機械学習による化学反応の予測と設計

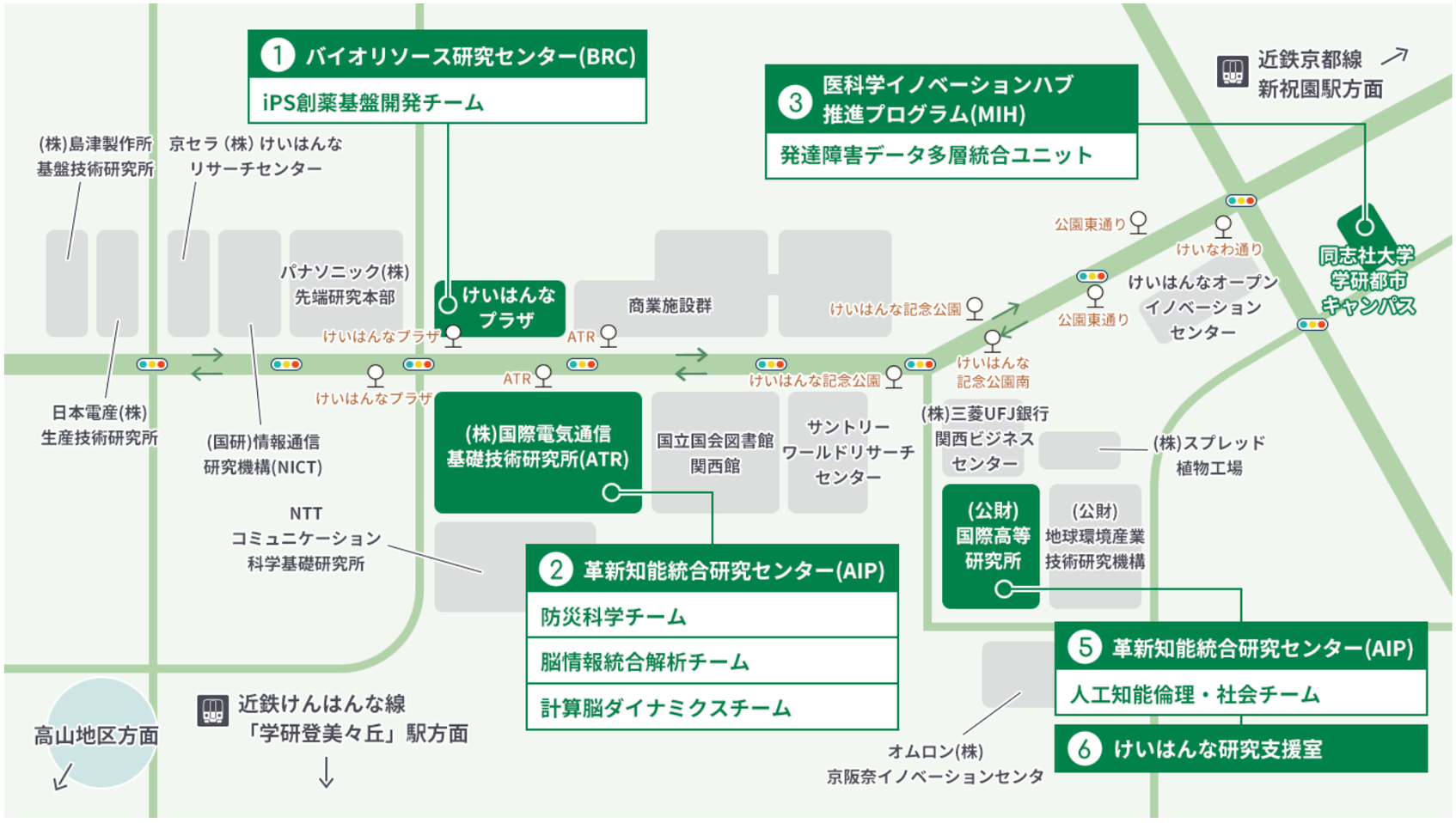
Machine Learning for Chemical Reaction Design and Discovery

瀧川 一学

- 理化学研究所 革新知能統合研究センター@京阪奈
  iPS細胞連携医学的リスク回避チーム
- 北海道大学 化学反応創成研究拠点 (WPI-ICReDD)

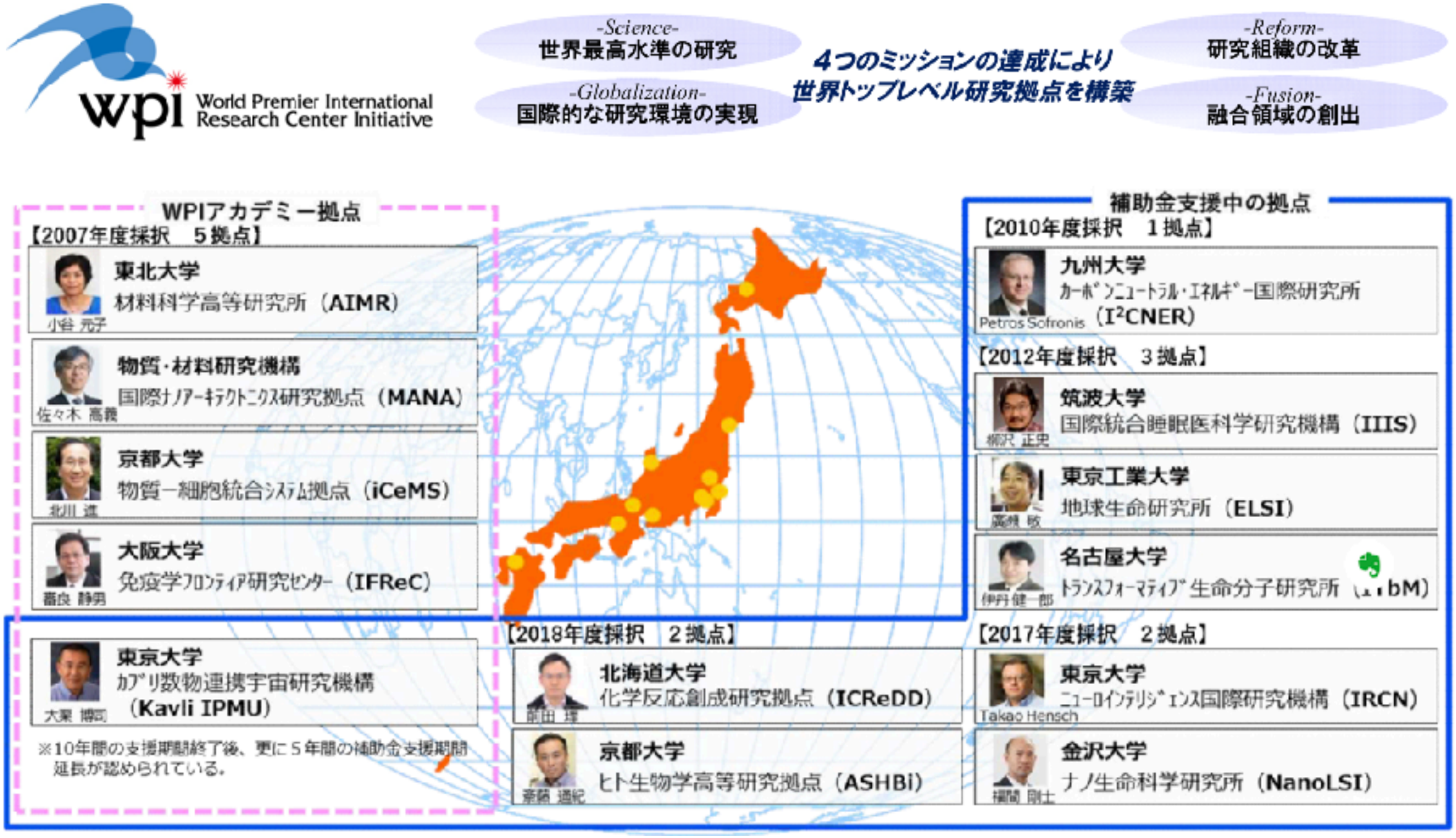
組織の紹介① 理化学研究所 革新知能統合研究センター(AIP)

- 文部科学省の**人工知能基盤技術**に関する研究拠点
- 2016年度設置、2017年活動開始
- 本部は東京・日本橋 (東京駅近く)
  COREDO日本橋のある日本橋一丁目三井ビルディング15階
- 所属チーム勤務地：京阪奈地区 ATR内



組織の紹介② 北海道大学 化学反応創成研究拠点

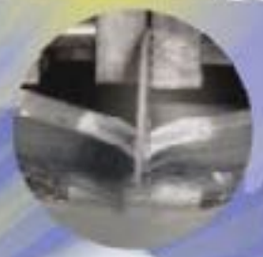
- 文部科学省の世界トップレベル研究拠点(WPI)プログラム拠点
- 2018年10月採択



High-value-added chemicals

New materials

Advanced medical technology



Chemical Reaction Design and Discovery (CReDD)

Acceleration of the development of chemical reactions

Seamlessly fusing three types of sciences

Computational science

Experimental science

Information science

ICReDD

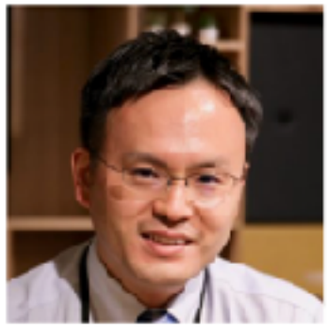
Institute for Chemical Reaction Design and Discovery

HOKKAIDO UNIVERSITY

Principal Investigators

https://www.icredd.hokudai.ac.jp/all-members

Computational Science



MAEDA, Satoshi

RUBINSTEIN, Michael

TAKETSUGU, Tetsuya

Experimental Science

GONG, Jian Ping

HASEGAWA, Yasuchika

INOKUMA, Yasuhide

ITO, Hajime

LIST, Benjamin

SAWAMURA, Masaya

TANAKA, Shinya

Information Science

KOMATSUZAKI, Tamiki

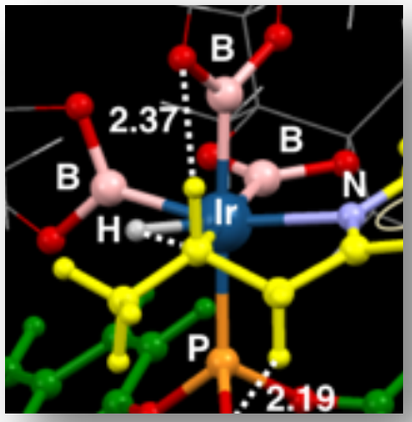
TAKIGAWA, Ichigaku

VARNEK, Alexandre

YOSHIOKA, Masaharu

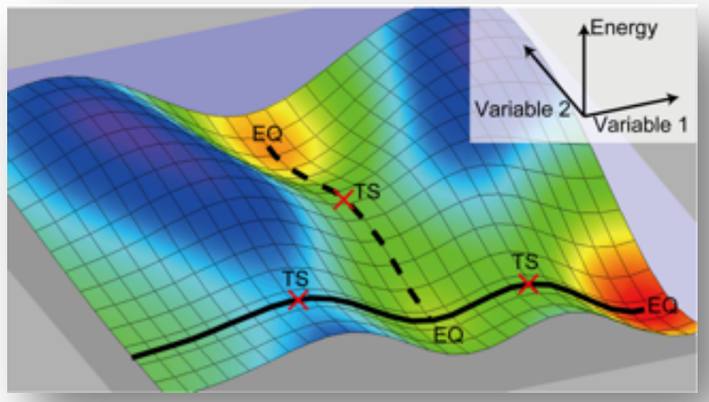
ICReDDのテクノロジー

Making full use of available computational / informatics tools, we establish chemical reaction design and discovery lead by computation / informatics

