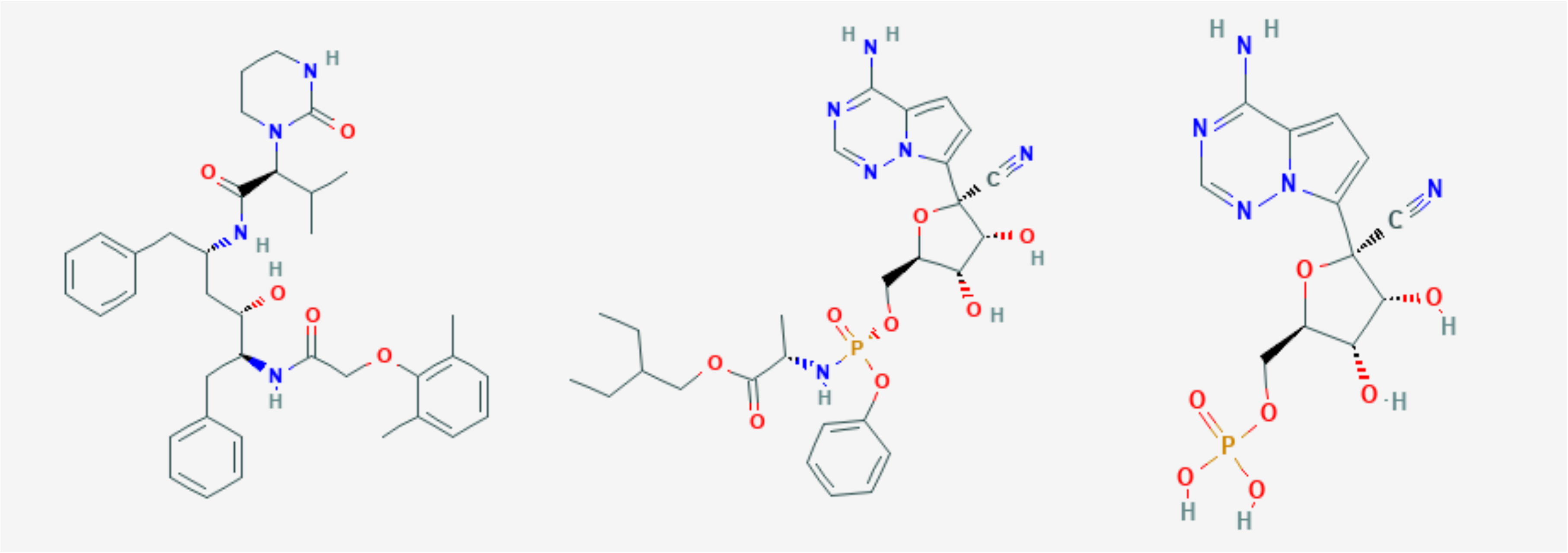
1-2 years
FDA Review & Approval

- Improve potency
- reduced off-target activities
- Reasonable physiochemical/metabolic properties *in vivo* pharmacokinetics

# Rational Drug Discovery

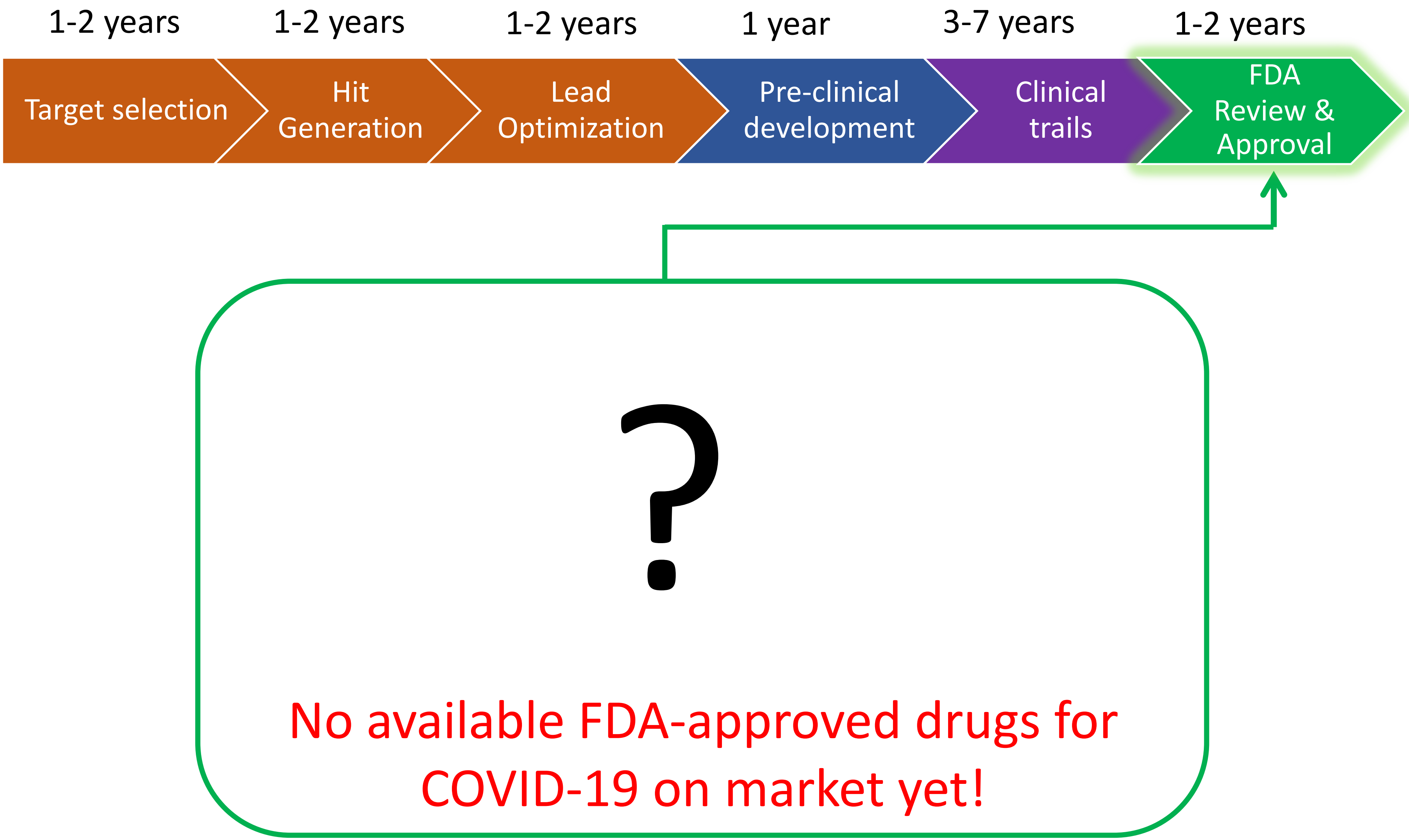


| 1-2 years | 1-2 years | 1-2 years | 1 year | 3-7 years | 1-2 years |
|---|---|---|---|---|---|
| Target selection | Hit Generation | Lead Optimization | Pre-clinical development | Clinical trails | FDA Review & Approval |



Important properties: binding affinity (IC50), toxicity, solubility, …

# Rational Drug Discovery



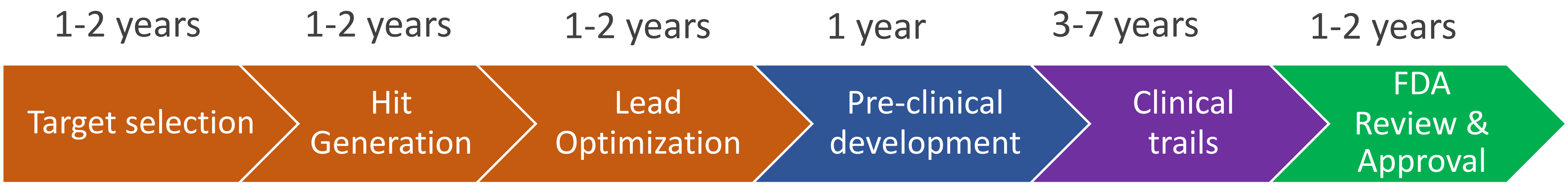
1-2 years | 1-2 years | 1-2 years | 1 year | 3-7 years | 1-2 years

Target selection → Hit Generation → Lead Optimization → Pre-clinical development → Clinical trails → FDA Review & Approval

?

No available FDA-approved drugs for COVID-19 on market yet!

# Rational Drug Discovery



1-2 years — Target selection

1-2 years — Hit Generation

1-2 years — Lead Optimization

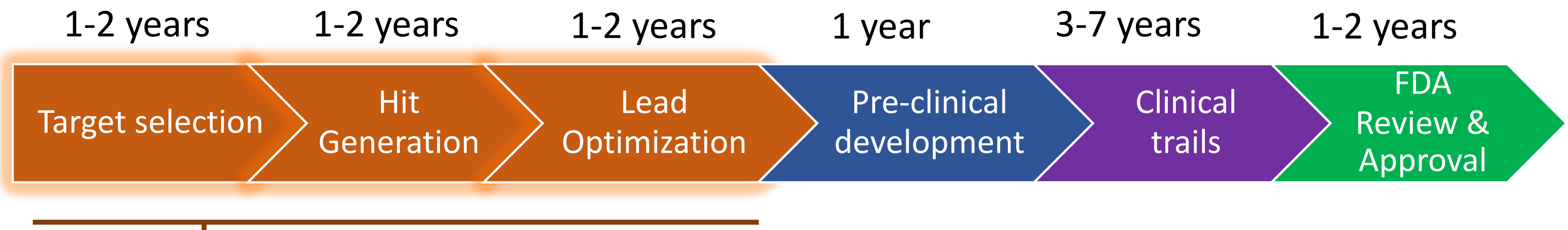
1 year — Pre-clinical development

3-7 years — Clinical trails

1-2 years — FDA Review & Approval

- ✗ Lengthy process ( > 10 years)
- ✗ Expensive ( > $2.6 billion)
- ✗ High failure rate

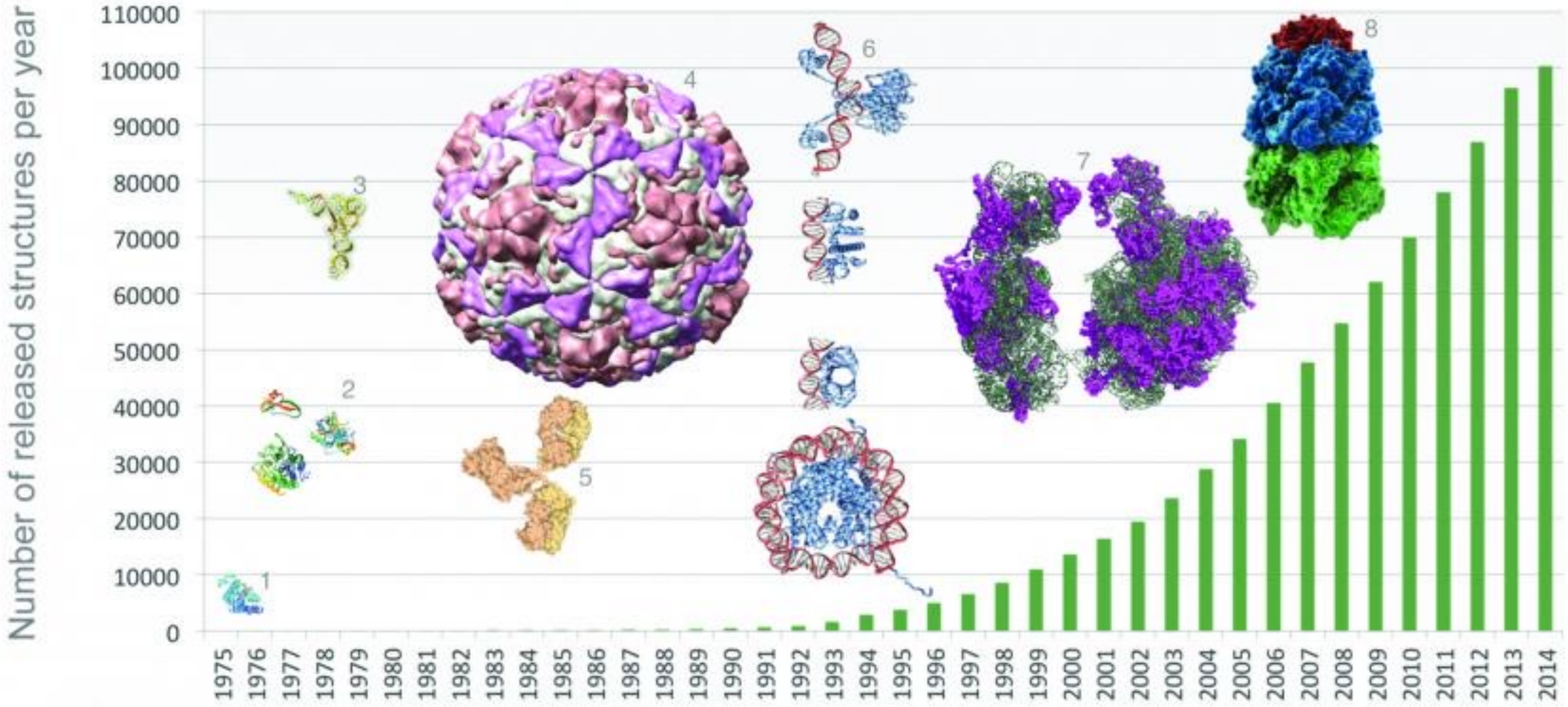
# Rational Drug Discovery



- ☒ Lengthy process ( > 10 years)
- ☒ Expensive ( > $2.6 billion)
- ☒ High failure rate

