分子のグラフ表現と機械学習の最近

第33回 理研AIP Open Seminar
2021年7月14日

瀧川 一学
ichigaku.takigawa@riken.jp

理化学研究所 革新知能統合研究センター
iPS細胞連携医学的リスク回避チーム

RIKEN AIP ICReDD Institute for Chemical Reaction Design and Discovery HOKKAIDO UNIVERSITY

たきがわ　いちがく
瀧川 一学

https://itakigawa.github.io/

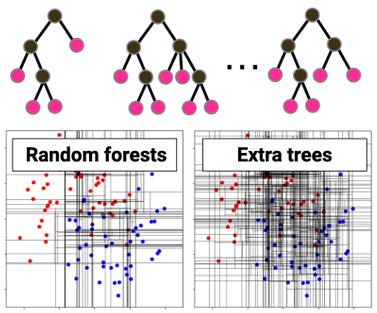
機械学習＋幹細胞生物学

理化学研究所 革新知能統合研究センター

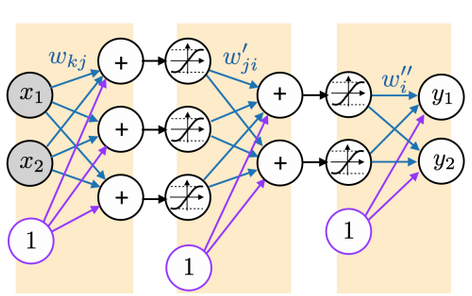
機械学習＋化学

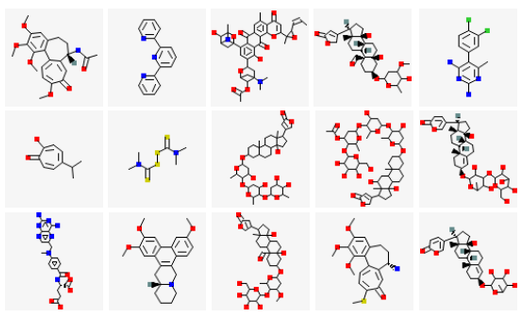
北海道大学 化学反応創成研究拠点

研究の関心： 離散構造・組合せ構造を伴う機械学習 ⟷ 機械学習＋自然科学

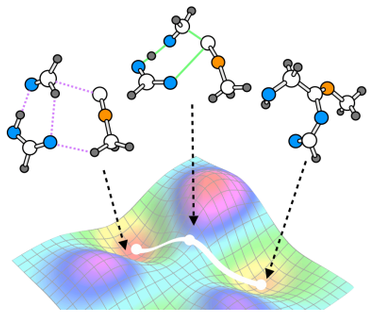


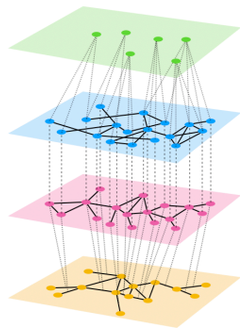
モデルが離散構造

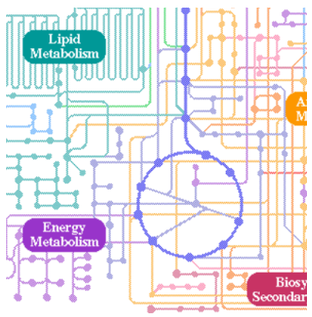




対象が離散構造







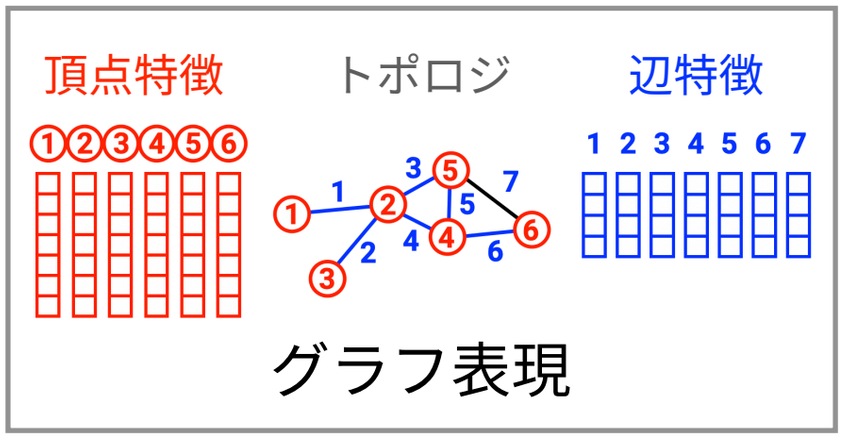
対象間の関係が離散構造

- 科学研究費補助金
  - 基盤研究(C) グラフ表現学習の転移性・構成性の獲得とその実践 (瀧川 一学)
  - 基盤研究(A) 離散構造処理系に基づく列挙と最適化の統合的技法の研究 (湊 真一)
  - 学術変革領域(A) 新しい概念に基づいたアルゴリズム・最適化の問題創出とその効率的求解方法の研究 (宇野毅明)
  - 基盤研究(S) 非平衡過程の実空間観察手法の転換：TEMによる溶液からの核生成過程の解明 (木村勇気)
  - 挑戦的研究(開拓) 化学における外挿探索を可能とする機械学習手法の開発と実証 (鳥屋尾 隆)
- 世界トップレベル研究拠点(WPI)プログラム 化学反応創成研究拠点 (前田 理)
- JST-CREST 学習/数理モデルに基づく時空間展開型アーキテクチャの創出と応用 (本村真人)
- JST-CREST 触媒インフォマティクスの創成のための実験・理論・データ科学研究 (清水研一)

本日の話題：分子のグラフ表現＋機械学習

Q. 分子のどんな情報をどんな表現で機械学習へ入力すれば良いのか？

分子の環境/条件/標的/相互作用等の情報



Representation Learning

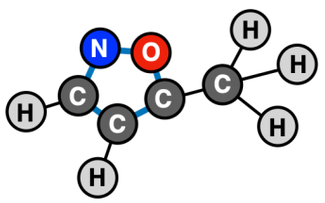
特徴ベクトル

Task-Specific Head

様々な下流タスク

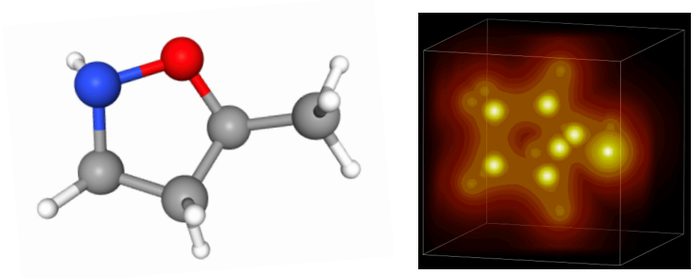
- 分類
- 回帰
- 生成

化学構造/骨格/官能基

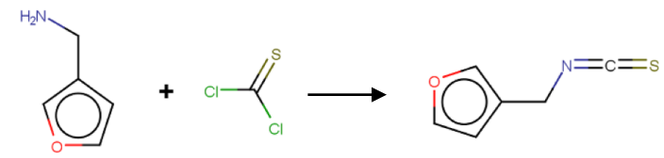


CC1CCNO1

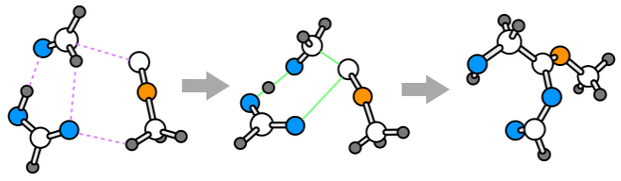
立体配座/電子状態



化学反応=構造の組替え



NCc1ccoc1.S=(Cl)Cl>>[RX_5]S=C=NCc1ccoc1



なぜグラフか？ 分子表現の組合せ論的側面 1/3

- 有機化合物は限られた元素(主にC/H/O/N/S/P/ハロゲン)の**規則的な組合せ**
  - “グラフ”という用語は数学者Sylvesterの化学構造の代数的列挙で初登場 (初登場以来の長年の因縁)

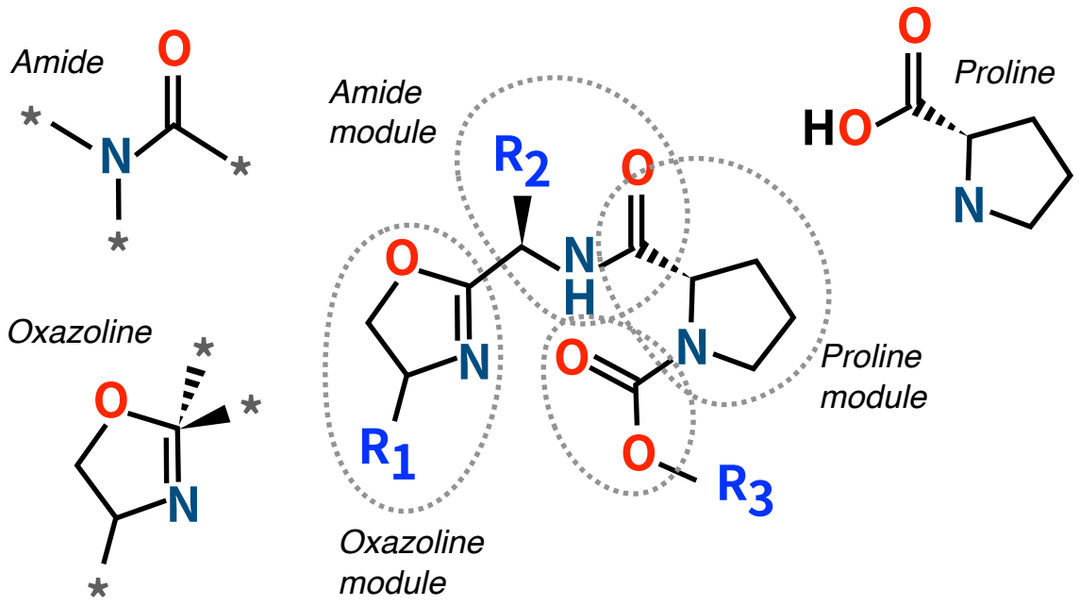
    J.J.Sylvester, Chemistry and Algebra, *Nature*, 17:284 (1878).
  - ReymondグループのChemical Spaceの列挙研究

    GDB-11 (Fink+, *JCIM* 2007) C,N,O,F,による11原子以下の分子の全列挙 (2640万)

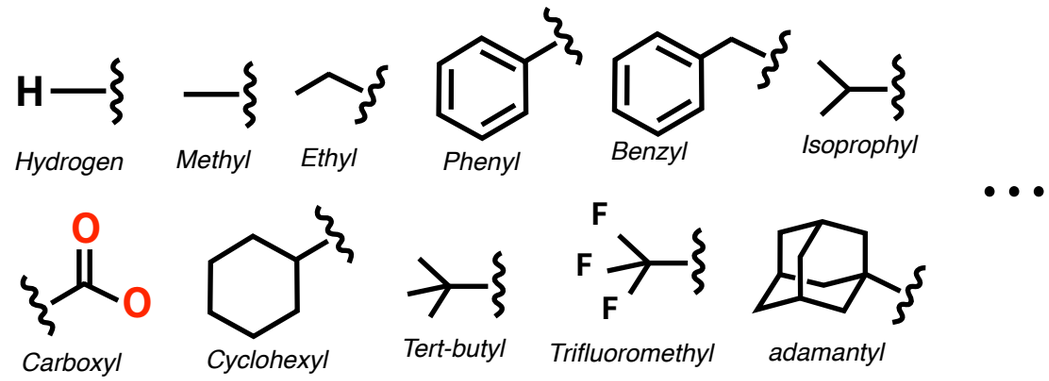
    GDB-13 (Blum+, *JACS* 2009) C,N,O,S,Clによる13原子以下の分子の全列挙 (9億7700万)

    GDB-17 (Ruddigkeit+, *JCIM* 2012) C,N,O,S,ハロゲンによる17原子以下の分子の全列挙 (1,664億)

- 明確な**構成性/モジュール性** (言語-文法的？= 要素の組合せで複雑なものを生成)



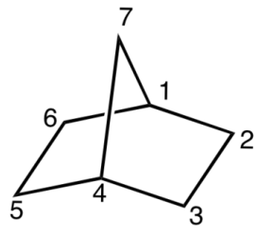
R_1 R_2 R_3 には様々な**置換基**が入れられる

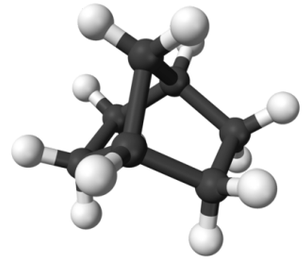


なぜグラフか？ 分子表現の組合せ論的側面 2/3

- 部分構造検索・部分一致マッチングのニーズ

https://pubchem.ncbi.nlm.nih.gov/#query=C12CC(C(CC1)C2)C





Norbornane (ノルボルナン)

PubChem About Blog Submit Contact

SEARCH FOR

C12CC(C(CC1)C2)C

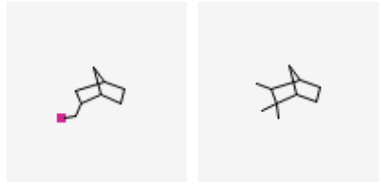
Treating this as a structure search for a SMILES identifier. Switch to **SMARTS**. Edit Structure

Identity (1) | Similarity (710) | Substructure (>1,000) | Superstructure (886) | 3D Similarity (>827) | Settings

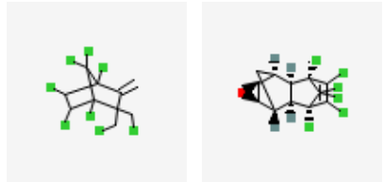
Find structures very closely related to the input, comparing chemical connectivity, and optionally stereoisomers and isotopes.

厳密一致検索
クエリ構造と同じ

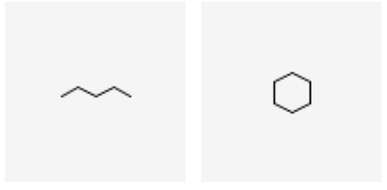
類似性検索
クエリと類似



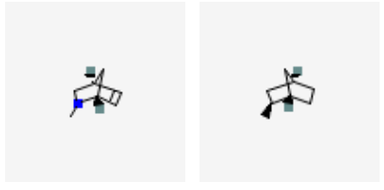
部分構造検索
クエリ構造を含む



上部構造検索
クエリ構造に含まれる

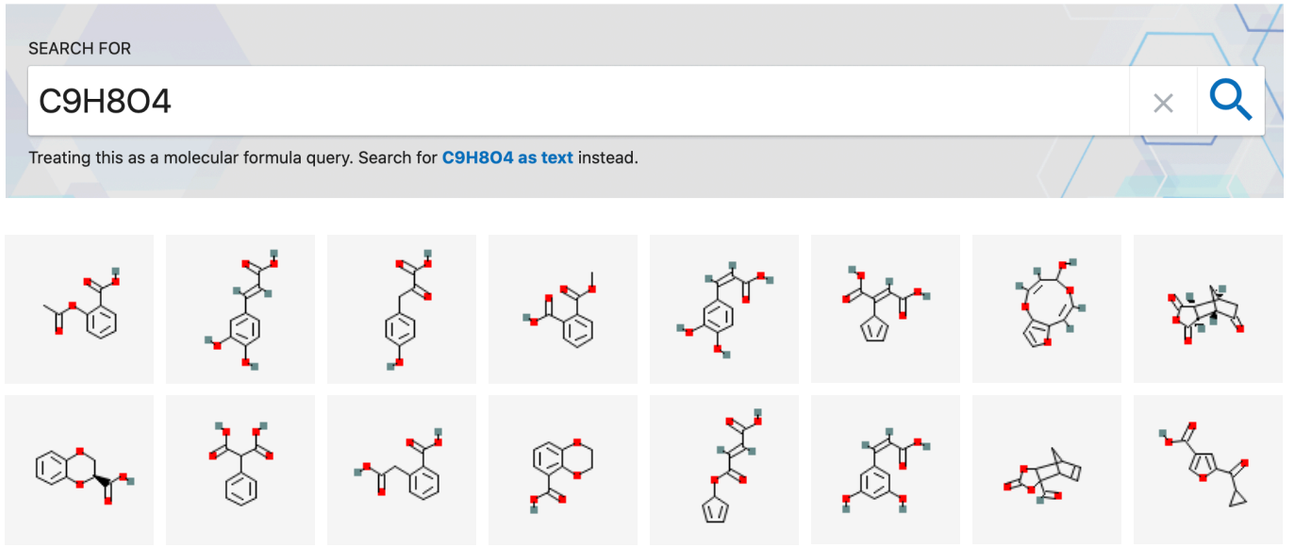


立体類似性検索
クエリ構造と立体的に類似

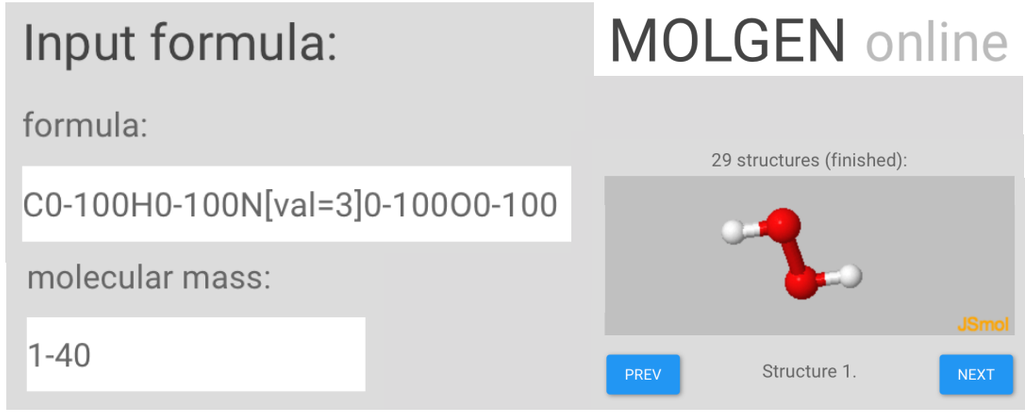


なぜグラフか？ 分子表現の組合せ論的側面 3/3

- 異性体とルールベースの分子生成

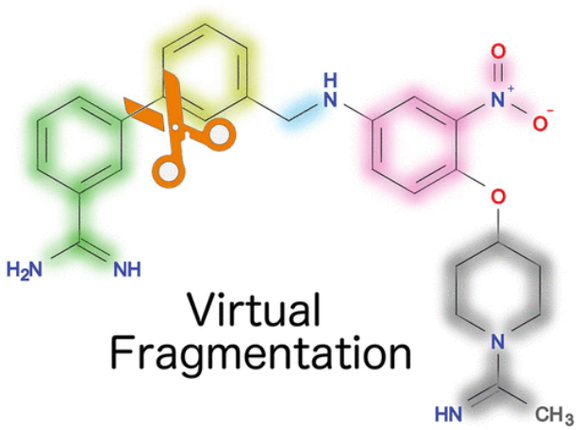


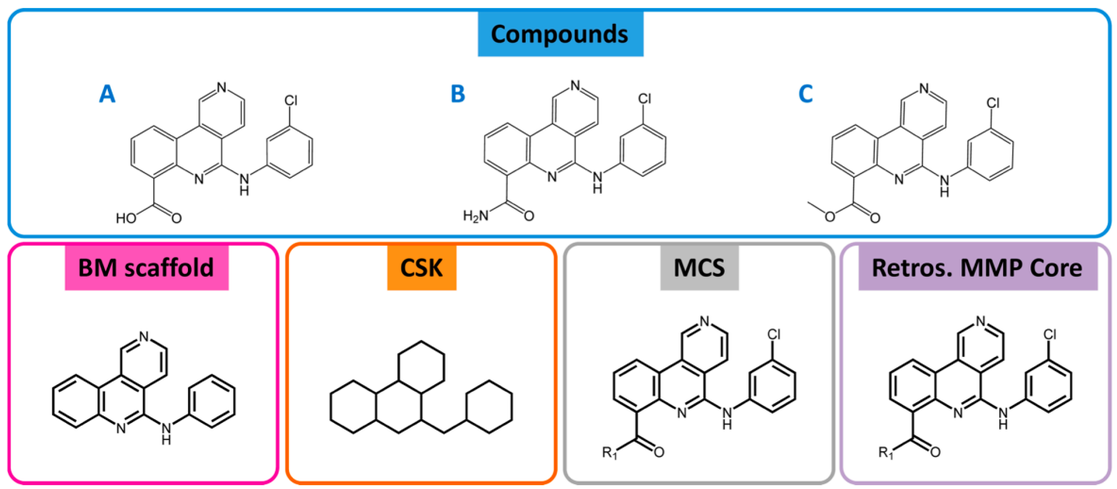
https://www.molgen.de/online.html

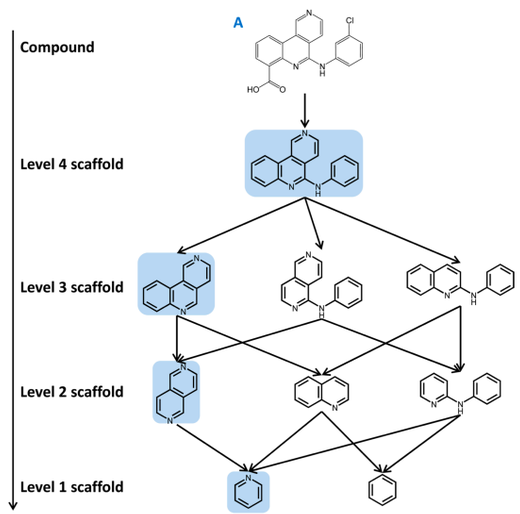


EXAMPLE 5: Generate all (theoretically possible) structures of mass ≤ 40 with elements, C, H, N^3, O

