nature chemistry

# Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning

David F. Nippa, Kenneth Atz, Remo Hohler, Alex T. Müller, Andreas Marx, Christian Bartelmus, Georg Wuitschik, Irene Marzuoli, Vera Jost, Jens Wolfard, Martin Binder, Antonia F. Stepan, David B. Konrad*, Uwe Grether*, Rainer E. Martin*& Gisbert Schneider*

Nature Chemistry
https://doi.org/10.1038/s41557-023-01360-5

**Bin Yang**

**2024.01.13**

nature chemistry

**Prof. Gisbert Schneider**

**Research interest:**

**Integration of artificial intelligence into pharmaceutical research and chemical biology**

- **Full professor, Department of Chemistry and Applied Bioscience, ETH Zurich;**
- **Director, Singapore-ETH Centre (SEC)**

- **Graduate from the Freie Universität Berlin, Germany;**
- **Pharmaceuticals Division at Roche, Switzerland;**
- **Goethe-University in Frankfurt, Germany (Beilstein Endowed Chair for Chem and Bioinformatics);**
- **Current position at the ETH**

https://sec.ethz.ch/people/management/gisbert-schneider.html

nature chemistry

David B. Konrad

ORGANIC CHEMISTRY

# A concise synthesis of tetrodotoxin

David B. Konrad[1]‡, Klaus-Peter Rühmann[2]‡, Hiroyasu Ando[3], Belinda E. Hetzler[2], Nina Strassner[4], Kendall N. Houk[4], Bryan S. Matsuura[2]*‡, Dirk Trauner[2]*§

Tetrodotoxin (TTX) is a neurotoxic natural product that is an indispensable probe in neuroscience, a biosynthetic and ecological enigma, and a celebrated target of synthetic chemistry. Here, we present a stereoselective synthesis of TTX that proceeds in 22 steps from a glucose derivative. The central cyclohexane ring of TTX and its α-tertiary amine moiety were established by the intramolecular 1,3-dipolar cycloaddition of a nitrile oxide, followed by alkynyl addition to the resultant isoxazoline. A ruthenium-catalyzed hydroxylactonization set the stage for the formation of the dioxa-adamantane core. Installation of the guanidine, oxidation of a primary alcohol, and a late-stage epimerization gave a mixture of TTX and anhydro-TTX. This synthetic approach could give ready access to biologically active derivatives.

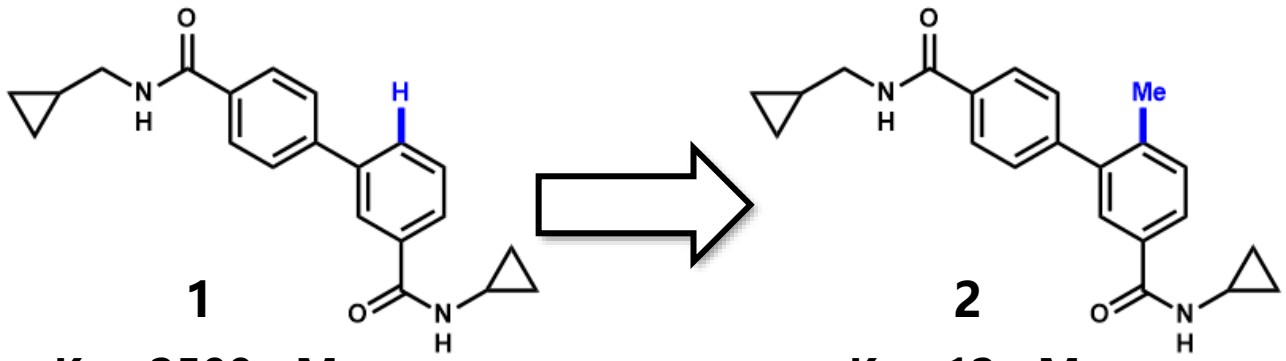
**Research interest:**

- **Medicinal Chemistry**
- **Chemical Methodology**
- **Cancer Cell Biology**
- **Chemo Proteomics**

https://konrad.cup.uni-muenchen.de/research/

# Late-stage Functionalization (LSF)

## p38α MAP Kinase Inhibitor

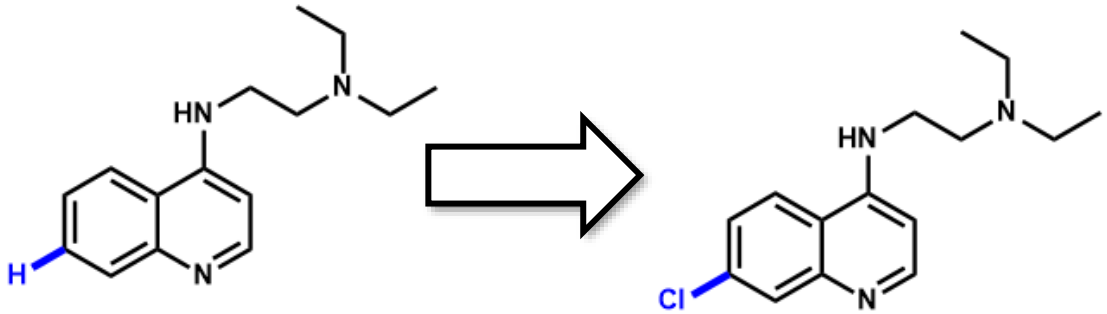



**1** $K_i$ > 2500 nM

**2** $K_i$ = 12 nM

**"Methyl Effect"**

## Antimalarial




**3** $IC_{50}$ = 1060 nM

**4 (Chloroquine)** $IC_{50}$ = 37 nM

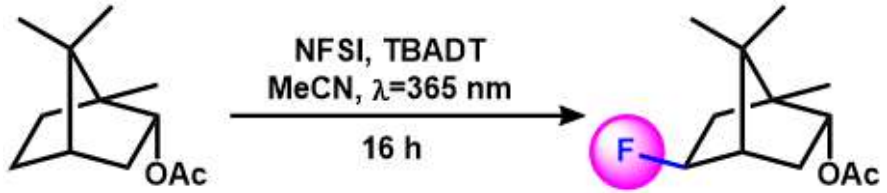
**"Chloro Effect"**

*J. Med. Chem.* 2012, 55, 9, 4489
*J. Med. Chem.* 2023, 66, 8, 5305

# Late-stage C-H Functionalization

## (A) Fluorination

*Angew. Chem.* 2014,126, 4778

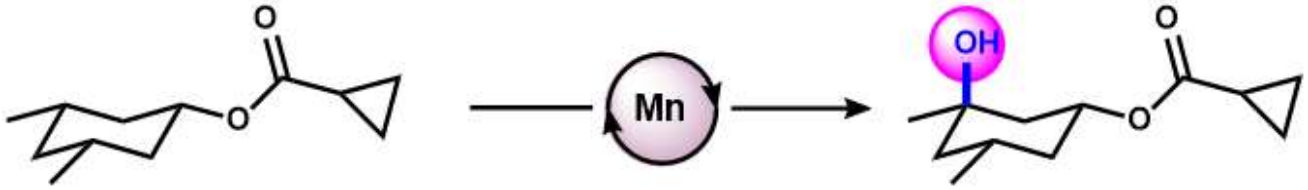



**5** **6**

yield 40%

## (B) Oxidation

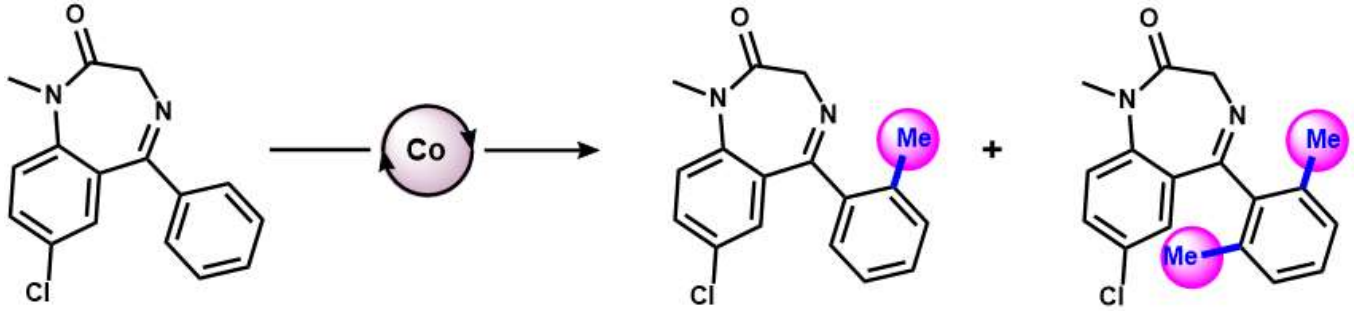
*J. Am. Chem. Soc.* 2023, 145, 29, 15742




**7** **8**

yield 90%
ee 94%

## (C) Methylation



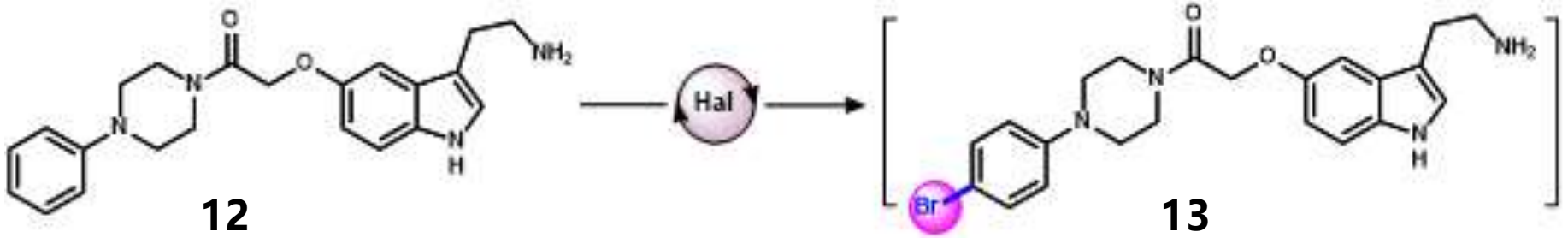

**9** **10** **11**

yield 31%

yield 55%

*Nat. Chem.* 2020, 12, 511

## (D) Halogenation



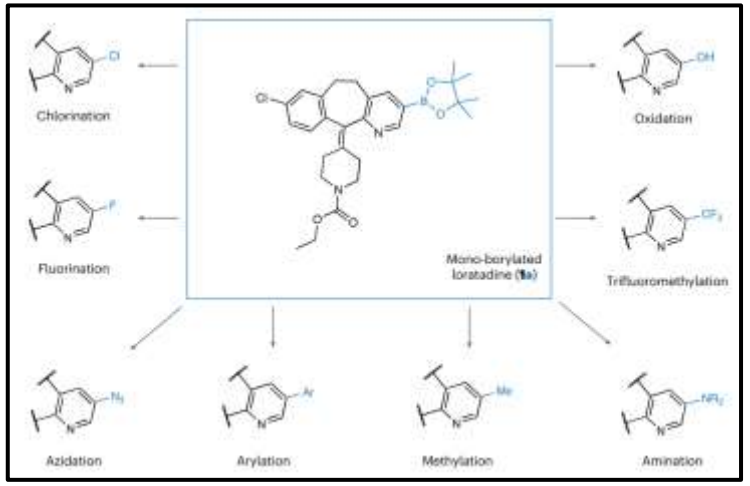

**12** **13**

*Nat. Catal.* 2021, 4, 385

# Late-stage C-H Functionalization

- Rare reports
- One single reaction type

**Limitations**

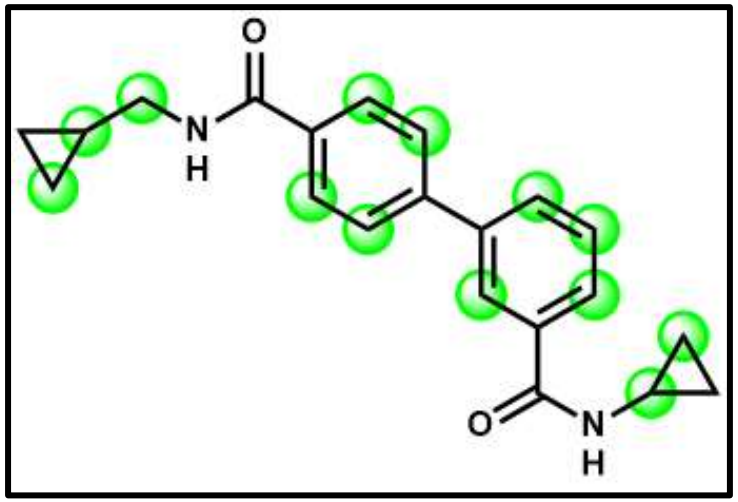



**C-H Borylation**

**Guidelines?**

- Various types of C-H bonds
- Different bond strengths
- Electronic properties
- Steric properties
- FG environment