Great opportunities for Math and AI

# Protein Data Bank, as of 2021: 174,994 structures



230 million compounds

ZINC

1.9 million compounds

ChEMBL

# Molecular Representations

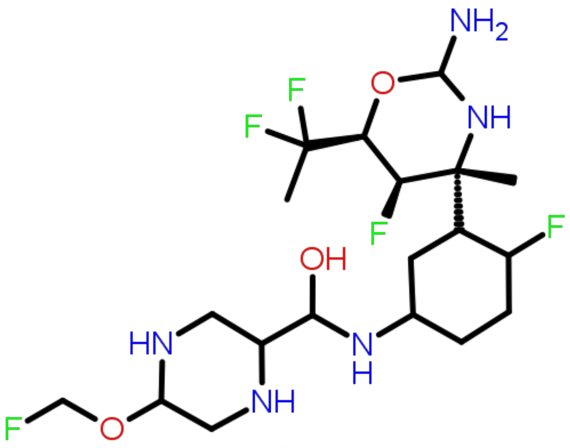
## SMILES String

CC(F)(F)[C@H]1OC(N)=N[C@](C)(c2cc(NC(=O)c3cnc(OCF)cn3)ccc2F)[C@H]1F

$\mathcal{F}_1$

Fingerprint

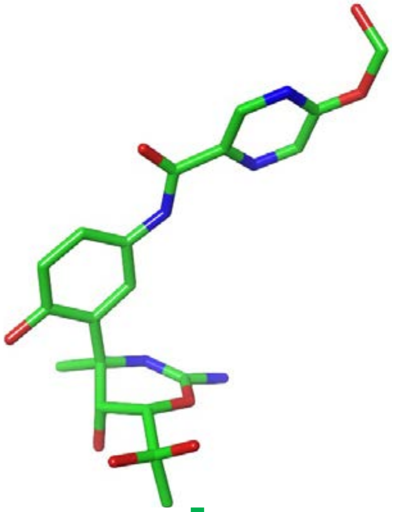
## 2D Diagram



$\mathcal{F}_2$

Fingerprint

## 3D

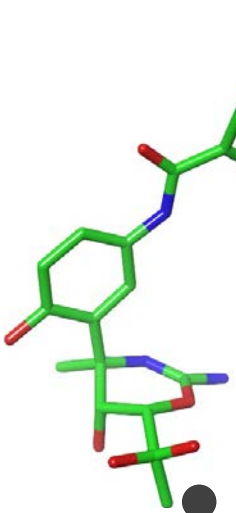


$\mathcal{F}_3$

Fingerprint

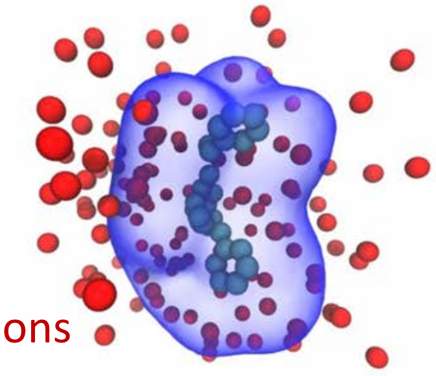
# Molecular Representations

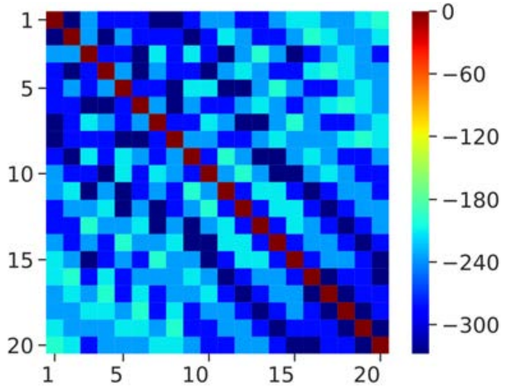
3D



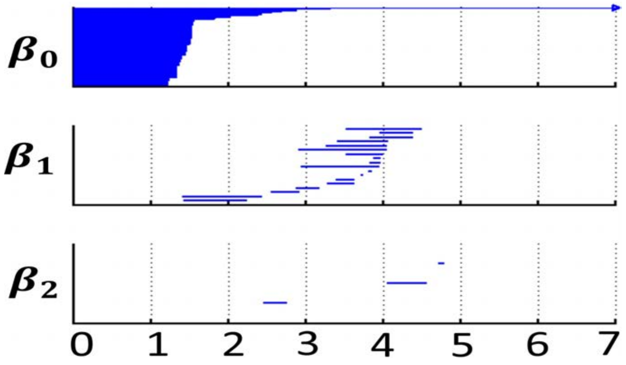
Mathematical Representations

(Nguyen, Cang, Wei, PCCP 2020, …)

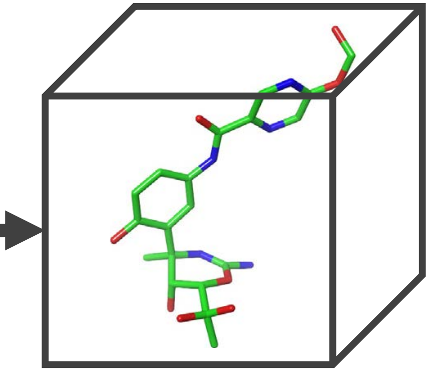




(Meng and Xia, Sci. Adv. 2021, …)



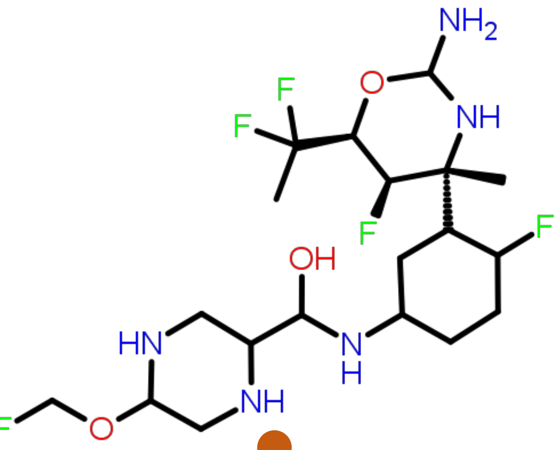
3D-Image Like Representations



Jimenez et. al. (2018),
Hassan-Harrirou et. al. (2020), …

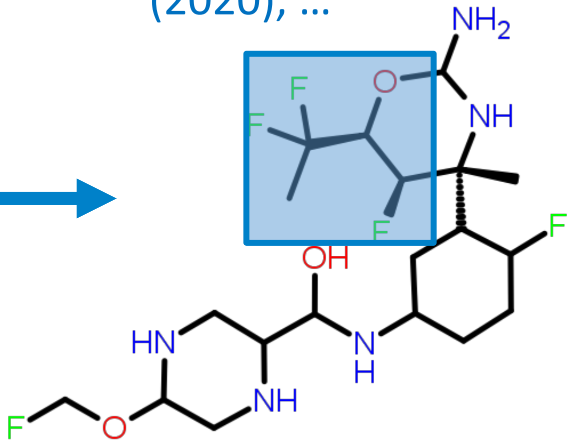
# Molecular Representations

## 2D Diagram

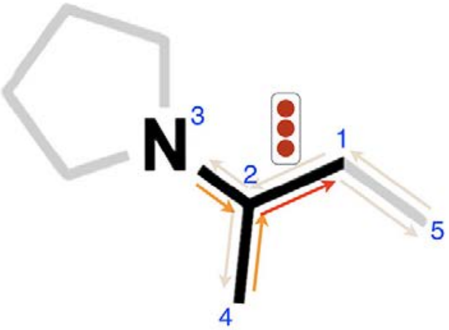


2D Image

Goh et. al. (2018), Rifaioglu et. al. (2020), …



Graph Neural Network



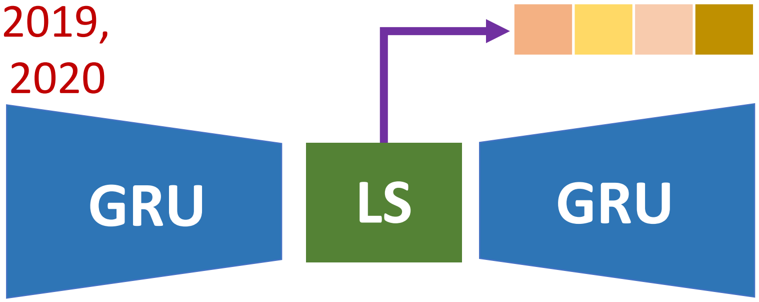
Duvenaud et. al. (2015), Kearnes et. al. (2016), Wu el. al. (2018), Yang el. al. (2019), Withnall (2020), …

# Molecular Representations

SMILES String

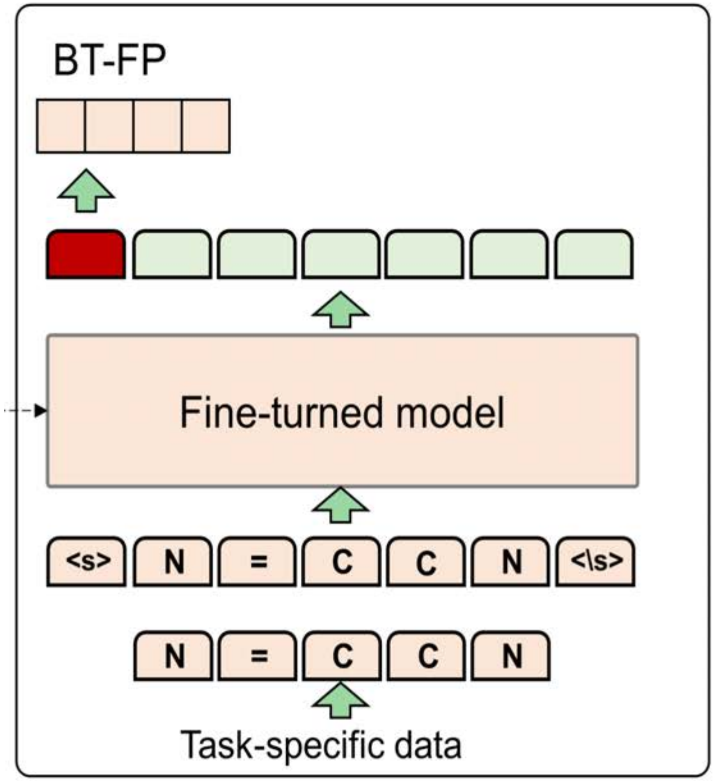
CC(F)(F)[C@H]1OC(N)=
N[C@](C)(c2cc(NC(=O)c
3cnc(OCF)cn3)ccc2F)[C
@H]1F

Grow et. al., CIS 2019,
Gao et. al., JCIM 2020



Deep Learning's
Hidden States

(Dong Chen et. al., Nat. Com. 2021)



Language Embedding
(Word2vec, ...)

2D Fingerprints

FP2, Daylight, MACCS,
Estate, ECPF4,
Pharm2D, ERG ...

(Gao et. al., PCCP 2020)

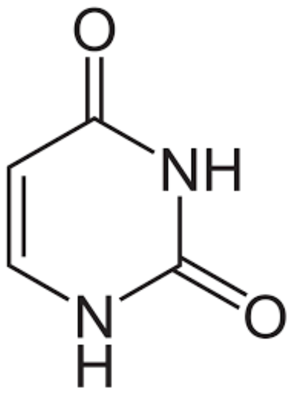
Rogers and Hahn (2010), ...

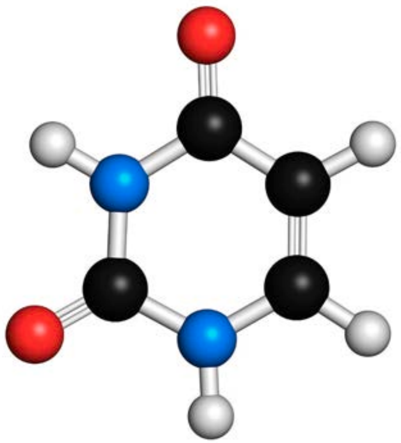
Jaeger et. al. (2018)
Goh et. al. (2018), ...