- 分子のコアやパーツの類似性：分子骨格(Scaffold)とフラグメント





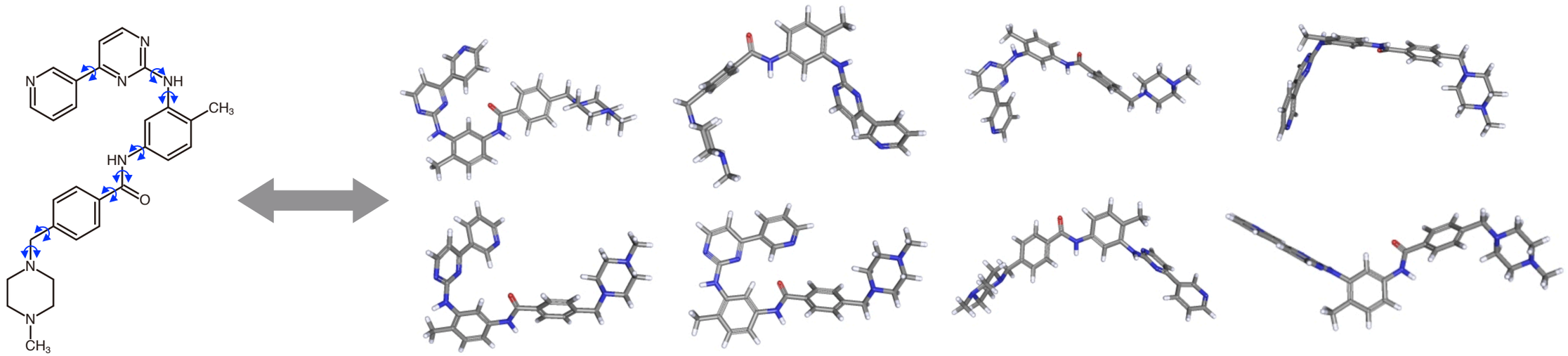


Liu, Naderi, Alvin, Mukhopadhyay, Brylinski, *JCIM* (2017) https://doi.org/10.1021/acs.jcim.6b00596

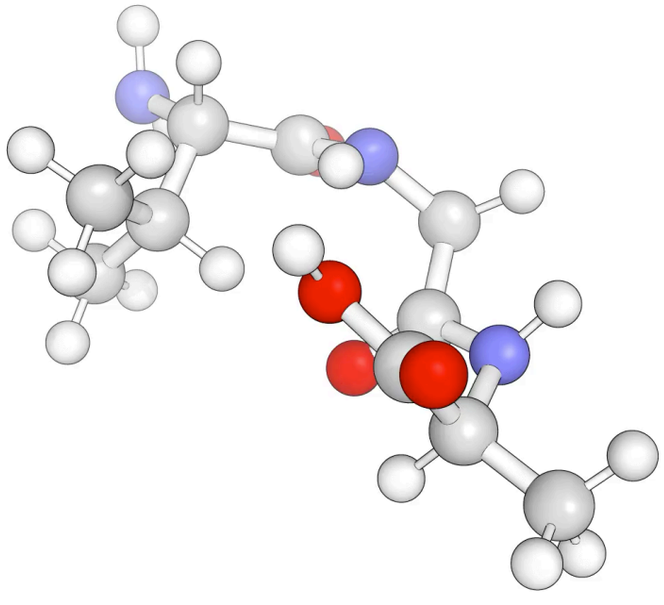
Hu, Stumpfe, Bajorath, *J Med Chem* (2016) https://doi.org/10.1021/acs.jmedchem.5b01746

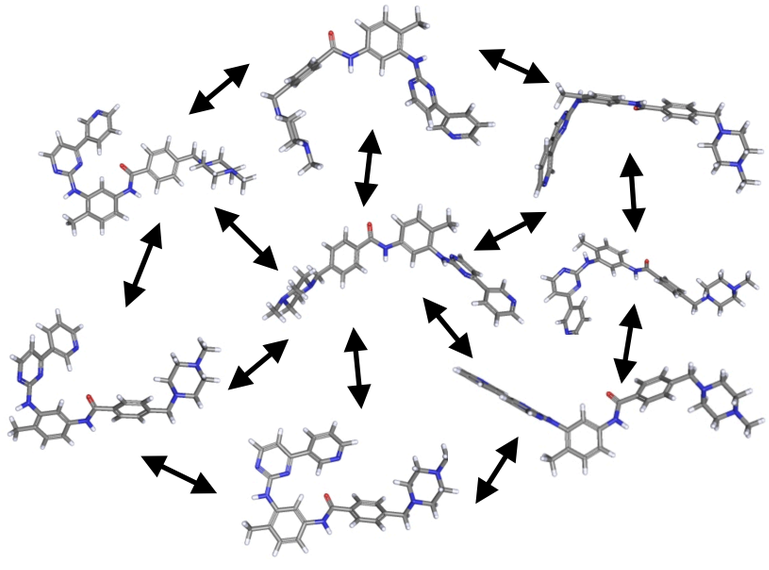
Reality Bites！ 明らかな非組合せ論的側面 1/2

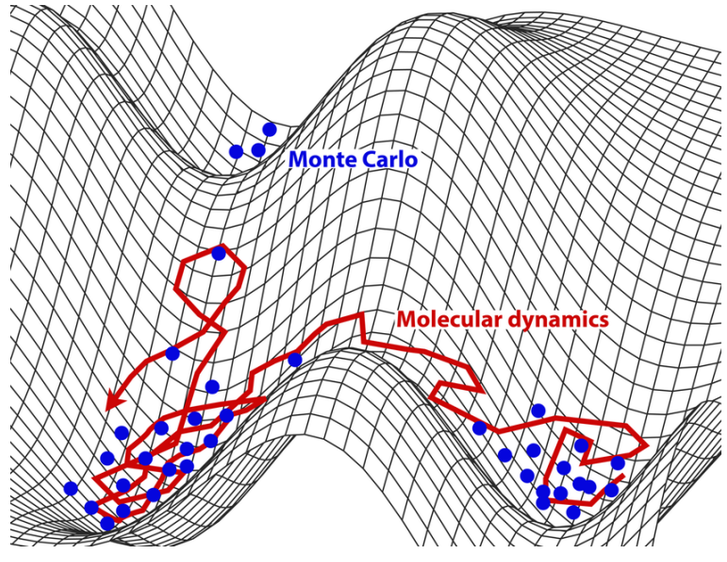
- 実際は3次元的な形を取る + 単結合の自由な回転で立体配座に多様性がある



- 構造は剛体的ではなく熱力学的なダイナミクスで変動・振動



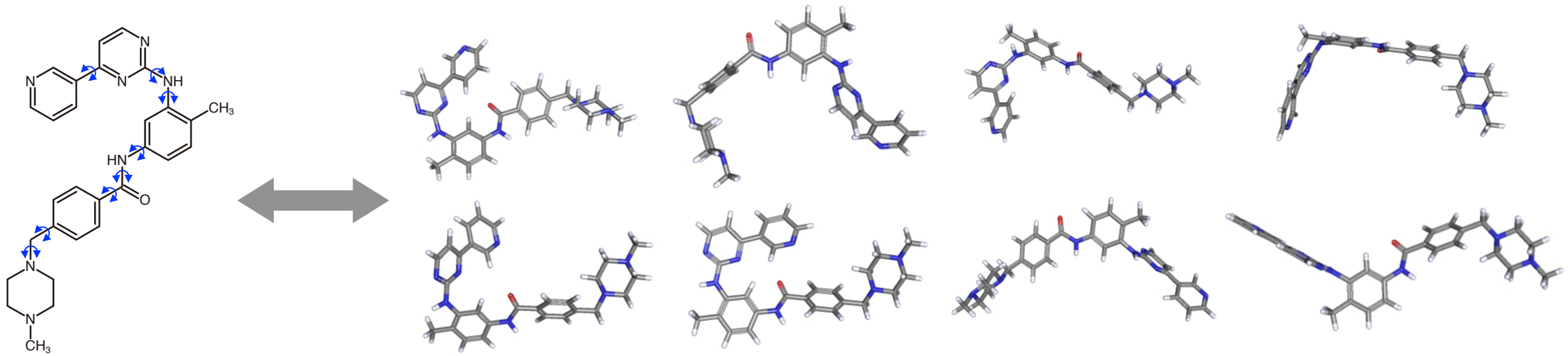




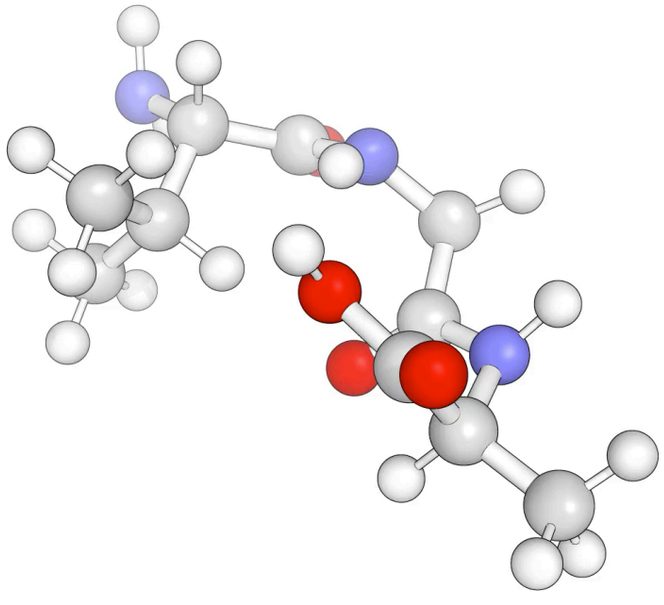
https://en.wikipedia.org/wiki/Molecular_dynamics

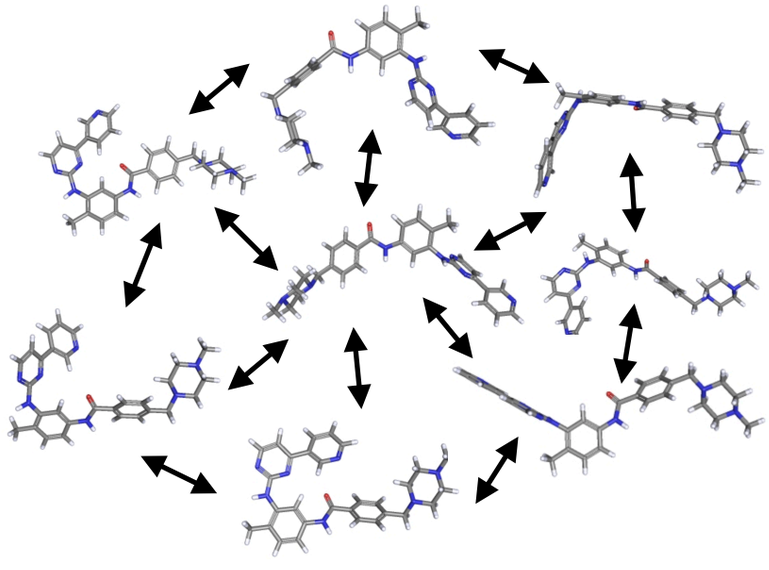
Reality Bites！ 明らかな非組合せ論的側面 1/2

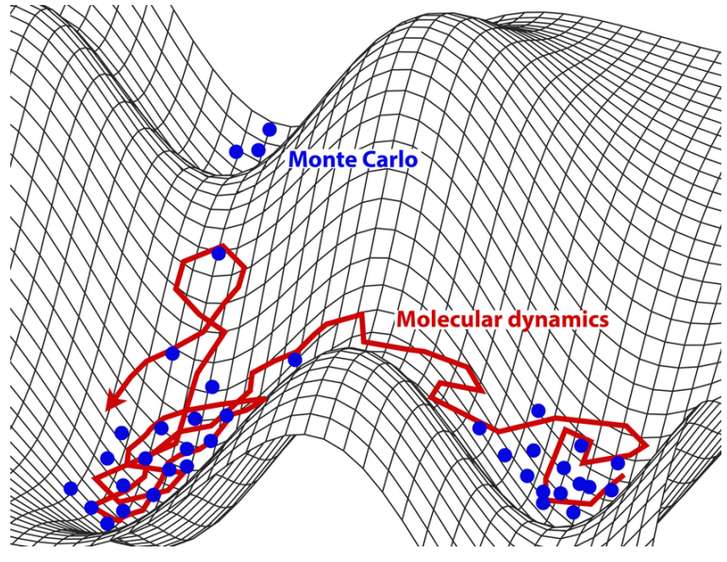
- 実際は3次元的な形を取る + 単結合の自由な回転で立体配座に多様性がある



- 構造は剛体的ではなく熱力学的なダイナミクスで変動・振動



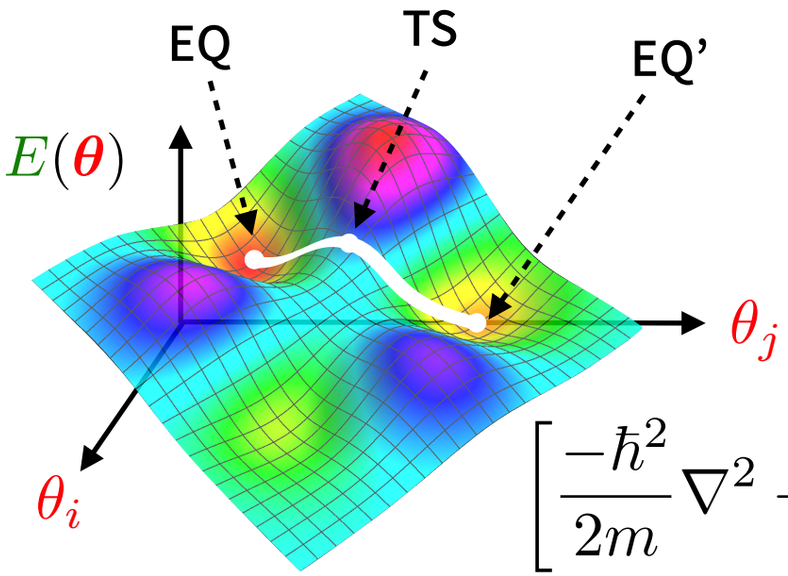


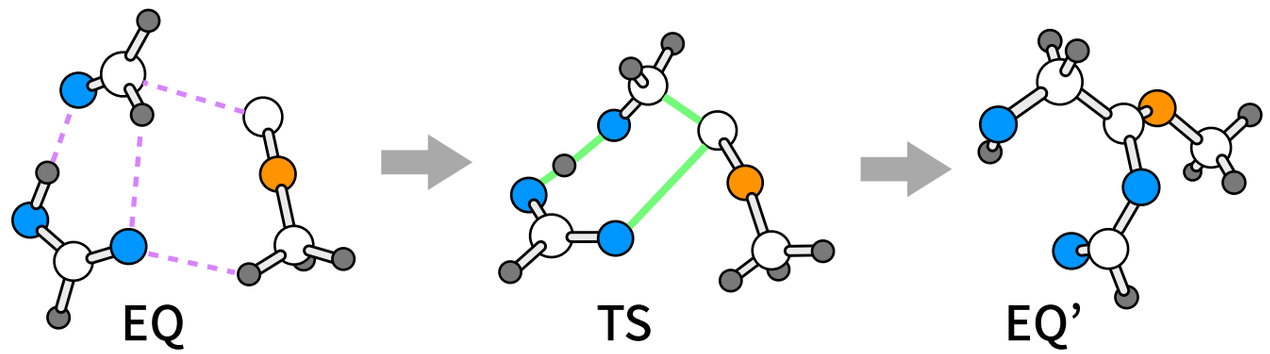


https://en.wikipedia.org/wiki/Molecular_dynamics

Reality Bites！ 明らかな非組合せ論的側面 2/2

- 第一原理(Schrödinger方程式)と量子化学的な電子状態

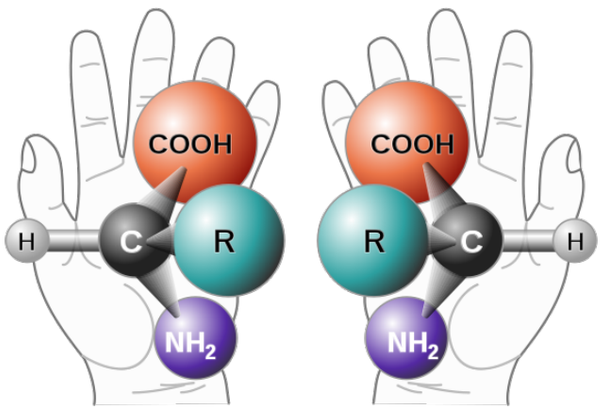




$$\left[\frac{-\hbar^2}{2m}\nabla^2 + V(\boldsymbol{\theta})\right]\Psi = E(\boldsymbol{\theta})\,\Psi$$

- グラフパターンの統計分析では掴みにくい対象・現象多数

不斉合成/エナンチオマー(鏡像異性体)

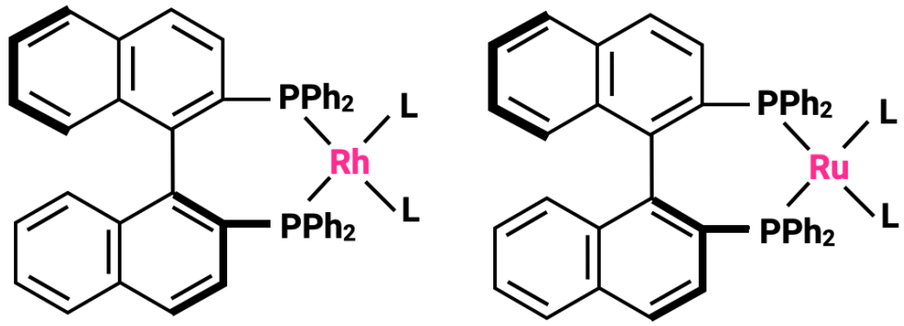


原子距離行列・物質量・エネルギーおよび物理的性質は**全く同じ**だが他の分子と相互作用するとき性質が違いうる
(生体内分子や有機化学の主対象)

https://en.wikipedia.org/wiki/Chirality_(chemistry)