**Reaction features**

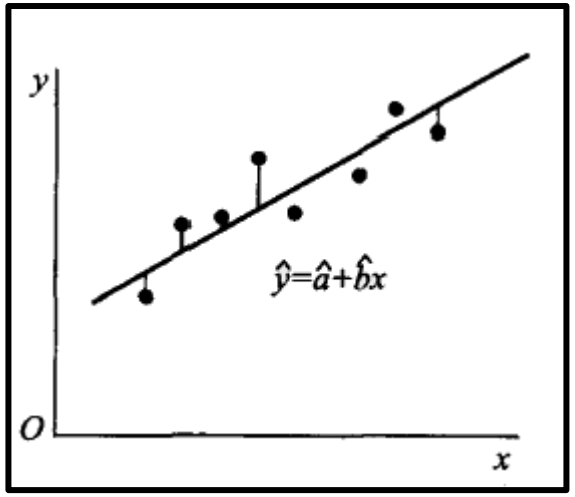


# Deep Learning & High-throughput Experimentation (HTE)

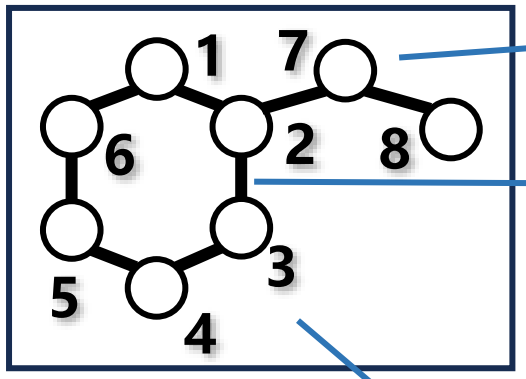
Guidelines ⇨ Deep Learning

Regression & Prediction

⇨ Samples HTE




## Learning Method——Graph Neuro Network (GNN)




Nodes

Edges

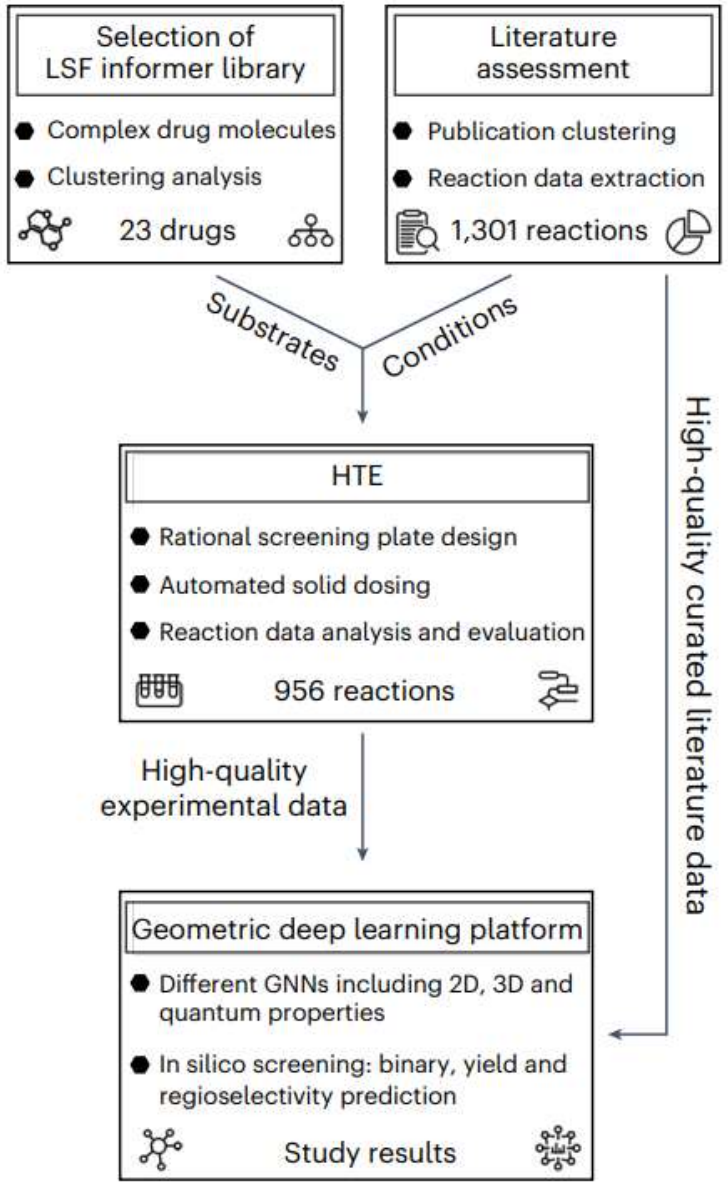
Graph

⇨

**Adjacency Matrix**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

# Overview of this study



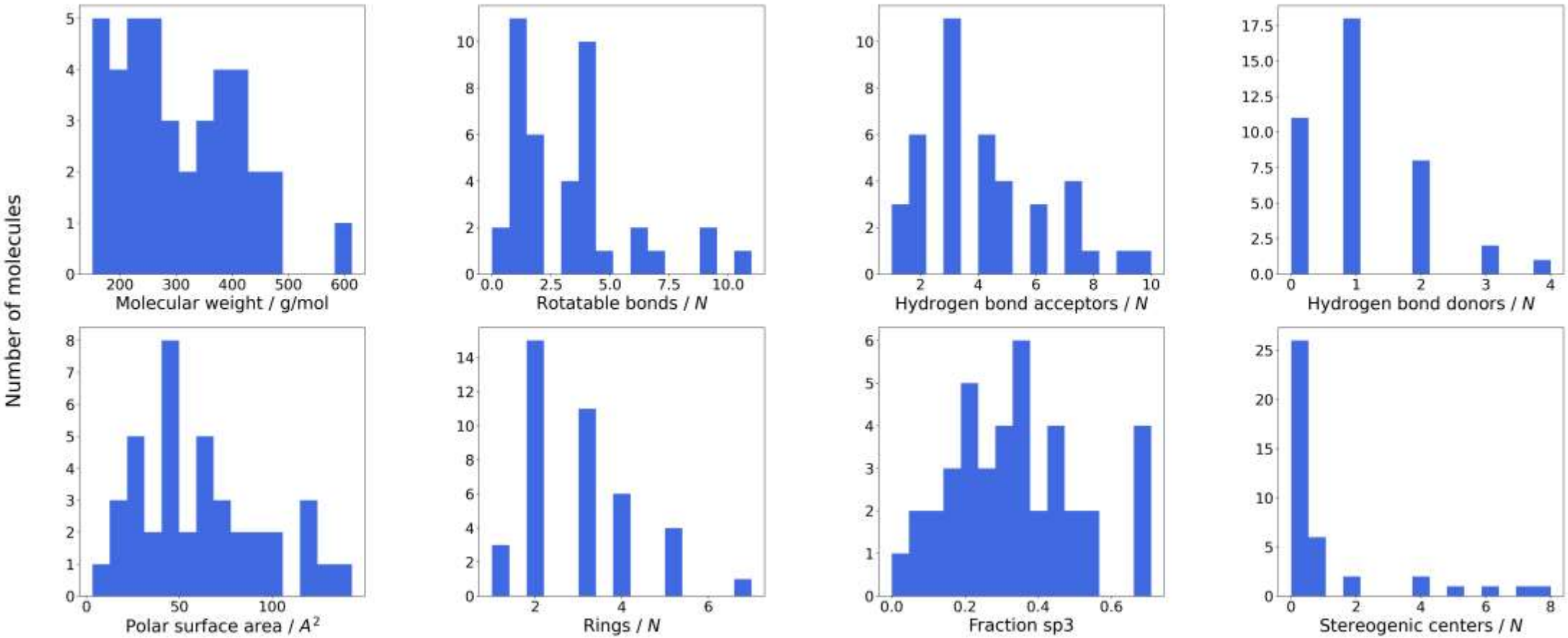

**Process:**

- **Select conditions;**
- **Select substrates;**
- **HTE;**
- **Model fitting**

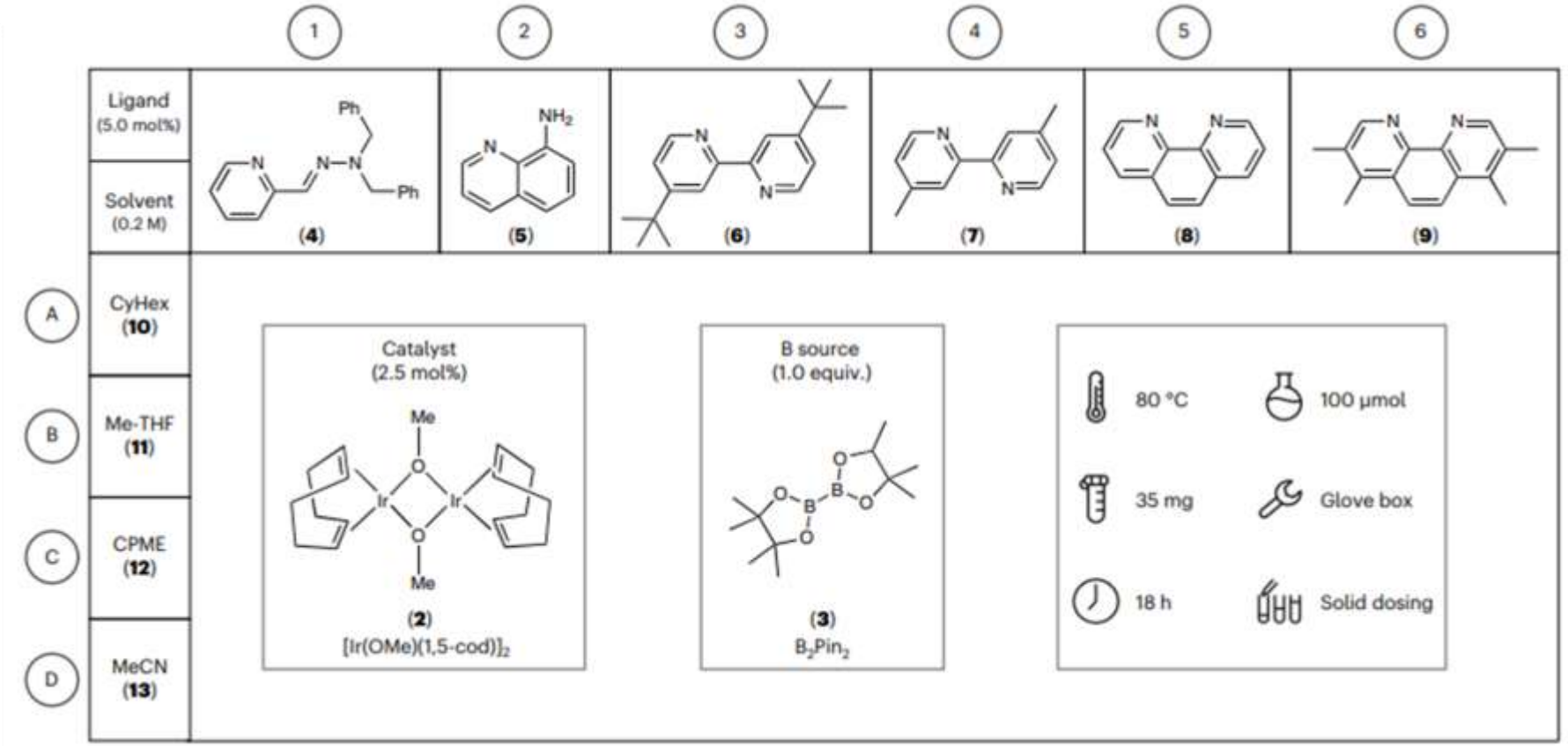
**Three questions:**

- **Reaction or not?**
- **Yield?**
- **Regioselectivity?**

# Method——Substrate & Condition selection



# Method——HTE




- **Operation in a glove-box**
- **Operation in 24- or 96-well plates**
- **Detection by LC-MS**

→

- **Reaction or not?**
- **Yield?**
  - **Reference products?**
  - **New products?**

# Deep learning

## (A) Outcome & yield

| | Reaction yield *r* value | Reaction yield m.a.e. (%) | Binary reaction outcome (random split), AUC (%) | Binary reaction outcome (substrate split), AUC (%) |
|---|---|---|---|---|
| **GTNN2D** | 0.896±0.006 | 4.53±0.09 | **91.8±2.1** | 52±2 |
| **GNN2D** | 0.866±0.005 | 5.61±0.06 | 87.5±1.0 | 51±2 |
| **GTNN3D** | 0.884±0.01 | 4.51±0.11 | 91.4±0.7 | 58±4 |
| **GNN3D** | 0.877±0.001 | 5.33±0.34 | 89.4±0.8 | 65±5 |
| **GTNN2DQM** | **0.898±0.003** | 4.41±0.17 | 90.9±1.5 | 53±5 |
| **GNN2DQM** | 0.876±0.01 | 5.41±0.10 | 89.0±1.1 | 59±5 |
| **GTNN3DQM** | 0.890±0.01 | **4.23±0.08** | **91.8±0.9** | **67±2** |
| **GNN3DQM** | 0.890±0.006 | 4.88±0.24 | 89.1±0.9 | 64±4 |
| **ECFP4NN** | 0.885±0.0006 | 4.55±0.14 | 89.3±1.3 | 52±3 |