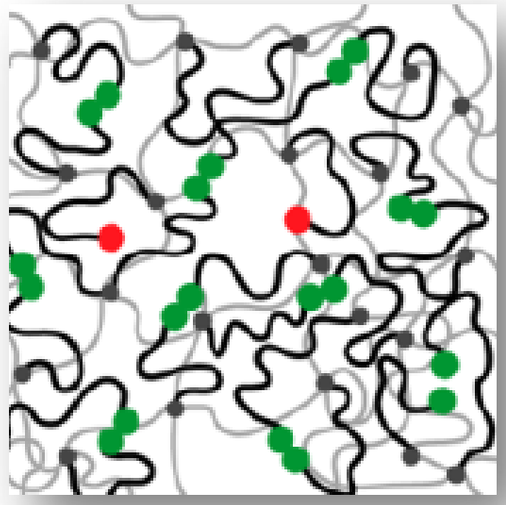


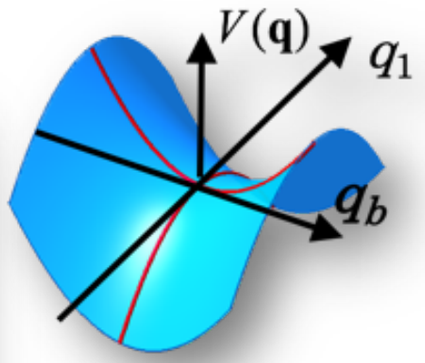
DFT for transition state calculation

AFIR for automated reaction path searching



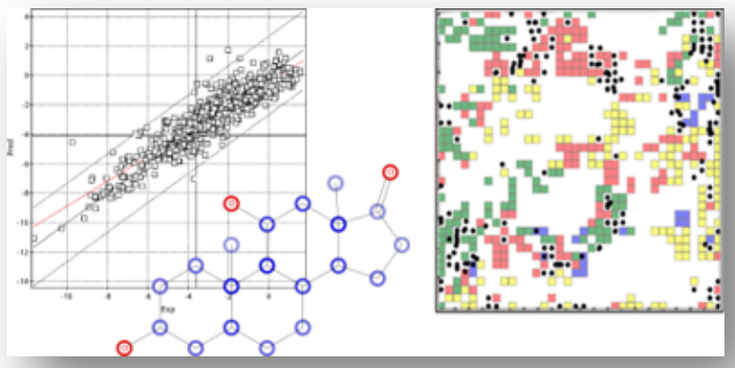
Large-scale simulation

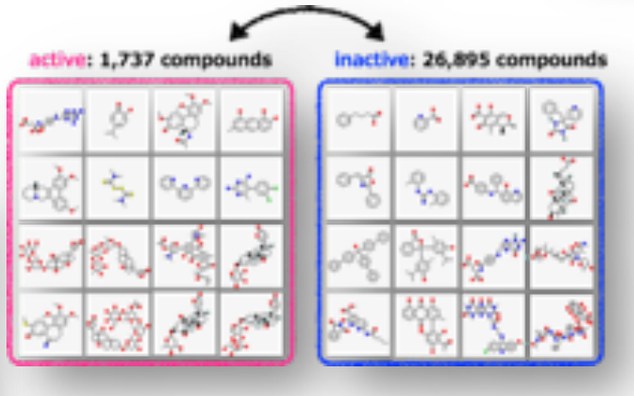




Mathematical modeling

Chemoinformatics





Machine-learning

AIMD for dynamical simulation of chemical reactions

自己紹介：瀧川 一学 (たきがわ・いちがく)

https://itakigawa.github.io/

専門：機械学習・データマイニングとその自然科学での利活用
「データからの学習」をどう問題解決に活用できるのか？

10年 北大 **統計的信号処理とパターン認識 (工学研究科)**
(1995〜2004) "劣決定信号源分離のL1ノルム最小解の理論分析"

7年 京大 **バイオインフォマティクス (化学研究所)**
(2005〜2011) **ケモインフォマティクス (薬学研究科)**

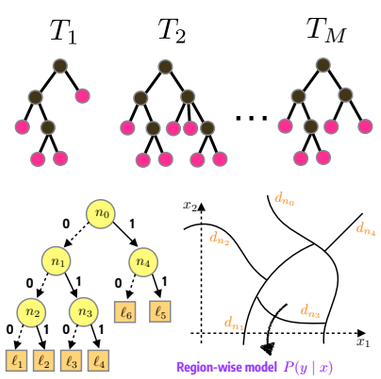
7年 北大 **データ駆動科学・ 離散構造を伴う機械学習**
(2012〜2018) **(情報科学研究科)**
+ JSTさきがけ: 材料インフォマティクス

？年 理研(京都) **AIPセンター iPS細胞連携医学的リスク回避チーム**
(2019〜) **(北大 化学反応創成研究拠点とクロアポ)**

最近の関心：「機械学習」+「離散構造」for「自然科学」

機械学習
－学習と知能－

木構造アンサンブル

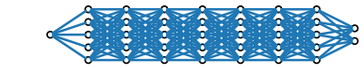


dmlc XGBoost / scikit learn / R

LightGBM (Microsoft)

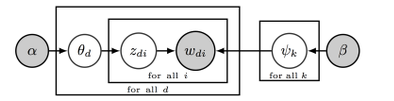
CatBoost (Yandex)
TFBoost (Google)
TenscentBoost (Tenscent)
Sherwood decision forests
See5/C5.0 & Cubist (RuleQuest)
CART® MARS® TreeNet®
Random Forests® (Salford Systems)
FAB/HME (NEC)

深層学習/計算グラフ



Microsoft Cognitive Toolkit / TensorFlow / Caffe2 / Caffe / theano / dmlc mxnet / K / nervana SYSTEMS / dy/net / PYTORCH / Chainer / torch / GLUON

確率的プログラミング

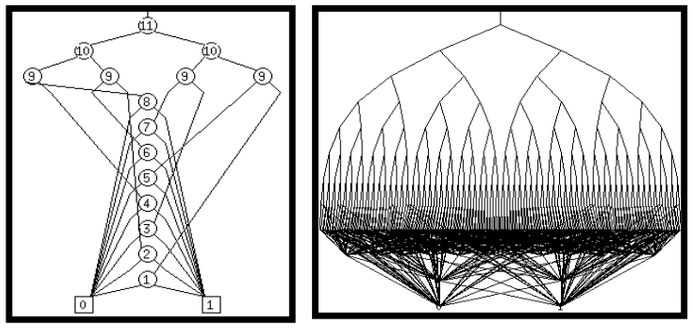


Stan / PyMC3 / BUGS

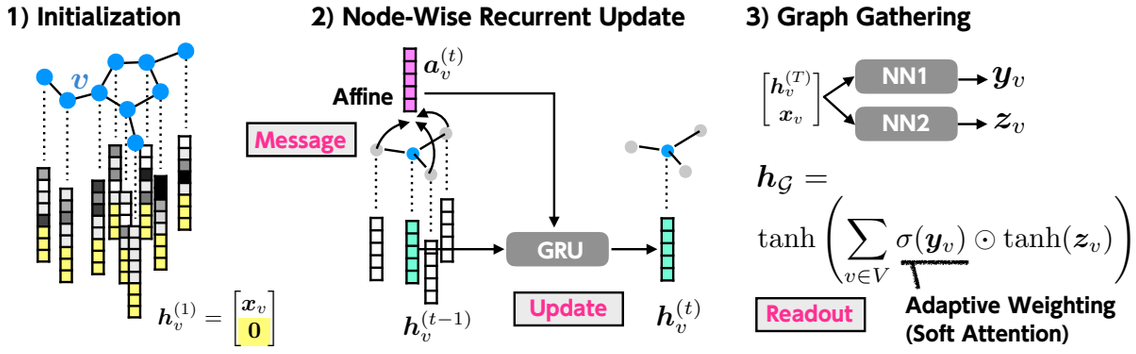
Edward / Prob Torch / Pyro / BayesFlow

離散構造
－構造と知識－

離散構造の表現と構成法

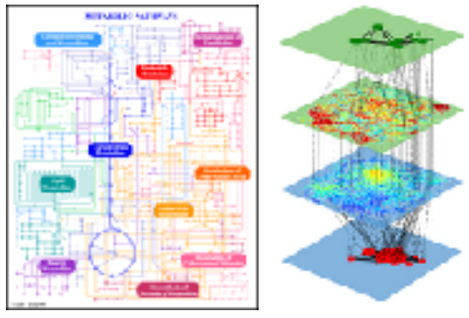


離散構造を入力・制約とする機械学習

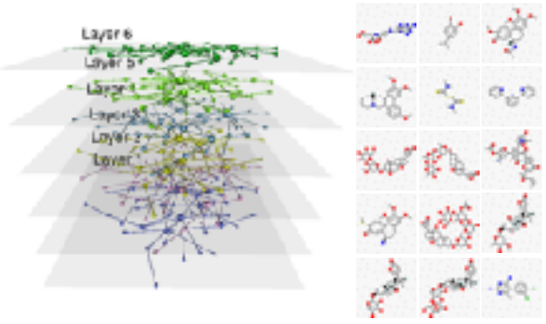


データ駆動科学
－情報と科学－

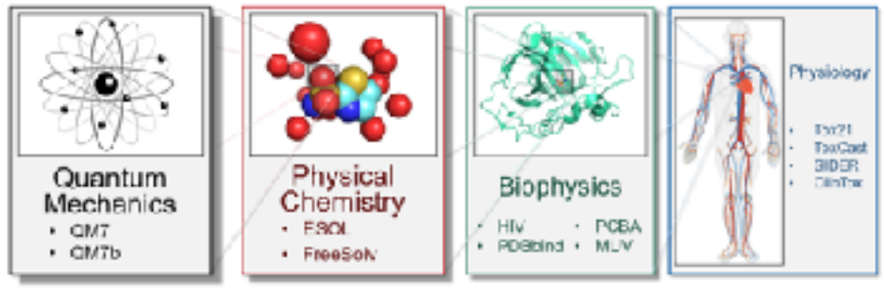
生命科学/医・薬・生物



化学(量子化学・触媒化学・生化学・有機化学)