# Mathematical Graphs for Molecules

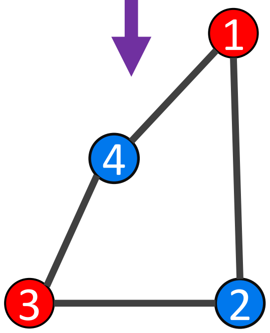
🔴 Oxygen ⚫ Carbon 🔵 Nitrogen ⚪ Hydrogen

Uracil





Subgraphs: $\mathcal{G}_{O,C}, \mathcal{G}_{O,N}, \mathcal{G}_{O,H}, \mathcal{G}_{C,N}, \mathcal{G}_{C,H}$



$\mathcal{G}_{O,N}$

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Adjacency Matrix

$$A = \begin{bmatrix} 0 & w_{12} & 0 & w_{14} \\ w_{21} & 0 & w_{23} & 0 \\ 0 & w_{32} & 0 & w_{34} \\ w_{41} & 0 & w_{42} & 0 \end{bmatrix}$$

Weighted Adjacency Matrix

- $w_{ij}(d) = e^{-\left(\frac{d}{\eta}\right)^k}$
- $w_{ij}(d) = \dfrac{1}{1 + \left(\frac{d}{\eta}\right)^v}$

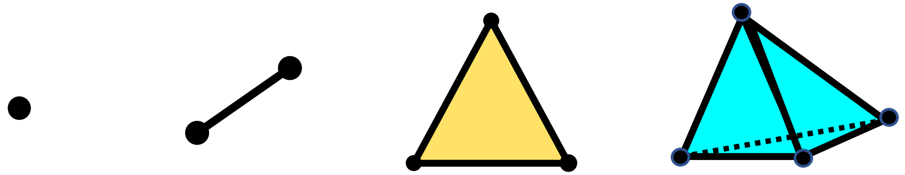
$$L = D - A = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

Laplacian Matrix

(Nguyen, Wei, JCIM 2019)

# Persistent Spectral Graph: Graph + Topology

- Simplexes:



***0*-simplex** $\sigma_0$ ***1*-simplex** $\sigma_1$ ***2*-simplex** $\sigma_2$ ***3*-simplex** $\sigma_3$

- $q$-chain: $\sum_{j} w_j \sigma_q^j,\quad w_j \in \mathbb{Z}_2, \sigma_q^j \in K$

Wang, Nguyen, Wei (IJNMBE, 2020)

Meng and Xia, (Sci. Adv. 2021)

- $q$-chain Group: $C_q(K) = \left(\left\{\sum_{j} w_j \sigma_q^j\right\}, +\right)$
- Boundary operator:

$$\partial_q : C_q(K) \to C_{q-1}(K) \quad , \partial_q\, \sigma_q = \sum_{j=0}^{q} (-1)^j \langle v_0, v_1, \ldots, \widehat{v_j}, \ldots, v_q \rangle$$

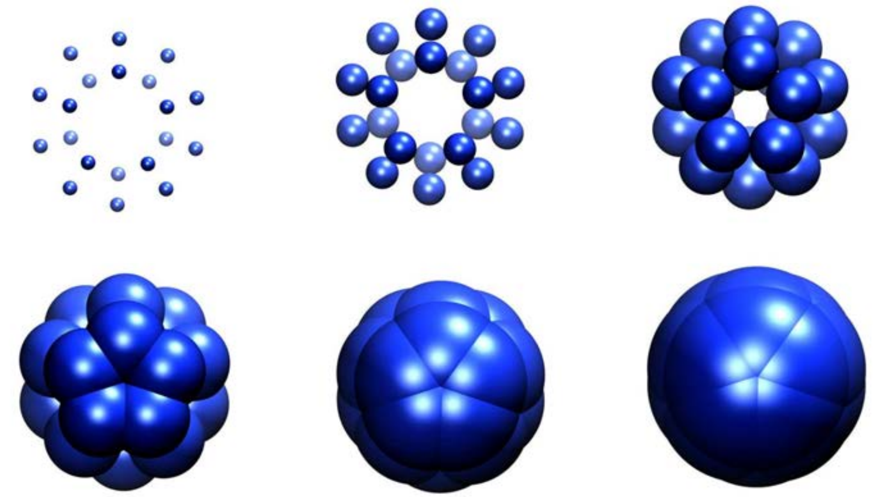
- Adjoint boundary operator: $\partial_q^* : C_{q-1}(K) \to C_q(K)$
- $q$-combinatorial Laplacian operator: $\Delta_q = \partial_{q+1}\partial_{q+1}^* + \partial_q^*\partial_q$
- $q$-combinatorial Laplacian matrix: $\mathcal{L}_q = \mathcal{B}_{q+1}\mathcal{B}_{q+1}^T + \mathcal{B}_q^T\mathcal{B}_q$
- Betti numbers: $\beta_q$ =# of zeros eigenvalues of $\mathcal{L}_q(K)$

# Applications to Biology

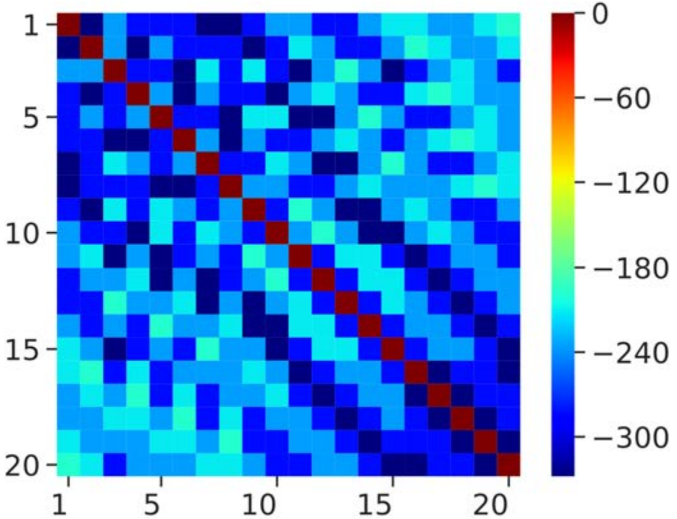
(Nguyen, Xia, Wei, JCP 2016)

Wang, Nguyen, Wei (IJNMBE, 2020)

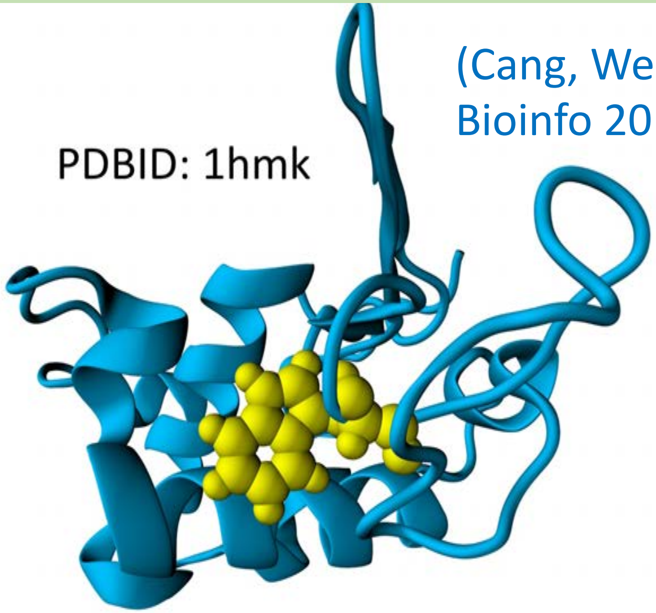
## Filtration on fullerene $C_{20}$



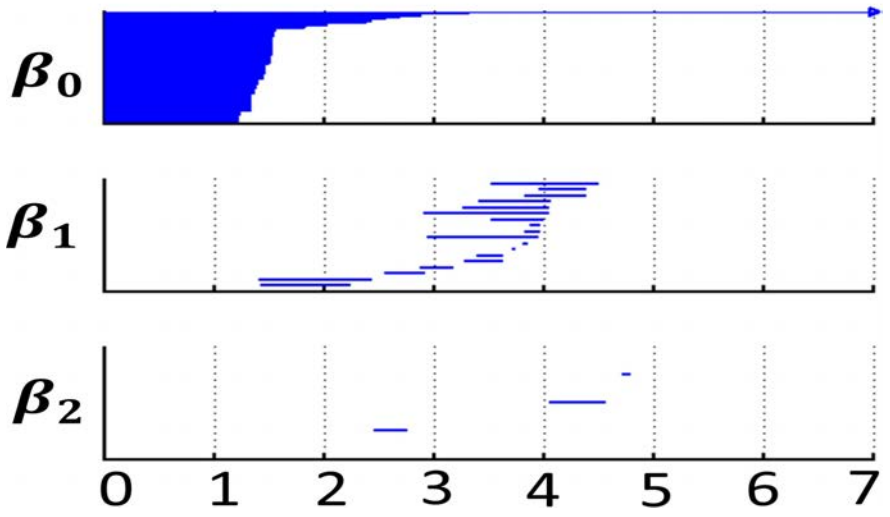
## Accumulated combinatorial Laplacian matrix for $C_{20}$



## Barcodes of protein-ligand



(Cang, Wei, Bioinfo 2017)

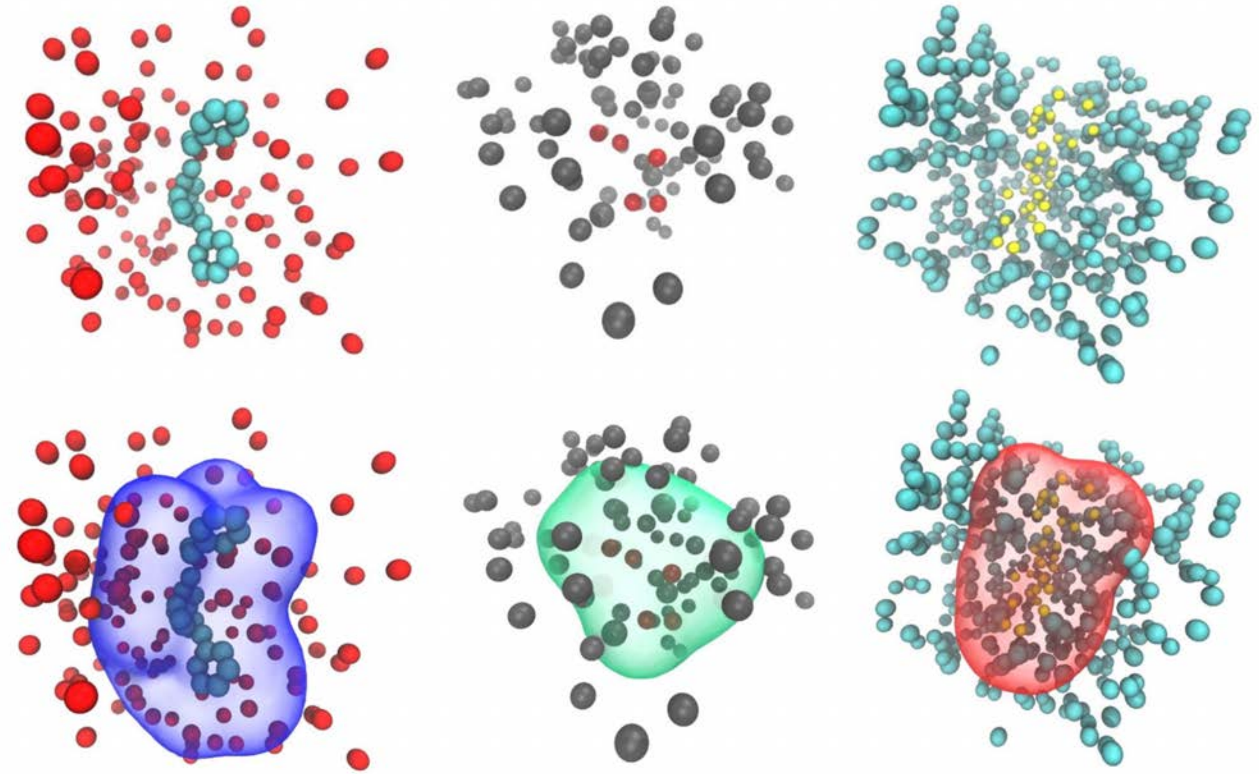


# Differential Geometry

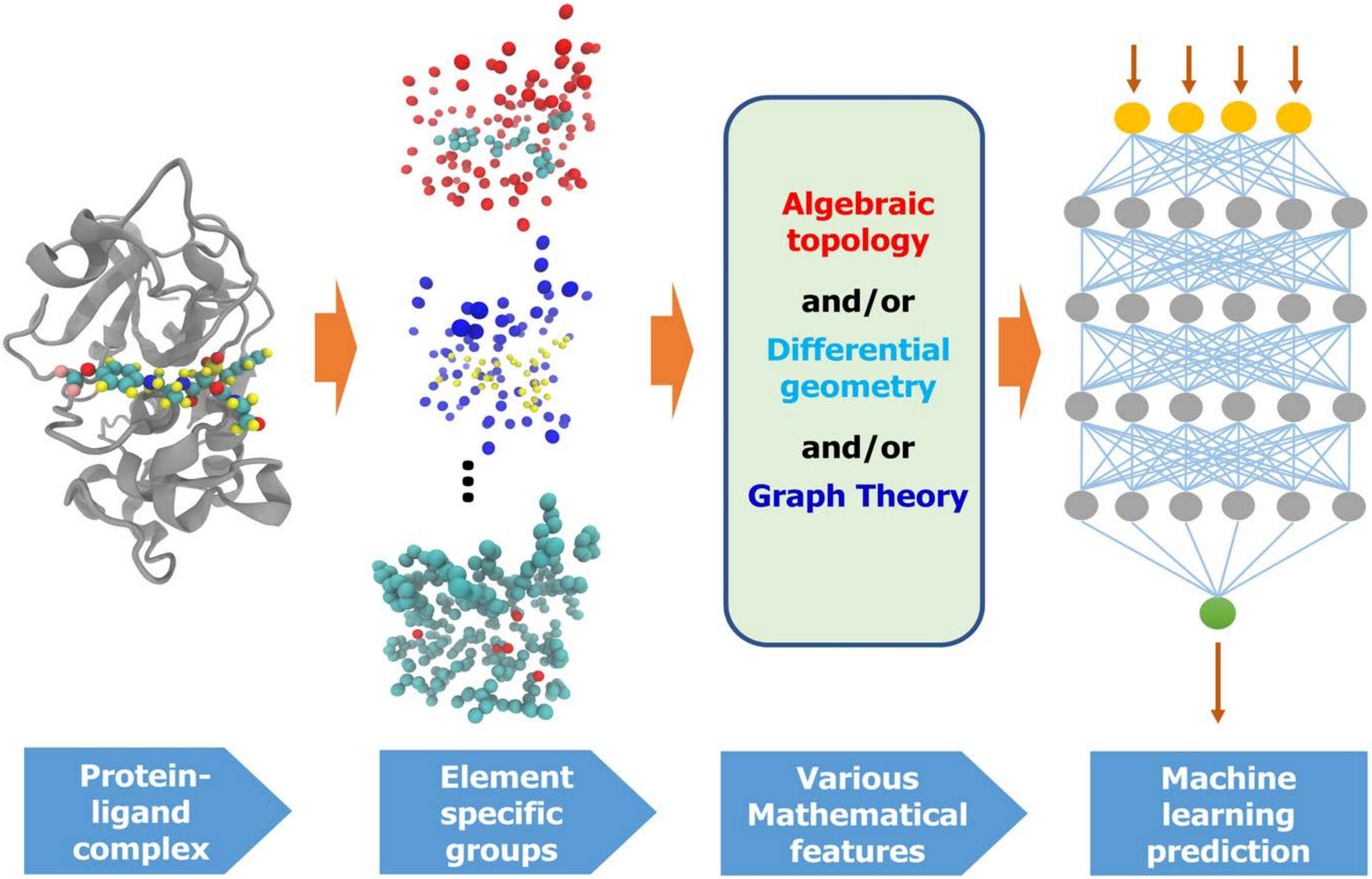
(Nguyen, Wei, IJNMBE 2018) (Rana, Nguyen, 2021)