**2 models**

- **GNN using sum pooling (GNN)**
- **Graph Transformer Neural Network (GTNN)**

**+ 1 baseline model:**

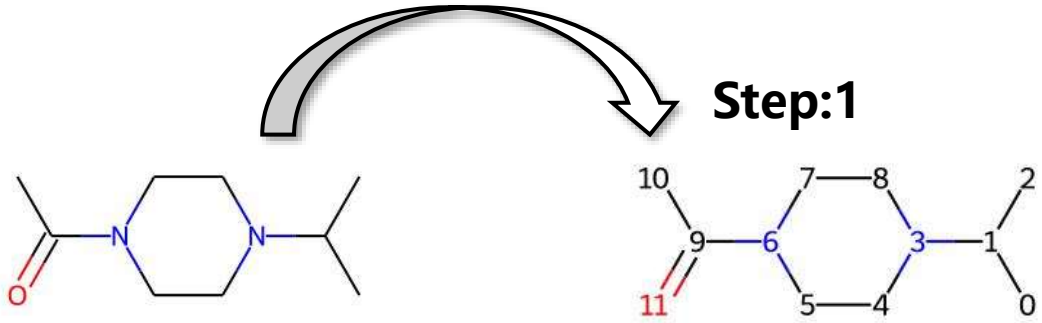
**ECFP4-based Neuro Network (ECFP4NN)**

**4 input types**

- **2D**
- **3D**
- **2DQM**
- **3DQM**

# Deep learning

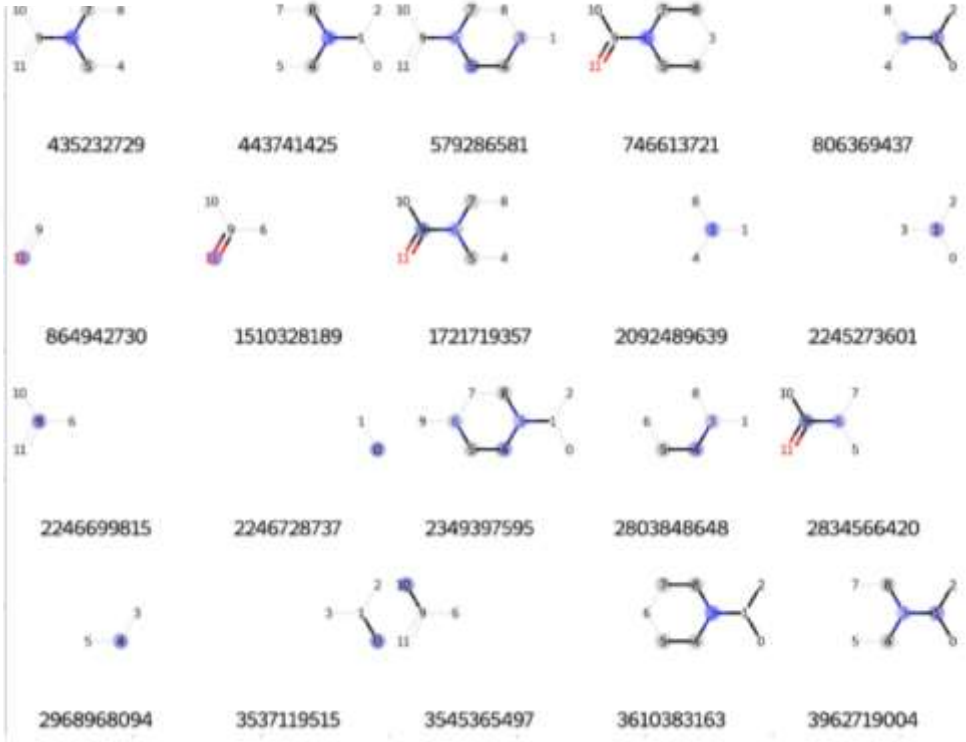
## Molecular fingerprint (ECFP4)



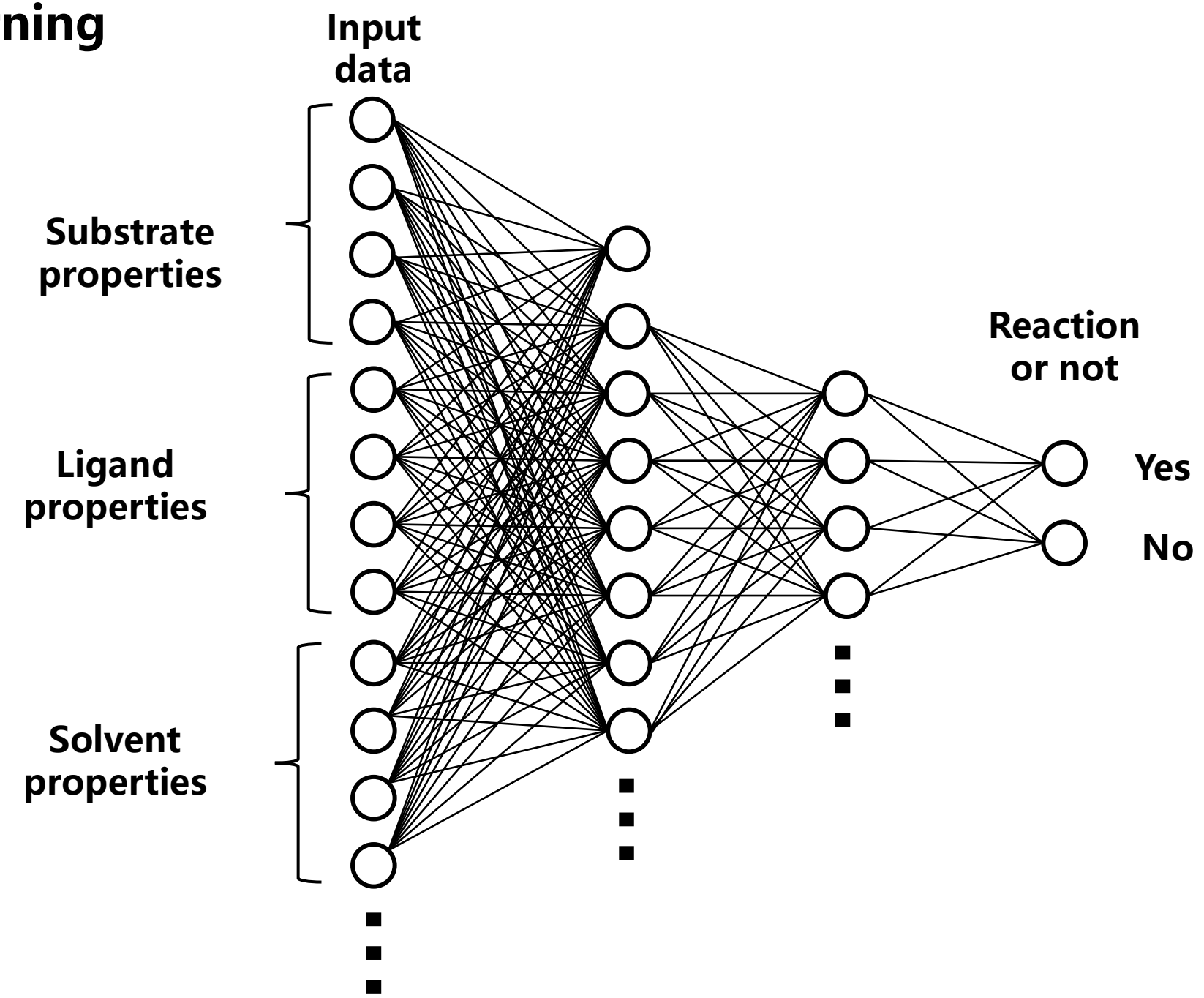

```
1.py ×
16
17      '''Step3: 计算ECFP4分子指纹'''
18      ECFP_bitinfo = {}
19      ECFP = AllChem.GetMorganFingerprint(mol, radius=2, bitInfo=ECFP_bitinfo, useFeatures=False)
20      # print(ECFP)
21      '''一个分子指纹是一个特殊的bit数组，其中囊括的信息就是一个第提取阵'''
22
23
24      '''查看每个有效信息的bit，'''
25      for v in ECFP_bitinfo.values():
26          print(v)
27      '''每个bit的信息(x,y)代表，第x个原子的半径y个化学键的信息，半径表信息一致'''
```

```
Run:  1 ×
D:\Download\envs\rdkit-env-2\python.exe C:/Users/lenovo/PycharmProjects/untitled/1.py
((6, 1),)
((3, 1),)
((5, 2), (7, 2))
((6, 2),)
((1, 1),)
((11, 0),)
((11, 1),)
```

Step:2

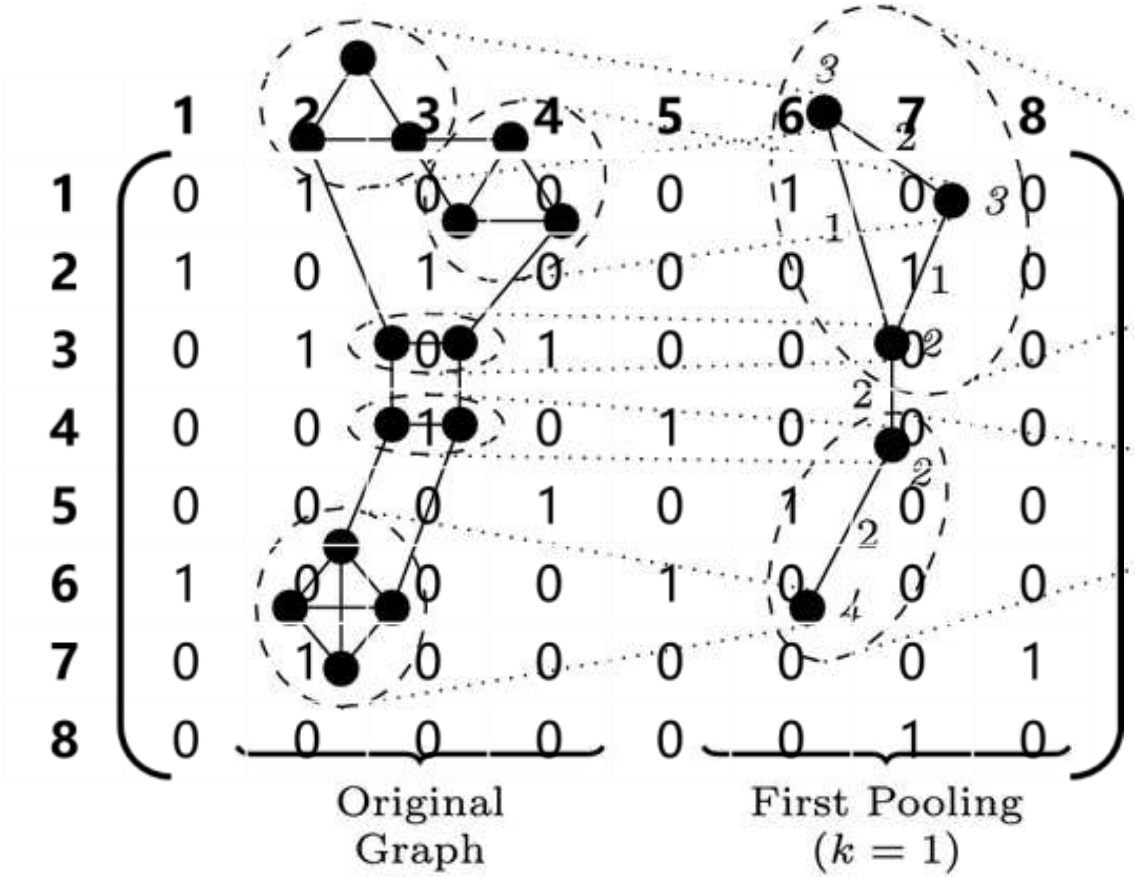


# Deep learning

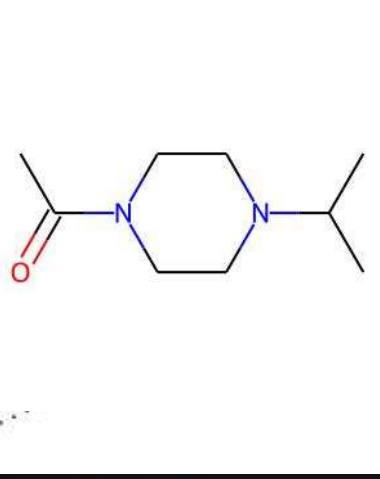


# Deep learning

## Sum pooling




## Why pooling







https://www.researchgate.net/figure/Two-examples-sum-pooling-on-the-same-graph-where-phv1documentclass12ptminimal_fig2_353839563

# Deep learning

## GNN input in this article:

### Nodes:

```
atomids = []
qml_atomids = []
is_ring = []
hyb = []
arom = []
crds_3d = []

AllChem.EmbedMolecule(mol, randomSeed=randomseed)
AllChem.UFFOptimizeMolecule(mol)

for idx, i in enumerate(mol.GetAtoms()):
    atomids.append(ATOMTYPE_DICT[i.GetSymbol()])
    qml_atomids.append(QML_ATOMTYPE_DICT[i.GetSymbol()])
    is_ring.append(IS_RING_DICT[str(i.IsInRing())])
    hyb.append(HYBRIDISATION_DICT[str(i.GetHybridization())])
    arom.append(AROMATOCITY_DICT[str(i.GetIsAromatic())])
    crds_3d.append(list(mol.GetConformer().GetAtomPosition(idx)))
```

- Atom ID
- Hybridization
- Coordinates (for 3D conformation)
- Ring status
- Aromatic status

### Edges:

```
edge_dir1 = []
edge_dir2 = []
for idx, bond in enumerate(mol.GetBonds()):
    a2 = bond.GetEndAtomIdx()
    a1 = bond.GetBeginAtomIdx()
    edge_dir1.append(a1)
    edge_dir1.append(a2)
    edge_dir2.append(a2)
    edge_dir2.append(a1)

edge_2d = torch.from_numpy(np.array([edge_dir1, edge_dir2]))
```