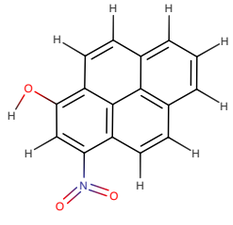
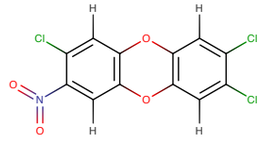
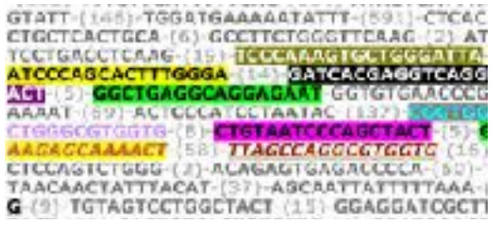
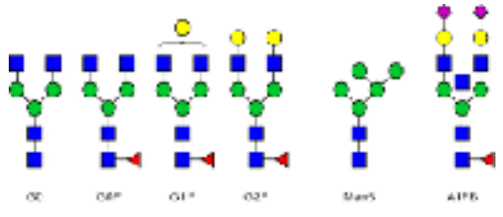
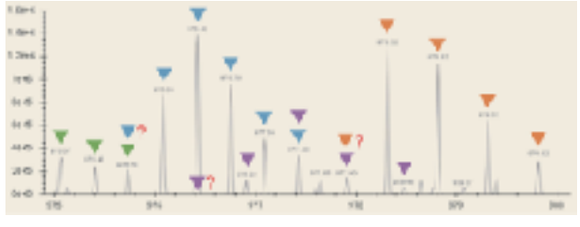
物質・材料科学



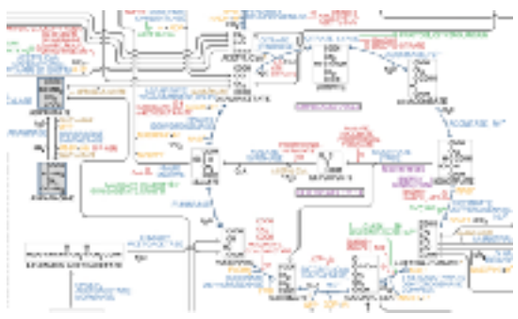
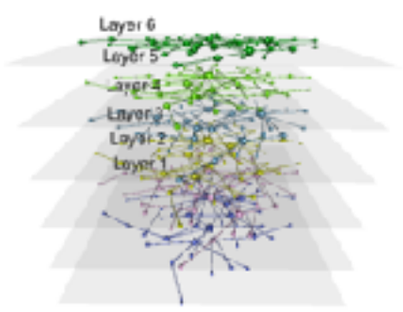
「離散構造」を伴う機械学習

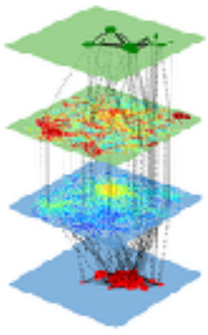
- **対象が「離散構造」を持つ**

論理、集合、関係、系列、組合せ、置換、分岐(木)、ネットワーク(グラフ)、代数系、…

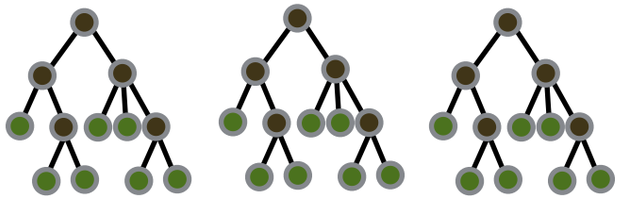
- **対象の関係が「離散構造」を持つ**

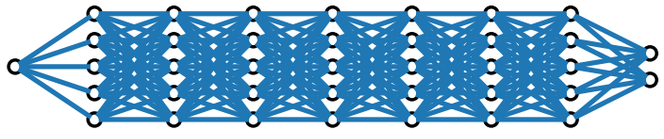



- **モデルが「離散構造」を持つ**

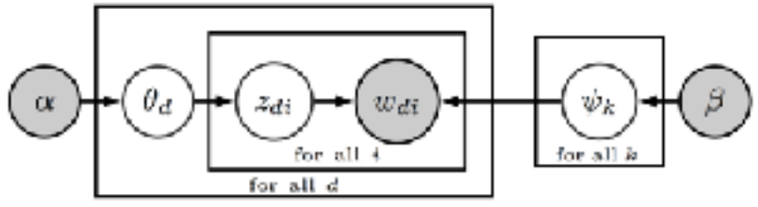
決定木・決定DAG



ニューラルネット



確率的プログラミング



北海道大学 化学反応創成研究拠点 (ICReDD)

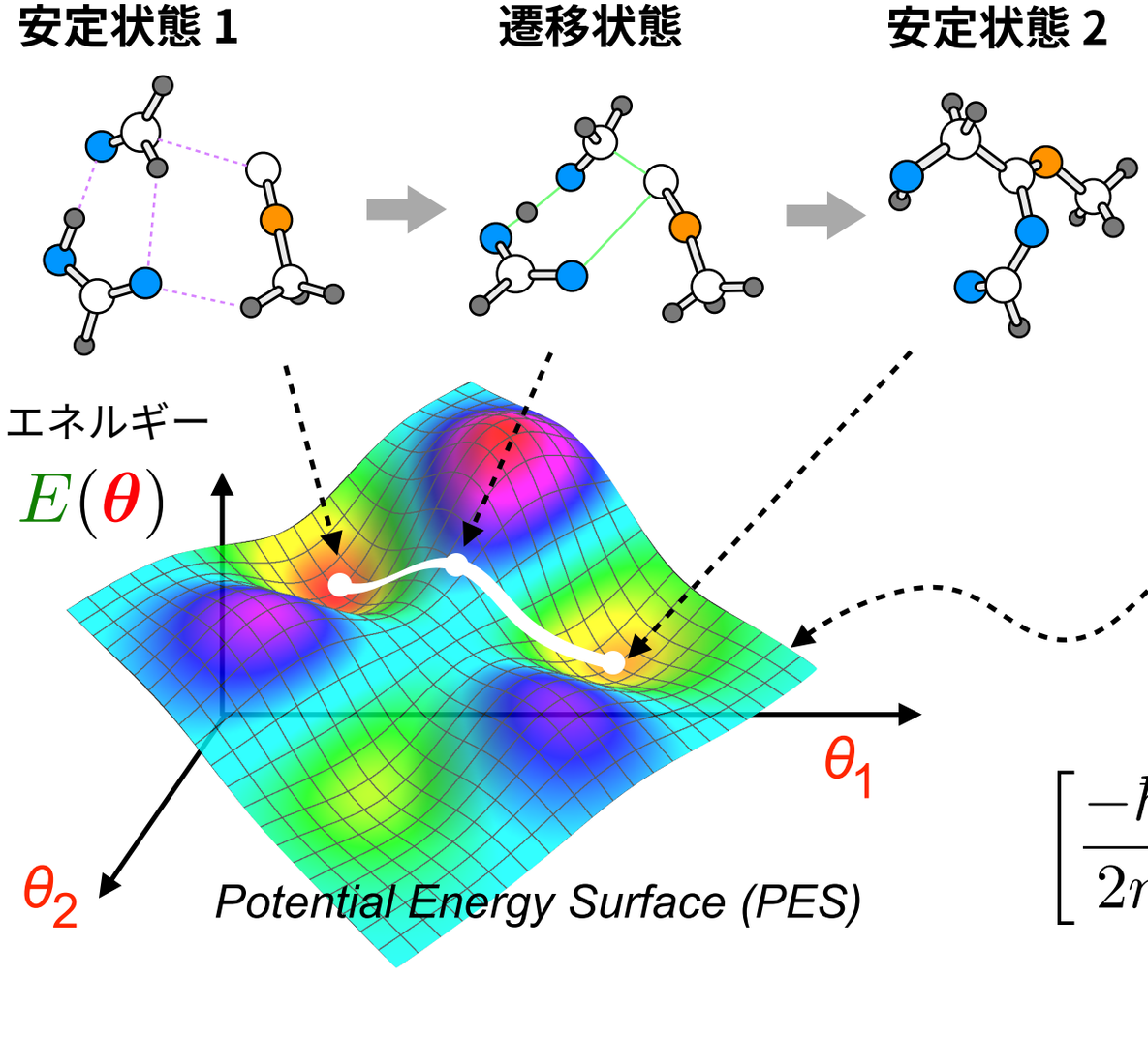
アイクレッド

Institute for Chemical Reaction Design and Discovery

2018年10月～ 文科省:世界トップレベル研究拠点(WPI)プログラム

World Premier International Research Center Initiative

物質の変換を司る化学反応の自在な設計と制御



- **入口出口は"離散組合せ的"対象**

  化合物 = 高々100数種類の元素
  (現在118種)がなす**膨大な**組合せ

- **化学反応は(自然法則が定める)**
  原子や結合の組み替え過程

エネルギーの谷→峠→谷の遷移だが
このエネルギー面 $E(\boldsymbol{\theta})$ を求めるには
毎点で電子に関する方程式の求解が必要

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

Schrödinger equation

展望：Theory-driven vs Data-drivenの解消と融合

参考) 人工知能の基本問題に関する総説(Open Access) http://id.nii.ac.jp/1004/00010296/

Theory-driven 【合理論】

- 対象現象の複雑化
- シミュレーション技法も複雑化
- "経験的に決める"パラメタや初期値
- 汎関数、交換相関項の設計

(人工知能分野)

- 知識ベースと論理推論(記号AI)の限界
- 厳密推論や探索の計算爆発(NP困難性)
- 大量データの知識化の問題
- 制約プログラミングや組合せ最適化

データ同化、模倣学習、論理合成、etc

モデルベース最適化、強化学習、メタ学習、ドメイン適応、生成モデル、etc

→ 新たな方法論へ？

Data-driven 【経験論】

- 小サンプル・低カウントの問題
- 外挿の不可能性の問題
- 帰納バイアスのモデルエンコード
- Blackbox性・解釈性の問題

(人工知能分野)

- Data-Driven手法(機械学習)と人間の論理的思考との大きなギャップ
- Dataがない領域の探索や「ひらめき」
- モデル適用範囲と信頼性・安全性

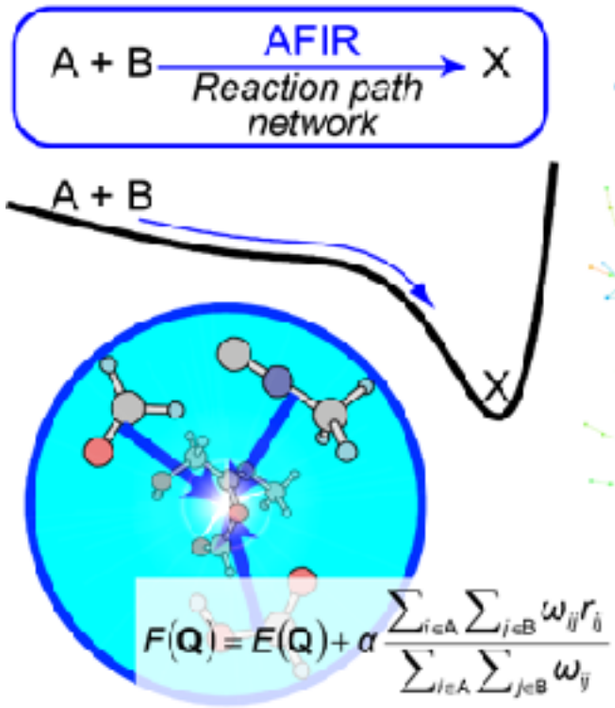
化学の分野で言うと...