- Element interactive density: $\rho_{kk'}(\mathbf{r}, \eta_{kk'}) = \sum_j w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{kk'})$, with $\|\mathbf{r} - \mathbf{r}_j\| > r_i + r_j + \sigma$
- Element interactive manifolds (EIMs): $\rho_{kk'}(r, \eta_{kk'}) = c\rho_{\max}, 0 \leq c \leq 1$
- Element interactive curvatures, element interactive areas, …

Element Interactive Manifolds



# Mathematics based Deep Learning Models (MathDL)



Protein-ligand complex

Element specific groups

Algebraic topology

and/or

Differential geometry

and/or

Graph Theory

Various Mathematical features

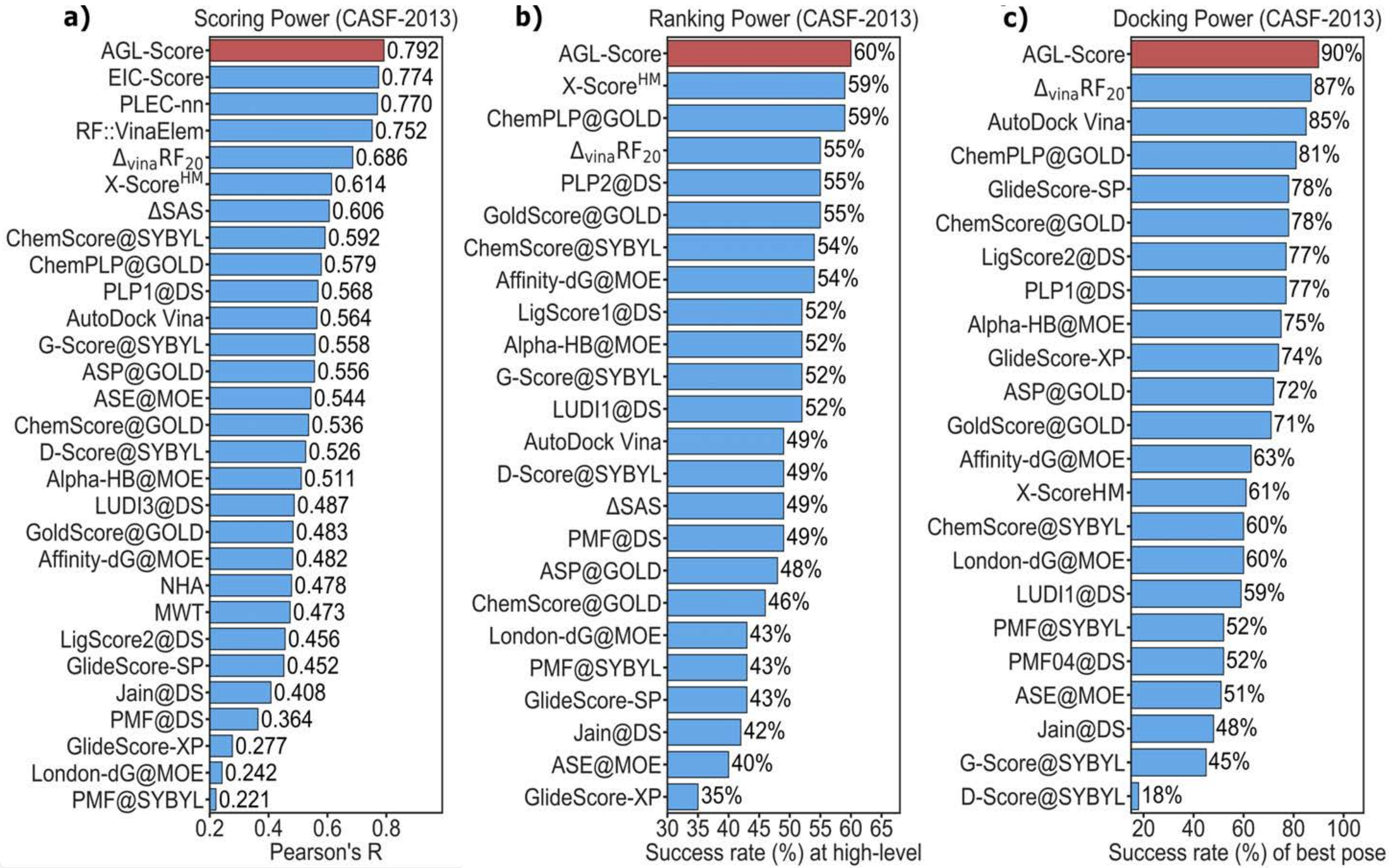
Machine learning prediction

(Nguyen et. al., JCAMD 2019) (Patent No.: US 2019 / 0304568 A1)

# Performance of MathDL in Virtual Screening, Docking, Affinity Ranking

**Red Color** is our Model

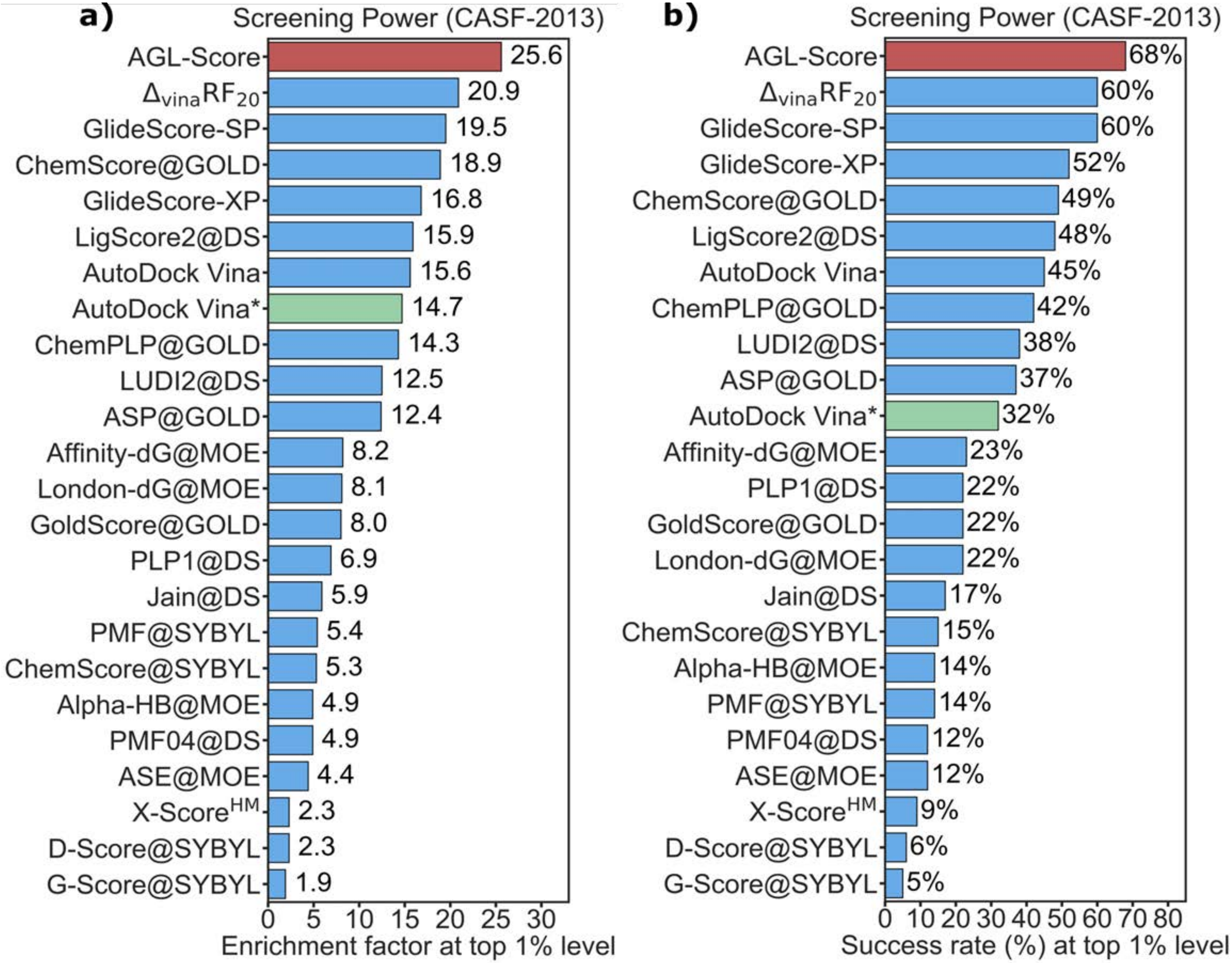
(Nguyen, Wei, JCIM 2019)



# Performance of MathDL in Virtual Screening, Docking, Affinity Ranking

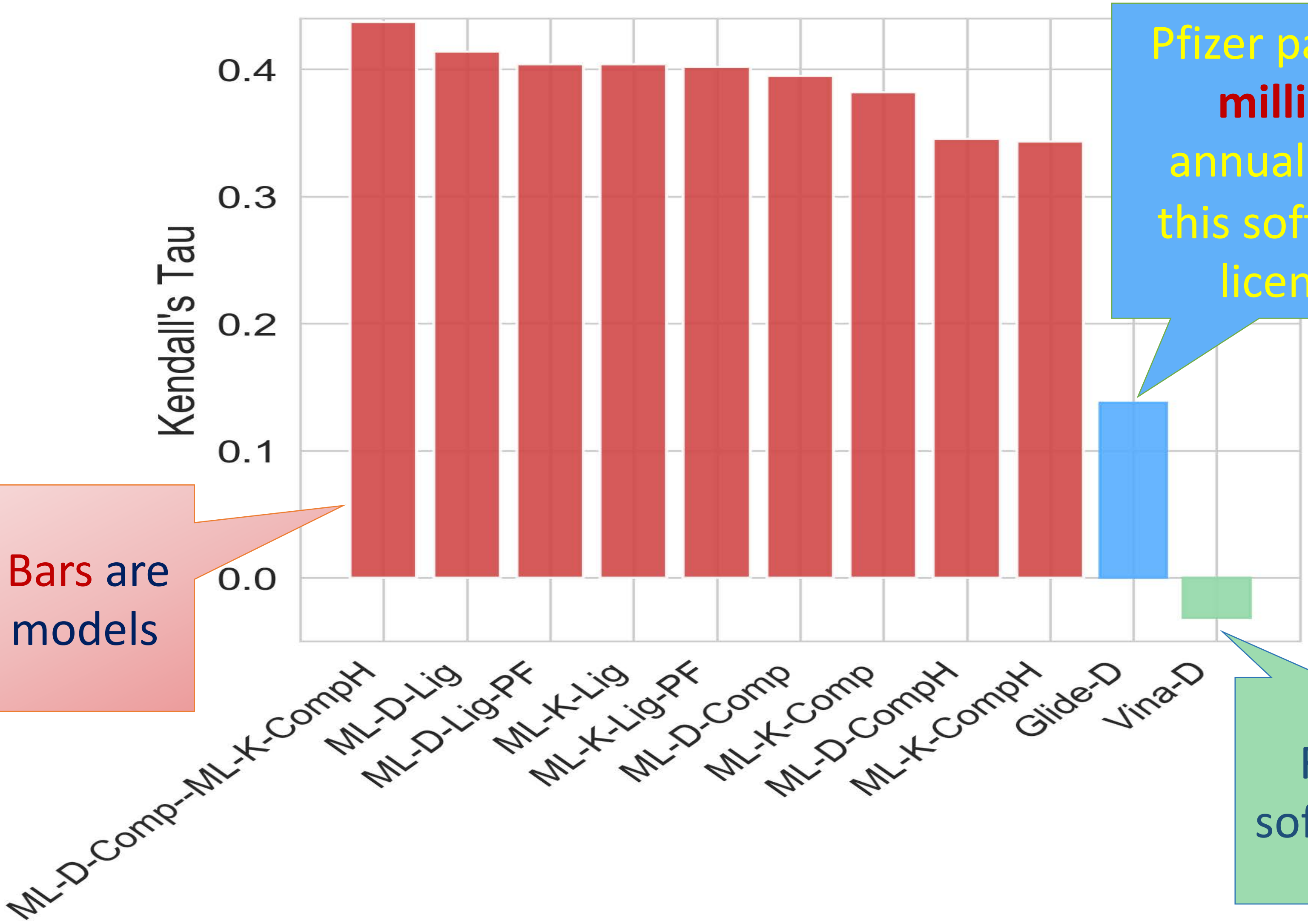
**Red Color** is our Model

(Nguyen, Wei, JCIM 2019)