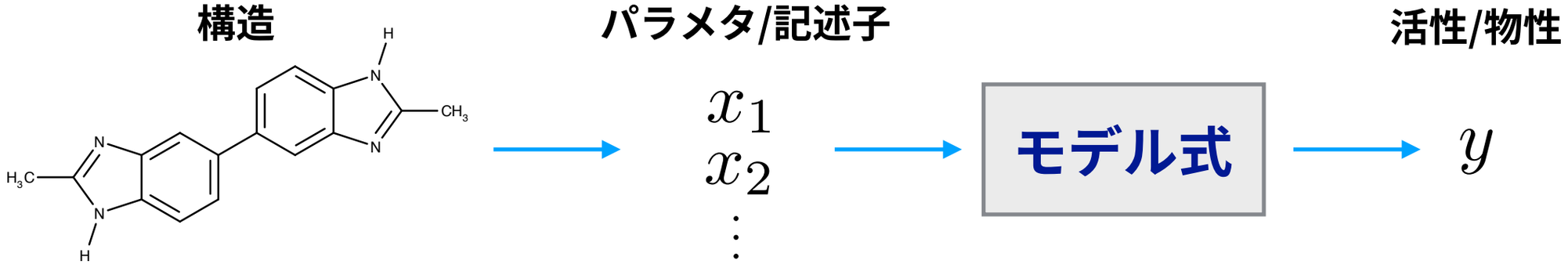
BINAP錯体

錯体など金属の原子が入る場合これは部分構造云々の問題ではなく元素特性の考慮が必要…



機械学習に求められること：経験則の合理化・精緻化？

例：構造活性相関(SAR) 「形/構造」から生物活性や物性を予測する



Hansch-Fujita QSAR (Hammet則の生物学版)

$$\log(1/C) = \beta_1(\log P)^2 + \beta_2 \log P + \beta_3\sigma + \beta_4 E_s + \text{Const.}$$

生物活性や濃度

疎水性の度合い (分配係数)

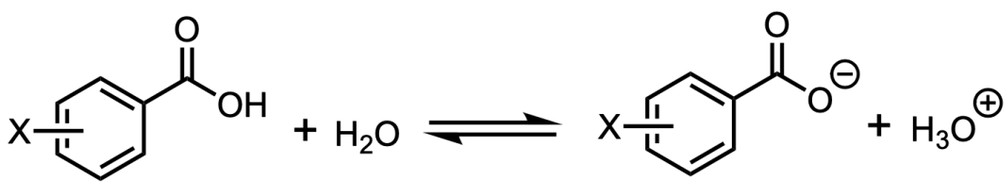
電子を出すか取るか度合い (Hammett定数)

立体効果の度合い (Taft定数)

Hammett則

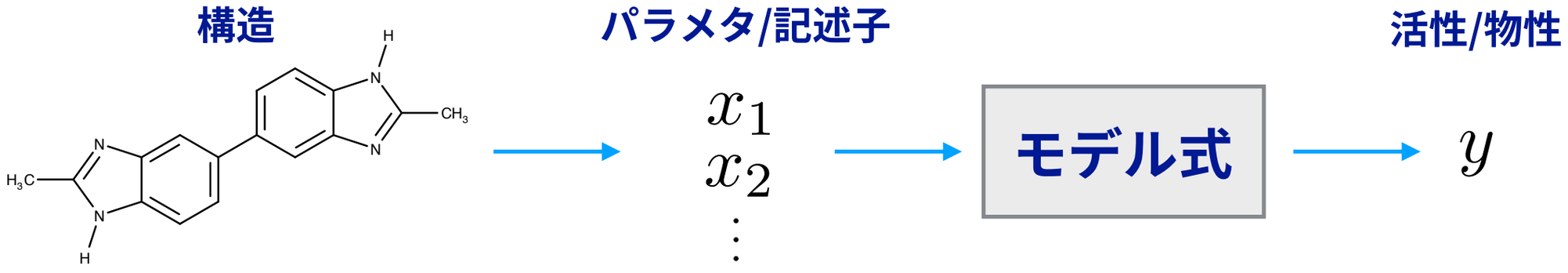
$$\log(K/K_H) = \rho\sigma$$

“Linear Free Energy Relationships (LFERs)”



X=H (安息香酸)のときとXに何か別の置換基を入れた時の反応速度比が一次式になる経験則

注意すべきこと (Open Problem)

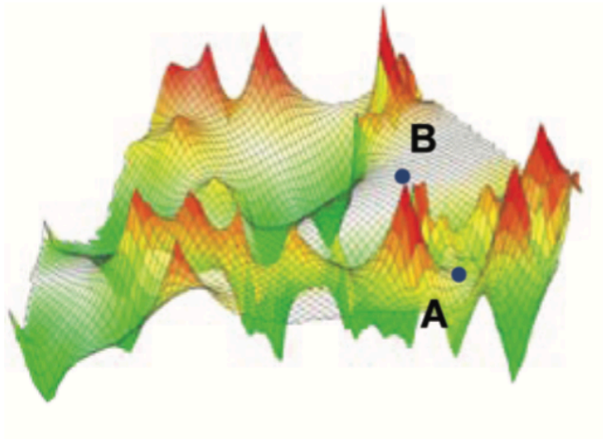


- 「活性 ≈ *f*(記述子)」としてケモインフォマティクスの主問題の一つに
- 現在では分子記述子を数千種類計算してくれるソフトウェアなどもある
- **応用統計学の基本のキ：何でもかんでも説明変数にいれてはいけない**
  - 係数を解釈したい(疎水性/電子効果/立体効果どれが効いたか?)が、相関が強い変数が複数あると(機械学習的にはその中のどれか一つでほぼ良いので)どれが効いたと言えなくなる、いわゆる**「マルチコ(多重共線形性)」**問題
  - **キッチンシンク(理論的根拠なしに何でもかんでも説明変数に加える)回帰**をするとoverfitするだけ！！サンプル数より変数が多いなんてアホなの！？

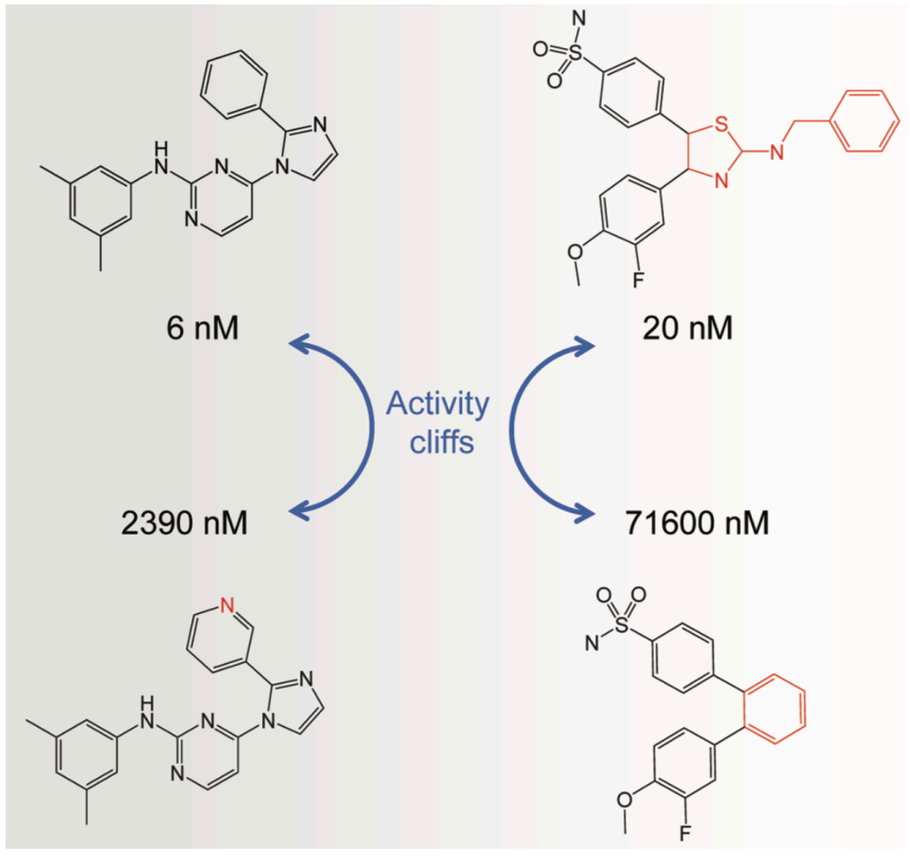
(最近の)機械学習の考え方：正則化やデータ設計(training/validation/test)でoverfitやleakageのリスクが適切に回避されていればOKなのでは？

構造が似ていても活性/選択性が似ていない例も多い…

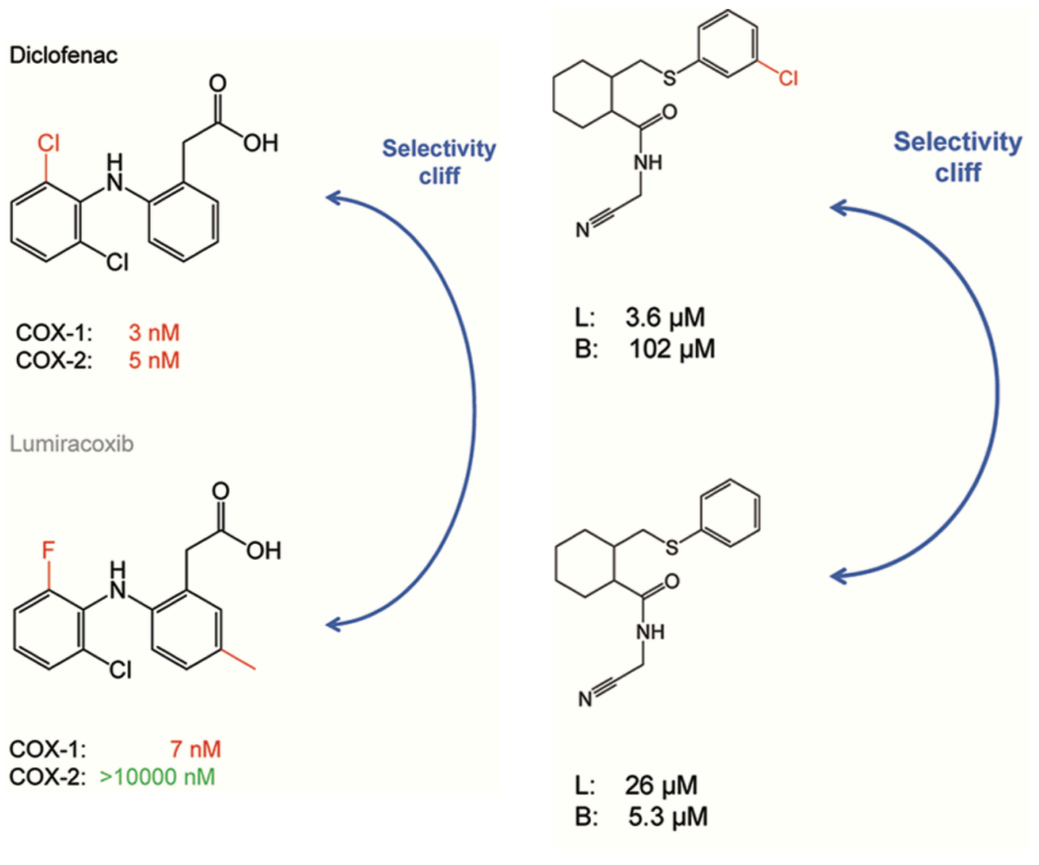
実際には構造がほんの少し変わると活性や選択性が大きく変わってしまう例が多く存在する



Activity cliffs

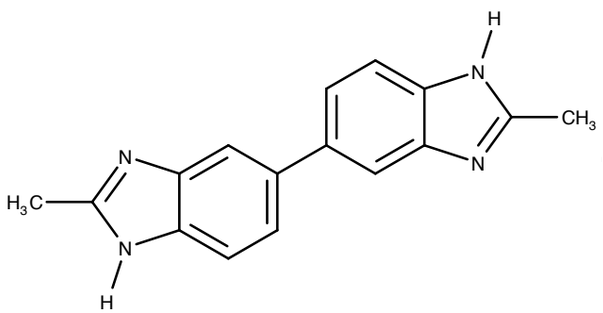


Selectivity cliffs



Stumpfe and Bajorath, *J Med Chem* (2012) https://doi.org/10.1021/jm201706b

機械学習からはこう見えがち…

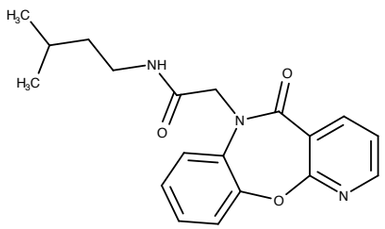


コンピュータプログラム

50%阻害濃度

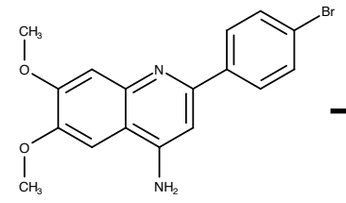
0.739

見本例からのプログラム生成

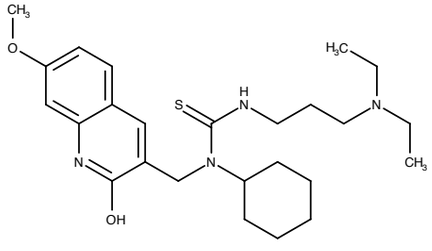


1.399

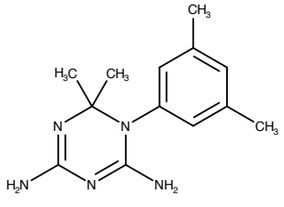
入出力の
見本例



-1.753



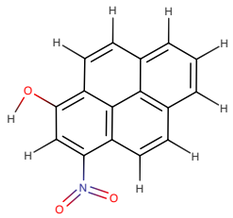
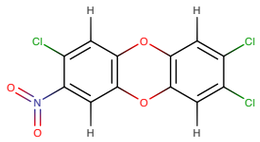
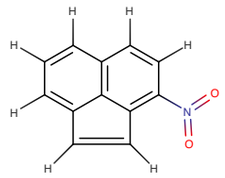
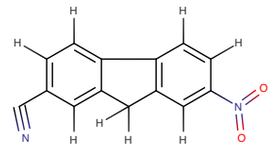
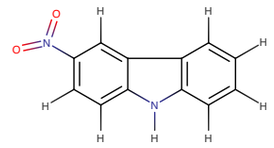
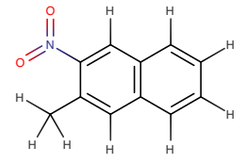
1.394



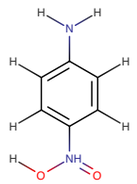
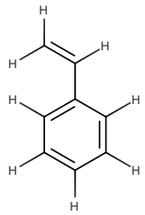
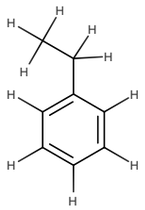
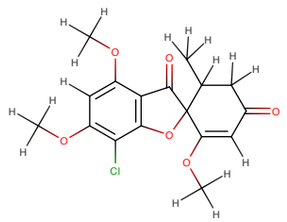
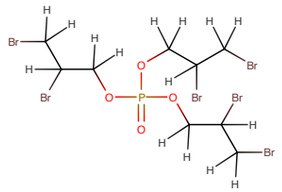
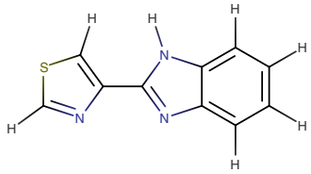
-3.171

実際のデータセットの様子

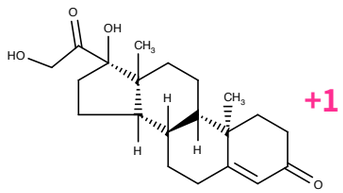
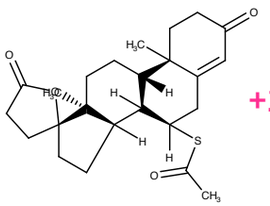
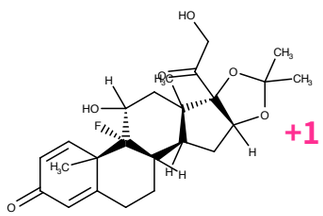
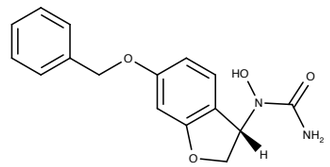
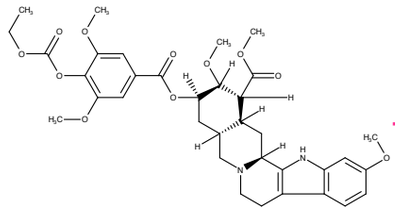
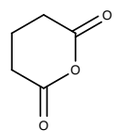
- **Mutagenic potency**


+1

-1

+1

+1

-1

-1

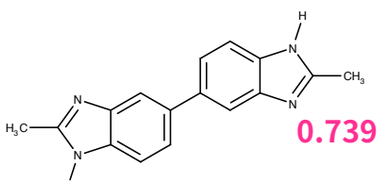
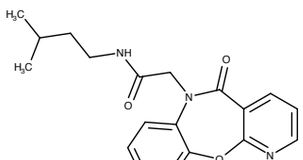
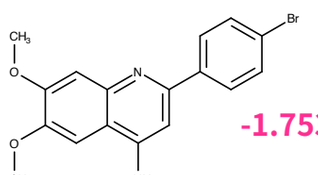
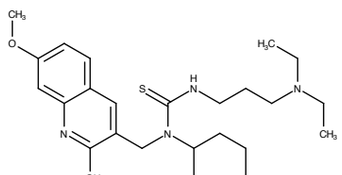
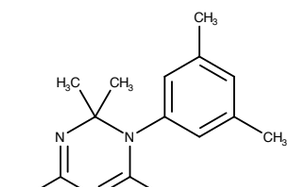
- **Carcinogenic potency**


+1

+1

-1

-1

+1

+1

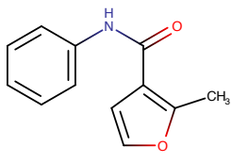
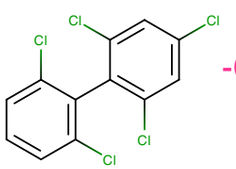
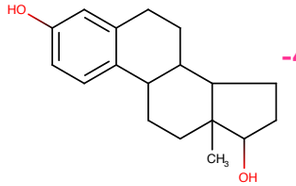
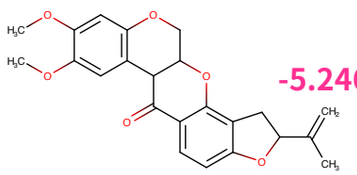
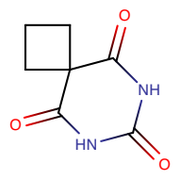
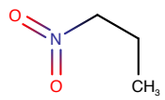
- **Endocrine disruption**


+1

+1

+1

-1

-1

-1

- **Growth inhibition**


0.739

1.399

-1.753

1.394

-3.171

- **Aqueous solubility**


-2.885

-6.545

-4.138

-5.246

-0.527

-0.816

現在までの機械学習の歩みを駆け足で振り返る！

離散ラベルつきグラフ表現：部分構造の有無や数からの予測

事前に定義された部分構造特徴

- MDL MACCS Keys
- PubChem Fingerprint

グラフカーネル＋カーネル法

- Kashima Kernel
- Weisfeiler-Lehman Kernel

データに生起する部分構造特徴

- ECFP (Circular Fingerprint)
- 頻出部分グラフ構造

部分グラフ探索＋モデル学習

- 部分グラフ列挙木上での線形学習・決定木学習
- gLARS/gBoost

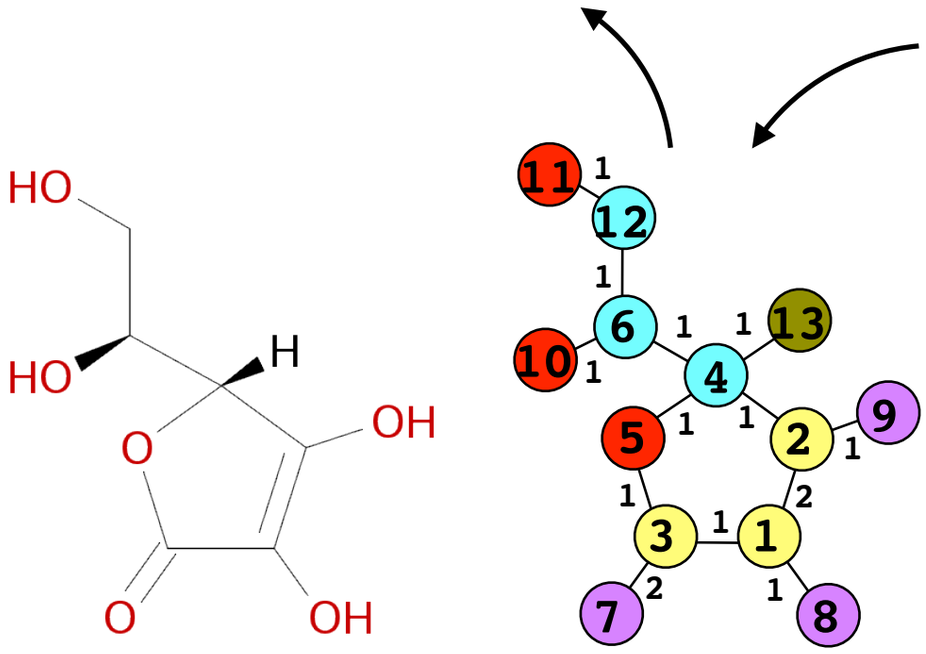
離散ラベルを超えて：表現学習とグラフニューラルネットワーク(GNN)