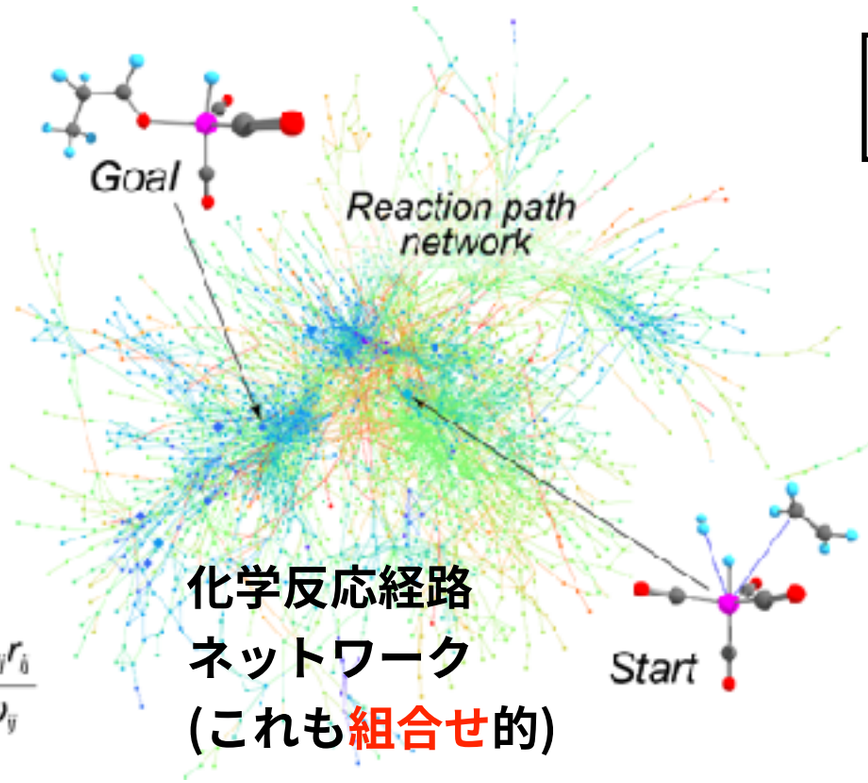
- 理論計算技術と計算機ハードの進歩により理論計算はかなり実践的な研究でも活用されるように。
- しかし主として実験的事実はスキルの高い実験化学者が極めて経験的に発見し、理論はその事実の解釈に活用されるのが現状
  - **通常こうした発見には実際の化学実験の長い「経験」が必要**
  - **当該分野の論文・テキストの経験的知見も混合された「勘」**

Key Question: 理論や計算は"化学的発見"を先導できるのか?

Takeaway: We also need 'data-driven' bridges!
first principles are not enough for us to throw away these *empirical* things; **data-driven approaches (ML)** play a complementary role!

計算科学・実験科学・情報科学による化学反応設計と探索





化学反応経路
ネットワーク
(これも組合せ的)

拠点のコア技術

"計算化学"
人工力誘起反応(AFIR)法による
(第一原理)反応経路自動探索

＋

(多様な)"実験化学"

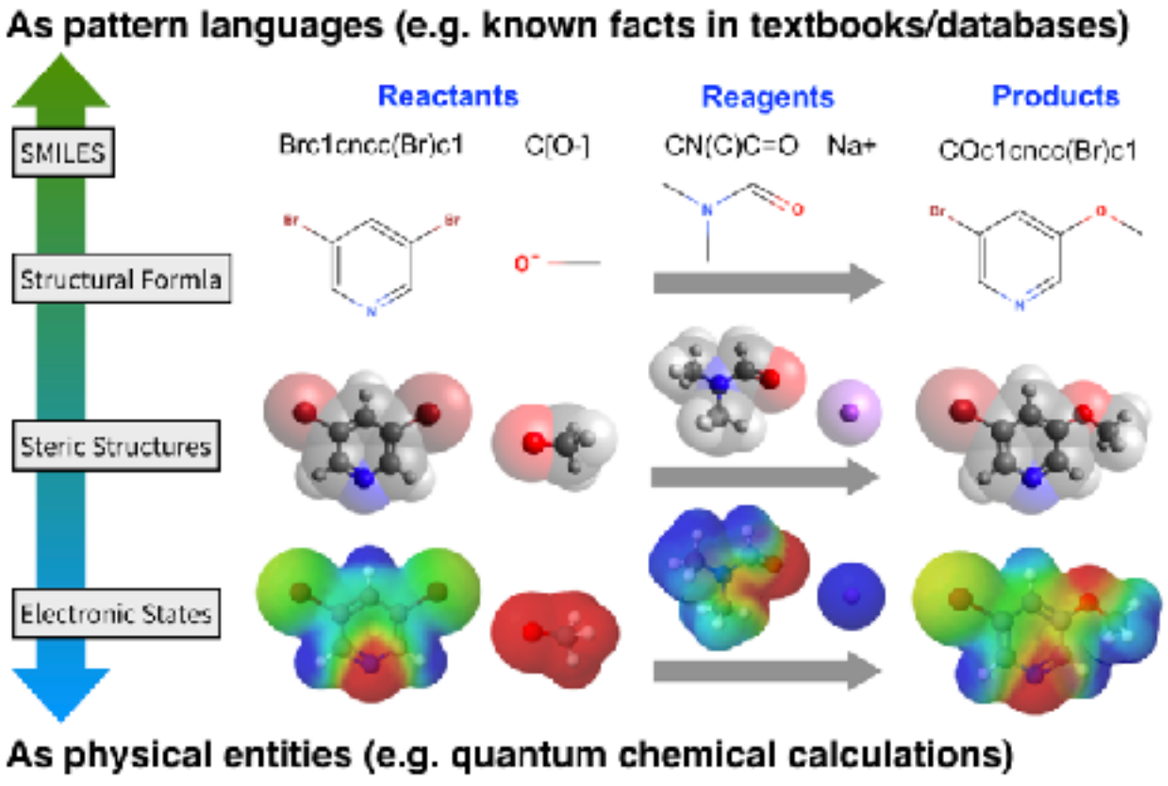
ただし理論計算予測だけでは難しい問題がいろいろある... → 「情報科学」への期待
(補完的な"第三の矢"？)

- 探索空間が組合せ的に巨大: 理論的に可能な経路の探索も組合せ爆発を起こす
- 計算時間・リソースが大きく計算できる系が限られる: 現実の系では何か妥協が必要
- 現実の化学反応の複雑さと不確定さ: 理論計算に入らない多様な要因が影響
- 現状の理論モデルの単純な仮定や不完全さ: 多体問題の近似や理論の例外の存在

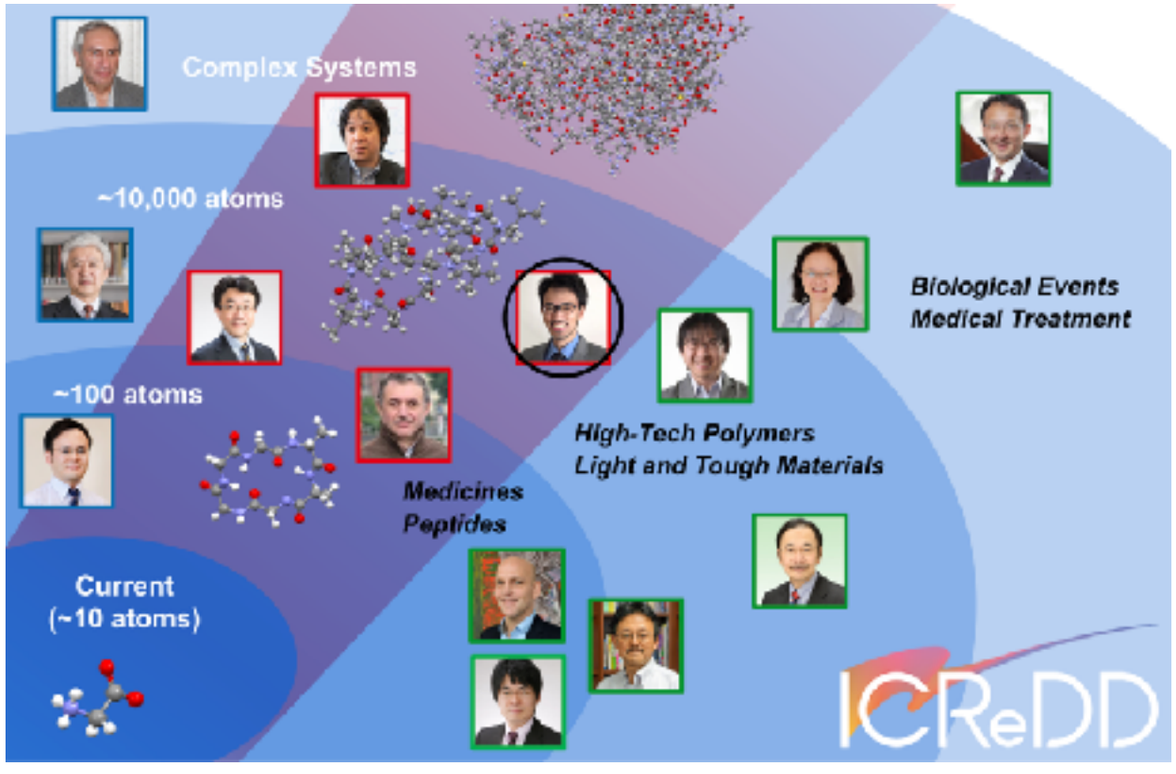
化学反応の予測

化学反応をどのような表現・レベルで捉えるか？

１分子レベルでの表現の多様性



量子・電子系 ～ 複雑系・生体系



① Theory-driven (Quantum Chem)