# Collaboration work with Pfizer

## Binding affinity ranking of 362 compounds (fully blind)



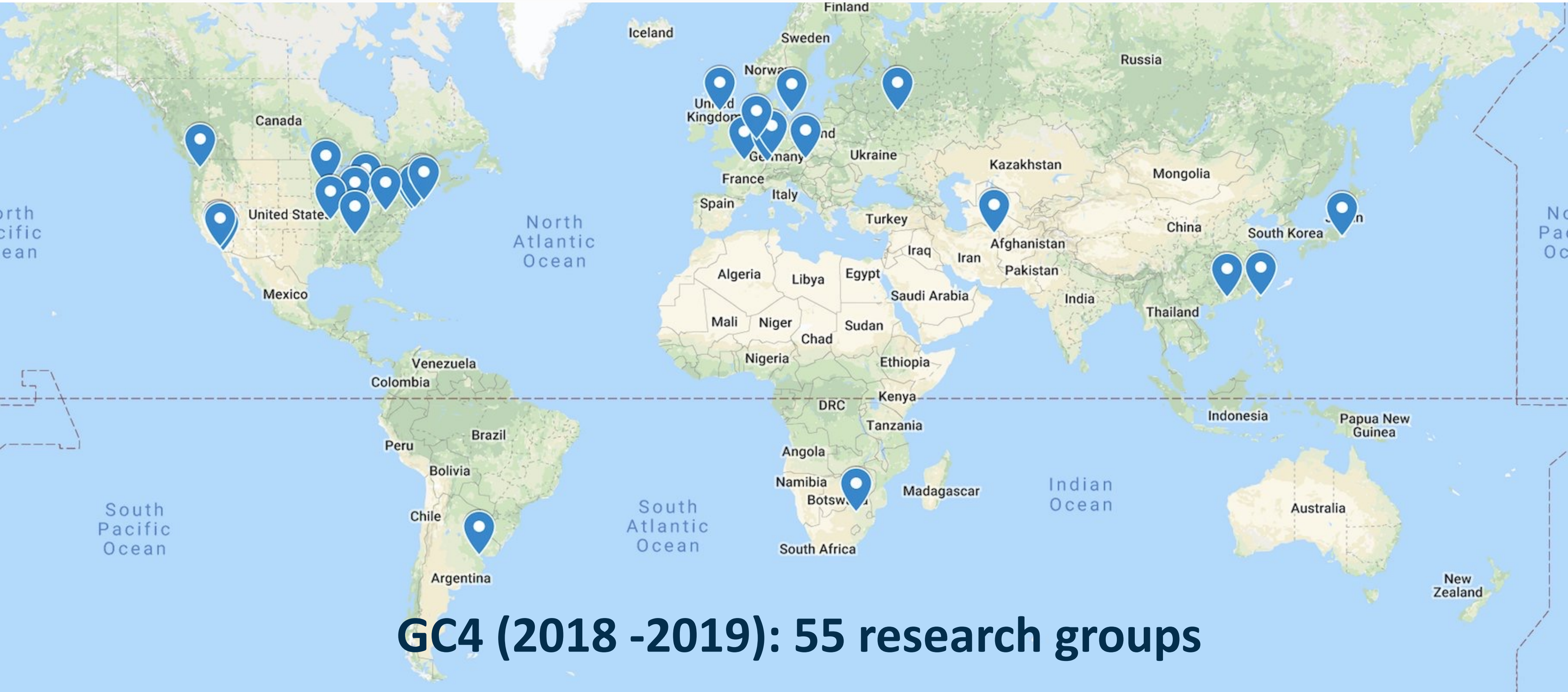
Pfizer pays **$1 million** annually for this software license

Red Bars are our models

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**

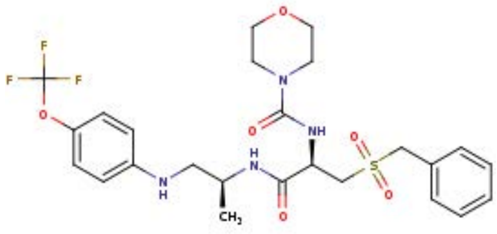
# Drug Design Data Resource (D3R) Grand Challenges

Given Data

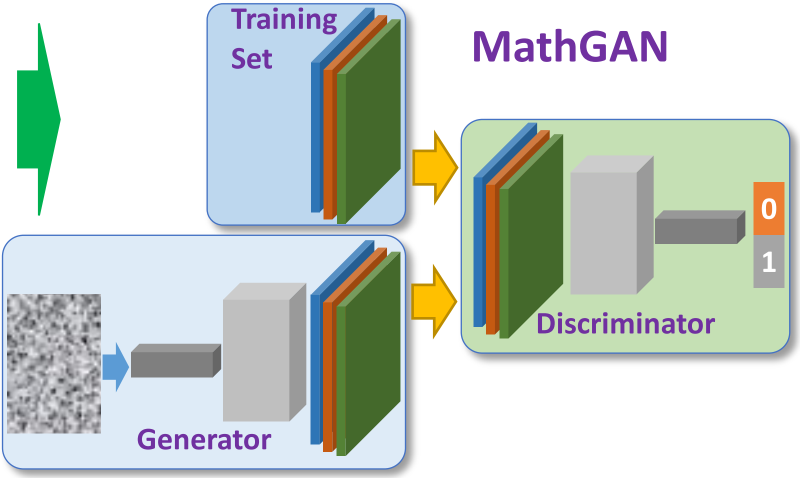
Protein Sequence

WAFSAVGALEALPDSV···

2D Ligands

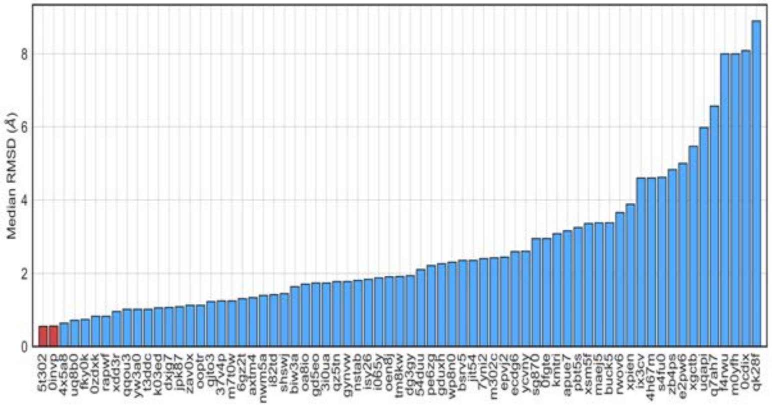


Pose Prediction

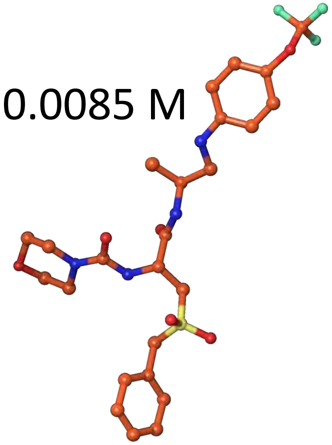


Predicted Complexes

Predictions and Evaluations



IC50=0.0085 M



MathDL

Predicted Affinity

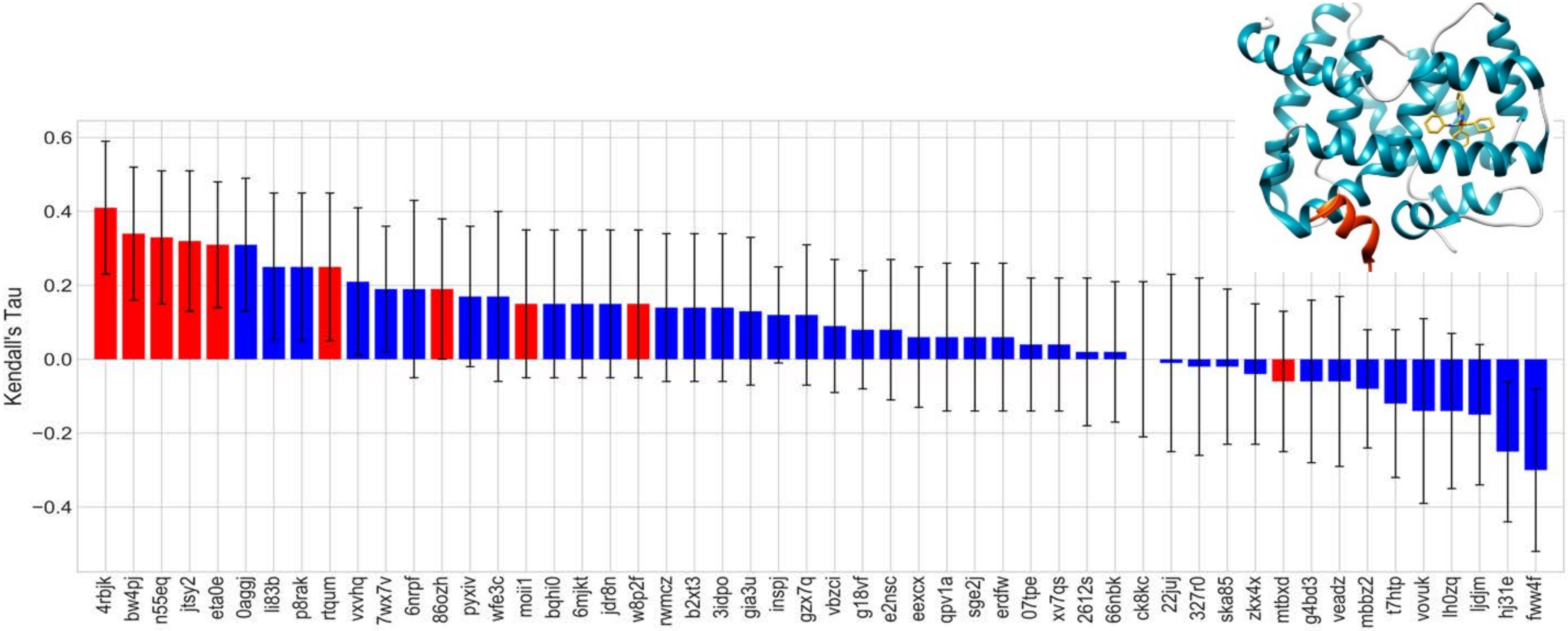
# D3R Grand Challenge 2 (2016-2017)

**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)



(Nguyen et. al., JCAMD 2018)

# 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 SC3**

Scoring

Free Energy Set 🥇🥈🥉

**JAK2 SC2**

Scoring (partials)

**TIE2**

Scoring 🥇🥈

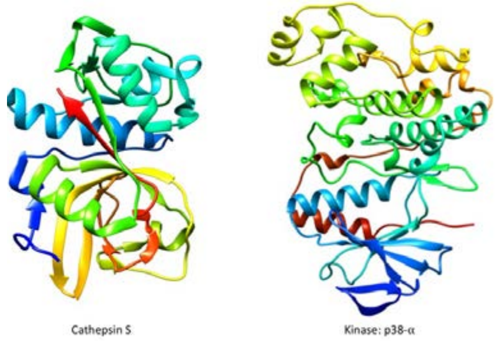
Free Energy Set 2 🥇🥈🥉

**p38-α**

Scoring

**ABL1**

Scoring (partials) 🥇🥈🥉



### Active / Inactive Classification

**VEGFR2**

Scoring (partials)

**JAK2 SC3**

Scoring

Free Energy Set 🥇🥈🥉

**JAK2 SC2**

Scoring (partials)

**TIE2**

Scoring (partials) 🥇🥈🥉

Free Energy Set 1 🥇🥈🥉

**p38-α**

Scoring (partials)

**ABL1**

Scoring (partials)

### Affinity Rankings for Cocrystalized Ligands

**Cathepsin Stage 1**

Scoring (partials)

Free Energy Set 🥇🥈

**Cathepsin Stage 2** 🥇🥈

Scoring (partials)

Free Energy Set 🥇🥈🥉

(Nguyen et. al., JCAMD 2018)