- GNNとWeisfeiler-Lehmanグラフ同型検査(WL-1)
- ECFPとNeural Fingerprint
- MPNNとD-MPNN/ChemProp
- GATとTransformer型GNN
- 分子表現の事前学習と転移学習
- 分子表現の生成
- 幾何的GNNと量子化学計算の高精度高速近似

離散ラベルつきグラフとしての表現

Canonical SMILES

OC[C@H](O)[C@H]1OC(=O)C(=C1O)O



Standard InChI Key

CIWBSHSKHKDKBQ-JLAZNSOCSA-N

Standard InChI

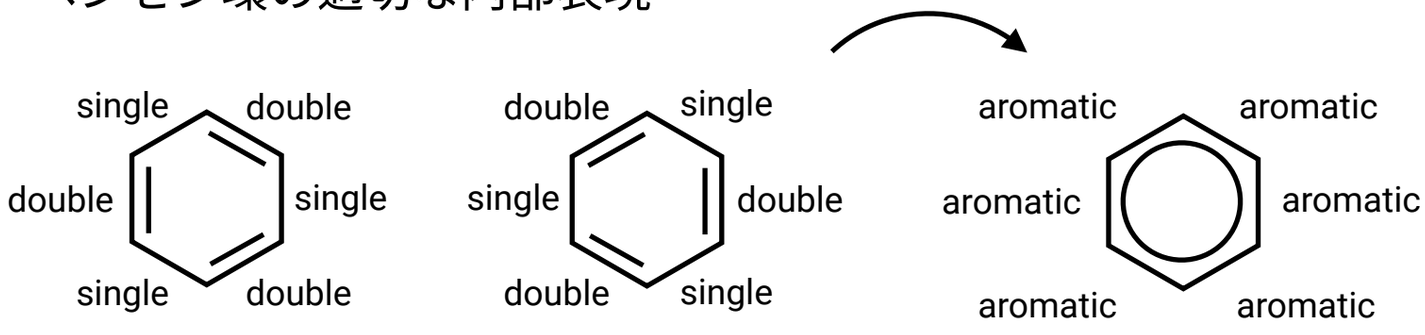
InChI=1S/C6H8O6/
c7-1-2(8)5-3(9)4(10)6(11)12-5/
h2,5,7-10H,1H2/t2-,5+/m0/s1

MOL2 Format

```
=@<TRIPOS>MOLECULE
*****
 13 13 0 0 0
SMALL
GASTEIGER

@<TRIPOS>ATOM
      1 C          -2.5458   -9.4750    0.0000 C.2     1  UNL1         0.3080
      2 C          -3.3708   -9.4750    0.0000 C.2     1  UNL1         0.2529
      3 C          -2.2875   -8.6917    0.0000 C.2     1  UNL1         0.3838
      4 C          -3.6208   -8.6917    0.0000 C.3     1  UNL1         0.2067
      5 O          -2.9583   -8.2042    0.0000 O.3     1  UNL1        -0.4441
      6 C          -4.3583   -8.3125    0.0000 C.3     1  UNL1         0.2245
      7 O          -1.5000   -8.4375    0.0000 O.2     1  UNL1        -0.2412
      8 O          -2.0583  -10.1417    0.0000 O.2     1  UNL1        -0.2764
      9 O          -3.8500  -10.1417    0.0000 O.2     1  UNL1        -0.2843
     10 O          -5.0500   -8.7542    0.0000 O.3     1  UNL1        -0.2164
     11 O          -3.6958   -7.0417    0.0000 O.3     1  UNL1        -0.2174
     12 C          -4.3958   -7.4875    0.0000 C.3     1  UNL1         0.2185
     13 H          -4.2083   -9.2667    0.0000 H       1  UNL1         0.0853
@<TRIPOS>BOND
     1     2     1     2
     2     3     1     1
     3     4     2     1
     4     5     3     1
     5     6     4     1
     6     7     3     2
     7     8     1     1
     8     9     2     1
     9     6    10     1
    10    11    12     1
    11    12     6     1
    12     4    13     1
    13     5     4     1
```

ベンゼン環の適切な内部表現

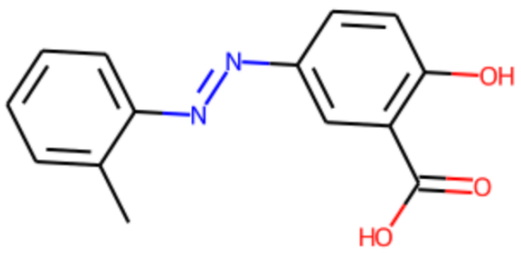


Kekulé Form

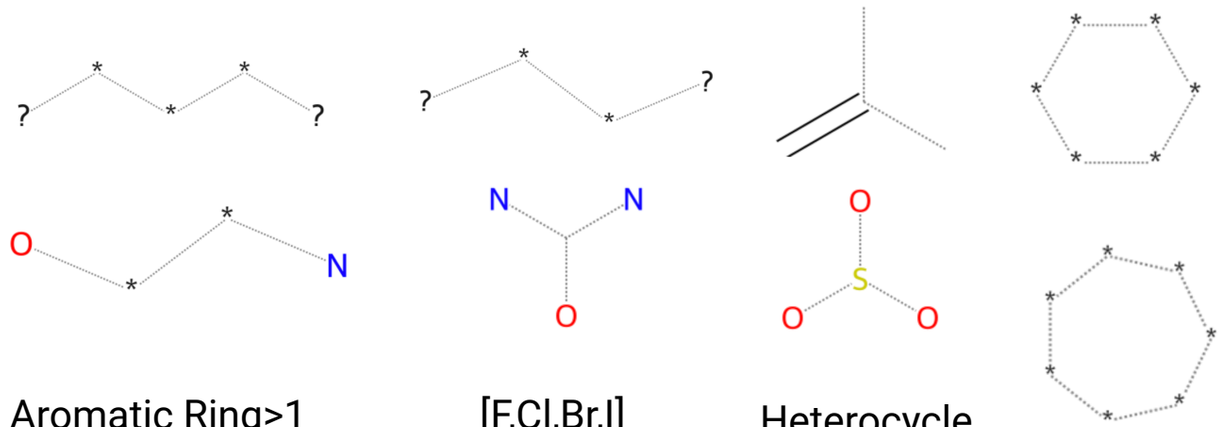
Fingerprint(分子指紋)と部分構造特徴

MDL MACCS Keys　事前に定義された166個の構造特徴の有無

より複雑な構造特徴パターンを網羅したPubChem Fingerprint (881bit)なども



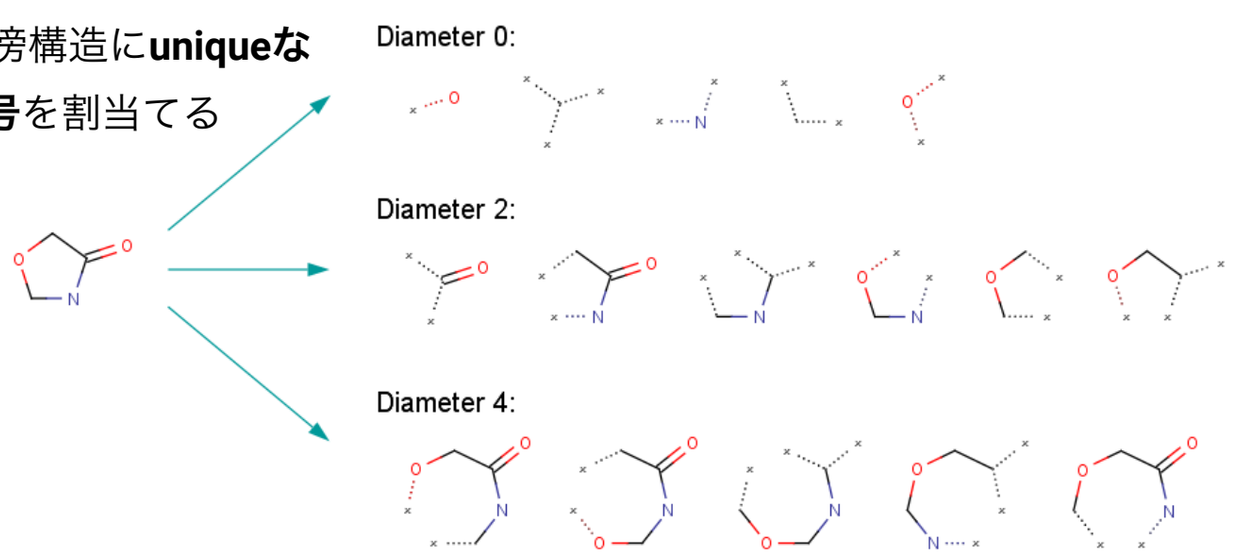
各構造特徴の有無を
各bit(0/1)に




```
00000000000000000000000000000000000000000000000001100
00000000000000000000000000000010000100000000000000000
01000001000111010001010100010011111000101010111111110
```

ECFP (Extended Connectivity Fingerprint)　半径rの頂点近傍構造特徴を効率的に全列挙

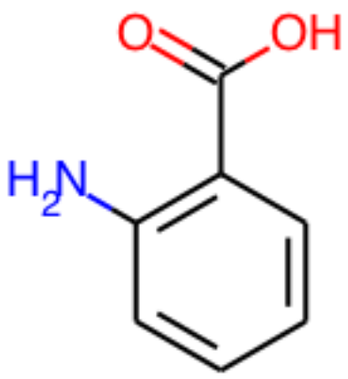
各近傍構造に**unique**な
id番号を割当てる



探索半径の上限とビット長はパラメタ

https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md

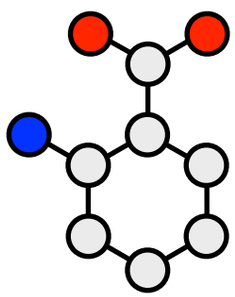
ECFPで使われる分子グラフ表現



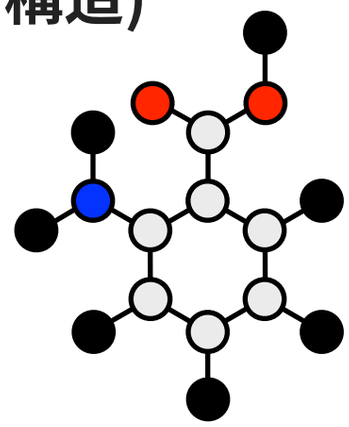
Structural Formula

分子グラフへ
エンコーディング

トポロジ (グラフ構造)



Implicit Hydrogens

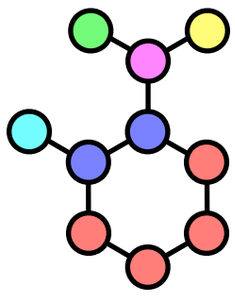


Explicit Hydrogens

各頂点(や各辺や各頂点対)に**「不変量(原子ごとに固有の多変量)」**を付与しそれを**離散ラベル**として扱う

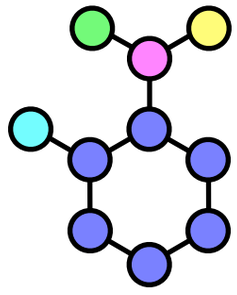
Atomic Invariants

ECFP ≈ RDkitの原子不変量1 (useFeatures=False)



atomic number	8	8	7	6	6	6
total degree	2	1	3	3	3	3
#Hs	1	0	2	0	0	1
formal charge	0	0	0	0	0	0
isotope	0	0	0	0	0	0
in Ring?	0	0	0	1	0	1

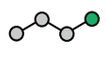
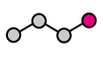
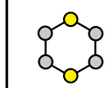
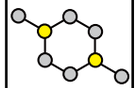
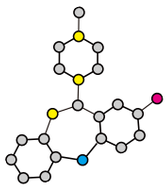
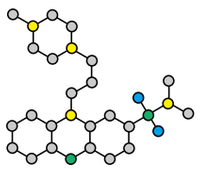
FCFP ≈ RDkitの原子不変量2 (useFeatures=True)



Donor	1	0	1	0	0
Acceptor	0	1	1	0	0
Aromatic	0	0	0	1	0
Halogen	0	0	0	0	0
Basic	0	0	1	0	0
Acidic	0	0	0	0	1

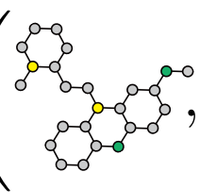
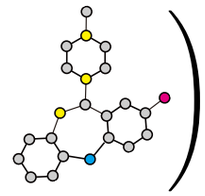
部分構造特徴の有無による特徴ベクトル化

クエリ部分構造として何をもってくるかが異なる
(ワイルドカード文字を許容するかなども含めて)

y	入力構造							...
0.1		0	0	1	1	1	0	...
0.7		1	0	0	0	0	1	...
0.9		1	1	0	1	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
1.2		1	0	1	1	1	0	...

グラフカーネル＋カーネル法

陽に特徴ベクトル化せず、内積値のみを
ダイレクトに計算 (カーネルトリック)し
SVMなど内積のみで計算可能な手法へ帰着

$k\left($  $,$  $\right) =$ 共通する部分グラフの数

部分グラフ探索＋モデル学習

データセットに頻出する部分構造を**陽に列挙**

↓ vs カーネル (常に全変数を考慮してしまう)

すべての生起部分構造の中から候補を探索し
所与タスクに寄与しうるものだけを選択的に
変数として機械学習モデルへ追加していく

グラフカーネル＋カーネル法

Kashima Kernel (Kashima+ 2003)

グラフ上の**すべての**ラベル列上の文字列カーネルの**期待値**として定義される代表的グラフカーネル

$$K(G,G') = \sum_{\mathbf{h}} \sum_{\mathbf{h}'} K_z(\mathbf{h},\mathbf{h}') p(\mathbf{h}|G) p(\mathbf{h}'|G')$$

グラフ上のランダムウォークで生成されるラベル列

Marginalized Kernel (Tsuda+ 2002)

$$K(x,x') = \sum_{h\in\mathcal{H}} \sum_{h'\in\mathcal{H}} p(h|x) p(h'|x') K_z(z,z')$$

観測変数 $x \in \mathcal{X}$
隠れ変数 $h \in \mathcal{H}$
$+\ z = (x,h), z' = (x',h')$

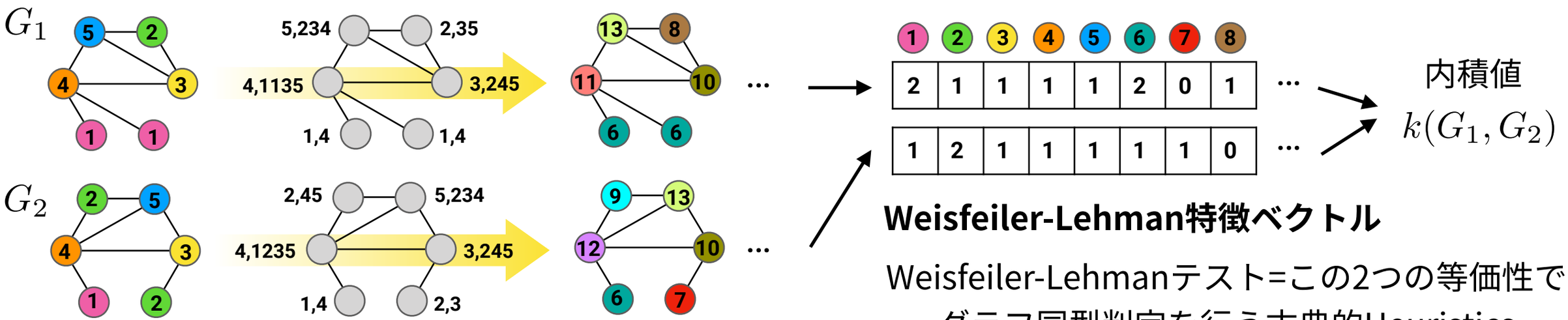
R-Convolution Kernel (Haussler 1999)

$$K(x,y) = \sum_{\vec{x}\in R^{-1}(x), \vec{y}\in R^{-1}(y)} \prod_{d=1}^{D} K_d(x_d, y_d)$$

関係*R*で*D*個の部分へ分解 $\vec{x} = x_1,\ldots,x_D$
$\vec{y} = y_1,\ldots,y_D$

Weisfeiler-Lehman Kernel (Shervashidze+ 2011)

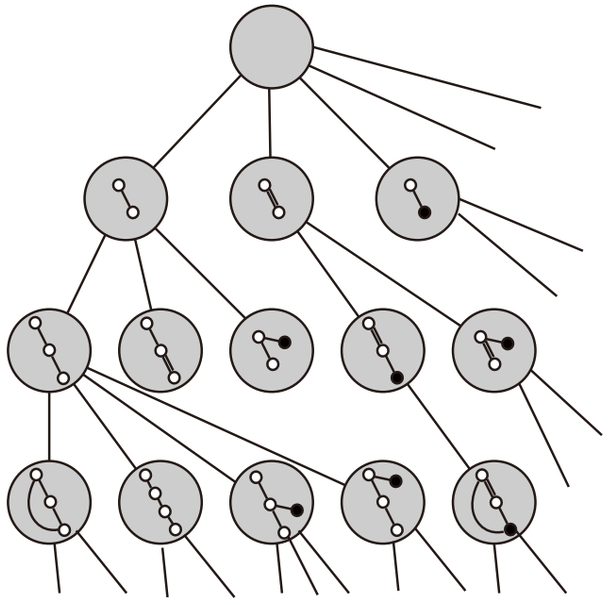
近傍集約＋再ラベリング (変化がなくなるまで繰り返す)



Weisfeiler-Lehman特徴ベクトル

Weisfeiler-Lehmanテスト=この2つの等価性でグラフ同型判定を行う古典的Heuristics

部分グラフ探索＋モデル学習



gSpanの木状の探索空間 (列挙木)

グラフ集合 $\{G_i\}_{i=1}^n$ に生起するすべての部分グラフを探索

例: gSpan (Yan and Han, ICDM 2002)で頻出部分グラフを列挙

Table 19.2. SAR performance of different descriptors.

Datasets	fp	ECFP	MK	FS	GF
NCI1	0.30	0.32	0.29	0.27	**0.33**
NCI109	0.27	**0.32**	0.24	0.26	**0.32**
NCI123	0.25	**0.27**	0.24	0.23	**0.27**
NCI145	0.30	0.35	0.28	0.30	**0.37**
NCI167	0.06	0.06	0.04	0.06	**0.07**
NCI220	**0.33**	0.28	0.26	0.21	0.29
NCI33	0.26	0.31	0.26	0.25	**0.33**
NCI330	0.34	**0.36**	0.31	0.24	**0.36**
NCI41	0.25	**0.36**	0.28	0.30	**0.36**
NCI47	0.26	**0.31**	0.26	0.24	**0.31**
NCI81	0.27	**0.28**	0.25	0.24	**0.28**
NCI83	0.26	**0.31**	0.26	0.25	**0.31**

Wale, Ning, Karypis (2010)

- fp (Hashed fingerprint w/ paths+cycles)
- ECFP
- MK (MACCS Keys)
- FS (頻出部分グラフ)
- GF (一定サイズ以下の全生起部分グラフ)

頻出だけは良くない..