### 2D: Start atom and end atom

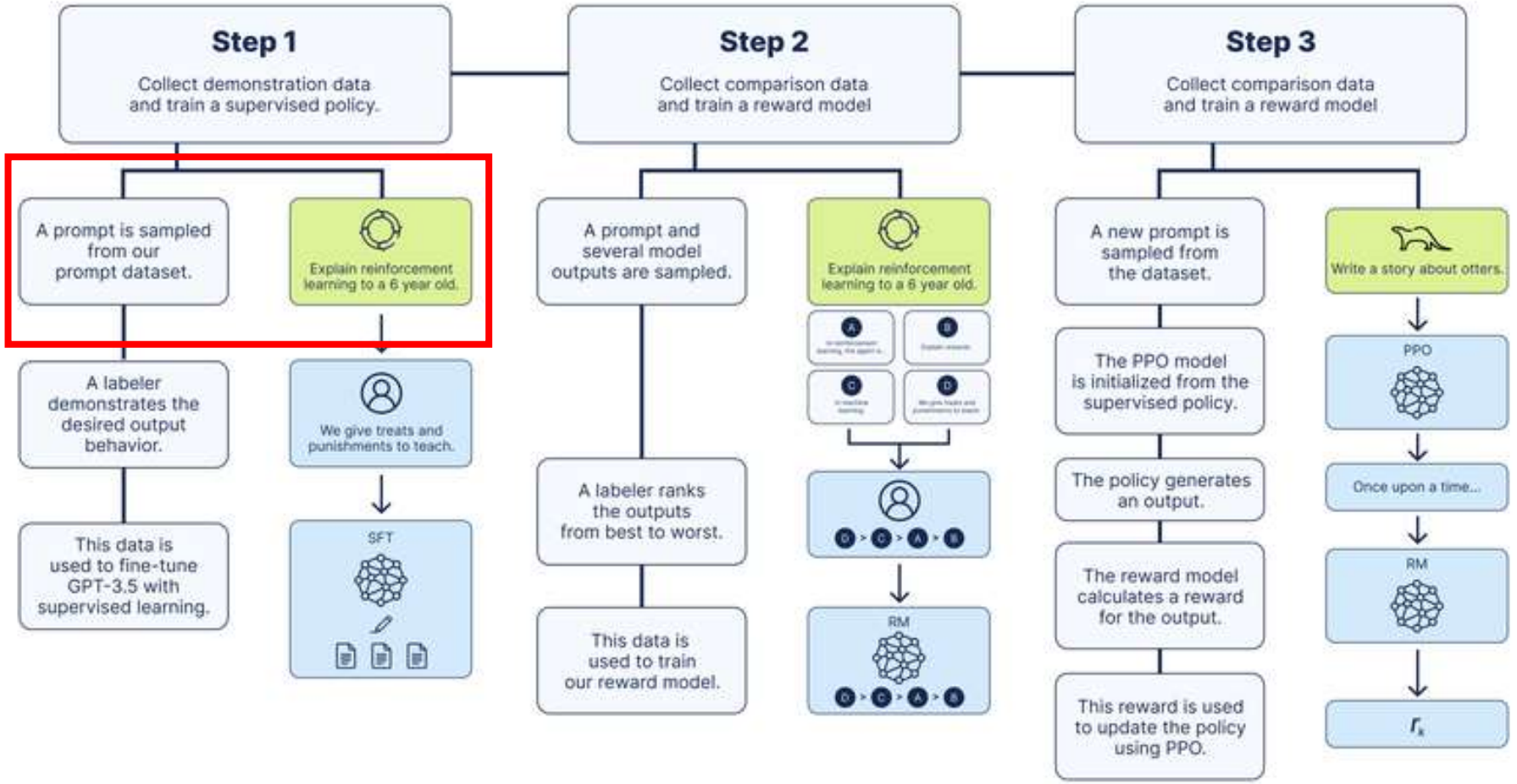
```
# 3D graph for qml and qml prediction
qml_atomids = torch.LongTensor(qml_atomids)
xyzs = torch.FloatTensor(crds_3d)
edge_index = np.array(nx.complete_graph(qml_atomids.size(0)).edges())
edge_index = to_undirected(torch.from_numpy(edge_index).t().contiguous())
edge_index, _ = add_self_loops(edge_index, num_nodes=crds_3d.shape[0])

qml_graph = Data(
    atomids=qml_atomids,
    coords=xyzs,
    edge_index=edge_index,
    num_nodes=qml_atomids.size(0),
)

charges = QMLMODEL(qml_graph).unsqueeze(1).detach().numpy()

# Get edges for 3d graph
distance_matrix = squareform(pdist(crds_3d))
np.fill_diagonal(distance_matrix, float("inf"))  # to remove self-loops
edge_3d = torch.from_numpy(np.vstack(np.where(distance_matrix <= radius)))
```

### 3D: Tensor

# Deep learning

## Graph transformer

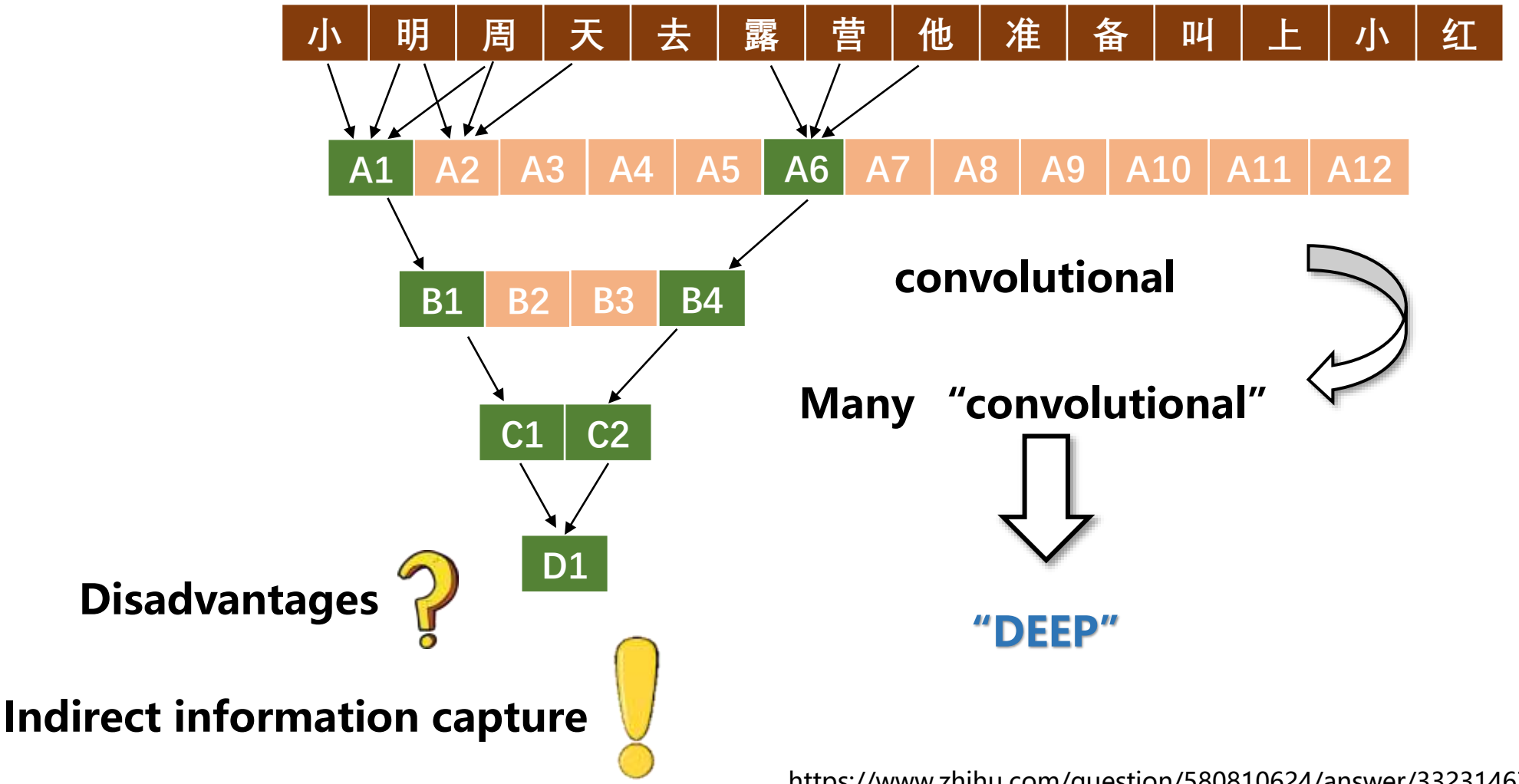
## What is "transformer"