# D3R Grand Challenge 4 (2018-2019)

## Pose Predictions

### BACE Stage 1A

Pose Predictions (Partials) 🥇 🥈

### BACE Stage 1B

Pose Prediction (Partials) 🥈 🥉

## Affinity Predictions

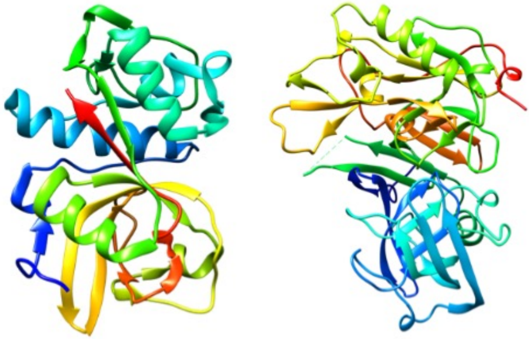
### Cathepsin Stage 1

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

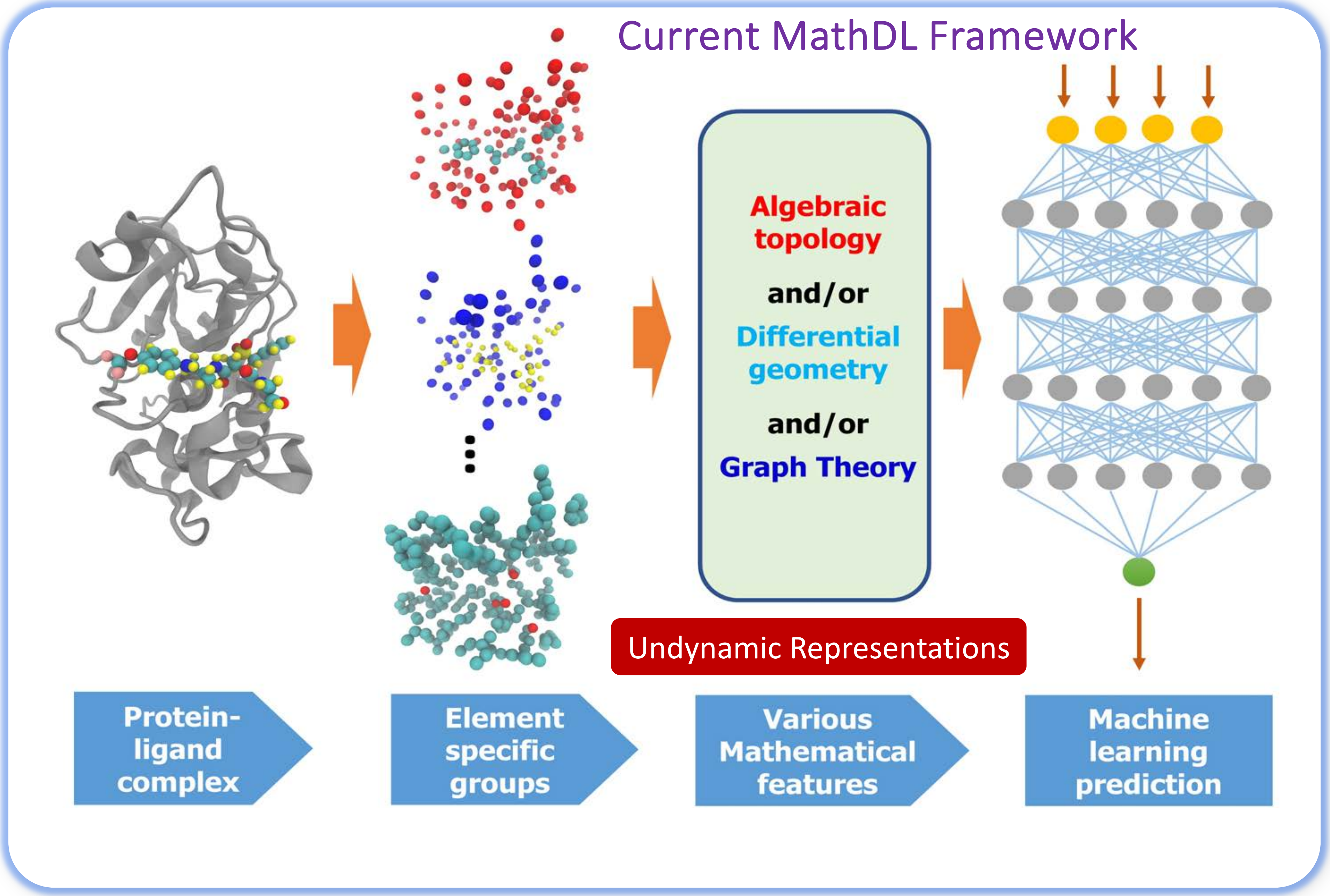
Ligand Based Scoring (No participation)

Structure Based Scoring (Partials)

Free Energy Set 🥈 🥉

(Nguyen et. al., JCAMD 2019)

# Moving Beyond MathDL



Current MathDL Framework

Algebraic topology

and/or

Differential geometry

and/or

Graph Theory

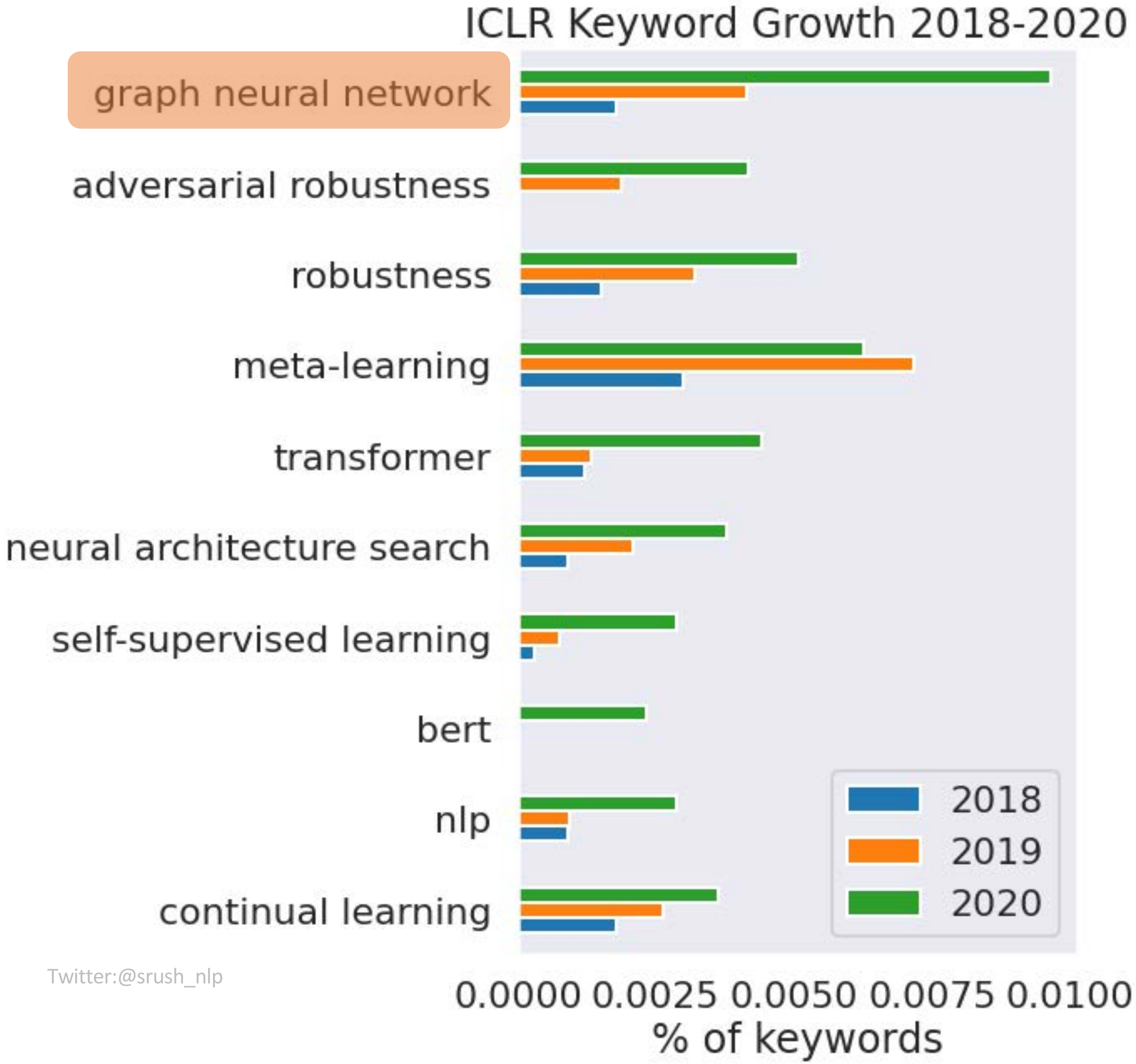
Undynamic Representations

Protein-ligand complex

Element specific groups

Various Mathematical features

Machine learning prediction



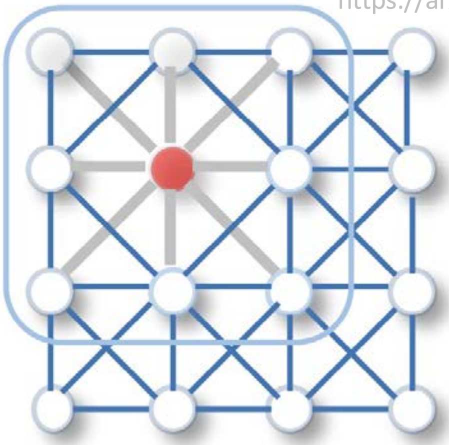
Twitter:@srush_nlp

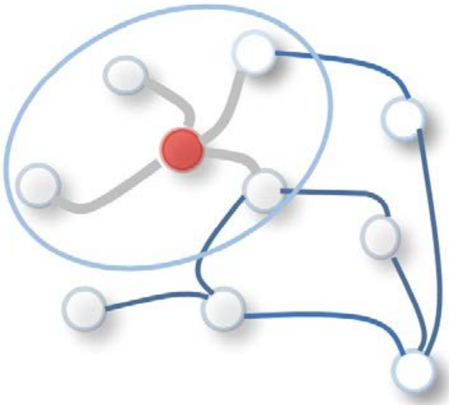
# Graph Neural Network (GNN)

Convolutional Layer

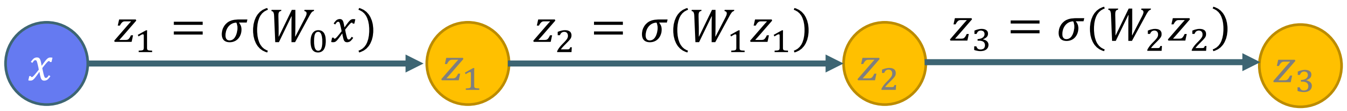
Graph Convolutional Layer

https://arxiv.org/pdf/1901.00596.pdf





Deep Neural Network

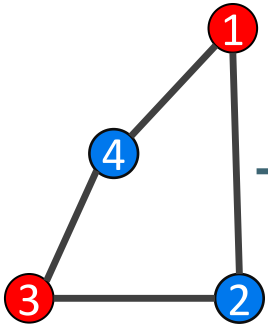


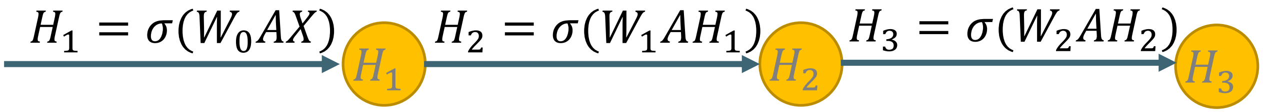
$$z_1 = \sigma(W_0 x) \qquad z_2 = \sigma(W_1 z_1) \qquad z_3 = \sigma(W_2 z_2)$$

Node

3 hidden layers

Graph Neural Network





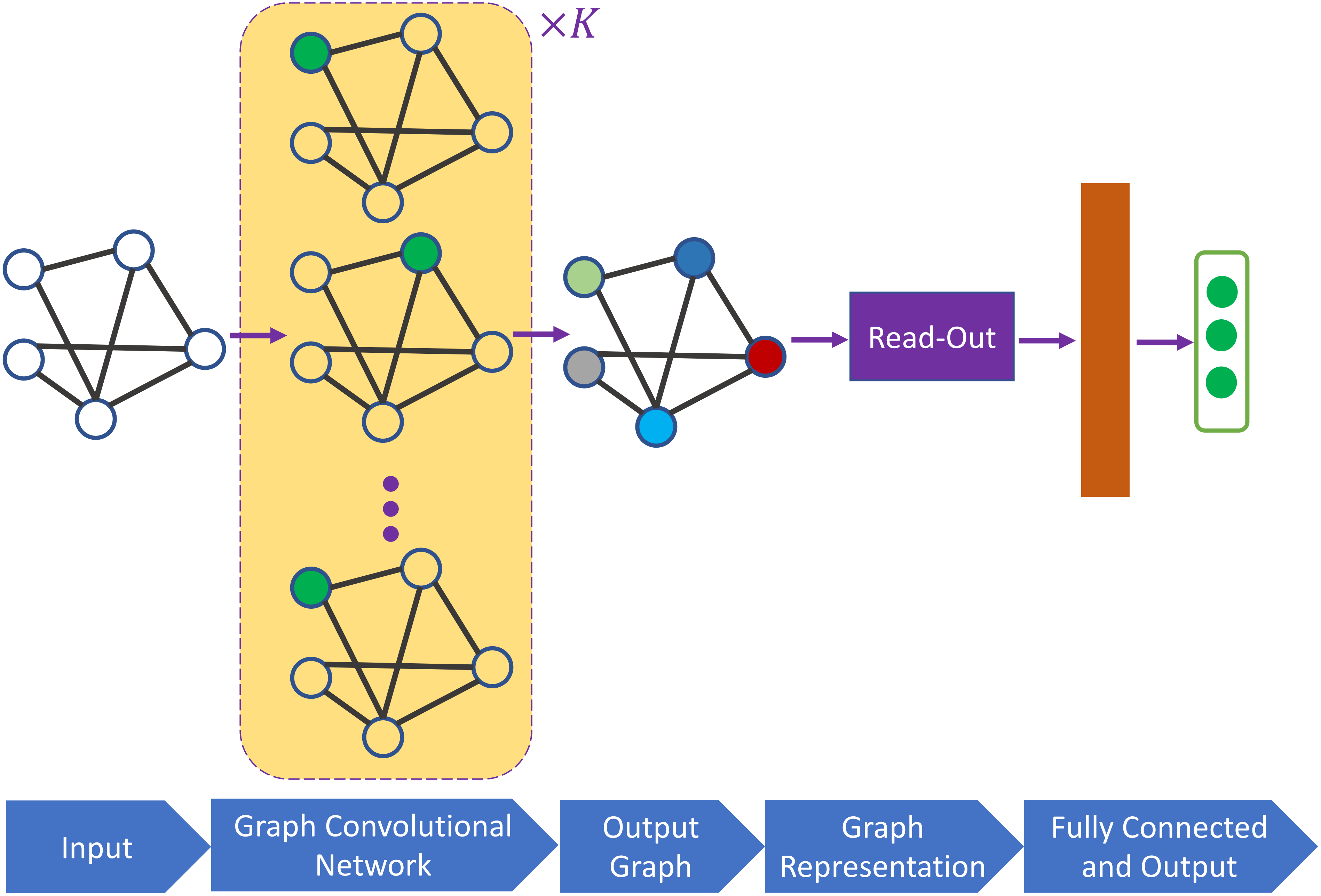
$$H_1 = \sigma(W_0 A X) \qquad H_2 = \sigma(W_1 A H_1) \qquad H_3 = \sigma(W_2 A H_2)$$