Wale N., Ning X., Karypis G. (2010) Trends in Chemical Graph Data Mining. In: Aggarwal C., Wang H. (eds) Managing and Mining Graph Data. Advances in Database Systems, vol 40. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-6045-0_19

Branch & Boundで必要な部分グラフ特徴を学習しながら同時に機械学習モデルを構築

生起部分グラフから有効なものを逐次的に選択し、同時にモデルを学習

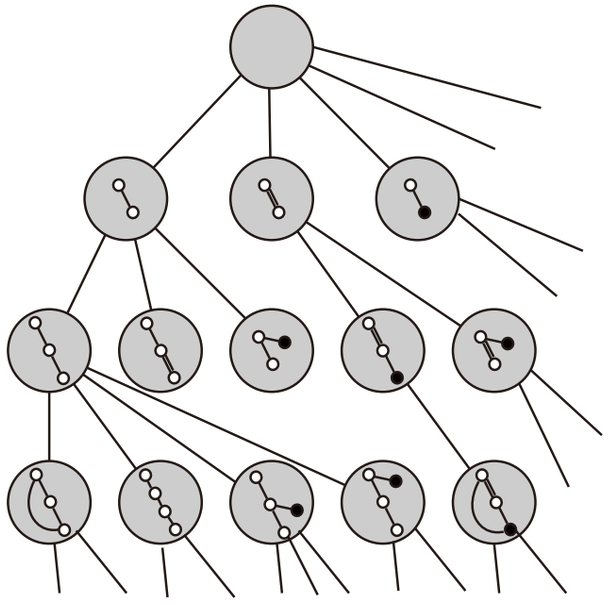
生起部分グラフの列挙木上での線形学習

- Adaboost (Kudo et al, NIPS 2005)
- gLARS (Tsuda, ICML 2007), gPLS (Saigo et al, KDD 2008), gBoost (Saigo et al, Mach Learn 2009)
- Elastic-Net正則化つき一般の線形学習への一般化 (Takigawa and Mamitsuka, TPAMI 2017)

生起部分グラフの列挙木上での非線形学習

- 決定木(分類木・回帰)を学習 ＋ 木アンサンブル学習 (Shirakawa et al, MLG 2018@KDD)

部分グラフ探索＋モデル学習



gSpanの木状の探索空間 (列挙木)

グラフ集合 $\{G_i\}_{i=1}^n$ に生起するすべての部分グラフを探索

例: gSpan (Yan and Han, ICDM 2002)で頻出部分グラフを列挙

Table 19.2. SAR performance of different descriptors.

Datasets	fp	ECFP	MK	FS	GF
NCI1	0.30	0.32	0.29	0.27	**0.33**
NCI109	0.27	**0.32**	0.24	0.26	**0.32**
NCI123	0.25	**0.27**	0.24	0.23	**0.27**
NCI145	0.30	0.35	0.28	0.30	**0.37**
NCI167	0.06	0.06	0.04	0.06	**0.07**
NCI220	**0.33**	0.28	0.26	0.21	0.29
NCI33	0.26	0.31	0.26	0.25	**0.33**
NCI330	0.34	**0.36**	0.31	0.24	**0.36**
NCI41	0.25	**0.36**	0.28	0.30	**0.36**
NCI47	0.26	**0.31**	0.26	0.24	**0.31**
NCI81	0.27	**0.28**	0.25	0.24	**0.28**
NCI83	0.26	**0.31**	0.26	0.25	**0.31**

Wale, Ning, Karypis (2010)

- fp (Hashed fingerprint w/ paths+cycles)
- ECFP
- MK (MACCS Keys)
- FS (頻出部分グラフ)
- GF (一定サイズ以下の全生起部分グラフ)

頻出だけは良くない..

Wale N., Ning X., Karypis G. (2010) Trends in Chemical Graph Data Mining. In: Aggarwal C., Wang H. (eds) Managing and Mining Graph Data. Advances in Database Systems, vol 40. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-6045-0_19

Branch & Boundで必要な部分グラフ特徴を学習しながら同時に機械学習モデルを構築

生起部分グラフから有効なものを逐次的に選択し、同時にモデルを学習

QSARでの予測精度はECFP+Random Forestと同程度…!?

生起部分グラフの列挙木上での線形学習

- Adaboost (Kudo et al, NIPS 2005)
- gLARS (Tsuda, ICML 2007), gPLS (Saigo et al, KDD 2008), gBoost (Saigo et al, Mach Learn 2009)
- Elastic-Net正則化つき一般の線形学習への一般化 (Takigawa and Mamitsuka, TPAMI 2017)

生起部分グラフの列挙木上での非線形学習

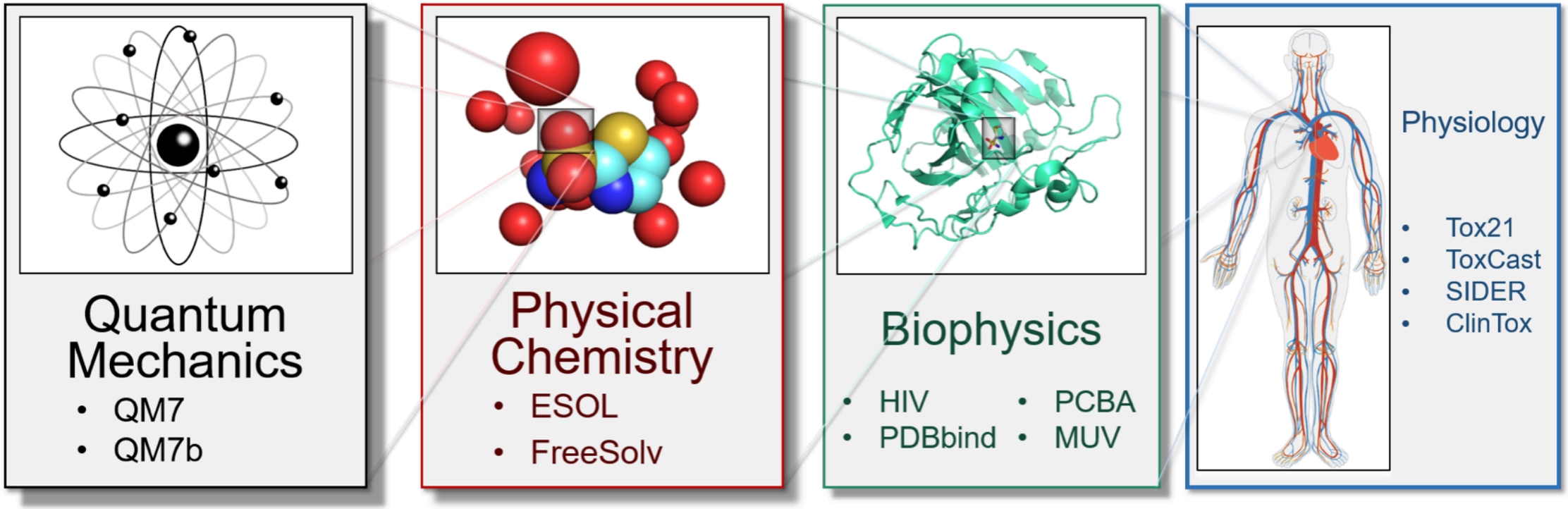
- 決定木(分類木・回帰)を学習 ＋ 木アンサンブル学習 (Shirakawa et al, MLG 2018@KDD)

分子タスクのレベルと汎用の分子表現の学習

- 部分グラフパターンの有無や数では捉えられない非組合せ論的側面を考慮できるか？
- 広いレベルで使える汎用の分子表現をデータから学習できるか？

分子自体の特性として定まる (シミュレーション/理論計算が可能)

他の分子との相互作用や環境条件など多因子的に定まる

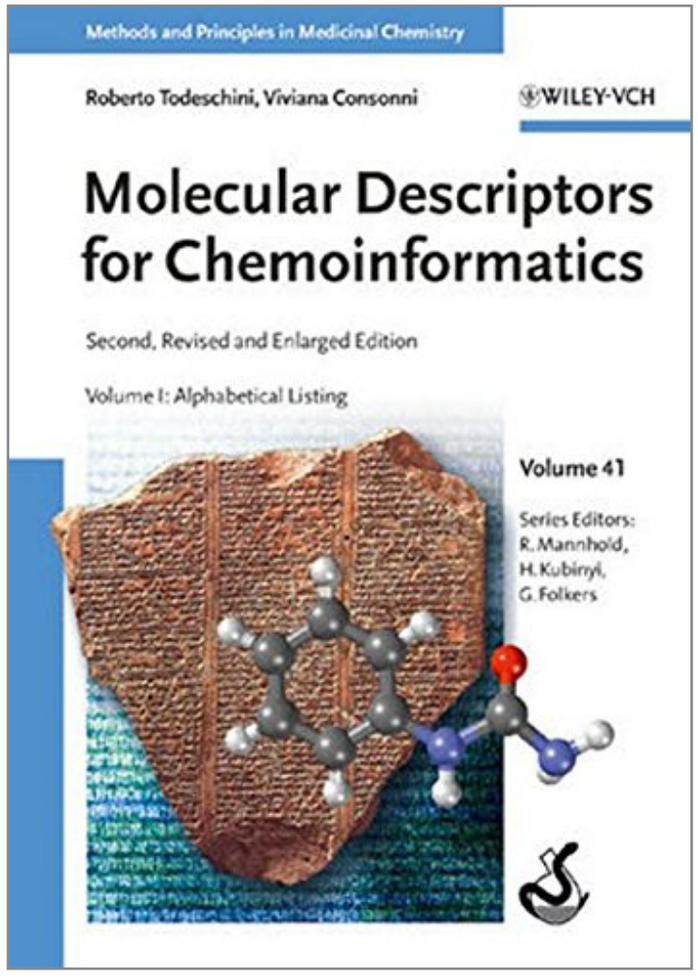


Wu et al, MoleculeNet: a benchmark for molecular machine learning, *Chem Sci* (2017)
https://doi.org/10.1021/jm201706b
http://moleculenet.ai

個別に人手でデザインされた分子記述子は多数…

- 実験的な計測量
- 計算的な記述子
  - 0D 記述子
    - constitutional descriptors
    - count descriptors
  - 1D 記述子
    - list of structural fragments
    - fingerprints
  - 2D 記述子
    - graph invariants
  - 3D 記述子
    - 3D MoRSE, WHIM, GETAWAY, ...
    - quantum-chemical descriptors
    - size, steric, surface, volume, etc.
  - 4D 記述子
    - GRID, CoMFA, Volsurf, ...

...more than 3,300 descriptors



Todeschini and Consonni, *Molecular Descriptors for Chemoinformatics.* Wiley-VCH, 2009. https://doi.org/10.1002/9783527628766

5,270 descriptors

DRAGON 7.0

kode chemoinformatics

商用の記述子ソフトウェア

RDKit
Open-Source Cheminformatics and Machine Learning

rdkit.Chem
- Descriptors
- Descriptors3D
- GraphDescriptors
- Fingerprints
- ChemicalFeatures
- ChemicalForceFields

rdkit.ML.Descriptors

オープンソースフレームワーク

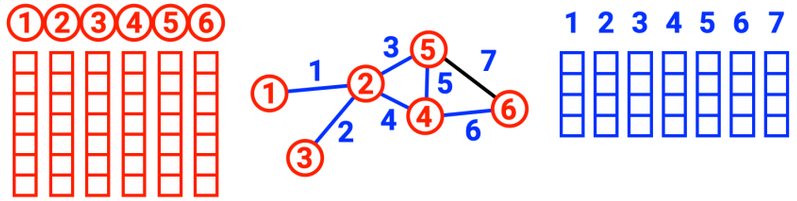
Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. Chem Rev, **2012**, *112* (5), pp 2889–2919

表現学習の試み：Graph Neural Networks (GNNs)

PyTorch geometric　DGL DEEP GRAPH LIBRARY　Spektral

分子の環境/条件/標的/相互作用等の情報

頂点特徴　トポロジ　辺特徴



グラフ表現

Representation Learning

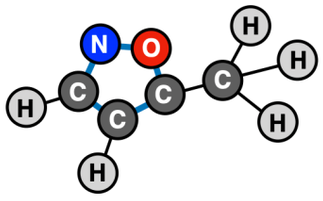
特徴ベクトル

Task-Specific Head

様々な下流タスク

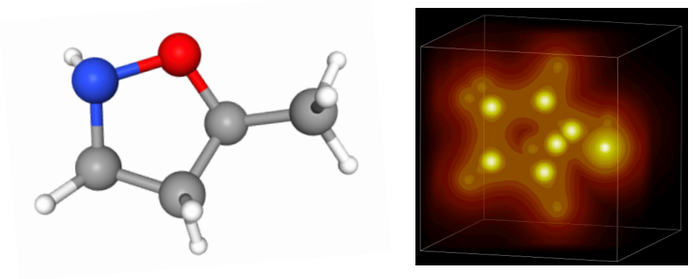
- 分類
- 回帰
- 生成

化学構造/骨格/官能基

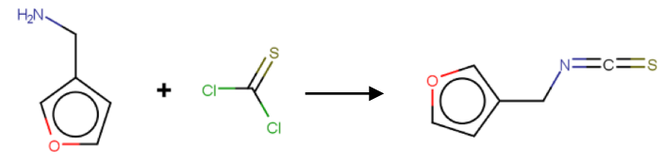


CC1CCNO1

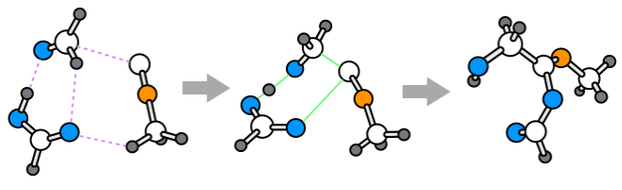
立体配座/電子状態



化学反応=構造の組替え

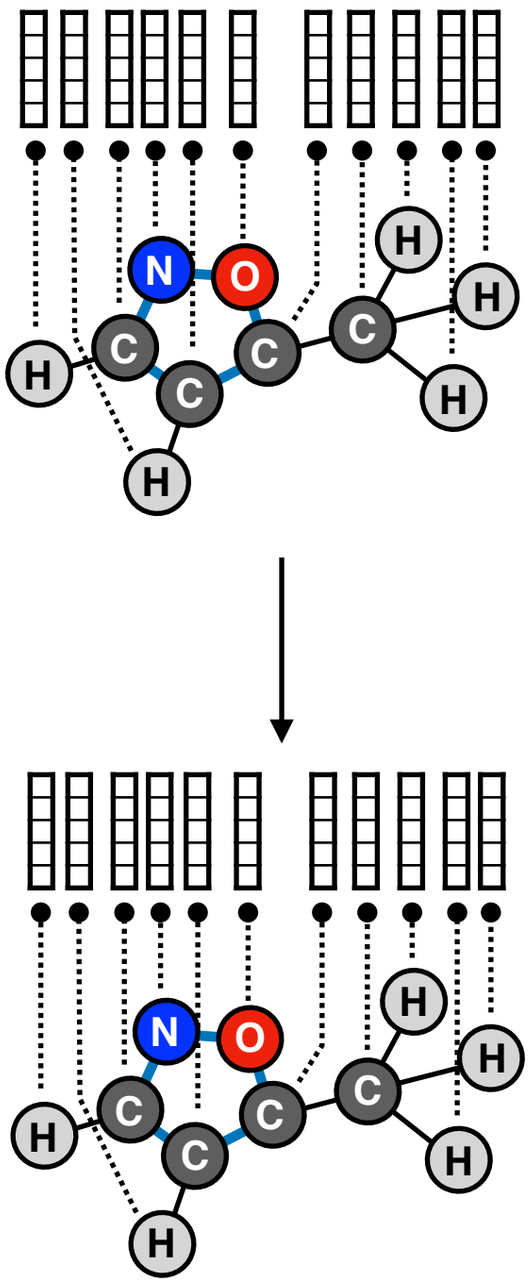


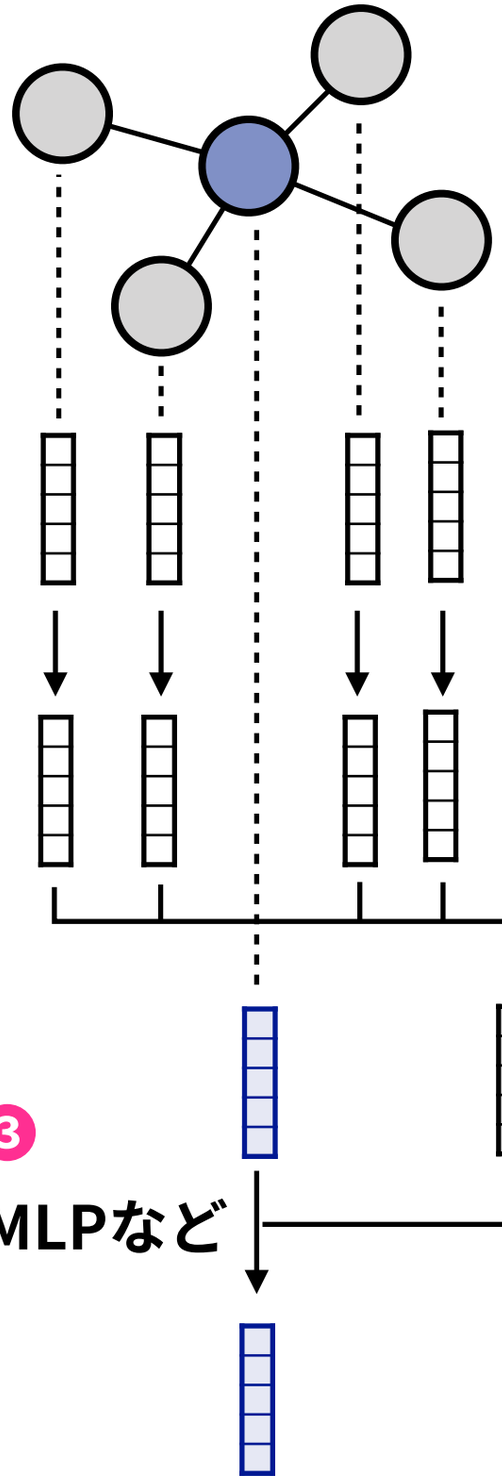
NCc1ccoc1.S=(Cl)Cl>>[RX_5]S=C=NCc1ccoc1



Message Passing Neural Networks (MPNNs)

各特徴ベクトルの更新





❶
MLPなど

置換**同変**
(equivariant)

❸
MLPなど

$$\boldsymbol{h}_i \leftarrow \psi\left(\boldsymbol{h}_i, \bigoplus_{j\in\mathcal{N}_i} \phi(\boldsymbol{h}_i, \boldsymbol{h}_i, \boldsymbol{e}_{ij})\right)$$

❸ ❷ ❶

❷
順番や数に依存しない集約操作
(sum, mean or max)
+ attention

置換**不変**
(invariant)

GNNの基本的性質についての理解

- Weisfeiler-Lehmanテストとの関係

Algorithm 1: WL-1 algorithm (Weisfeiler & Lehmann, 1968)

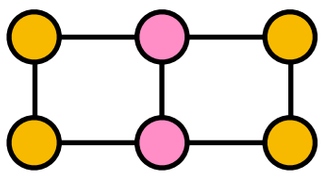
Input: Initial node coloring $(h_1^{(0)}, h_2^{(0)}, ..., h_N^{(0)})$
Output: Final node coloring $(h_1^{(T)}, h_2^{(T)}, ..., h_N^{(T)})$
$t \leftarrow 0$;
repeat
  for $v_i \in \mathcal{V}$ **do**
    $h_i^{(t+1)} \leftarrow \text{hash}\left(\sum_{j \in \mathcal{N}_i} h_j^{(t)}\right)$;
  $t \leftarrow t + 1$;
until *stable node coloring is reached*;

順番や数に依存しない集約操作

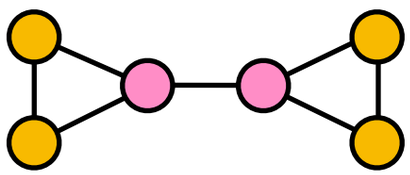
$$h_i^{(l+1)} \leftarrow \phi\left(h_i^{(l)}, \bigoplus_{j \in \mathcal{N}_i} c_{ij}\psi(h_j^{(l)})\right)$$

hash操作を微分可能な演算にするとGCNに
(Kipf and Welling, ICLR 2017)

- グラフ同型判定の意味ではGNNはWL-1 Algorithmと表現力が同じ



G_1



G_2

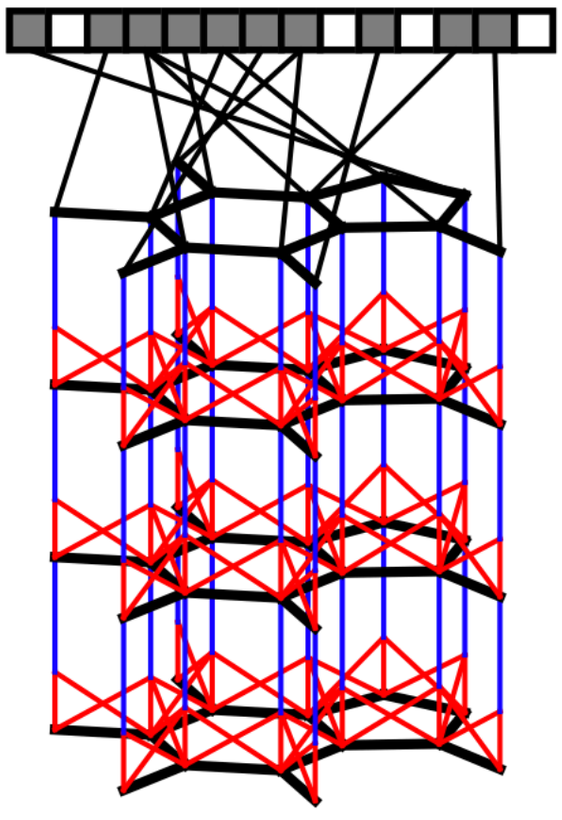
二つのグラフは同型ではないが
WL-1やGNNでは区別できない

- 集約操作 $\bigoplus$ の選択で表現力が変わる： sum > mean > max

Kipf and Welling, Semi-supervised classification with graph convolutional networks. *ICLR* (2017)
Xu, Hu, Leskovec, Jegelka, How powerful are graph neural networks? *ICLR* (2019)

ECFPとNeural Graph Fingerprint

- Neural Graph Fingerprint: 最初期に提案されたGNNの一つ
- Graph Convolutionを用いたGNNの一種とみなせる
- ECFP(Circular Fingerprint)のFingerprint計算をパラメタを持つ微分可能な演算で書き直すことで得られる**学習可能なFingerprint**という位置づけ



Algorithm 1 Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \dots, \mathbf{r}_N]$ ▷ concatenate
    $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$ ▷ hash function
    $i \leftarrow \text{mod}(r_a, S)$ ▷ convert to index
    $\mathbf{f}_i \leftarrow 1$ ▷ Write 1 at index
Return: binary vector $\mathbf{f}$

Algorithm 2 Neural graph fingerprints

Input: molecule, radius R, hidden weights $H_1^1 \dots H_R^5$, output weights $W_1 \dots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

Figure 2: Pseudocode of circular fingerprints (*left*) and neural graph fingerprints (*right*). Differences are highlighted in blue. Every non-differentiable operation is replaced with a differentiable analog.