ChatGPT

# Deep learning

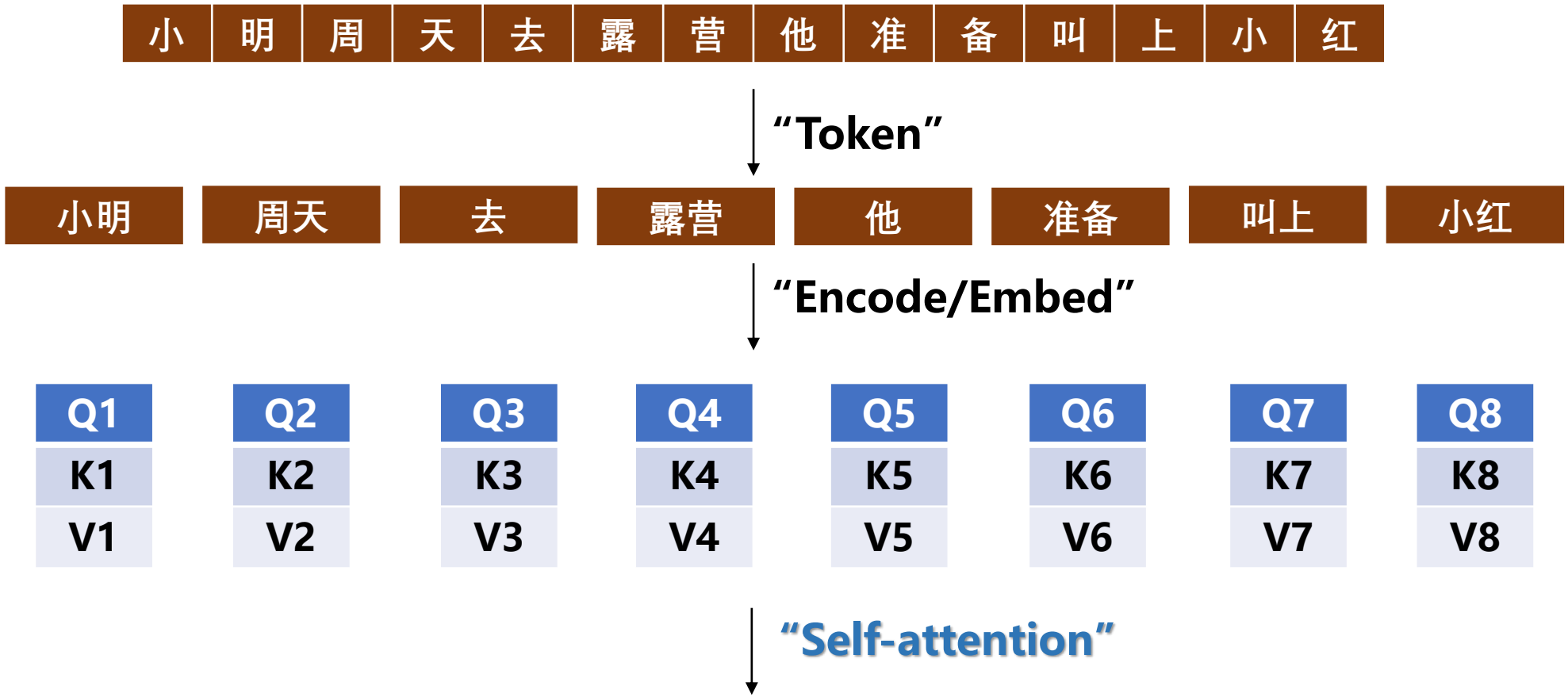
## CNN VS Transformer in NLP




https://www.zhihu.com/question/580810624/answer/3323146786

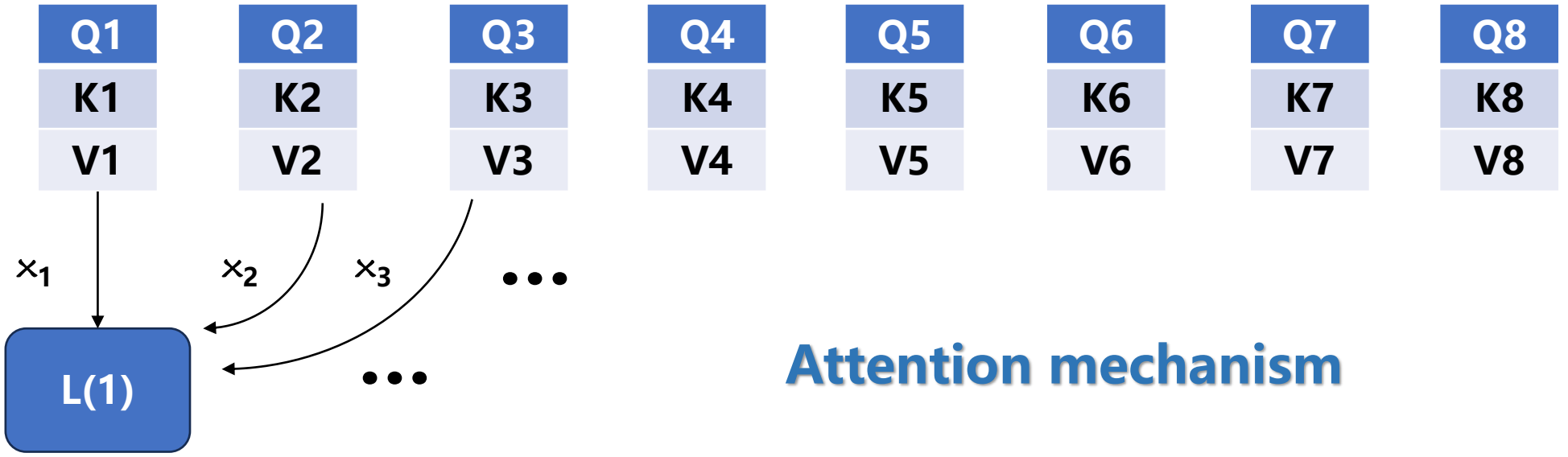
# Deep learning

## CNN VS Transformer in NLP



# Deep learning

## CNN VS Transformer in NLP

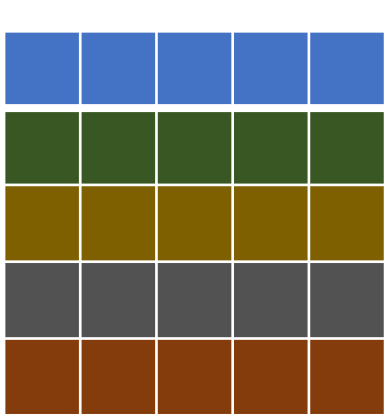



**Attention mechanism**

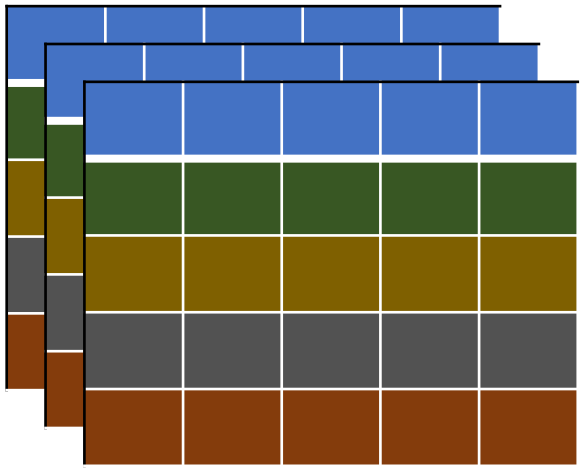
**This framework can be applied to many other tasks**

# Deep learning

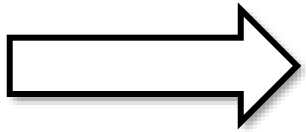
## How was ChatGPT trained?



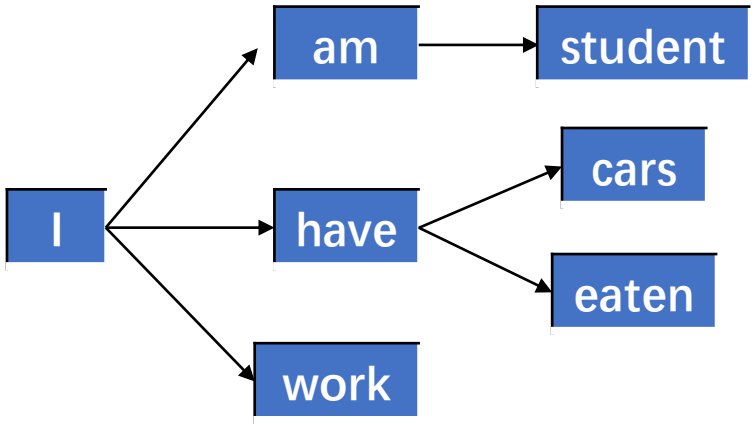
Enormous text input



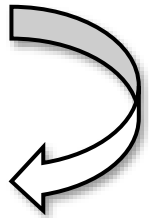
Many L(i)
(what words are related to each other)



Decode




**Speak what humans speak**

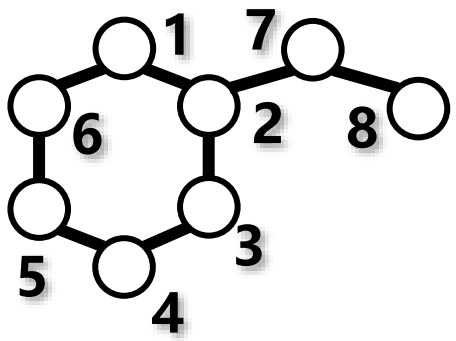


**Speak what humans understand**

https://www.stackbuilders.com/blog/inside-the-brain-of-chatgpt/
https://www.nolibox.com/creator_articles/principle_of_ChatGPT.html

# Deep learning

## What is Graph Transformer?




$$L(A2) = x1*H(A1) + x2*H(A2) + x3*H(A3) + \ldots$$


## Comparison to normal GNN?

**Normal GNN:**
**Manually identification of Nodes & Edges before learning**

**GTNN:**
**trap key features of nodes & Edges directly for learning**

# Deep learning

## (A) Outcome & yield

| | Reaction yield *r* value | Reaction yield m.a.e. (%) | Binary reaction outcome (random split), AUC (%) | Binary reaction outcome (substrate split), AUC (%) |
|---|---|---|---|---|
| GTNN2D | 0.896±0.006 | 4.53±0.09 | **91.8±2.1** | 52±2 |
| GNN2D | 0.866±0.005 | 5.61±0.06 | 87.5±1.0 | 51±2 |
| GTNN3D | 0.884±0.01 | 4.51±0.11 | 91.4±0.7 | 58±4 |
| GNN3D | 0.877±0.001 | 5.33±0.34 | 89.4±0.8 | 65±5 |
| GTNN2DQM | **0.898±0.003** | 4.41±0.17 | 90.9±1.5 | 53±5 |
| GNN2DQM | 0.876±0.01 | 5.41±0.10 | 89.0±1.1 | 59±5 |
| GTNN3DQM | 0.890±0.01 | **4.23±0.08** | **91.8±0.9** | **67±2** |
| GNN3DQM | 0.890±0.006 | 4.88±0.24 | 89.1±0.9 | 64±4 |
| ECFP4NN | 0.885±0.0006 | 4.55±0.14 | 89.3±1.3 | 52±3 |

**Testing data from the experimental data**

***r*** **(Pearson coefficient):**
**预测值和观测值的相关系数。**
**r越接近1，相关性越强。**

**m.a.e. (Mean absolute error):**
**预测值和观测值之间绝对误差的平均值。**
**m.a.e.越小，预测越准确。**

**Reaction yield**

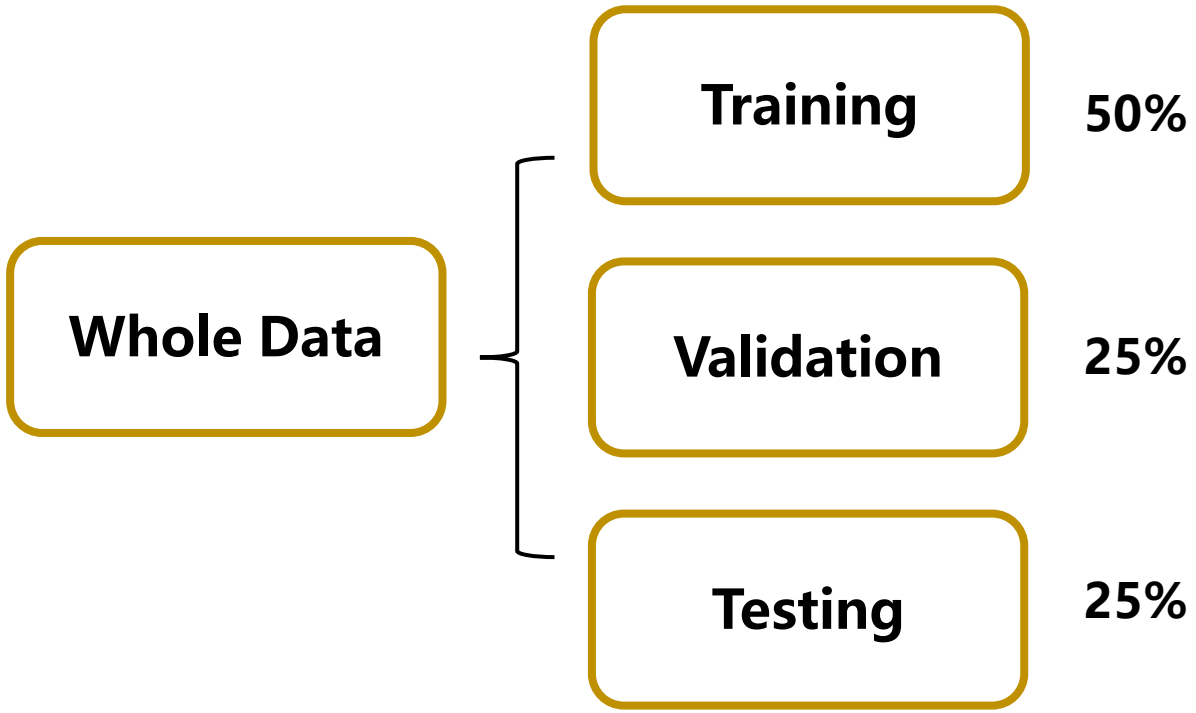
**AUC (Area under curve):**
**在二分类问题中，AUC越大，性能越好。**

→ **Reaction outcome**

# Deep learning

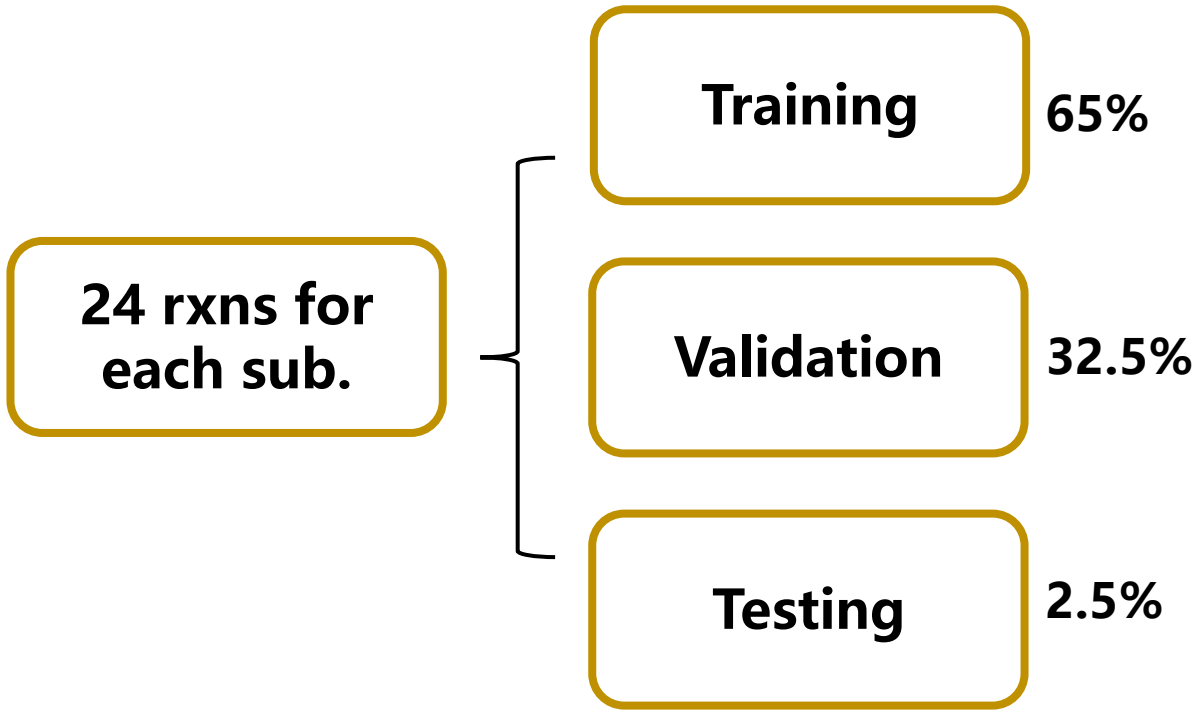
## (A) Outcome & yield

### Random split




Investigate the performance on **new conditions** for **known substrates**

### Substrate-based split




Investigate the performance on **unknown substrates** for **known conditions**

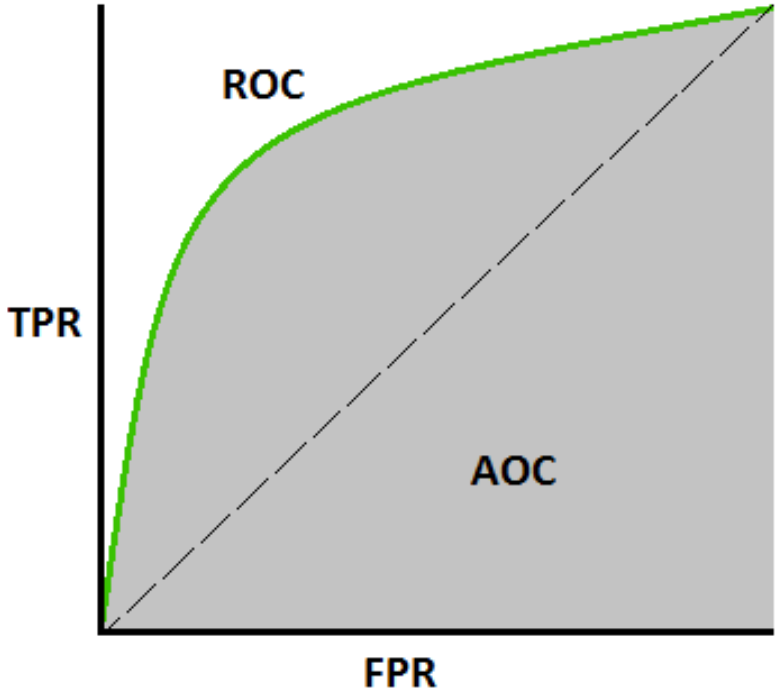
# Deep learning

## (A) Outcome & yield

## What is AUC in DL?

**TPR: 真正率，正确预测为正样本的样本数占真正样本的比例**

**FPR: 假正率，错误预测为正样本的样本数占真负样本的比例**




| | 预测 | | |
|---|---|---|---|
| 实际 | | 正样本 | 负样本 |
| | 正样本 | True Positive (TP) | False Negative (FN) |
| | 负样本 | False Positive (FP) | True Negative (TN) |