3 hidden layers

Graph

- $A$: adjacency matrix

# Standard Architecture of GNN



Input

Graph Convolutional Network

Output Graph

Graph Representation

Fully Connected and Output

# Standard Architecture of GNN

- **Aggregation:**

$$a_u^{(k)} = \mathrm{AGGREGATE}^{(k)}\left(\left\{\underbrace{h_v^{(k-1)}}_{\text{Node State}}, v \in \underbrace{\mathcal{N}(u)}_{\text{Neighbor of node } u}\right\}\right)$$

Example

$$a_u^{(k)} = \frac{1}{|\mathcal{N}(u)|} \sum_{v \in \mathcal{N}(u)} h_v^{(k-1)}$$

- **Combination/Updating:**

$$h_u^{(k)} = \mathrm{COMBINE}^{(k)}\left(h_u^{(k-1)}, a_u^{(k)}\right)$$

Example

$$h_u^{(k)} = W_0^{(k)} h_u^{(k-1)} + W_1^{(k)} a_u^{(k)}$$

- **Read-out (Graph Invariant):**

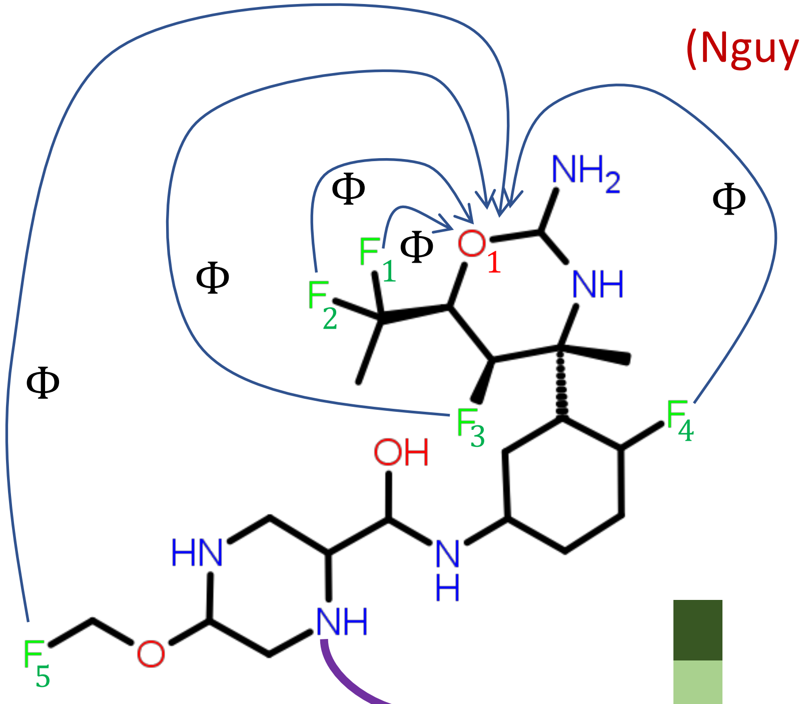
$$h_G = \mathrm{READOUT}\left(\left\{h_u^{(K)}, u \in G\right\}\right)$$

Example

$$h_G = \mathrm{MEAN}\left(\left\{h_u^{(K)}, u \in G\right\}\right)$$

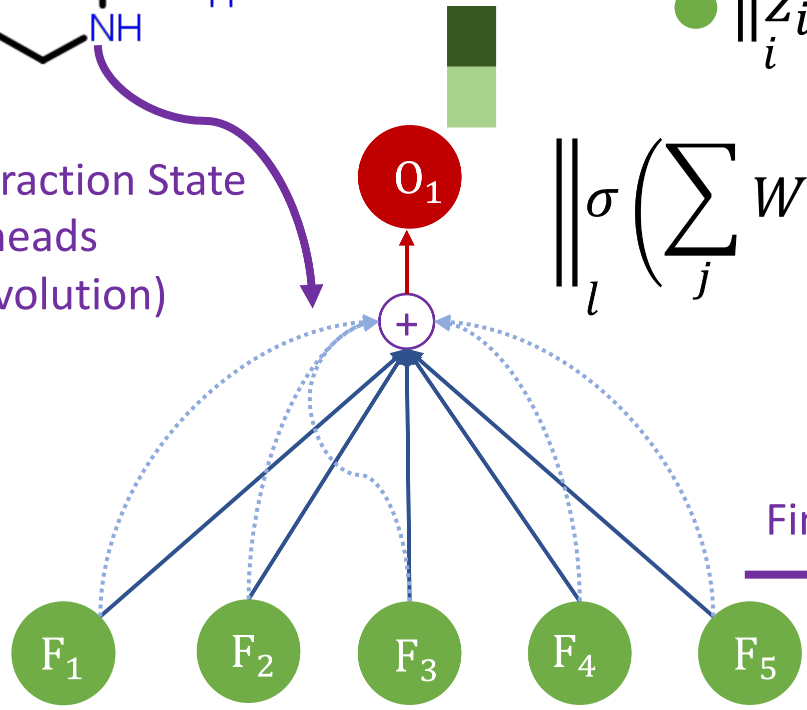
# Math-Based Deep Graph ($\text{MathDL}_{\text{GNN}}$)

(Nguyen et. al., 2021)



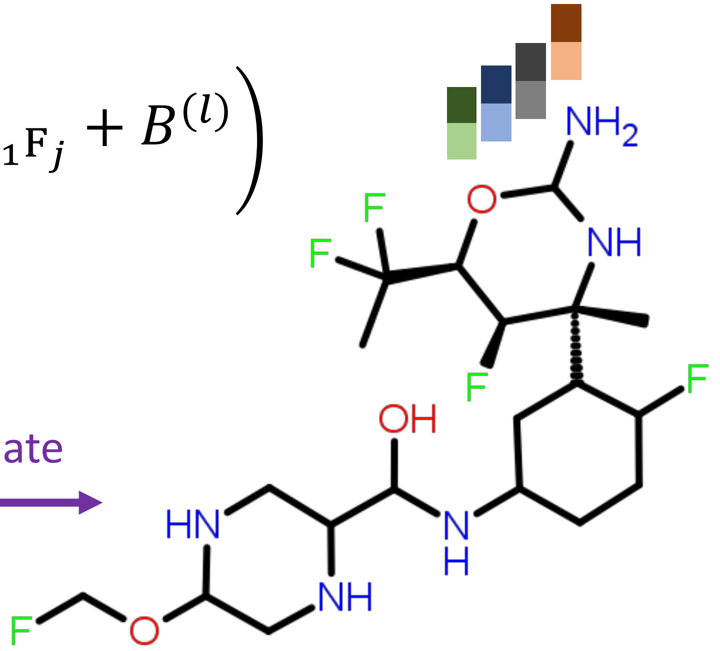
- Update Oxygen 1 ($O_1$) by its neighbors $F_i$
- $\Phi_{uv}$: Node interaction (radial basis function)
- $\Phi_{uv} = e^{-\left(\frac{\text{dist}(h_u,h_v)}{\eta}\right)^{\kappa}}$
- $\Vert_i z_i$: concatenate all vector $z_i$

Hidden Interaction State with multi heads (Graph Convolution)



$$\Big\Vert_{l} \sigma\left(\sum_{j} W^{(l)}\Phi_{O_1F_j} + B^{(l)}\right)$$

Final update

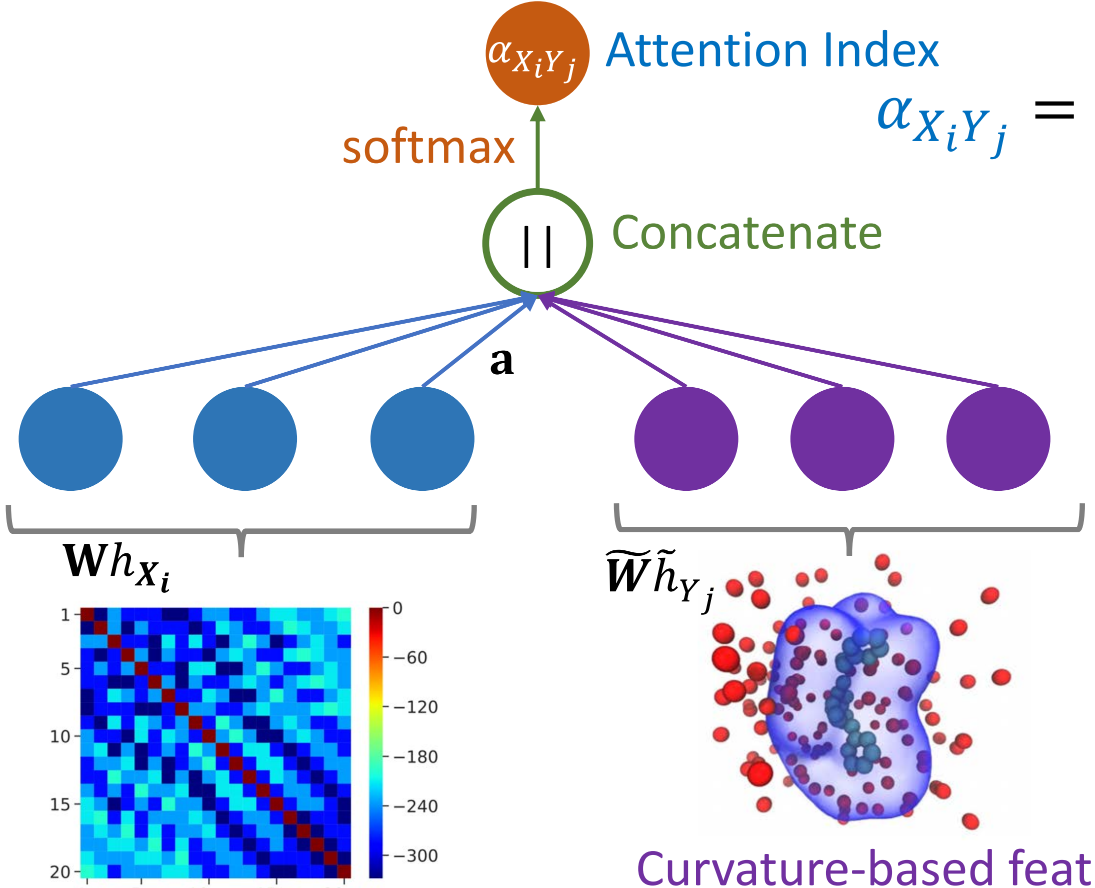


# Math-Based Deep Graph (MathDL$_{\mathbf{GNN}}$)

(Nguyen et. al., 2021)

- **Attention mechanism** with another Math Features

Vaswani et. al. (2017)



$$\alpha_{X_iY_j} = \frac{\exp\left(\sigma(\mathbf{a}\left[\mathbf{W}h_{\boldsymbol{X_i}}||\widetilde{\boldsymbol{W}}\tilde{h}_{Y_j}\right]\right)}{\sum_k \exp\left(\sigma(\mathbf{a}[\mathbf{W}h_{\boldsymbol{X_i}}||\widetilde{\boldsymbol{W}}\tilde{h}_{Y_k}]\right)}$$

Graph-based features

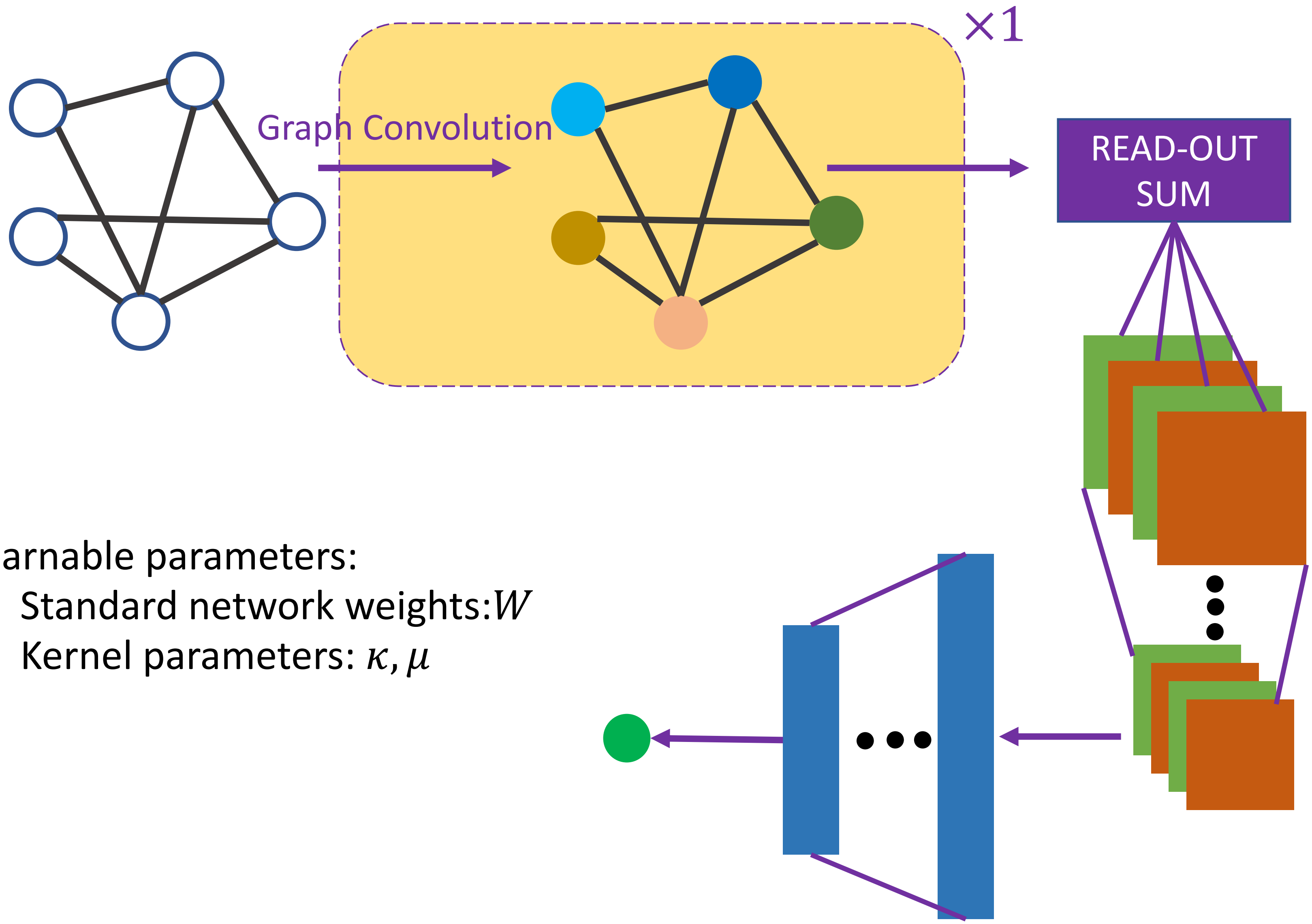
Curvature-based features
(volume, area, electrostatics,
other physical interactions …)