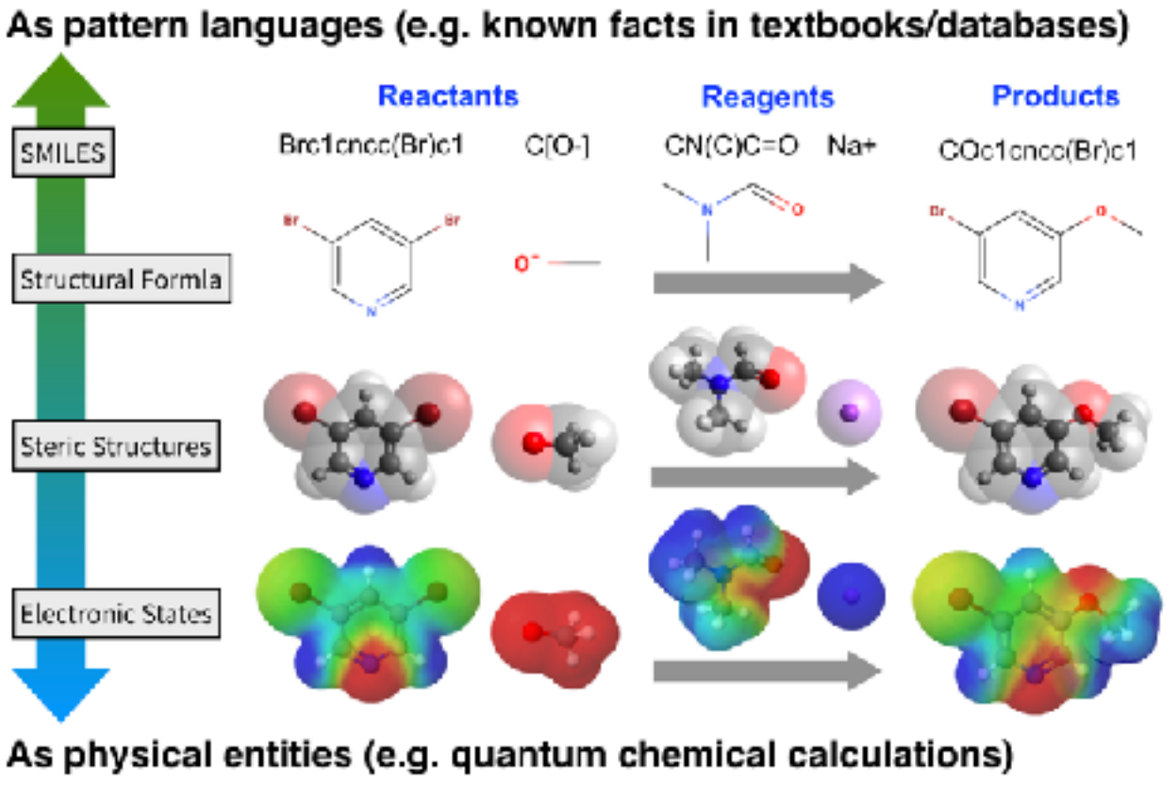
② Knowledge-driven (Knowledge Bases)

③ Data-driven (Machine Learning)

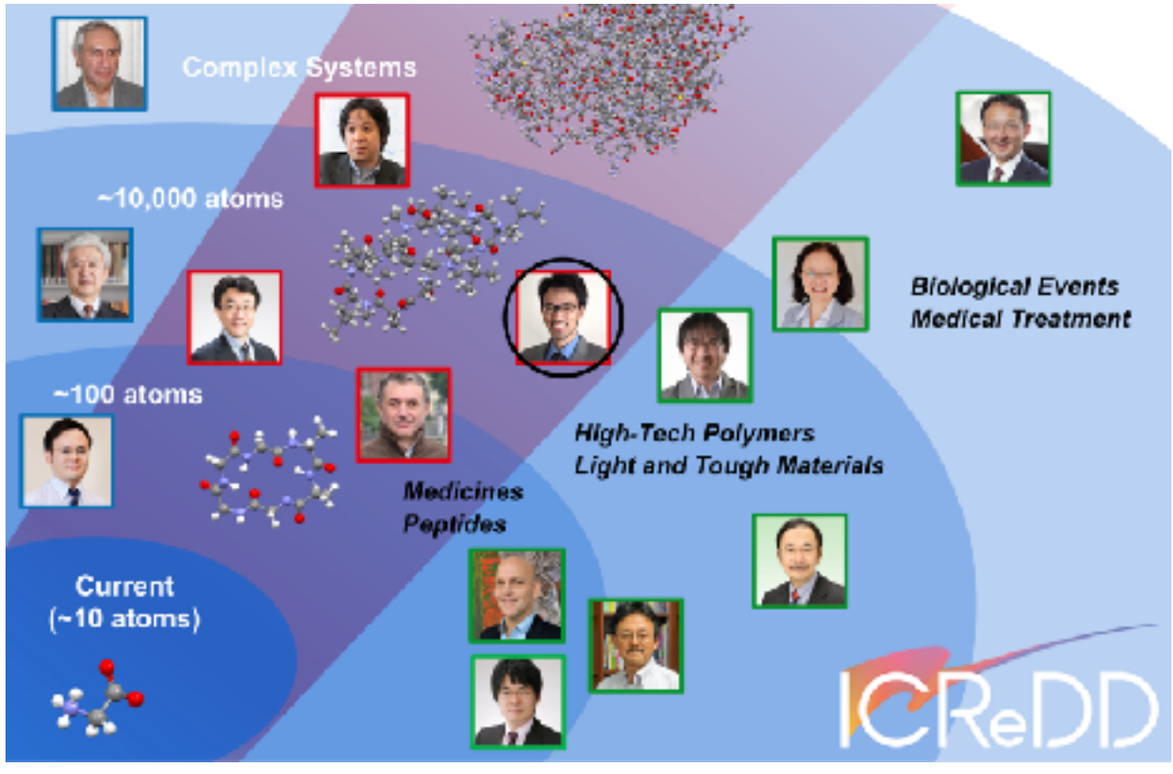
化学反応の予測

化学反応をどのような表現・レベルで捉えるか？

１分子レベルでの表現の多様性



量子・電子系 ～ 複雑系・生体系



① Theory-driven (Quantum Chem)

② Knowledge-driven (Knowledge Bases)

③ Data-driven (Machine Learning)

A traditional topic in chemoinformatics

Computer-assisted synthetic planning
(path search on knowledge bases)

or AI-Assisted Synthesis?
(with Machine Learning)

Reaxys® SciFinder® A CAS SOLUTION CHEMATICA

Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101
German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

*Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski**

Angewandte International Edition Chemie

Angew. Chem. Int. Ed. **2016**, 55, 5904 – 5937

VIP *AI-Assisted Synthesis* *Very Important Paper*

International Edition: DOI: 10.1002/anie.201912083
German Edition: DOI: 10.1002/ange.201912083

Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

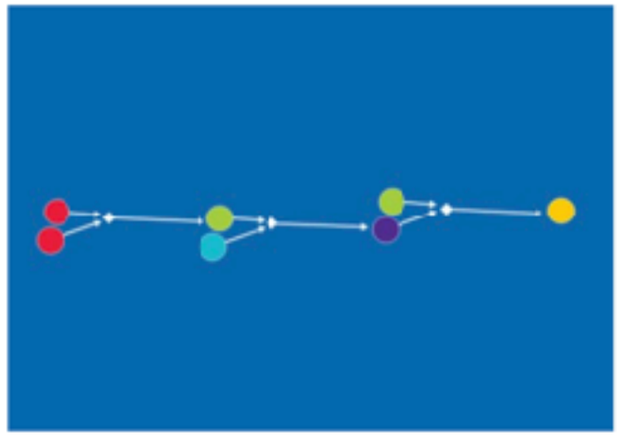
*Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski**

Angew. Chem. Int. Ed. **2019**, *58*, 1 – 7

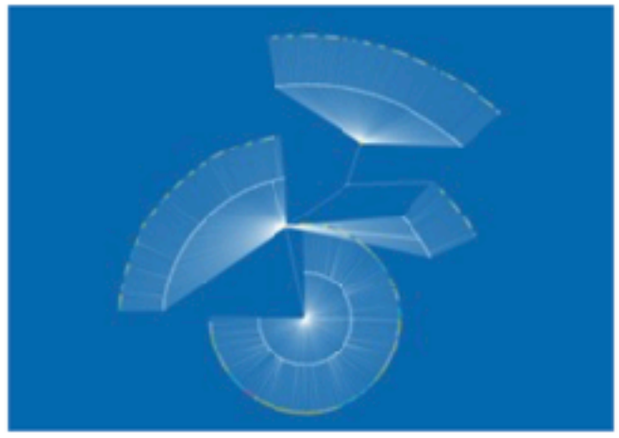
CHEMATICA

Synthia™
Retrosynthesis Software

Three design modules to amplify your approach:

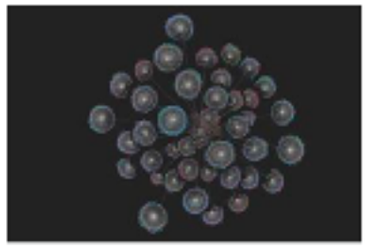


Automatic Retrosynthesis



Manual Retrosynthesis

Network of Organic Chemistry



RESEARCH

Program finds 5 million synthetic routes to complex chemicals

8 JANUARY 2020

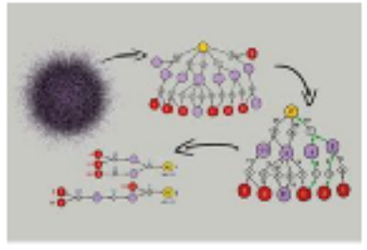
Only around five hundred 'tactical combinations' for advanced organic synthesis existed – until Chematica was let loose on the problem

RESEARCH

Retrosynthetic algorithm broadened to design similar, but different, molecules

26 AUGUST 2019

Chematica can now design efficient syntheses for large compound libraries



RESEARCH

Synthesis-searching software's superior scoring sharpens selections

12 MARCH 2019

Realistic costs and diverse suggestions make Chematica more insightful

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2019, **10**, 4640

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 16th December 2018
Accepted 24th February 2019

DOI: 10.1039/c8sc05611k

rsc.li/chemical-science

Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans†

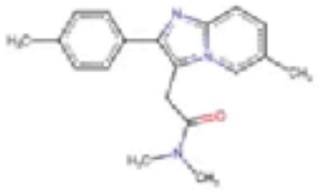
Tomasz Badowski,‡[a] Karol Molga ‡[a] and Bartosz A. Grzybowski *[ab]

As the programs for computer-aided retrosynthetic design come of age, they are no longer identifying just one or few synthetic routes but a multitude of chemically plausible syntheses, together forming large, directed graphs of solutions. An important problem then emerges: how to select from these graphs and present to the user manageable numbers of top-scoring pathways that are cost-effective, promote convergent vs. linear solutions, and are chemically diverse so that they do not repeat only minor variations in the same chemical theme. This paper describes a family of reaction network algorithms that address this problem by (i) using recursive formulae to assign realistic prices to individual pathways and (ii) applying penalties to chemically similar strategies so that they are not dominating the top-scoring routes. Synthetic examples are provided to illustrate how these algorithms can be implemented – on the timescales of ~1 s even for large graphs – to rapidly query the space of synthetic solutions under the scenarios of different reaction yields and/or costs associated with performing reaction operations on different scales.