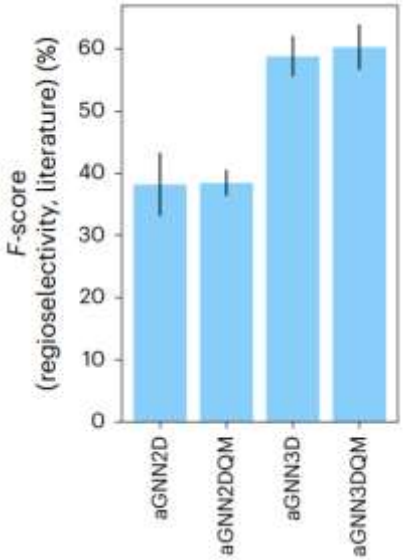
https://zhuanlan.zhihu.com/p/121036836
https://blog.csdn.net/qq_45769063/article/details/120008815

# Deep learning

## (B) Regioselectivity

| | *F*-score (%) | PVV (%) | TPR (%) | Accuracy (%) |
|---|---|---|---|---|
| **aGNN2D** | 38±5 | 56±1 | 30±6 | 88±1 |
| **aGNN2DQM** | 39±2 | 54±2 | 30±3 | 88±0.3 |
| **aGNN3D** | 59±3 | **62±2** | 56±4 | **90±1** |
| **aGNN3DQM** | **60±4** | **62±2** | **59±6** | **90±1** |

**Testing data from the experimental data**




**Testing data from literature data**

**a : atomic features (No pooling)**

**PPV (positive predictive value): 阳性预测值。识别为正确的样品中真正正确的比值。**

$$PPV = \frac{TP}{TP + FP}$$

**TPR (true positive rate): 阳性预测正确值。真正正确的样品中被识别为正的比值。**

$$TPR = \frac{TP}{TP + FN}$$

**F-score: PPV和TPR的加权平均，权衡两个指标。**

$$F - score = \frac{PPV * TPR * 2}{PPV + FPR}$$

**Accuracy: 被正确分类的样品占总体的比值。**

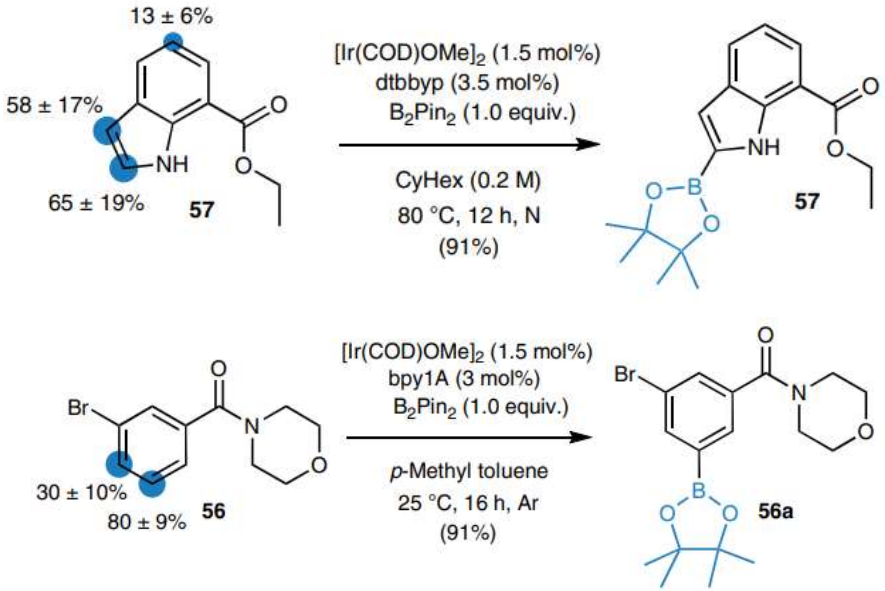
$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

| | 预测 | | |
|---|---|---|---|
| 实际 | | 正样本 | 负样本 |
| | 正样本 | True Positive (TP) | False Negative (FN) |
| | 负样本 | False Positive (FP) | True Negative (TN) |

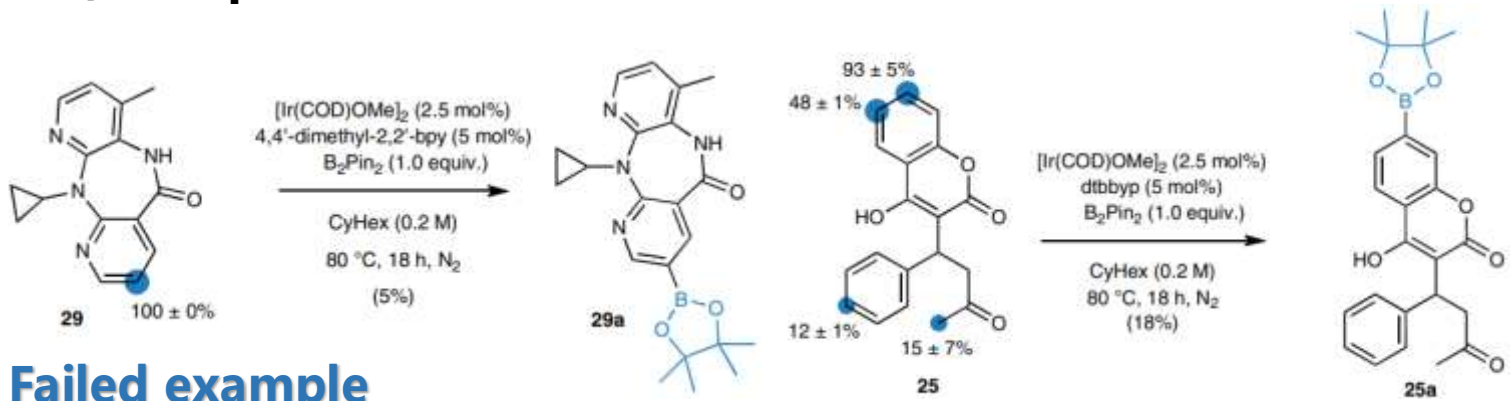
# Deep learning

## (B) Regioselectivity

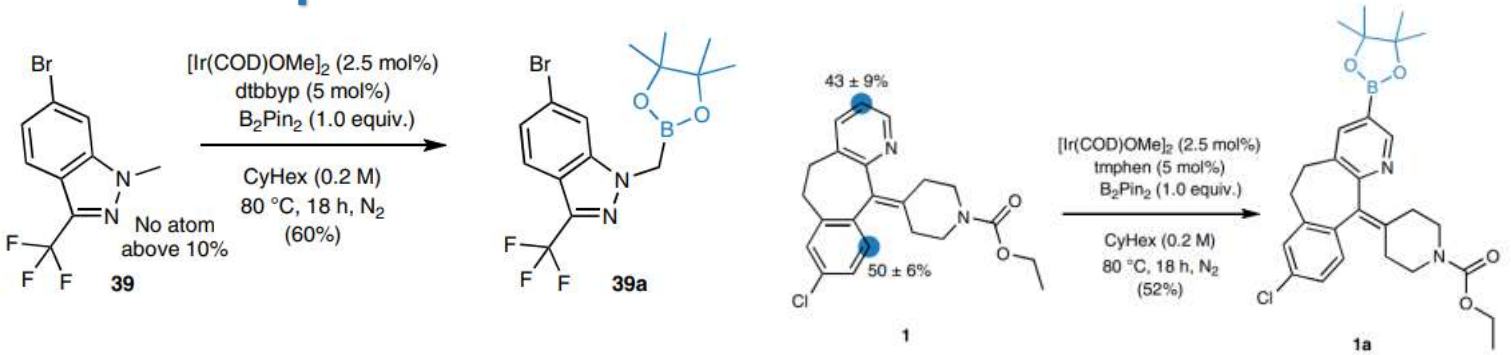
### 1) Retrospective from test data



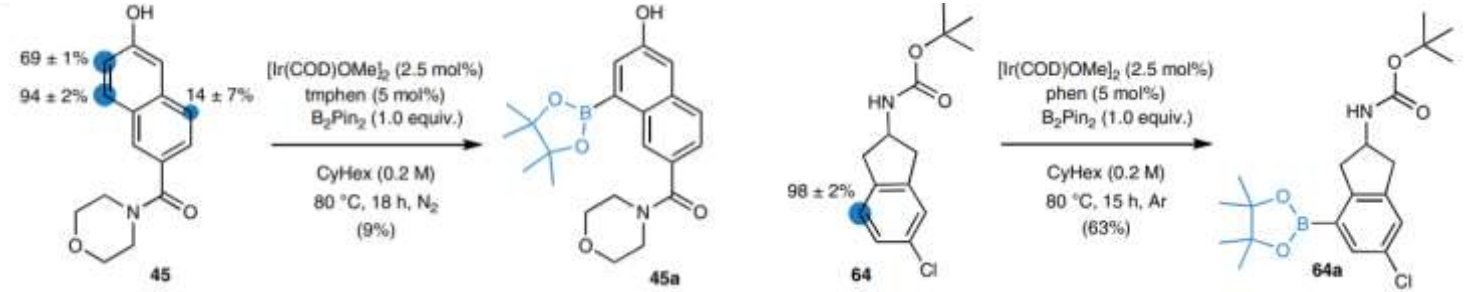

### 2) Prospective from literature data




**Failed example**



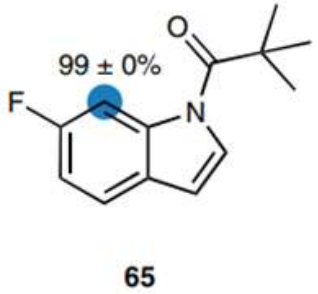

### 3) Retrospective from Roche dataset



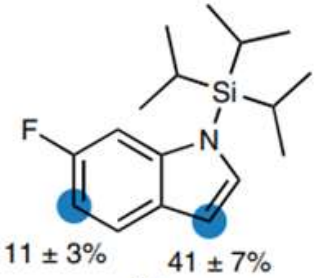
# Deep learning

## (B) Regioselectivity

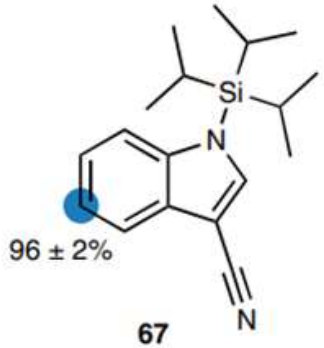
### 4) Identification of 3D properties




65

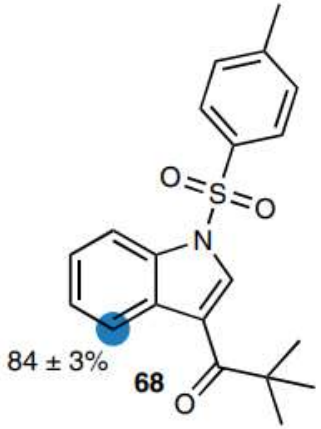



66

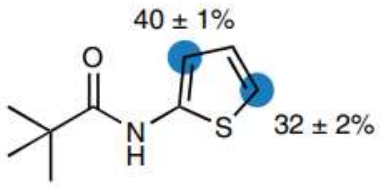



67

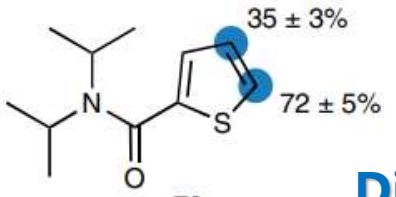
**Steric hindrance**




68




69




70

**Directing group**

**Unseen molecules**

# Discussion

## (1) HTE VS Literature

### Prediction in outcome & yield

| | Reaction yield *r* value | Reaction yield m.a.e. (%) |
|---|---|---|
| **GTNN2D** | 0.896±0.006 | 4.53±0.09 |
| **GNN2D** | 0.866±0.005 | 5.61±0.06 |
| **GTNN3D** | 0.884±0.01 | 4.51±0.11 |
| **GNN3D** | 0.877±0.001 | 5.33±0.34 |
| **GTNN2DQM** | **0.898±0.003** | 4.41±0.17 |
| **GNN2DQM** | 0.876±0.01 | 5.41±0.10 |
| **GTNN3DQM** | 0.890±0.01 | **4.23±0.08** |
| **GNN3DQM** | 0.890±0.006 | 4.88±0.24 |
| **ECFP4NN** | 0.885±0.0006 | 4.55±0.14 |