- New hidden state update

$$h_{X_i}^{(new)} = \Big\Vert \sigma\left(\sum_j \alpha_{X_iY_j} W^{(k)}\Phi_{X_iY_j} + B^{(k)}\right)$$

# AweGNN

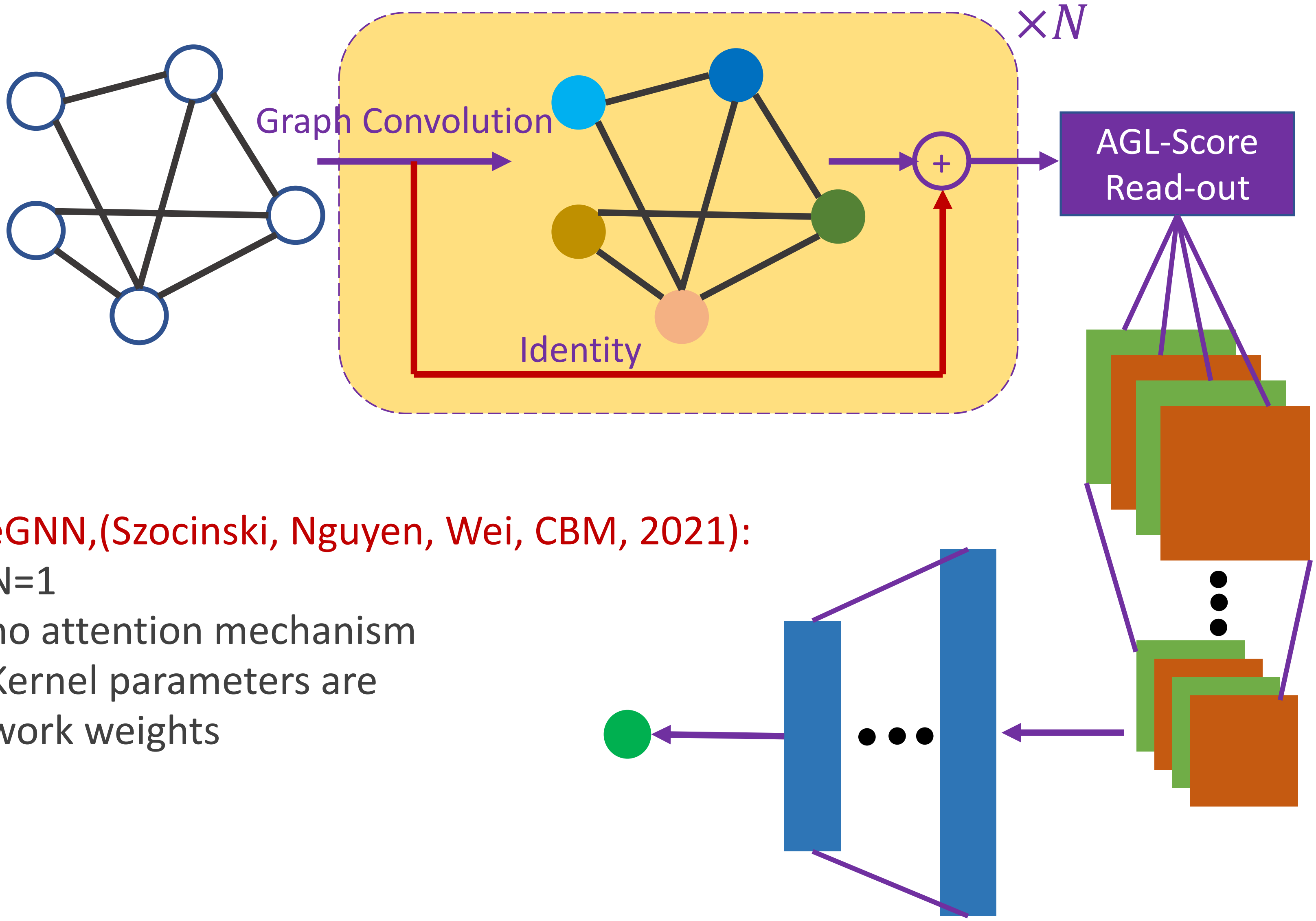
(Szocinski, Nguyen, Wei, CBM, 2021)



- Learnable parameters:
  - Standard network weights: $W$
  - Kernel parameters: $\kappa, \mu$

# Math-Based Deep Graph ($\mathrm{MathDL}_{\mathrm{GNN}}$)

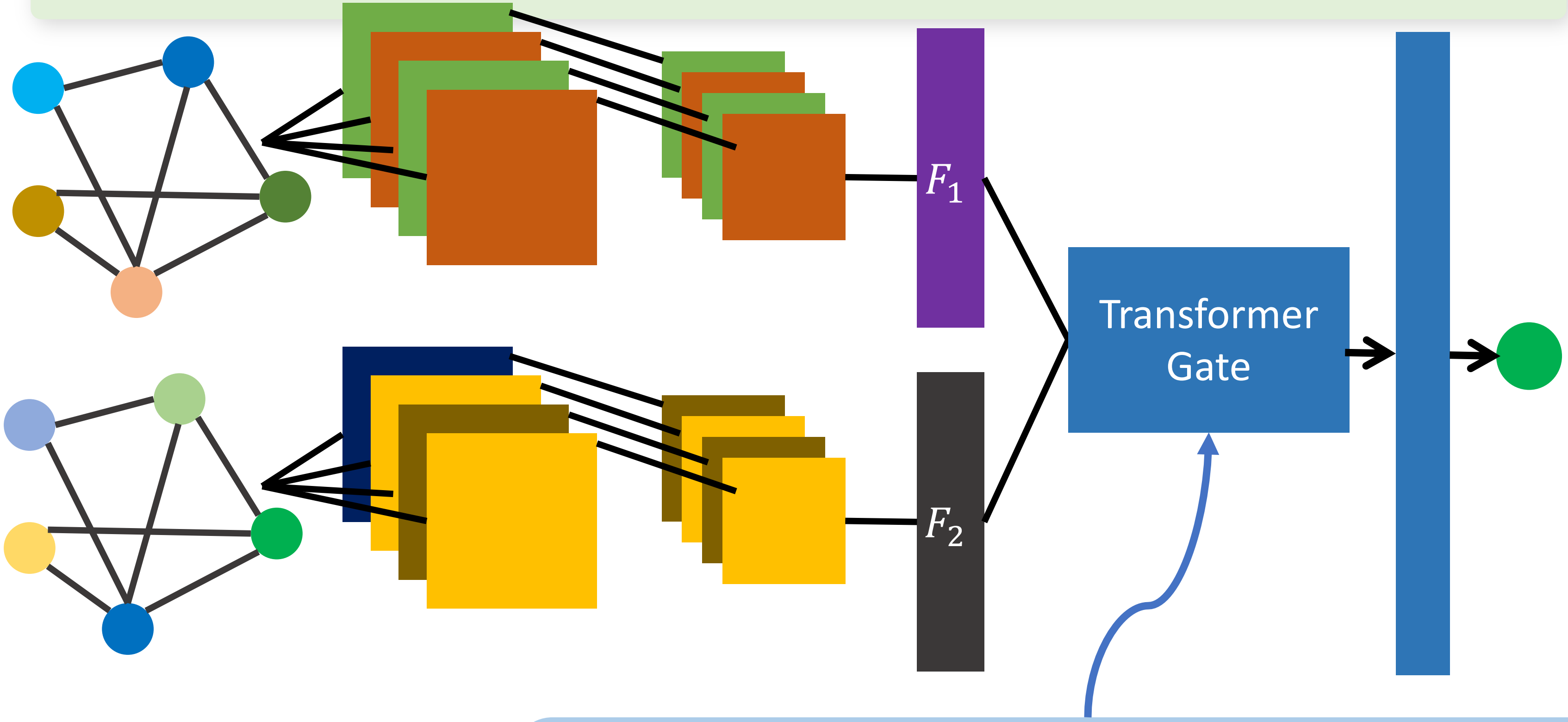
(Nguyen et. al., 2021)



- AweGNN,(Szocinski, Nguyen, Wei, CBM, 2021):
  - N=1
  - no attention mechanism
  - Kernel parameters are network weights

# Math-Based Deep Graph (MathDL$_{\text{GNN}}$)

- **Transformer gate:** combining different Math-based Graphs



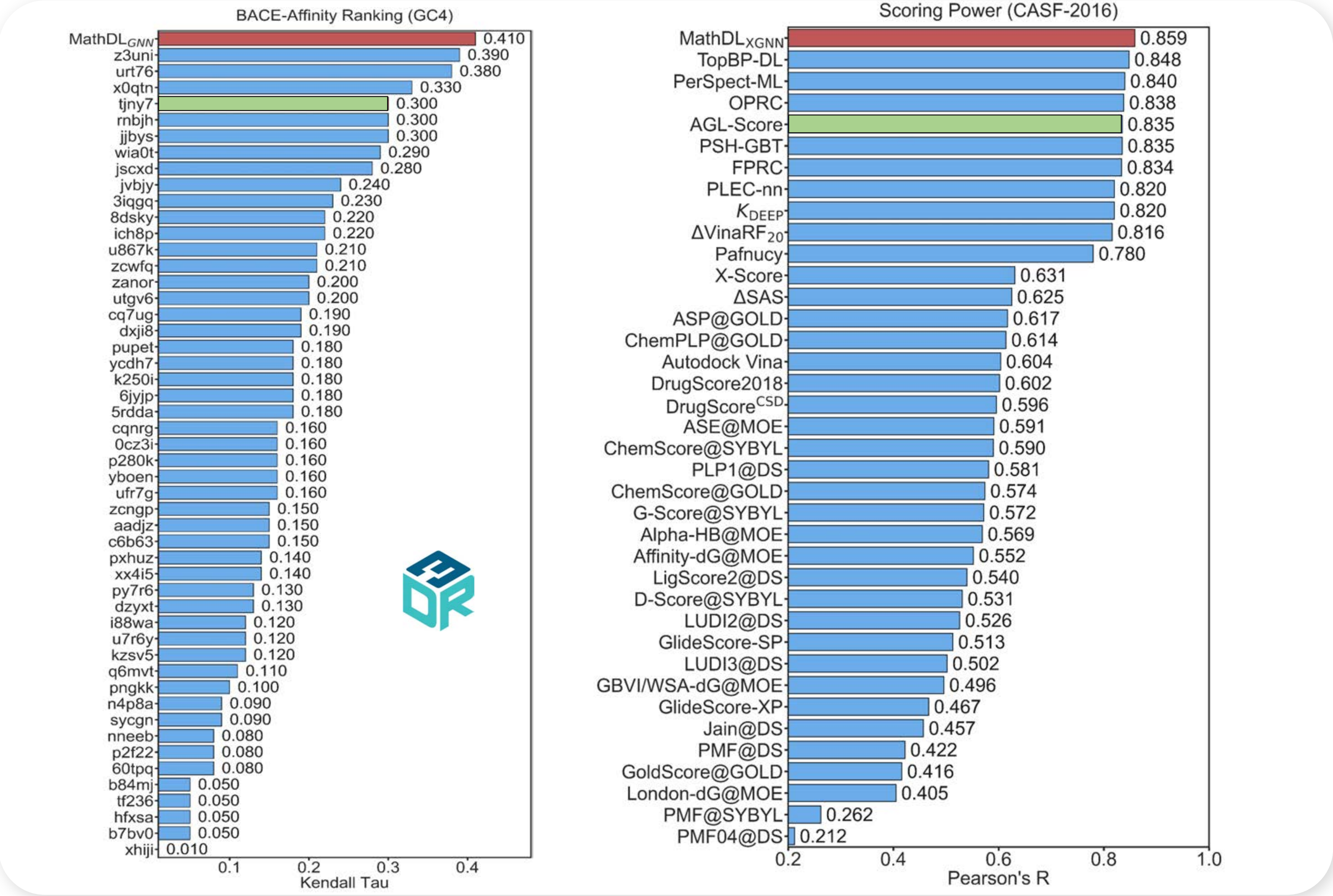
$$z = \sigma\left(\widetilde{\boldsymbol{W}}_1 F_1 + \widetilde{\boldsymbol{W}}_2 F_2 + b\right)$$

$$F_{\text{combined}} = z \odot \boldsymbol{W}_1 F_1 + (1 - z) \odot \mathbf{W}_2 F_2$$

(Nguyen et. al., 2021)

# Performance of MathDL in Scoring Power

(Nguyen et. al., 2021)



Thank you!