CHEMATICA

Synthia™
Retrosynthesis Software

ALL | FAVORITES | PENDING | RECENT



zolpidem

Algorithm: manual retrosynthesis

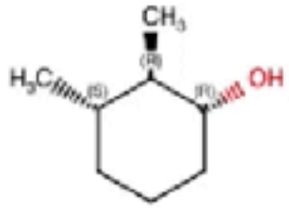
Computations:
Successfully completed - 2

SMILES: Cc1ccc(-c2nc3ccc(C)cn3c2CC(=O)N(C)C)cc1

Copy SMILES to clipboard

6/1/2018, 12:03 PM ☆ FAVORITE

Absolute



trans2-cis3-dimethyl-cyclohexanol

Algorithm: syntaurus

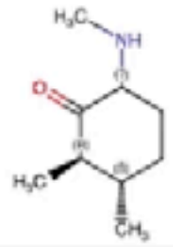
Computations:
Successfully completed - 1

SMILES: C[C@H]1[C@H](O)CCC[C@@H]1C

Copy SMILES to clipboard

4/11/2018, 10:20 AM ☆ FAVORITE

Absolute



(re)Analysis 1678

Algorithm: syntaurus

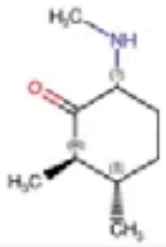
Computations:
Successfully completed - 1

SMILES: CNC1CC[C@H](C)[C@@H](C)C1=O

Copy SMILES to clipboard

4/11/2018, 10:01 AM ☆ FAVORITE

Absolute



Analysis 1678

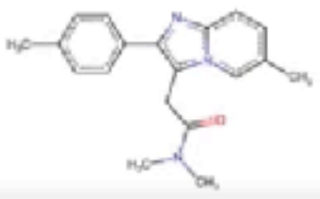
Algorithm: syntaurus

Computations:
Successfully completed - 1

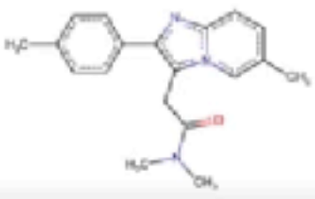
SMILES: CNC1CC[C@H](C)[C@@H](C)C1=O

Copy SMILES to clipboard

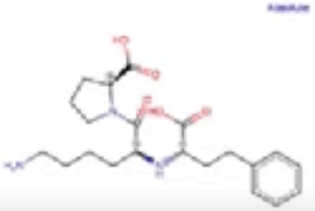
4/11/2018, 9:53 AM ☆ FAVORITE



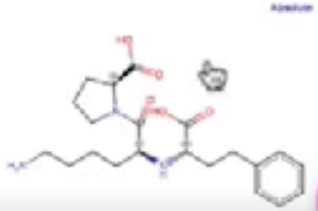
(re)zolpidem



zolpidem



(re)lisinopril



lisinopril

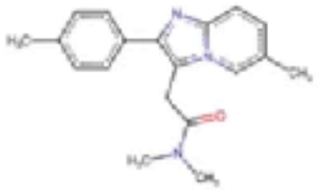
0:00 / 1:15

720p

CHEMATICA

Synthia™
Retrosynthesis Software

ALL | FAVORITES | PENDING | RECENT



zolpidem

Algorithm: manual retrosynthesis

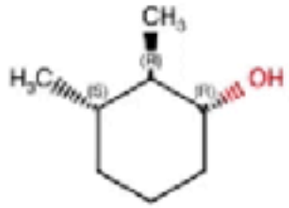
Computations:
Successfully completed - 2

SMILES: Cc1ccc(-c2nc3ccc(C)cn3c2CC(=O)N(C)C)cc1

Copy SMILES to clipboard

6/1/2018, 12:03 PM ☆ FAVORITE

Absolute



trans2-cis3-dimethyl-cyclohexanol

Algorithm: syntaurus

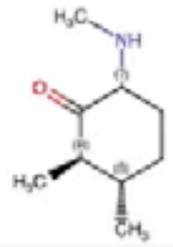
Computations:
Successfully completed - 1

SMILES: C[C@H]1[C@H](O)CCC[C@@H]1C

Copy SMILES to clipboard

4/11/2018, 10:20 AM ☆ FAVORITE

Absolute



(re)Analysis 1678

Algorithm: syntaurus

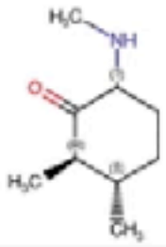
Computations:
Successfully completed - 1