**Validated by HTE results**

| Prediction error | Mean absolute error/ % |
|---|---|
| GTNN2D | 16.7 (±0.13) |
| GNN2D | 16.4 (±0.2) |
| GTNN3D | 16.4 (±0.24) |
| GNN3D | 16.2 (±0.14) |
| GTNN2DQM | **16.1 (±0.02)** |
| GNN2DQM | 16.3 (±0.04) |
| GTNN3DQM | 16.2 (± 0.16) |
| GNN3DQM | 16.2 (±0.14) |
| ECFP4NN | 18.2 (±0.05) |

**Validated by literature results**

**Why different** 

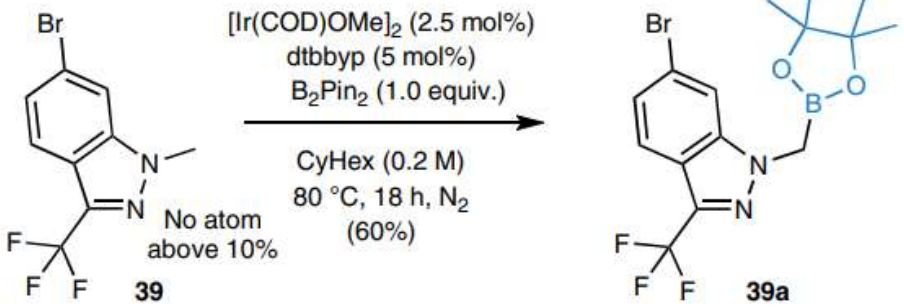
1. **HTE: All in the same standard**

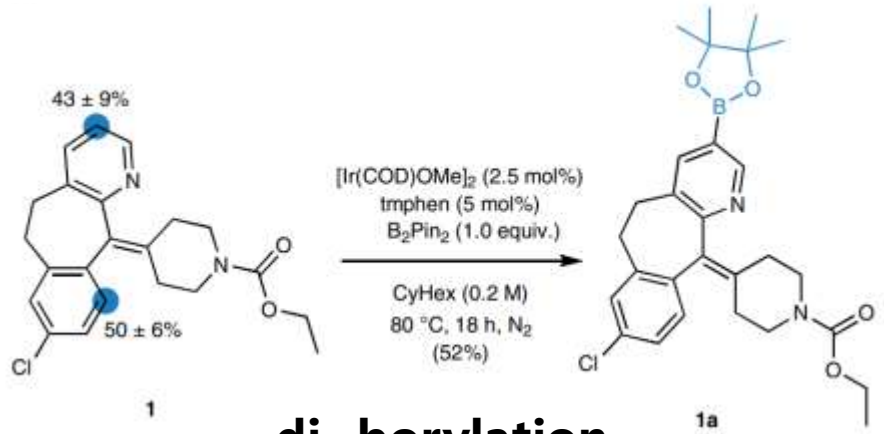
   **Lit.: standards variates**

2. **HTE covers a less diverse reaction parameter space**

# Discussion

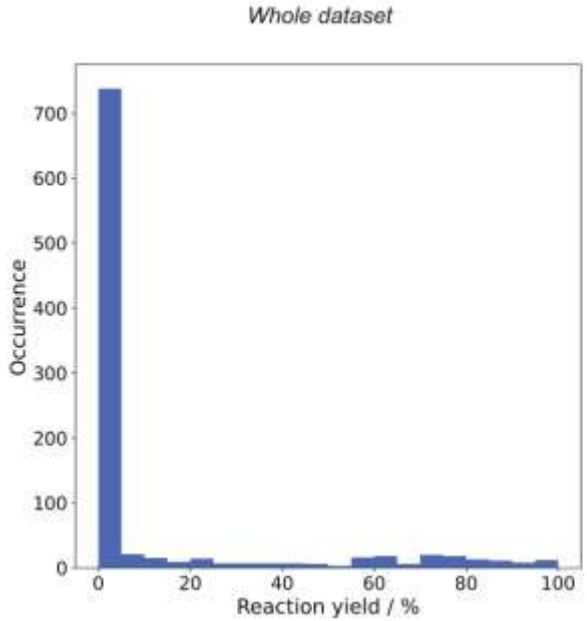
## (2) Deep learning greatly relies on the quality of input data!

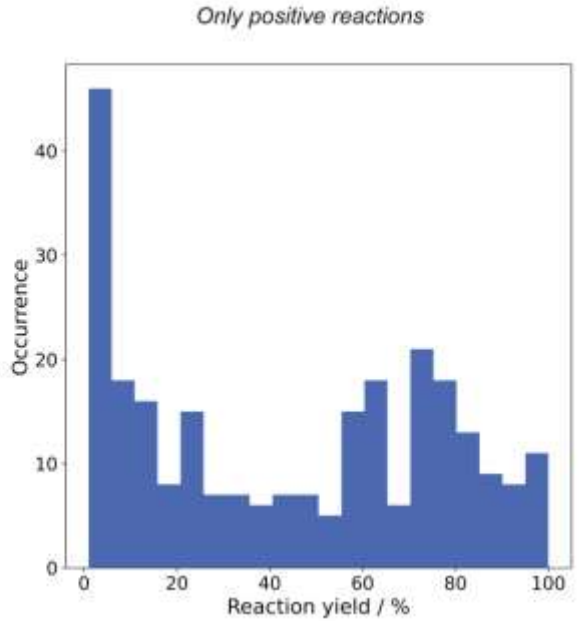



***sp³* borylation**

**di- borylation**



# Discussion

## (3) Does this research deserve Nature Chemistry?

**To be honest, I don´t think so…**

**Combining DL and HTS in synthetic chemistry is not a fancy idea now…**


**Highlights of this article: Application in C-H borylation, thus for LSF**

**However, no prospective application on out-of-dataset was reported**

**DL+HTS should have guidance significance to drug synthesis**

- **Tell what condition is the best?**
- **In combination with AI-promoted retro-synthesis analysis? (New route)**
- **In combination with virtual screening? (High-quality scaffold)**

# Thanks for your attention

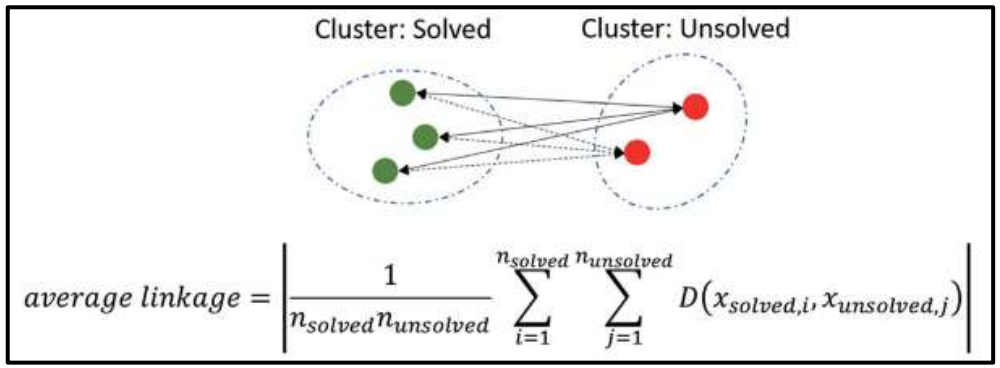
**Bin Yang**

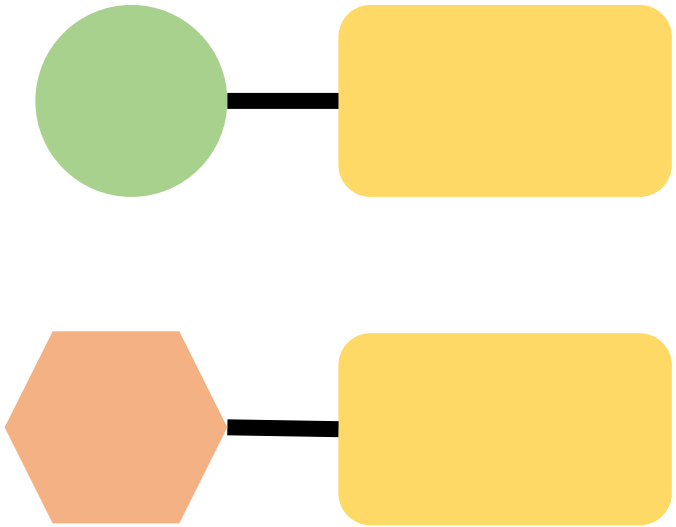
**2024.01.13**

# Input for Deep Learning

## (A) Molecular fingerprint

### ——Example: Retrosynthetic accessibility score



**Known retrosynthetic route (Solved)**

**RAscore**



**Unknown retrosynthetic route (Unsolved)**

Thakkar A, Chadimová V, Bjerrum E J, et al. Chem Sci, 2021, 12(9): 3339.

# Input for Deep Learning

## (B) Density Functional Theory calculation

The first principle of Kohn-Hohenberg:

只要知道基态电荷数密度，那么就能确定基态的能量

The second principle of Kohn-Hohenberg:

使整体能量最小的电荷密度就是真实电荷密度


## DFT = Solving the Schrödinger equation

### 自洽求解：

- S1: 猜一个电荷数密度$n_{trial}(r)$，代入方程，求得波函数$\Psi_i(r)$;
- S2: 将$\Psi_i(r)$代入方程，求得$n_{sol}(r)$;
- S3: 比较$n_{trial}(r)$和$n_{sol}(r)$，若误差小于给定精度，输出结果，否则用$n_{sol}(r)$取代$n_{trial}(r)$，循环往复。

DFT in DL:

1). ΔG in transition states

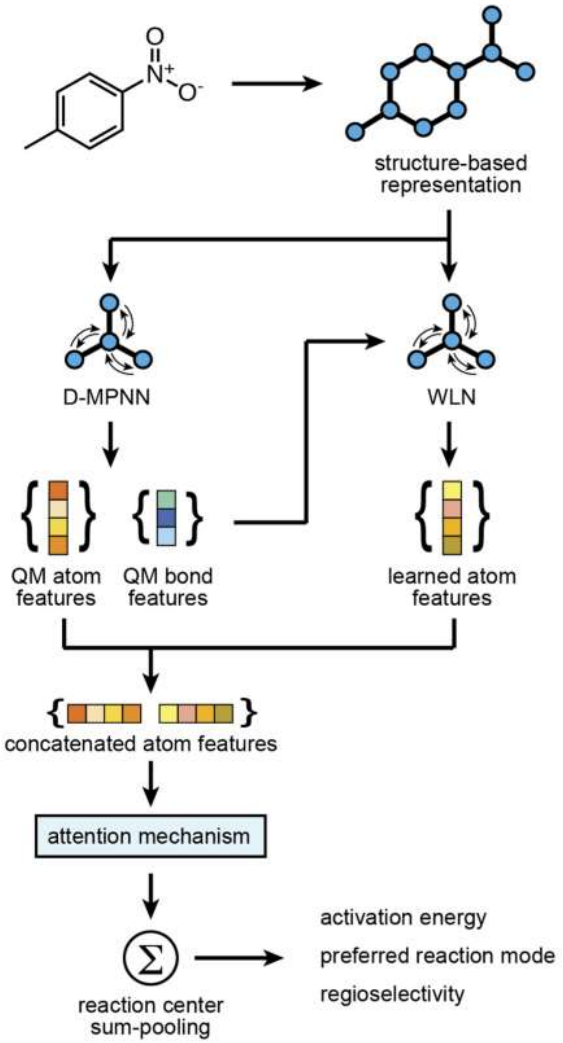
2). Atomic partial charges

Regioselectivity, Diasteroselectivity

# Input for Deep Learning

## (B) Density Functional Theory calculation

### ——Example: Reactivity prediction




**Atom features**: Atomic number, ring status, etc.

**QM bond features**: Bond order, bond length from QM calculation, etc.

**QM atom features**: Hirshfeld atomic charges, nucleo- and electrophilic Fukui functions, etc.

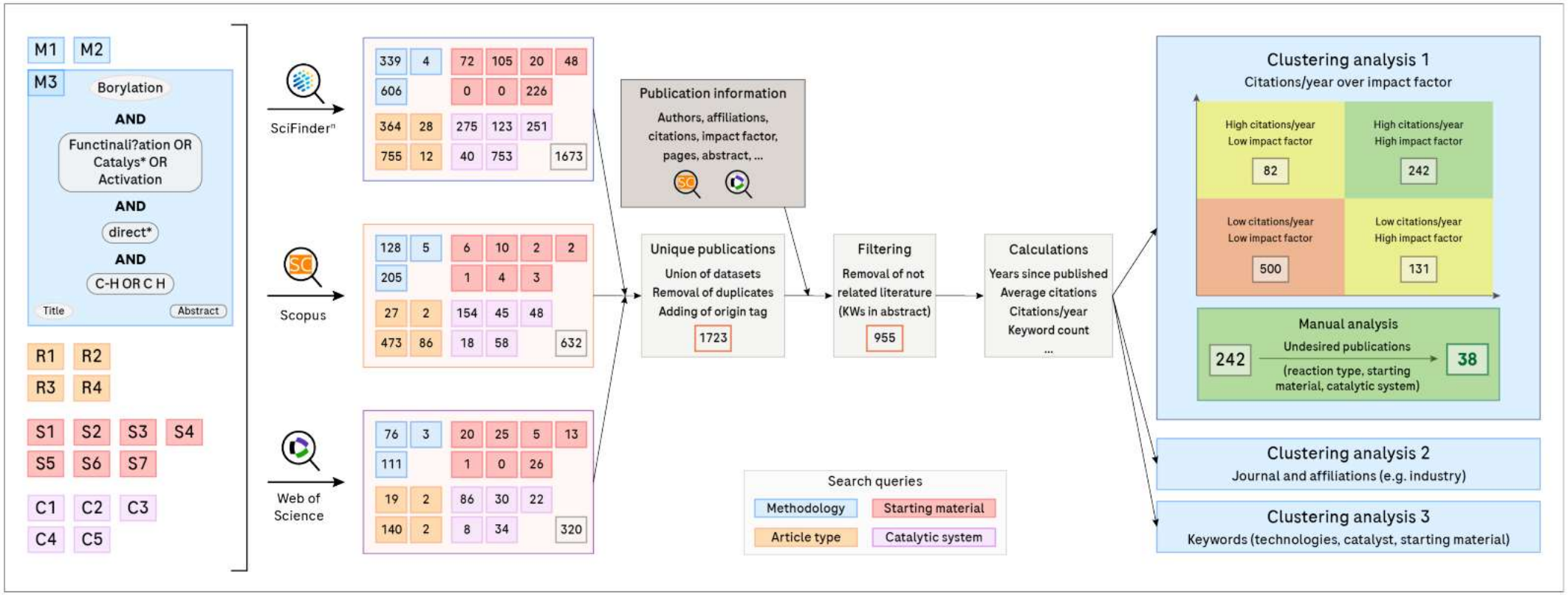
**Table Average MAE (kcal/mol) when predicting SN2 barrier heights**