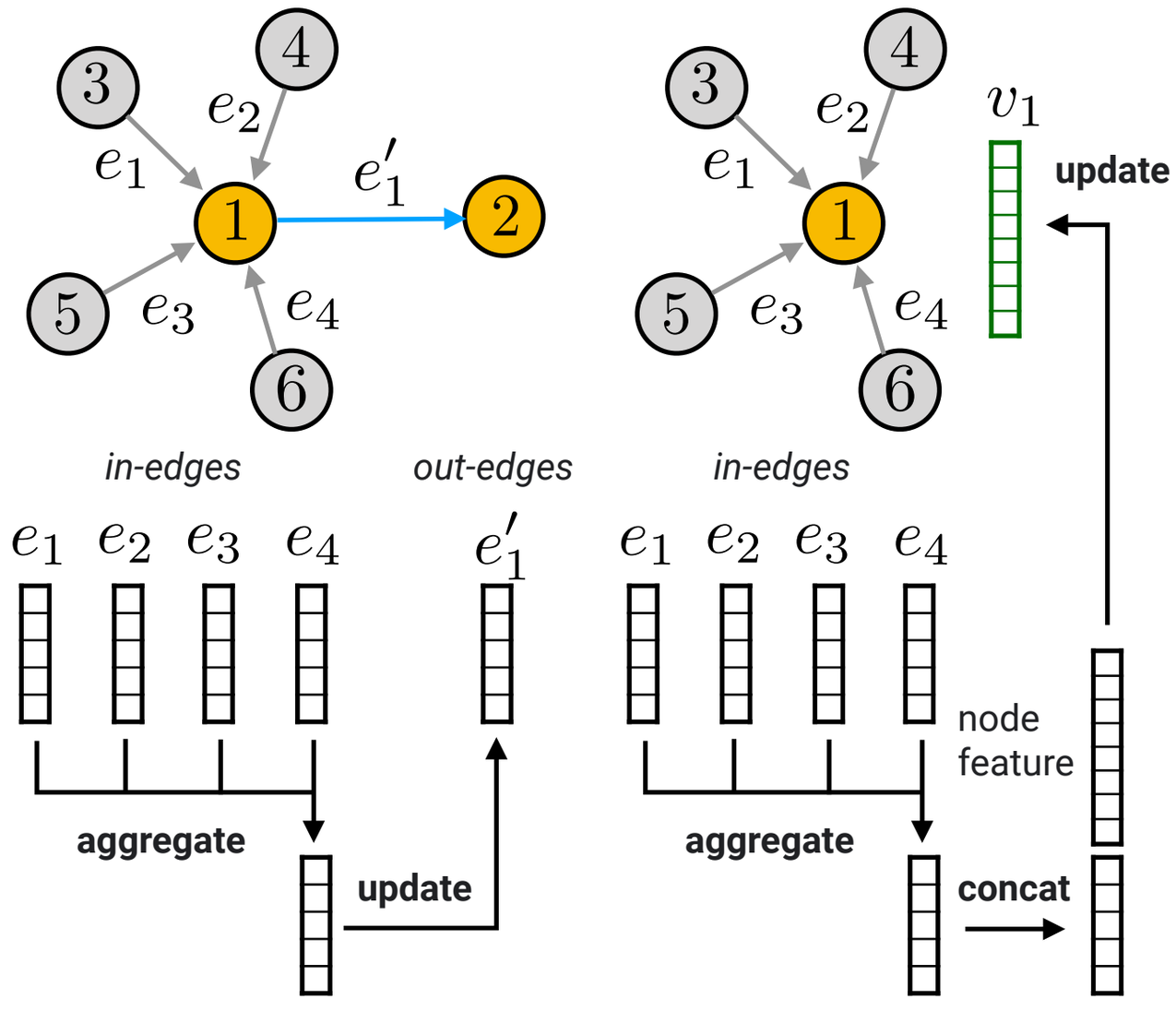
Duvenaud, Maclaurin, Aguilera-Iparraguirre, Gómez-Bombarell, Hirzel, Aspuru-Guzik, Adams, Convolutional networks on graphs for learning molecular fingerprints. *NIPS* (2015)

D-MPNN / ChemProp

Directed MPNN (Dai et al, ICML 2016) ⇢ ChemProp (Yang et al, JCIM 2019)

隠れ変数を頂点ではなく有向辺に対応づけて更新

辺の隠れ変数を更新 (T回) → 頂点の変数を更新



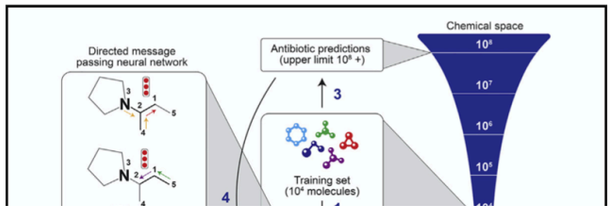
https://github.com/chemprop/chemprop

Machine Learning for Pharmaceutical Discovery and Synthesis Consortium @ MITが開発し、実際の抗生物質探索に用いられた有名なGNNの成功例

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors
Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence
regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

Stokes et al, *Cell* (2020) https://doi.org/10.1016/j.cell.2020.01.021

nature

NEWS | 20 February 2020

Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

Jo Marchant

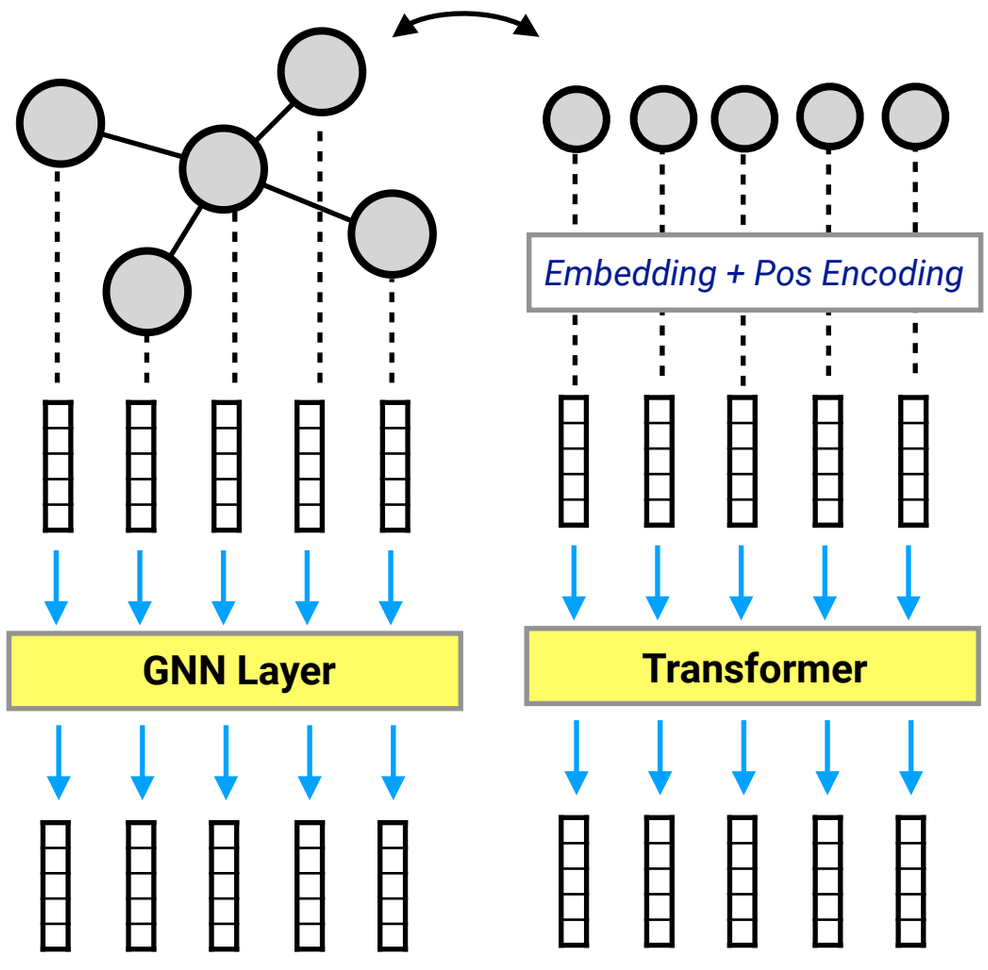
Marchant, *Nature* (2020) https://doi.org/10.1038/d41586-020-00018-3

GATとTransformer型GNN

- 各頂点の特徴ベクトルを更新する際にAttentionを入れたい
- Transformerはトポロジ制約のないGraph Attention Network (GAT)変種とみなせる

  Veličković, Cucurull, Casanova, Romero, Liò, Bengio, Graph Attention Networks (ICLR 2018) https://arxiv.org/abs/1710.10903
  Joshi, Transformers are Graph Neural Networks. (2020) https://graphdeeplearning.github.io/post/transformers-are-gnns/
- 逆にもちろんTransformer型のSelf-AttentionをGNNにもちこむこともできる

似ている…?

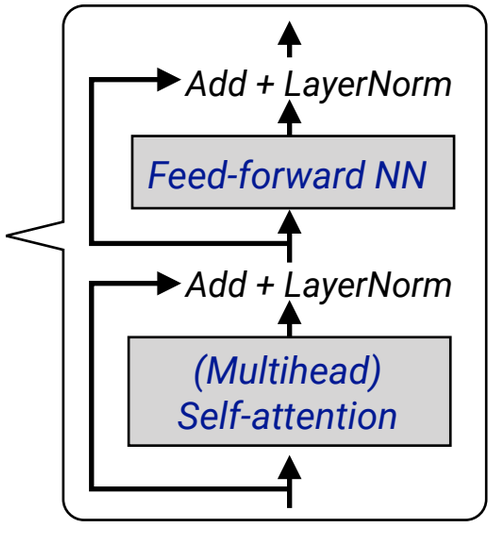


A Generalization of Transformer Networks to Graphs
Dwivedi & Bresson (2020) https://arxiv.org/abs/2012.09699

Do Transformers Really Perform Bad for Graph Representation?
Ying et al (2021) https://arxiv.org/abs/2106.05234

Communicative Representation Learning on Attributed Molecular Graphs
Song et al (2020) https://www.ijcai.org/proceedings/2020/0392.pdf

Graph-BERT: Only Attention is Needed for Learning Graph Representations
Zhang et al (2020) https://arxiv.org/abs/2001.05140



Ying et al (2021) のGraphormerは
KDDCup 2021のOpen Graph Benchmark Large-Scale Challenge(後述)のGraph-levelタスクの優勝モデルで使われた

大規模データならグラフでも
Transformerは有効…!?

分子表現の事前学習と転移学習

- Transformerへの関心は(Self-Supervisedな)大規模事前学習と転移への期待の現れ
- 分子タスクも現実の個別状況では小サンプルであることがほとんど
- もし汎用の分子表現を大規模事前学習により獲得しFew-shot/Zero-shot転移ができるのなら波及効果は計り知れない (cf. CVのImageNet-pretrain CNN, NLPのBERTやGPT)

Self-Supervised Graph Transformer on Large-Scale Molecular Data
Rong, Bian, Xu, Xie, Wei, Huang, Huang (NeurIPS 2020)
https://arxiv.org/abs/2007.02835

Strategies for Pre-training Graph Neural Networks
Hu, Liu, Gomes, Zitnik, Liang, Pande, Leskovec (ICLR 2020)
https://arxiv.org/abs/1905.12265

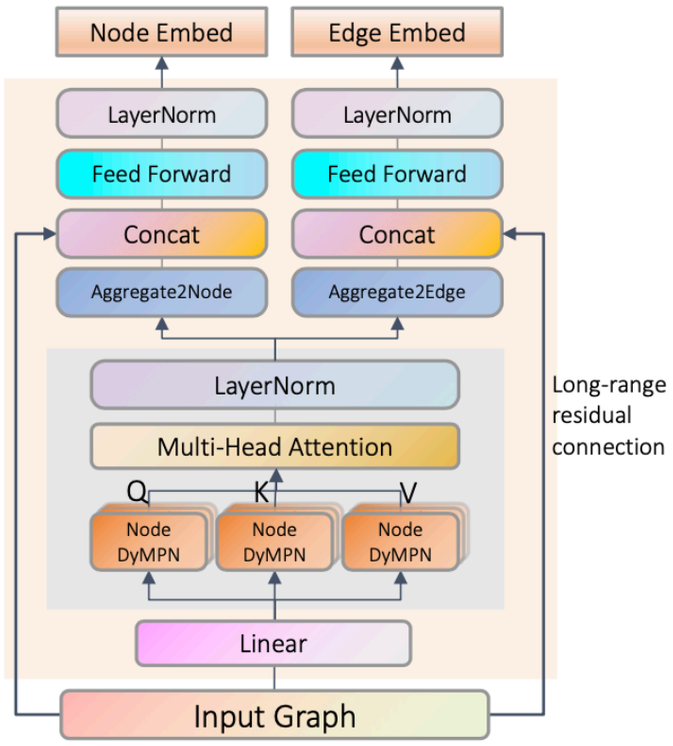


Figure 1: Overview of GTransformer.

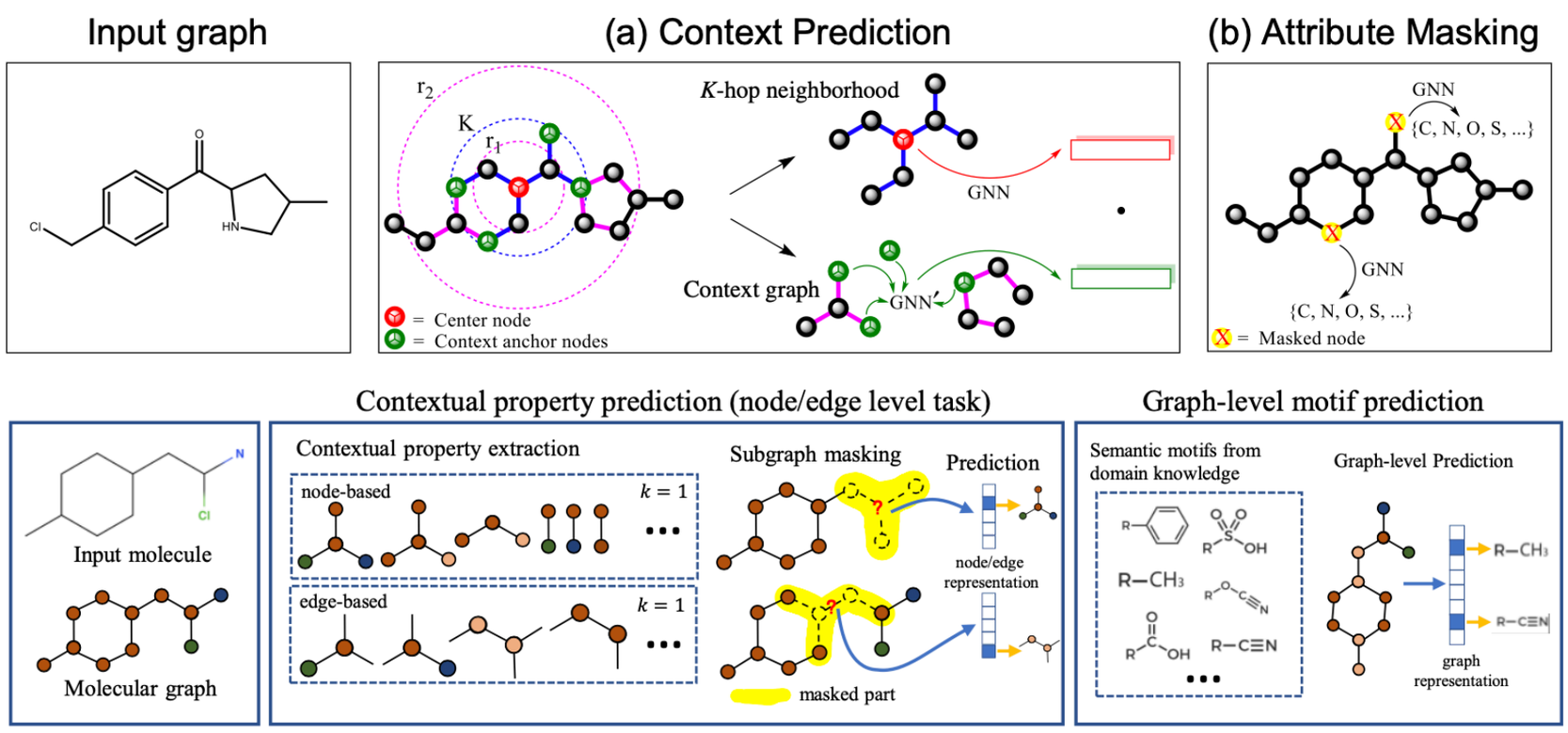


Figure 2: Overview of the designed self-supervised tasks of GROVER.

分子表現の生成

- もうひとつの分子の表現学習への期待は**分子グラフや分子構造の生成**
- 分子生成の場合は特にDecoderが非自明で構造的な処理を実現する必要がある
- 構成性/モジュール性や化学的ルールも考慮しないと意味のない出力になり得る
- 文字列表現(SMILES記法)からの生成は直接的なのでグラフ表現の優位性も要検証

Deep Graph Generators: A Survey

FAEZEH FAEZ[1], YASSAMAN OMMI[2], MAHDIEH SOLEYMANI BAGHSHAH[1], AND HAMID R. RABIEE[1], (Senior Member, IEEE)
[1]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[2]Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
Corresponding authors: Hamid R. Rabbiee and Mahdieh Soleymani Baghshah (e-mails: rabiee@sharif.edu , soleymani@sharif.edu).