SMILES: CNC1CC[C@H](C)[C@@H](C)C1=O

Copy SMILES to clipboard

4/11/2018, 10:01 AM ☆ FAVORITE

Absolute



Analysis 1678

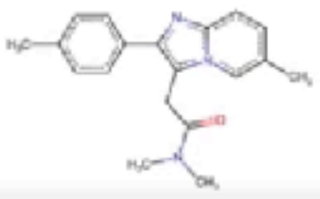
Algorithm: syntaurus

Computations:
Successfully completed - 1

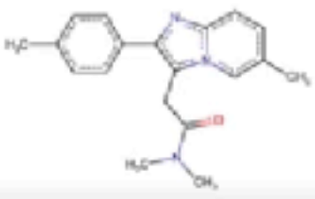
SMILES: CNC1CC[C@H](C)[C@@H](C)C1=O

Copy SMILES to clipboard

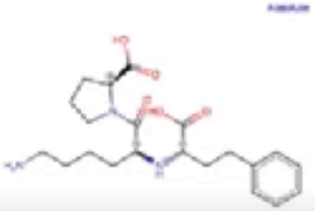
4/11/2018, 9:53 AM ☆ FAVORITE



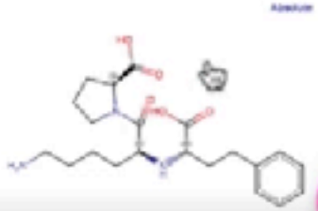
(re)zolpidem



zolpidem



(re)lisinopril



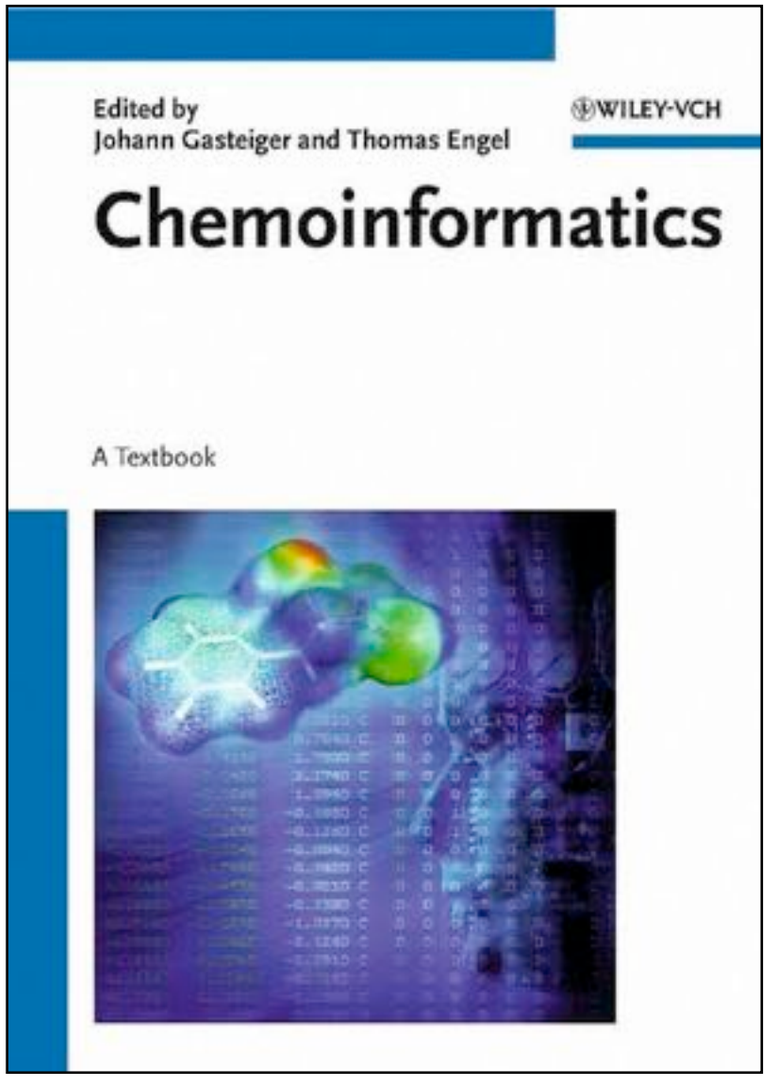
lisinopril

0:00 / 1:15

720p

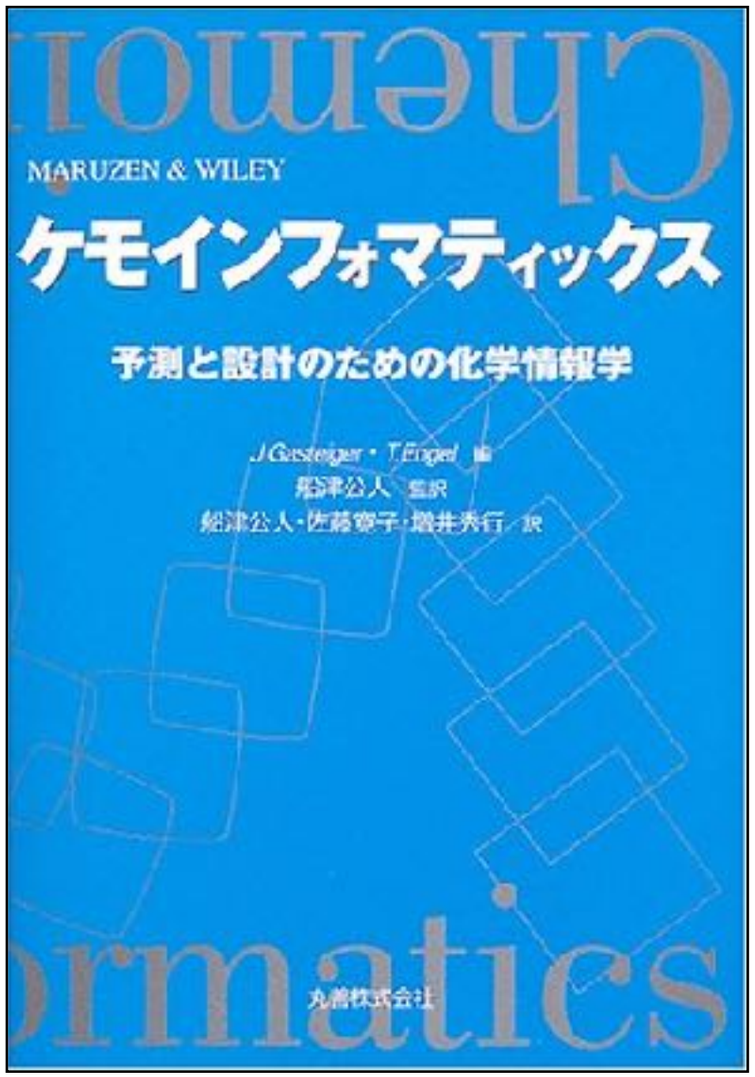
2000年代初頭にChemoinformatics分野が(ある程度)確立

Gasteiger & Engel



2003

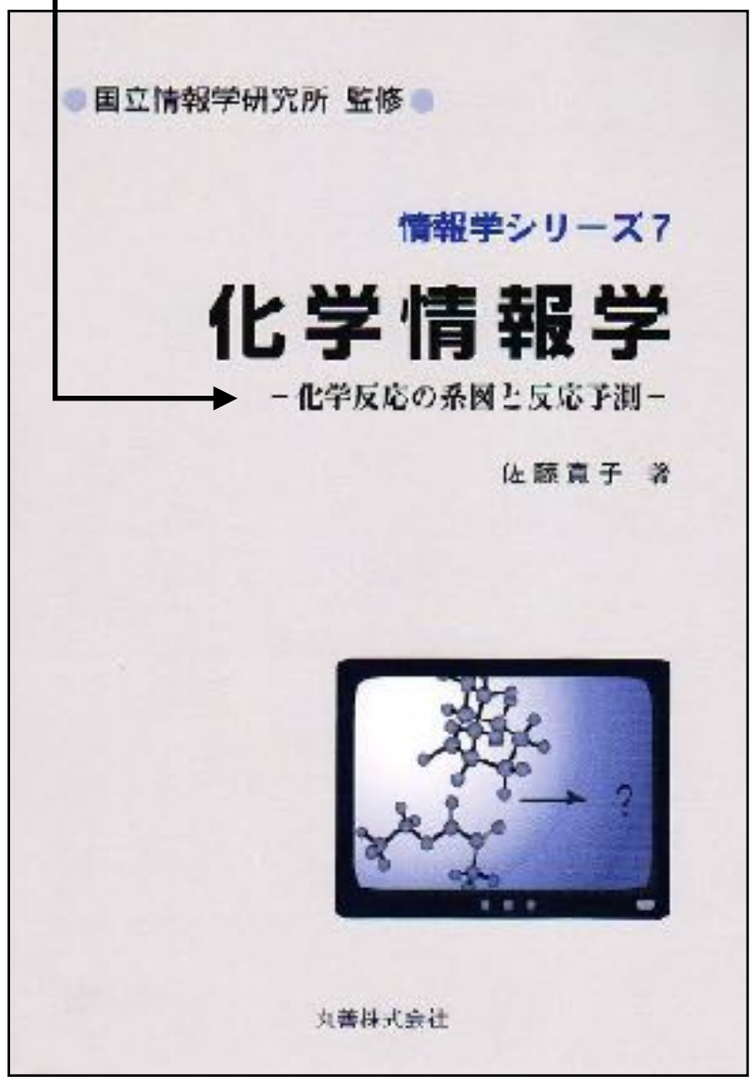
船津・佐藤・増井 訳



2005

化学反応の系図と反応予測

佐藤



2003

70年代から続く長い歴史

化学反応をデータベース化して反応経路を探索(検索)したい

e.g.) 化学反応の表現、化学経路の分類、化学反応の探索、…

Corey+ 1972

J Am Chem Soc (JACS), 94(2), 1972.

Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates

E. J. Corey,* Richard D. Cramer III, and W. Jeffrey Howe

Contribution from the Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138. *Received January 30, 1971*

Abstract: A classification of synthetic reactions is outlined which is suitable for use in a machine program to generate a tree of synthetic intermediates starting from a given target molecule. The generation of a particular intermediate by the program involves the search of appropriate data tables of synthetic processes, the search being driven by the information obtained by machine perception of the parent structure and certain basic strategies. Procedures have been developed for the evaluation of chemical interconversions which allow the effective exclusion of invalid or naive structures. The paper provides a view of the status of computer-assisted synthetic problem solving as of 1970.

The communication of chemical structural information to and from a digital computer by graphical methods has been discussed in detail in a foregoing paper,[1] as has the machine representation and perception of key features within structures,[2] as for example, functional groups and rings. This paper is concerned with the ways in which the structural information made available by the perception process can be utilized to and necessary control strategies, and also for eventual inclusion of a fairly complete collection of families. In the discussion which follows, the degree of implementation of each area of study will be cited.

A variety of rational schemes for creating families of synthetic reactions already exists. However, most of these depend on properties of the reactants,[5] and as such they are irrelevant to a computer program which

Ugi+ 1993

Angew Chem Int Ed Engl. 32, 202-227, 1993.

Computer-Assisted Solution of Chemical Problems— The Historical Development and the Present State of the Art of a New Discipline of Chemistry

By Ivar Ugi,* Johannes Bauer, Klemens Bley, Alf Dengler, Andreas Dietz, Eric Fontain, Bernhard Gruber, Rainer Herges, Michael Knauer, Klaus Reitsam, and Natalie Stein

Dedicated to Professor Karl-Heinz Büchel

The topic of this article is the development and the present state of the art of computer chemistry, the computer-assisted solution of chemical problems. Initially the problems in computer chemistry were confined to structure elucidation on the basis of spectroscopic data, then programs for synthesis design based on libraries of reaction data for relatively narrow classes of target compounds were developed, and now computer programs for the solution of a great variety of chemical problems are available or are under development. Previously it was an achievement when any solution of a chemical problem could be generated by computer assistance. Today, the main task is the efficient, transparent, and non-arbitrary selection of meaningful results from the immense set of potential solutions—that also may contain innovative proposals. Chemistry has two aspects, constitutional chemistry and stereochemistry,

Gasteigerの主関心が化学反応経路の予測や合成支援(たぶん)

教科書の「反応」関連の章もGasteiger自身が執筆

Gasteiger+ 1990

Analytica Chimica Acta, 235 (1990) 65–75
Elsevier Science Publishers B.V., Amsterdam – Printed in The Netherlands

Computer-assisted reaction prediction and synthesis design

J. GASTEIGER *, W D. IHLENFELDT, P ROSE and R. WANKE
Institute of Organic Chemistry, Technical University Munich, D-8046 Garching (F R G)

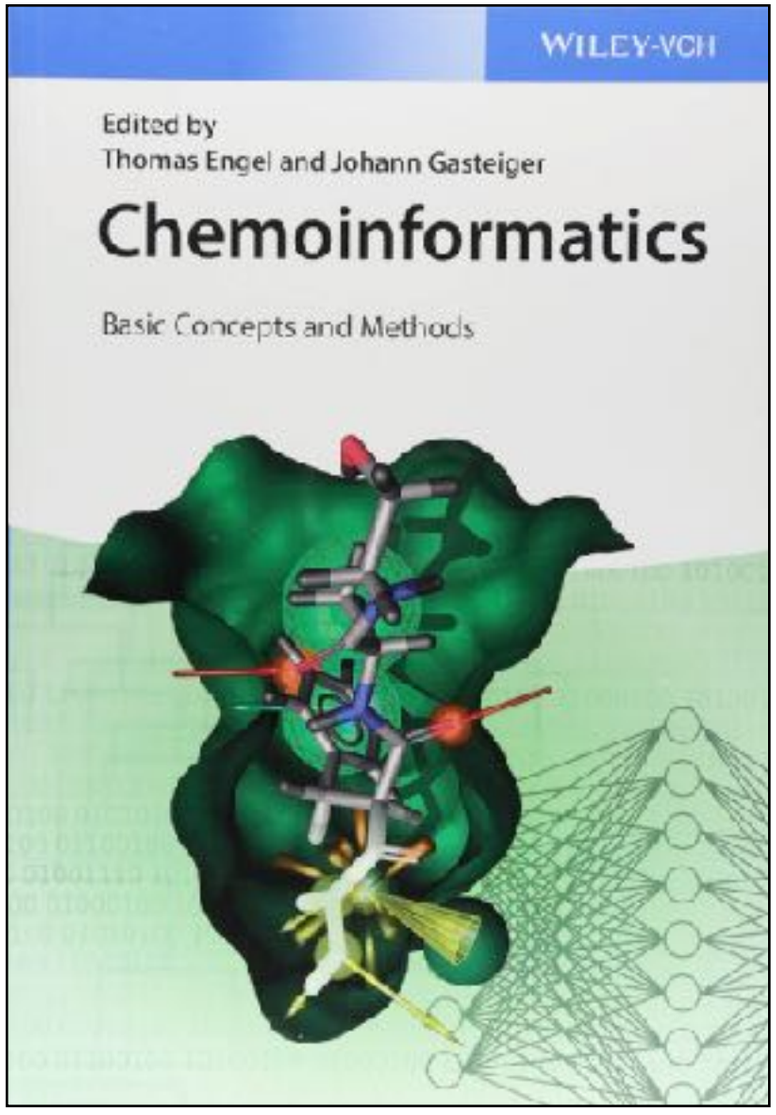
(Received 18th December 1989)

ABSTRACT

The design of organic syntheses requires deep insight into chemical reactivity. Methods have been developed to calculate important electronic and energy effects and use them for the modelling of reaction mechanisms The approach is illustrated by the haloform reaction. The procedures have been incorporated into different versions of the EROS system. A study of the synthesis of fredericamycin illustrates the use of these EROS procedures in synthesis design and reaction prediction. Recent work on synthesis design for aromatic compounds is highlighted and new definitions of the similarity of chemical compounds are discussed.

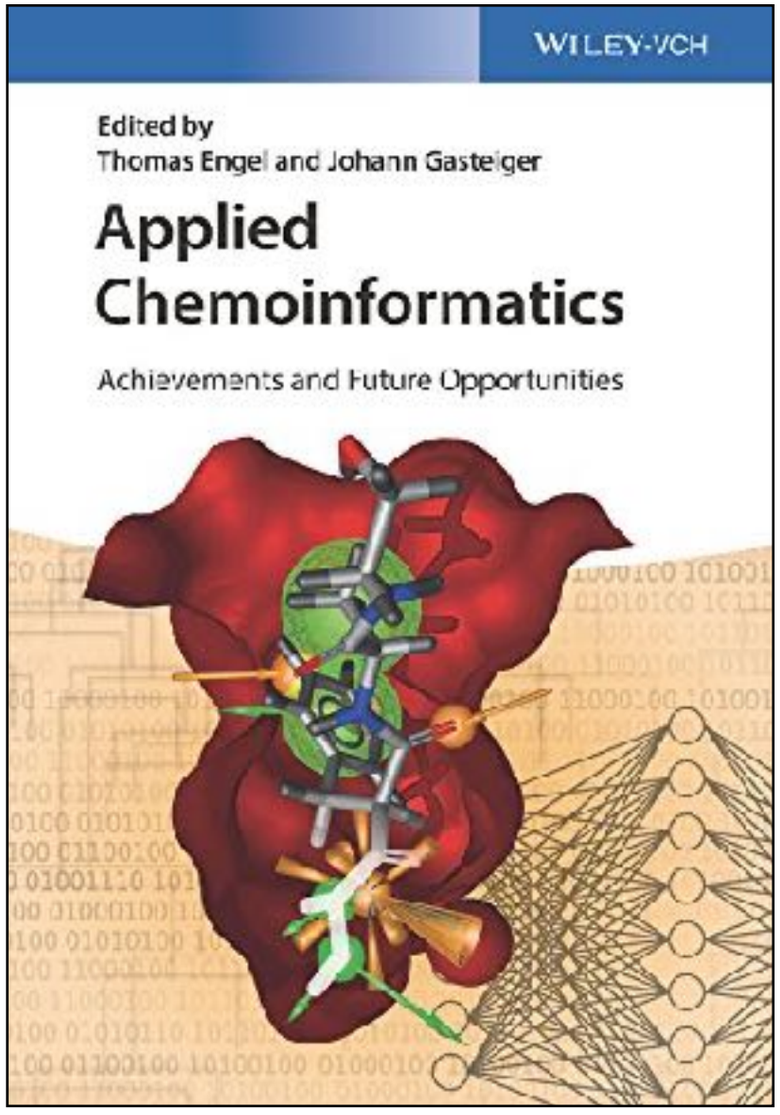
最近またChemoinformaticsがリバイバル?