| Labeled points | Baseline GNN | ml-QM-GNN | Chemprop | BoB[a] | SLATM[a] |
|---|---|---|---|---|---|
| 225 (180 + 45) | 9.07 ± 0.04 | 3.61 ± 0.14 | 6.71 ± 0.08 | 4.89 | 4.44 |
| 450 (360 + 90) | 8.89 ± 0.13 | 3.28 ± 0.04 | 4.01 ± 0.02 | 4.28 | 3.87 |
| 900 (720 + 180) | 8.61 ± 0.07 | 2.97 ± 0.03 | 3.23 ± 0.07 | 3.78 | 3.21 |
| 1800 (1440 + 360) | 8.49 ± 0.03 | 2.76 ± 0.01 | 2.85 ± 0.02 | 3.49 | 2.92 |

[a]Taken directly from the work of Heinen, von Rudorff, and von Lilienfeld.[40]

Stuyver T, Coley C W. J Chem Phys, 2022, 156(8).

# Method——Systematic literature analysis




**M: Different methodologies (Direct/Indirect, etc.)**

**S: Different substrates (aromatic/aliphatic, etc.)**

**R: Article types (review/article, etc.)**

**C: Catalytic systems (Ir/Rh, etc.)**

**Query keywords**

# Method——Substrate selection

**Cortellis Drug Discovery Intelligence (CDDI database)**

**DRUGS**

**1174 drugs (200<MW<800)**

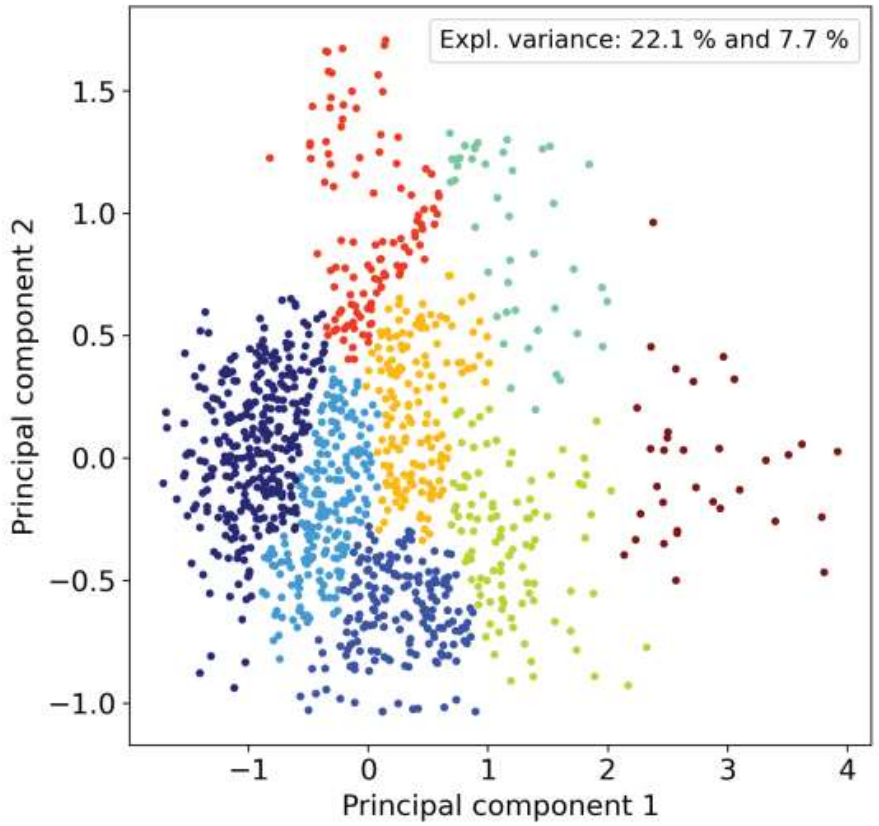
**(Based on ECFP4)**

**8 clusters**

**10 closest molecules to the cluster center**

**3/10 commercially available**
**Total sample: 23**

**Jaccard similarity:** $J = \frac{A \cap B}{A \cup B}$



# Method——Substrate selection

**Top 100 most popular ring assemblies in Roche´s database**

- **MW<300**
- **Heavy atoms < 20**
- **>1 g stock available**
- **Not involved in other projects**

**268 fragments**

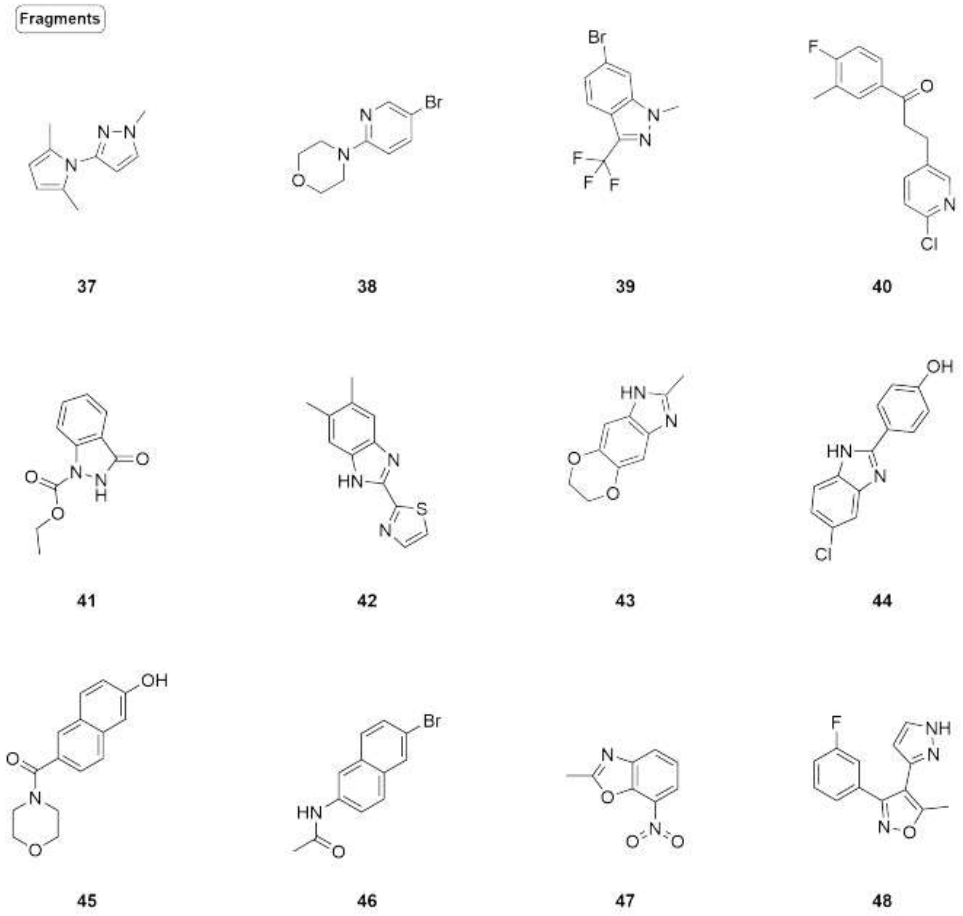
**Manually selection:**

- **Halogen / OH on aromatic rings**
- **Frequently used heterocycles**

**16 fragments**

**+ 5 frequently occurring literature substrates**

## FRAGMENTS



# Method——Substrate selection

# Input for Deep Learning

## (B) Density Functional Theory calculation

**What does DFT do?  An approximate solution for the Schrödinger equation**

**Schrödinger equation in describing electrons:**

$$H\Psi(r_1, \dots, r_N) = E\Psi(r_1, \dots, r_N)$$

$$\mathrm{E} = -\underbrace{\frac{h}{2m_e}\sum_i^N \nabla^2}_{\text{电子动能}} + \underbrace{\sum_i^N V_{ext}(r_i)}_{\text{电子与相对固定的原子核的互作}} + \underbrace{\sum_{i=1}^N \sum_{j=i+1}^N U(r_i, r_j)}_{\text{电子-电子的互作}} \quad \text{Eq (1)}$$

**Considering a system with N electrons, there will be 3N variables!** ⇒ **Almost impossible to accurately describe the situation of electrons!**

**All electrons as objects** ⇒ **Electron density for ONE electron as objects**

**Only 3 variables**

# Input for Deep Learning

## (B) Density Functional Theory calculation

**The first principle of Kohn-Hohenberg:**

**只要知道基态电荷数密度，那么就能确定基态的能量**

**How can we get n(r)** 

$$E[\{\Psi_i\}] = E_{known}[\{\Psi_i\}] + E_{XC}[\{\Psi_i\}] \quad \text{Eq (2)}$$

**Where:**

$$E_{known}[\{\Psi_i\}] = -\frac{h^2}{m_e}\sum_i \Psi_i^* \nabla^2 \Psi_i d^3r + \int V(r)\, n(r) dr^3 + \frac{e^2}{2}\iint \frac{n(r)n(r')}{|r-r'|} d^3r d^3r' \quad \text{Eq (3)}$$

**电子动能** | **电子与相对固定的原子核的互作** | **电子-电子的互作**

$E_{XC}[\{\Psi_i\}]$ **：交换关联泛函，DFT的关键，只能近似**

Luo et al., *Science* 381, 1072–1079 (2023)

**Fig. 4. Five proposed pathways for the oxidative addition of haloacetonitriles to ionic or neutral Cu(I) complexes and free energies of each species computed with DFT.** The calculated activation free energies for the oxidative addition of $BrCH_2CN$(**2-Br**) to the ionic Cu(I) complex $[Ph_4P]^+[Cu(CF_3)_2]^-$ (**1a**) in DMSO are given in blue, and those for oxidative addition of $ClCH_2CN$ (**2-Cl**) to the neutral [(bpy)Cu($CF_3$)] (**1b**) in DMF are given in red. The energies are in kilocalories per mole and indicate the relative free energies calculated at the PBE0-D3(BJ)/Def2-TZVP(SMD, solvent)//PBE0-D3(BJ)/Def2-SVP(SMD,solvent) level. SMD, solvation model density.

# Input for Deep Learning

## (B) Density Functional Theory calculation

**The second principle of Kohn-Hohenberg:**

**使整体能量最小的电荷密度就是真实电荷密度**

$$n(r) = 2\times \sum_i \Psi_i^*(r)\Psi_i(r) \quad \text{Eq (4)}$$


**自洽求解：**

- **S1: 猜一个电荷数密度$n_{trial}(r)$，代入KS方程与Eq.2，求得波函数$\Psi_i(r)$;**
- **S2: 将$\Psi_i(r)$代入Eq.4，求得$n_{sol}(r)$;**
- **S3: 比较 $n_{trial}(r)$ 和 $n_{sol}(r)$，若误差小于给定精度，输出结果，否则用$n_{sol}(r)$取代$n_{trial}(r)$，循环往复。**

**Calculation results in ML input:**

**1). ΔG in transition states**

**2). Atomic partial charges**

**Regioselectivity, Diasteroselectivity**

# Principal component analysis (PCA)

**Lipinski´s Rule of Five:**

- **MW < 500**
- **HBD < 5**
- **HBA < 10**
- **logP < 5**

Adv Drug Deliver Rev, 1997, 23(1-3), 3
J. Med. Chem. 2002, 45, 12, 2615
J. Med. Chem. 2001, 44, 12, 1841
J. Med. Chem. 2009, 52, 21, 6752

**Veber´s Rule:**

- **Rotable bonds < 10**
- **PSA < 140**

**Muegge´s Rule:**

- **Num Rings < 7**

**Fsp3:**

- **0.3 < Fsp3 < 0.5**