Category	Key Characteristic	Publications
Autoregressive DGGs	Adopting a sequential generation strategy, either node-by-node or edge-by-edge	[1]–[26]
Autoencoder-Based DGGs	Making the generation process dependent on latent space variables	[14]–[19], [27]–[39]
RL-Based DGGs	Utilizing reinforcement learning algorithms to induce desired properties in the generated graphs	[3], [20]–[26], [40]
Adversarial DGGs	Employing generative adversarial networks (GANs) [41] to generate graph structures	[20], [22], [38]–[40], [42]–[47]
Flow-based DGGs	Learning a mapping from the complicated graph distribution into a distribution mostly modeled as a Gaussian for calculating the exact data likelihood	[12], [13], [37], [48]

https://arxiv.org/abs/2012.15544

参考：グラフ的ではない特徴が依然有効な場合も

RESEARCH ARTICLE SUMMARY

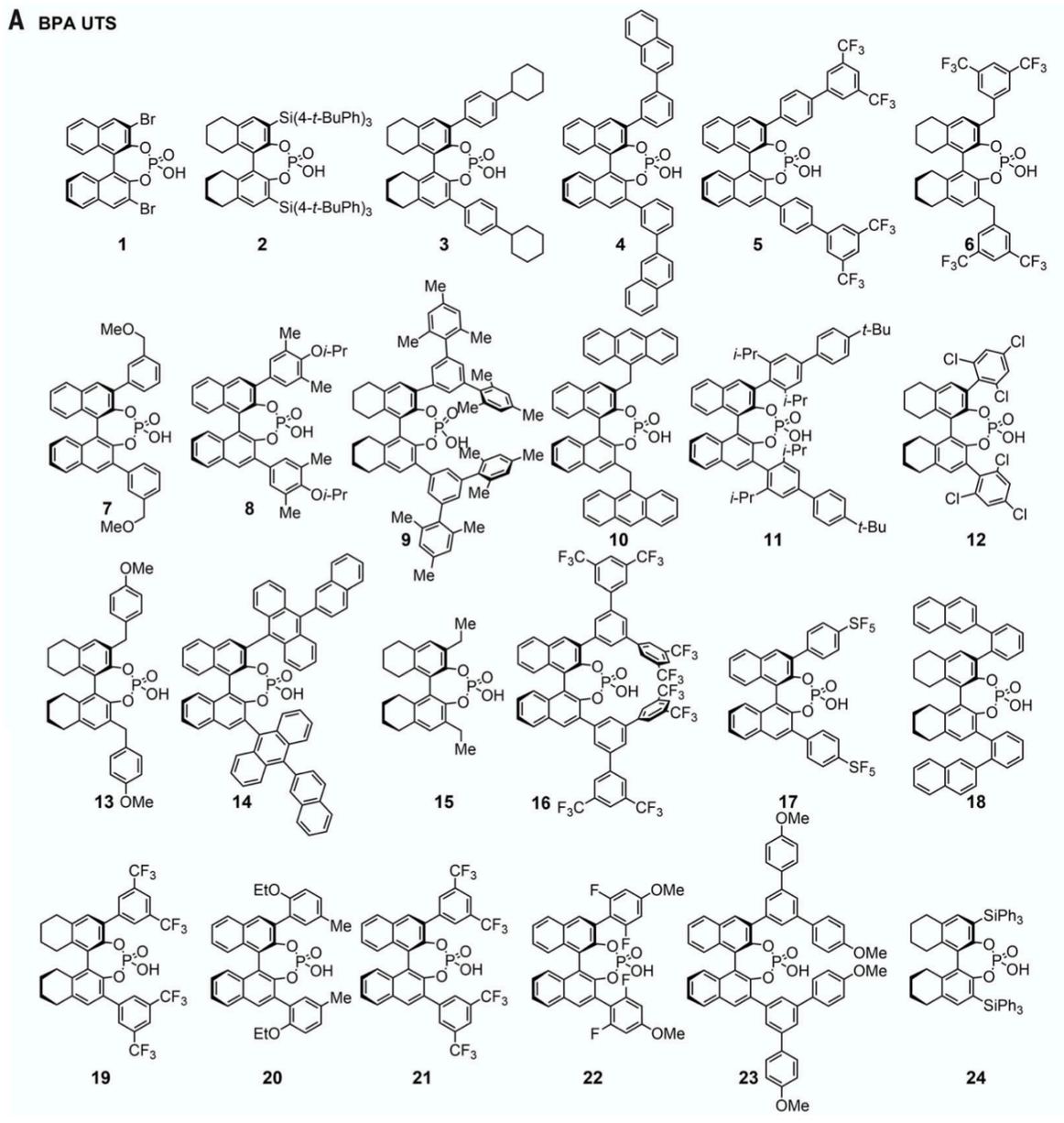
Science

ASYMMETRIC CATALYSIS

Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

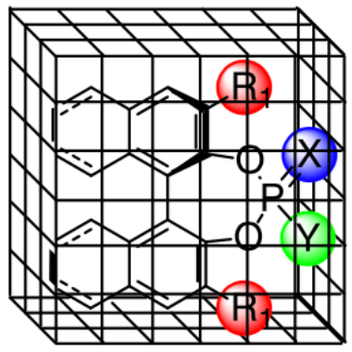
Andrew F. Zahrt*, Jeremy J. Henle*, Brennan T. Rose, Yang Wang, William T. Darrow, Scott E. Denmark†

Zahrt, Henle, Rose, Wang, Darrow, Denmark,
Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. ***Science***, 363(6424), 2019.
https://doi.org/10.1126/science.aau5631

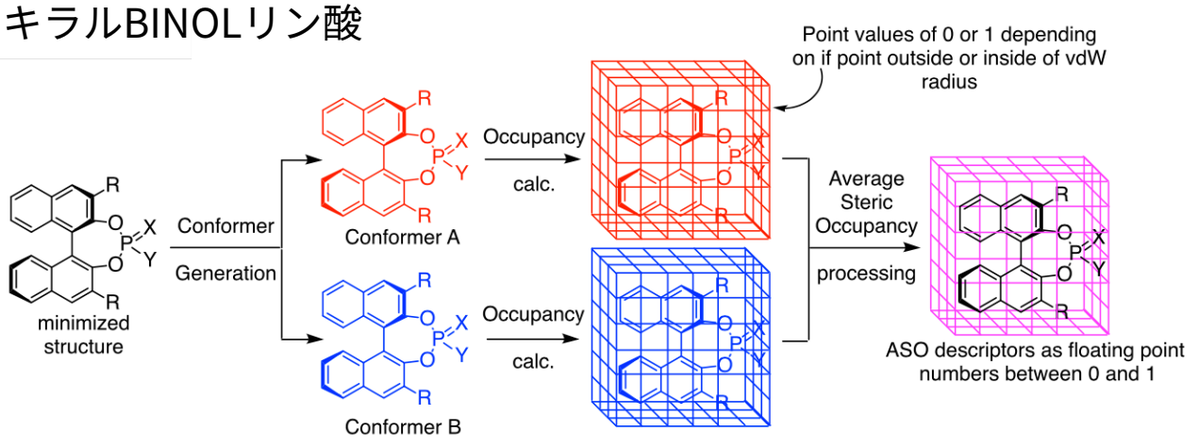


触媒設計の関心
(骨格は固定)

R_1 にどんな置換基を入れるか？



キラルBINOLリン酸



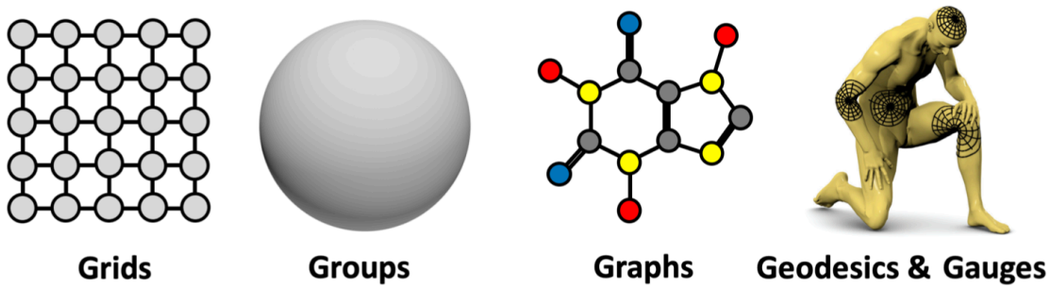
幾何的GNNと5G

https://arxiv.org/abs/2104.13478

[Submitted on 27 Apr 2021 (v1), last revised 2 May 2021 (this version, v2)]

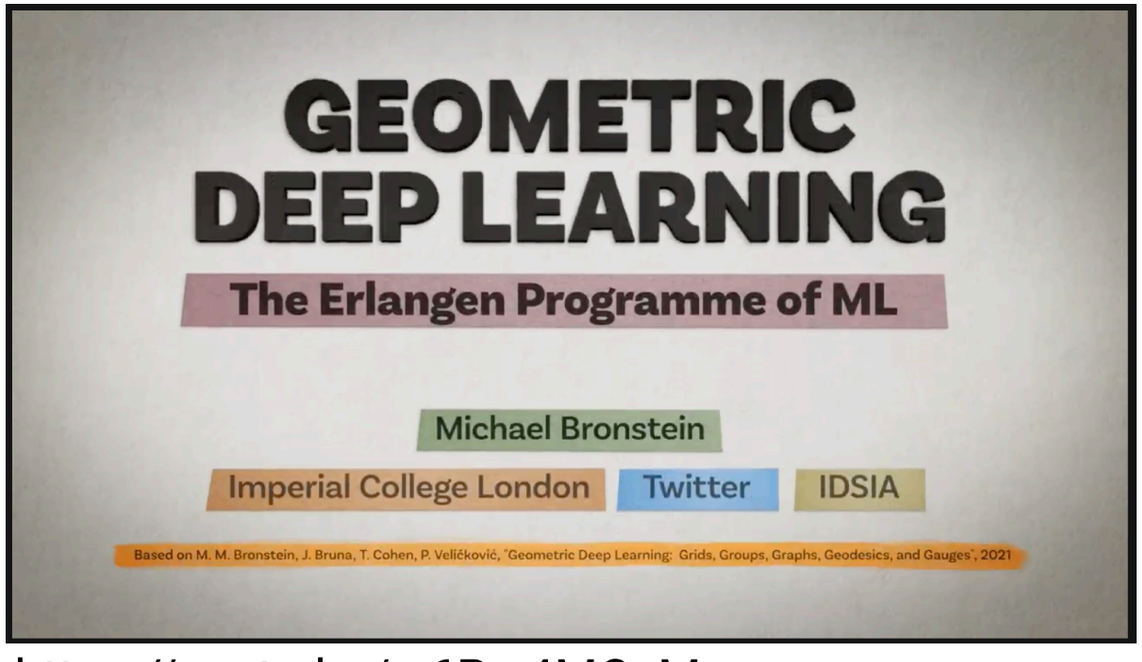
Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges

Michael M. Bronstein, Joan Bruna, Taco Cohen, Petar Veličković



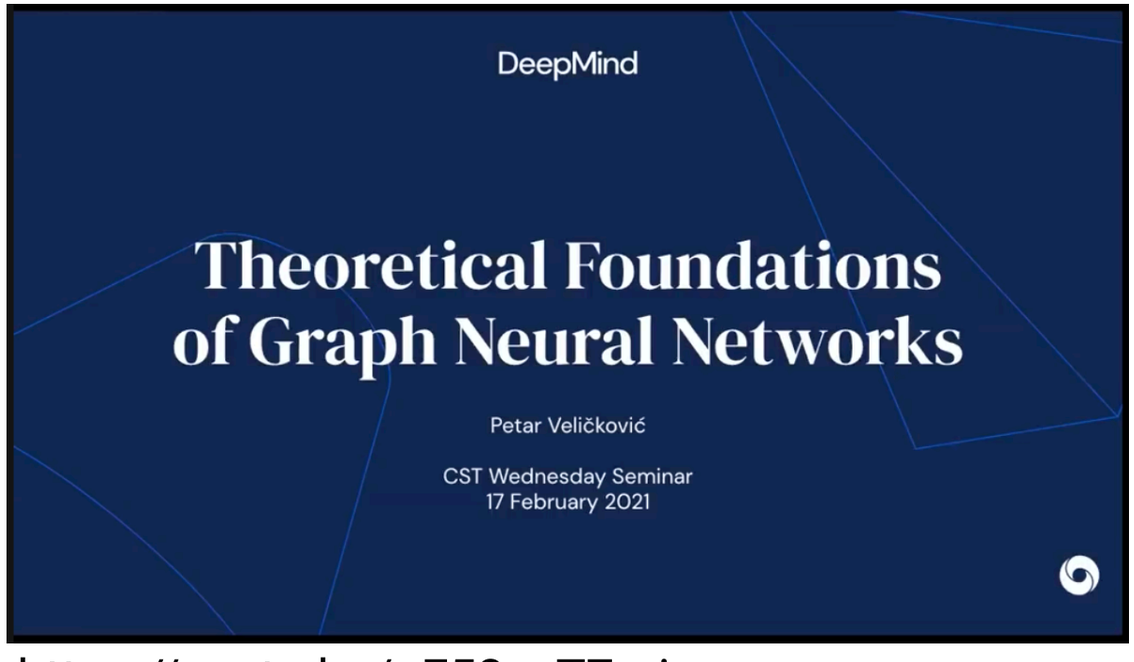
GNNは幅広い幾何構造を統一的に扱える枠組み
(機械学習のエルランゲン・プログラム!?)
5Gs: **G**rids, **G**roups, **G**raphs, **G**eodesics/**G**auges

ICLR 2021 Keynote (Michael Bronstein)



https://youtu.be/w6Pw4MOzMuo

Seminar Talk (Petar Veličković)

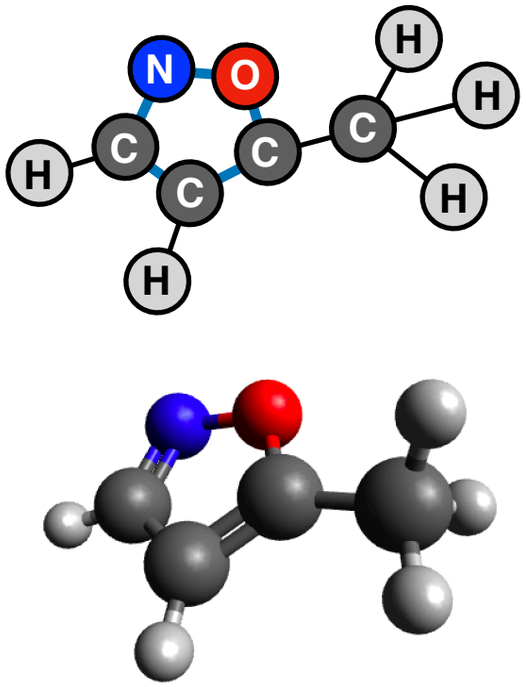


https://youtu.be/uF53xsT7mjc

新たな問題とニーズ: 量子化学計算の高精度高速近似

理論に基づく演繹的なシミュレーション計算を機械学習で代理させるニーズ

特に分子動力学計算(MD計算)などの用途で高精度な**原子間ポテンシャル**がほしい！
(与えられた原子の空間位置から原子系のポテンシャルエネルギーを計算する関数)



$\sim 10^3$ seconds

量子化学計算
密度汎関数法(DFT)計算

GNN

$\sim 10^{-2}$ seconds

0	3.119400	Dipole moment
1	48.060001	Isotropic polarizability
2	-6.876316	HOMO energy
3	-0.193201	LUMO energy
4	6.683115	Gap between HOMO and LUMO
5	483.779388	Electronic spatial extent
6	2.328750	Zero point vibrational energy
7	-7763.241699	Internal energy at 0K
8	-7763.098633	Internal energy at 298.15K
9	-7763.073242	Enthalpy at 298.15K
10	-7764.022949	Free energy at 298.15K
11	18.292999	Heat capavity at 298.15K
12	-47.811089	Atomization energy at 0K
13	-48.092346	Atomization energy at 298.15K
14	-48.349361	Atomization enthalpy at 298.15K
15	-44.743744	Atomization free energy at 298.15K
16	9.280260	Rotational constant A
17	3.564360	Rotational constant B
18	2.617160	Rotational constant C

ICML2017 https://arxiv.org/abs/1704.01212

Neural Message Passing for Quantum Chemistry

Justin Gilmer [1] Samuel S. Schoenholz [1] Patrick F. Riley [2] Oriol Vinyals [3] George E. Dahl [1]

JCTC 2017 https://doi.org/10.1021/acs.jctc.7b00577

JCTC Journal of Chemical Theory and Computation — Article — pubs.acs.org/JCTC

Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error

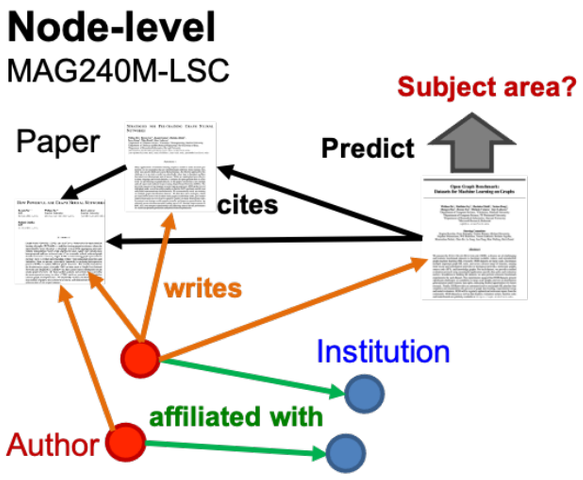
Felix A. Faber,[†] Luke Hutchison,[‡] Bing Huang,[†] Justin Gilmer,[‡] Samuel S. Schoenholz,[‡] George E. Dahl,[‡] Oriol Vinyals,[¶] Steven Kearnes,[‡] Patrick F. Riley,[‡] and O. Anatole von Lilienfeld*[,†]

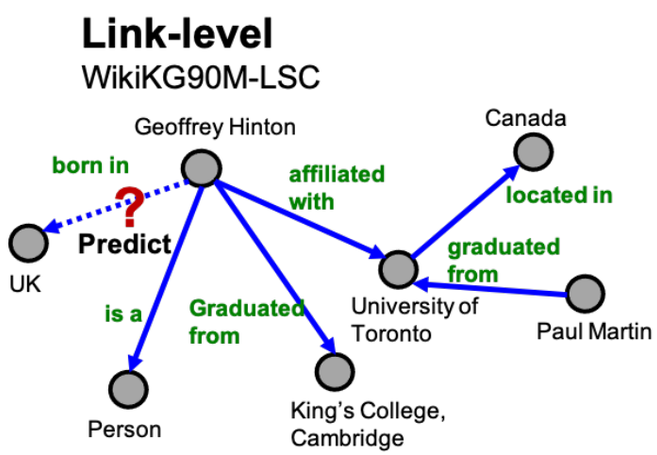
例：OGB Large-Scale Challenge (KDDCup2021)

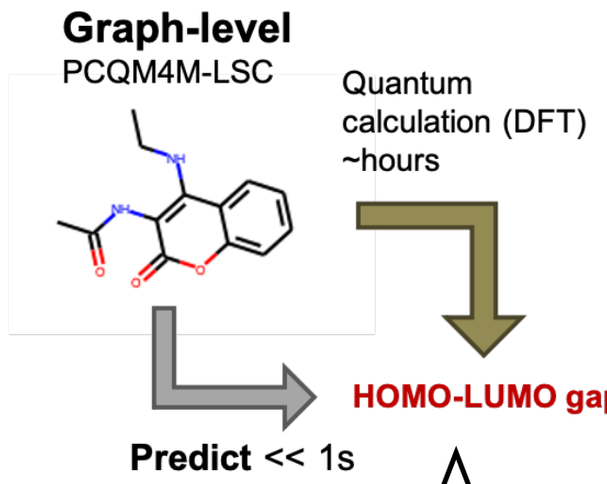
OGB-LSC — Large-Scale Challenge — @ KDD Cup 2021

OGB OPEN GRAPH BENCHMARK

https://ogb.stanford.edu/kddcup2021/







2Dの分子グラフから**量子化学計算(DFT計算)で求めたHOMO-LUMOギャップを予測**するタスク
データセット：PubChemQCから3,803,453グラフ (cf. QM9は133,885グラフ)

Results: https://ogb.stanford.edu/kddcup2021/results/#awardees_pcqm4m

1st place: Test MAE 0.1200 (eV) **10 GNNs (12-Layer Graphormer) + 8 ExpC*s (5-Layer ExpandingConv)**

2nd place: Test MAE 0.1204 (eV) **73 GNNs (11-Layer LiteGEMConv with Self-Supervised Pretraining)**

3rd place: Test MAE 0.1205 (eV) **20 GNNs (32-Layer GNN with Noisy Nodes)**

幾何的GNNで使われる分子表現

オリジナルのECFP原子不変量

- the number of immediate neighbors who are “heavy” (non-hydrogen) atoms
- the valence minus the number of hydrogens
- the atomic number
- the atomic mass
- the atomic charge
- the number of attached hydrogens

Daylight 原子不変量

- whether the atom is contained in at least one ring

オリジナルのFCFP原子不変量

- hydrogen-bond acceptor or not?
- hydrogen-bond donor or not?
- negatively ionizable or not?
- positively ionizable or not?
- aromatic or not?
- halogen or not?

Rogers and Hahn, *JCIM* (2005) https://doi.org/10.1021/ci100050t

MPNNによる量子化学計算近似で用いられた頂点・辺特徴

Table 1. Atom Features for the MG Representation[a]

feature	description
atom type	H, C, N, O, F (one-hot)
chirality	R or S (one-hot or null)
formal charge	integer electronic charge
ring sizes	for each ring size (3−8), the number of rings that include this atom
hybridization	*sp*, sp^2, or sp^3 (one-hot or null)
hydrogen bonding	whether this atom is a hydrogen bond donor and/or acceptor (binary values)
aromaticity	whether this atom is part of an aromatic system

Table 2. Atom Pair Features for the MG Representation[a]

feature	description
bond type	single, double, triple, or aromatic (one-hot or null)
graph distance	for each distance (1−7), whether the shortest path between the atoms in the pair is less than or equal to that number of bonds (binary values)
same ring	whether the atoms in the pair are in the same ring
spatial distance	the Euclidean distance between the two atoms

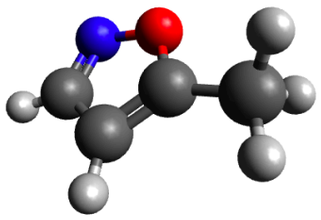
連続量ラベル

Faber et al, *JCTC* (2017) https://doi.org/10.1021/acs.jctc.7b00577

SchNetとContinuous-Filter Convolutions

各原子の核電荷とxyz座標値だけが与えられるときの代表的GNN (Schütt+ NeurIPS 2017)

Schütt, Kindermans, Sauceda, Chmiela, Tkatchenko, Müller,
SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. https://arxiv.org/abs/1706.08566

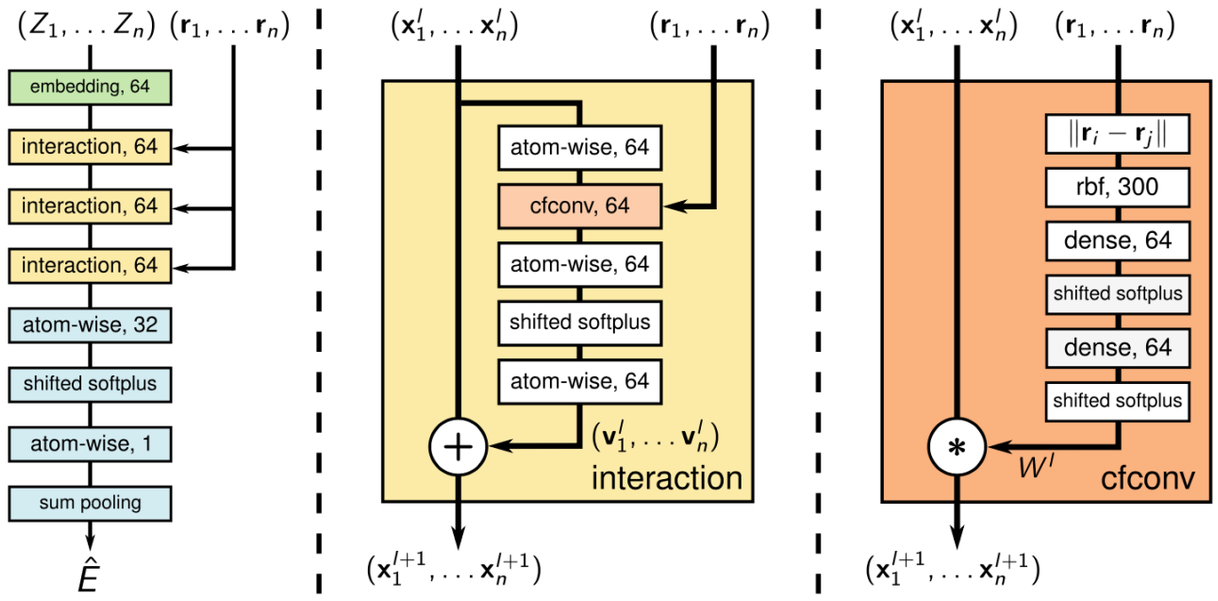


$(Z_1, \ldots Z_n)$ 各原子の核電荷

```
[6, 6, 6, 6, 7, 8, 1, 1, 1, 1, 1]
```

$(\mathbf{r}_1, \ldots \mathbf{r}_n)$ 各原子のxyz座標値

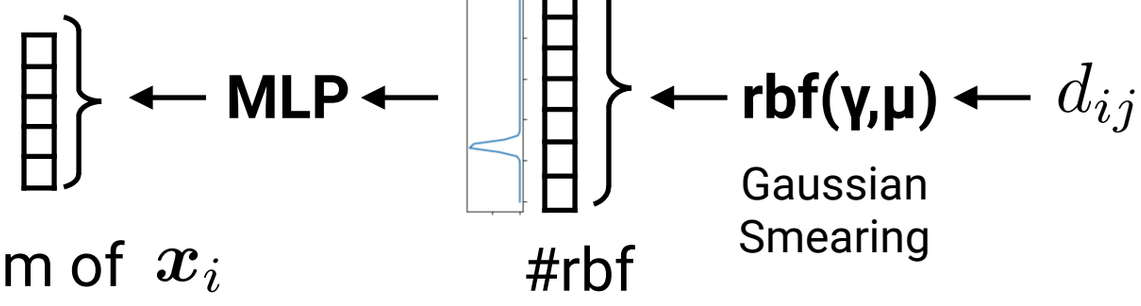
```
[[ 0.0264,  1.4944,  0.0287],
 [-0.0850,  0.0111,  0.0449],
 [-1.0957, -0.8588, -0.2308],
 [-0.5082, -2.1326, -0.0113],
 [ 0.7433, -2.0451,  0.3632],
 [ 1.0156, -0.6777,  0.3997],
 [ 0.8032,  1.8213, -0.6702],
 [-0.9230,  1.9401, -0.2734],
 [ 0.2923,  1.8782,  1.0190],
 [-2.1005, -0.6271, -0.5415],
 [-0.9559, -3.1113, -0.1136]]
```



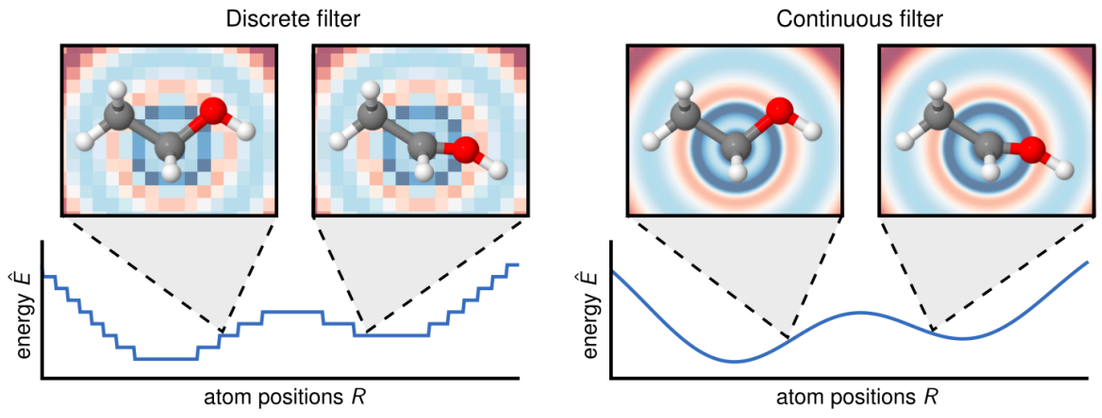
Continuous-Filter Convolutions (cfconv layers)

element-wise product

$$\boldsymbol{x}_i \leftarrow \sum_{j \in \mathcal{N}_i} \boldsymbol{x}_j \odot \phi(\exp(-\gamma(d_{ij} - \mu)))$$



画像など離散グリッドのConvolution演算の連続拡張になっている



ユークリッドの運動群に関する不変性・同変性

ユークリッド群 E(3) = 3Dの並進・回転対称性
特殊ユークリッド群 SE(3) = 3Dの並進・回転・鏡像対称性

集合 X, Y に作用する変換群 G に対して、写像 $f : X \to Y$ が変換 $g \in G$ に関して

不変 (invariat) $f(g \cdot x) = f(x)$

変換してもしないときと変わらない

同変 (equivariant) $f(g \cdot x) = g \cdot f(x)$

変換してから写像しても写像してから変換しても変わらない

幾何的GNNでは基本的な要請
(特に量子化学計算近似の場合)

E(3)不変
- Schütt et al, SchNet. (2017) https://arxiv.org/abs/1706.08566
- Unke et al, PhysNet. (2019) https://arxiv.org/abs/1902.08408
- Klicpera et al, DimeNet++. (2020) https://arxiv.org/abs/2011.14115

SE(3)同変
- Anderson et al, Cormorant. (2019) https://arxiv.org/abs/1906.04015
- Fuchs et al, SE(3)-Transformers. (2021) https://arxiv.org/abs/2006.10503

E(3)同変
- Thomas et al, Tensor Field Networks. (2018) https://arxiv.org/abs/1802.08219
- Köhler et al, Equivariant Flows (Radial Field). (2020) https://arxiv.org/abs/2006.02425
- Satorras et al, E(n) Equivariant Graph Neural Networks. (2021) https://arxiv.org/abs/2102.09844

化学でのホットトピックであるだけでなく…

Science

REVIEW

Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling[1] and Alán Aspuru-Guzik[2,3,4]*

Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning

Andrew F. Zahrt*, Jeremy J. Henle*, Brennan T. Rose, Yang Wang, William T. Darrow, Scott E. Denmark†

nature

REVIEW

Machine learning for molecular and materials science

Keith T. Butler[1], Daniel W. Davies[2], Hugh Cartwright[3], Olexandr Isayev[4]* & Aron Walsh[5,6]*

Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler[1,2], Mike Preuss[3] & Mark P. Waller[4]

Holistic prediction of enantioselectivity in asymmetric catalysis

Jolene P. Reid[1] & Matthew S. Sigman[1]*

Bayesian reaction optimization as a tool for chemical synthesis

Benjamin J. Shields[1], Jason Stevens[2], Jun Li[2], Marvin Parasram[1], Farhan Damani[3], Jesus I. Martinez Alvarado[1], Jacob M. Janey[2], Ryan P. Adams[3]✉ & Abigail G. Doyle[1]✉

nature reviews chemistry

REVIEWS

Synthetic organic chemistry driven by artificial intelligence

*A. Filipa de Almeida[1], Rui Moreira[1] and Tiago Rodrigues[2]**

PERSPECTIVES

Exploring chemical compound space with quantum-based machine learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko

機械学習分野のホットトピックでもある！

NeurIPS 2020

- *Self-Supervised Graph Transformer on Large-Scale Molecular Data*
- *RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist*
- *Reinforced Molecular Optimization with Neighborhood-Controlled Grammars*
- *Autofocused Oracles for Model-based Design*
- *Barking Up the Right Tree: an Approach to Search over Molecule Synthesis DAGs*
- *On the Equivalence of Molecular Graph Convolution and Molecular Wave Function with Poor Basis Set*
- *CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models*

ICML 2020, 2021

- *A Graph to Graphs Framework for Retrosynthesis Prediction*
- *Hierarchical Generation of Molecular Graphs using Structural Motifs*
- *Learning to Navigate in Synthetically Accessible Chemical Space Using Reinforcement Learning*
- *Reinforcement Learning for Molecular Design Guided by Quantum Mechanics*
- *Multi-Objective Molecule Generation using Interpretable Substructures*
- *Improving Molecular Design by Stochastic Iterative Target Augmentation*
- *A Generative Model for Molecular Distance Geometry*
- *GraphDF: A Discrete Flow Model for Molecular Graph Generation*
- *An End-to-End Framework for Molecular Conformation Generation via Bilevel Programming*
- *Equivariant message passing for the prediction of tensorial properties and molecular spectra*
- *Learning Gradient Fields for Molecular Conformation Generation*
- *Self-Improved Retrosynthetic Planning*

ICLR 2020, 2021

- *Directional Message Passing for Molecular Graphs*
- *GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation*
- *Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space*
- *A Fair Comparison of Graph Neural Networks for Graph Classification*
- *MARS: Markov Molecular Sampling for Multi-objective Drug Discovery*
- *Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design*
- *Learning Neural Generative Dynamics for Molecular Conformation Generation*
- *Conformation-Guided Molecular Representation with Hamiltonian Neural Networks*
- *Symmetry-Aware Actor-Critic for 3D Molecular Design*

課題と展望

- 量子化学計算などデータが計算由来の場合、機械学習で良い代理モデルを作るための周到なデータ獲得計画が課題 (計算がゲームの場合のように軽くない)
- 適切な帰納バイアスの設計が課題
  Chemical Spaceは広大すぎるので第一原理(量子化学)や熱力学、ConformerやDynamicsなど演繹的物理制約を用いてモデルを制約/正則化するのが有望?
- 幾何的GNNの得意タスクはKashimaカーネルやSOAPカーネル+GAP等でも解けるので実用レベルの表現学習の試金石は大規模事前学習＋few-shot/zero-shot転移の実現?
- 大規模なデータがあれば分子の表現学習にブレイクスルーは起こるのか?
  汎用モデルと帰納バイアス: CNNs vs Transformers vs GNNs vs MLPs


CHEMICAL REVIEWS

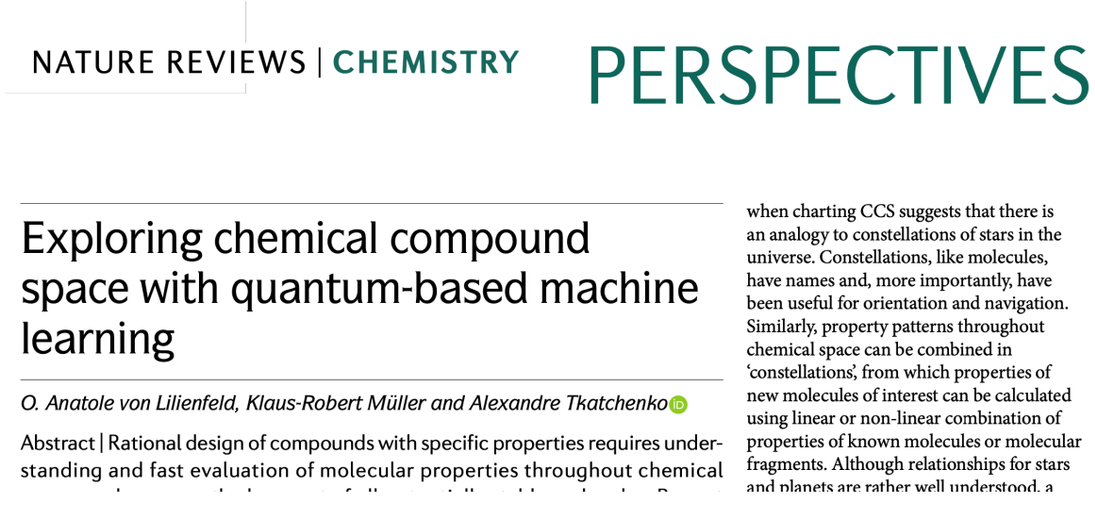
pubs.acs.org/CR Review

Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,* Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller,* and Alexandre Tkatchenko*

Cite This: https://doi.org/10.1021/acs.chemrev.1c00107 Read Online

https://arxiv.org/abs/2102.06321


NATURE REVIEWS | CHEMISTRY PERSPECTIVES

Exploring chemical compound space with quantum-based machine learning

O. Anatole von Lilienfeld, Klaus-Robert Müller and Alexandre Tkatchenko

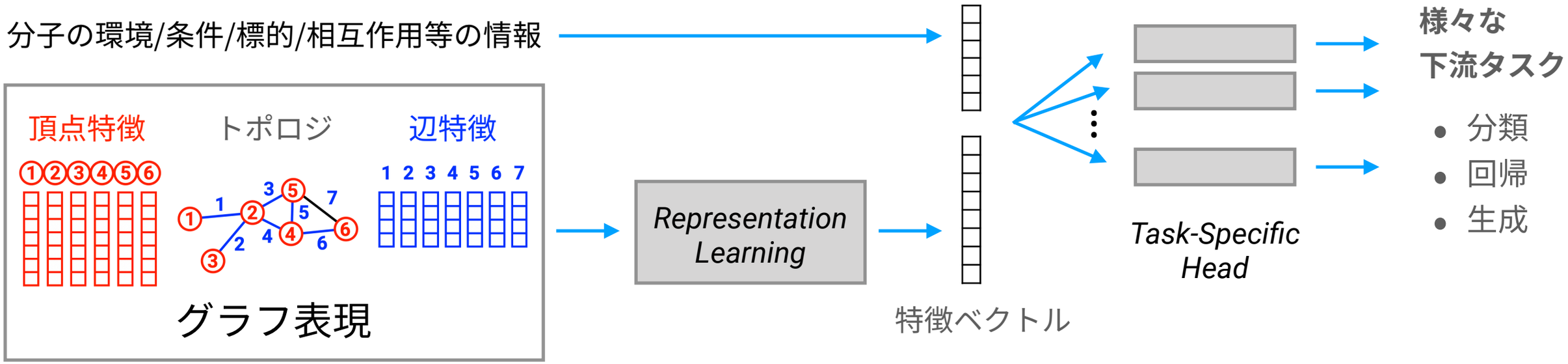
Abstract | Rational design of compounds with specific properties requires understanding and fast evaluation of molecular properties throughout chemical

when charting CCS suggests that there is an analogy to constellations of stars in the universe. Constellations, like molecules, have names and, more importantly, have been useful for orientation and navigation. Similarly, property patterns throughout chemical space can be combined in 'constellations', from which properties of new molecules of interest can be calculated using linear or non-linear combination of properties of known molecules or molecular fragments. Although relationships for stars and planets are rather well understood, a

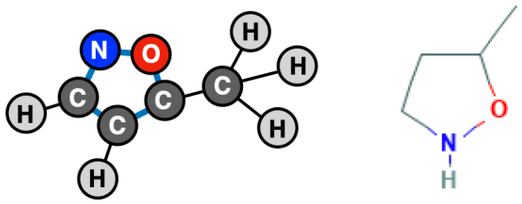
https://arxiv.org/abs/1911.10084

本日の話題：分子のグラフ表現＋機械学習

分子の環境/条件/標的/相互作用等の情報

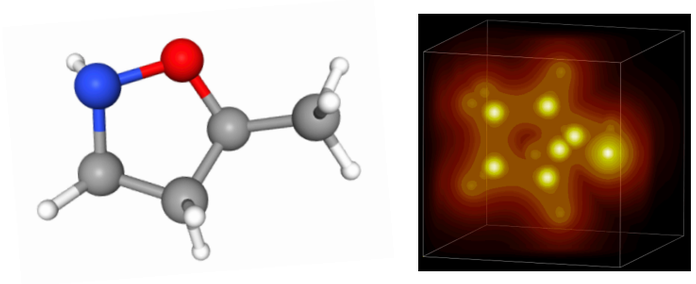


化学構造/骨格/官能基

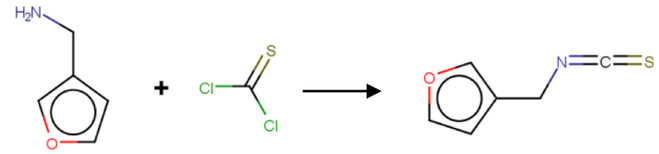


CC1CCNO1

立体配座/電子状態



化学反応=構造の組替え



NCc1ccoc1.S=(Cl)Cl>>[RX_5]S=C=NCc1ccoc1