Engel & Gasteiger



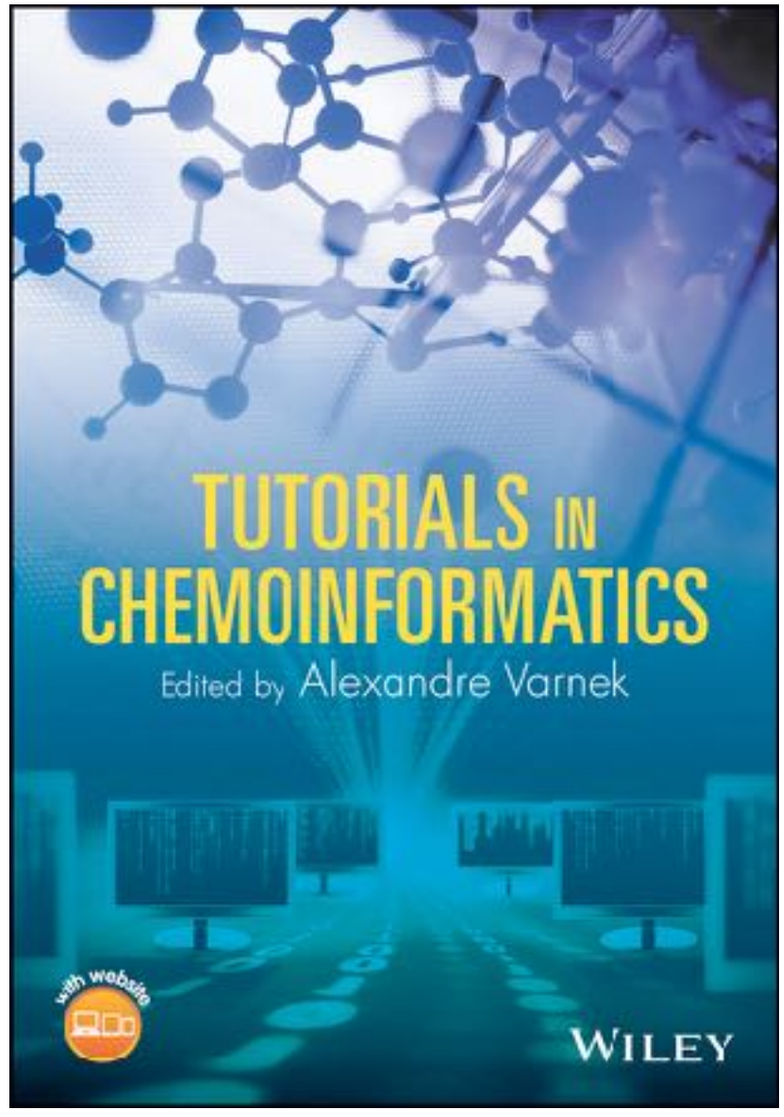
2018

Engel & Gasteiger



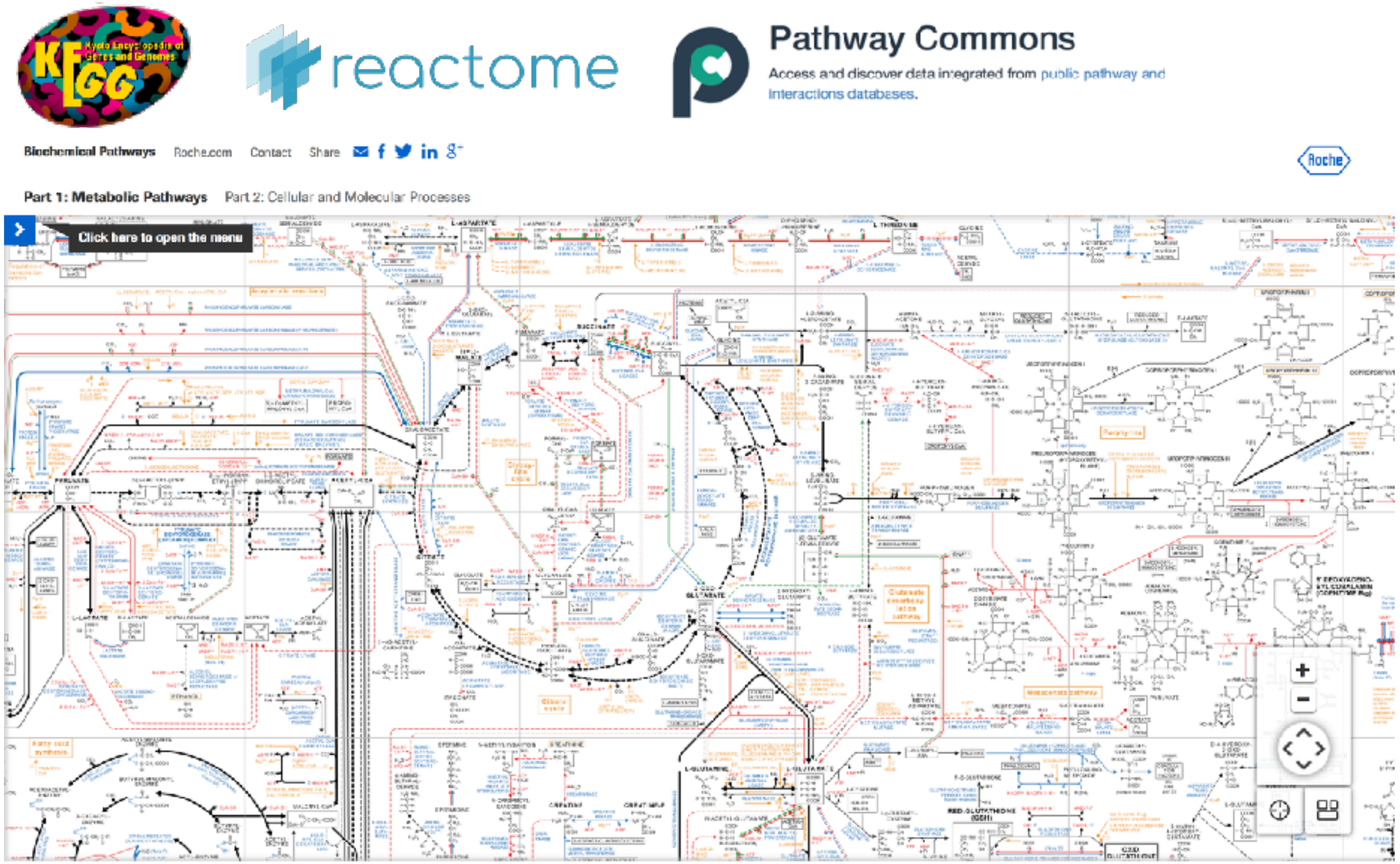
2018

Varnek



2017

事例) Life Scienceでの代謝反応系とゲノム情報



http://biochemical-pathways.com/#/map/1

KEGG REACTION: R00259

Help

Entry	R00259 Reaction
Name	acetyl-CoA:L-glutamate N-acetyltransferase
Definition	Acetyl-CoA + L-Glutamate <=> CoA + N-Acetyl-L-glutamate
Equation	C00024 + C00025 <=> C00010 + C00624
Reaction class	RC00004 C00010_C00024 RC00064 C00025_C00624
Enzyme	2.3.1.1
Pathway	rn00220 Arginine biosynthesis rn01100 Metabolic pathways rn01110 Biosynthesis of secondary metabolites rn01130 Biosynthesis of antibiotics rn01210 2-Oxocarboxylic acid metabolism rn01230 Biosynthesis of amino acids
Module	M00028 Ornithine biosynthesis, glutamate => ornithine M00845 Arginine biosynthesis, glutamate => acetylcitrulline => arginine
Orthology	KO0618 amino-acid N-acetyltransferase [EC:2.3.1.1] KO0619 amino-acid N-acetyltransferase [EC:2.3.1.1] KO0620 glutamate N-acetyltransferase / amino-acid N-acetyltransferase [EC:2.3.1.35 2.3.1.1] K11067 N-acetylglutamate synthase [EC:2.3.1.1] K14681 argininosuccinate lyase / amino-acid N-acetyltransferase [EC:4.3.2.1 2.3.1.1]

Arginine biosynthesis

[Pathway menu | Organism menu | Pathway entry | Download KGML | User data mapping]

Reference pathway (Reaction) 100%

ARGININE BIOSYNTHESIS

KEGG PATHWAY

KEGG COMPOUND

KEGG REACTION

KEGG REACTION Database

Knowledge base for predicting biodegradation and biosynthesis

Menu PATHWAY BRITE LIGAND COMPOUND GLYCAN REACTION ENZYME RModule

Search REACTION for Go

View chemical structure transformation pattern (see examples)

Reactant 1: C/D number or MOL file ファイルを選択 ファイル未選択
Reactant 2: C/D number or MOL file ファイルを選択 ファイル未選択
Go Clear

Chemical Reactions

KEGG REACTION is a database of chemical reactions, mostly enzymatic reactions, containing all reactions that appear in the KEGG metabolic pathway maps and additional reactions that appear only in the Enzyme Nomenclature. Each reaction is identified by the R number, such as R00259 for the acetylation of L-glutamate. Reactions are linked to enzyme KOs as defined by the KO database, enabling integrated analysis genomic (enzyme genes) and chemical (compound pairs) information.

- Enzymatic reactions
- IUBMB reaction hierarchy

PathSearch: Search Similar Reaction Paths

PathSearch PathComp PathPred KEGG2

About PathSearch

PathSearch computes similar paths for a query reaction path using the RCLASS database for reaction patterns matching to the query reactions. The results can be mapped to KEGG pathway diagrams using KEGG Mapper. The colors of the objects can be also specified (see Ex. 2 below).

Enter reaction path:

Ex.1 List of Reaction IDs or Rpair IDs.
R00089 R00191 R03004

Ex.2 Reaction or Rpair IDs followed by bgcolor, fgcolor.
R02740 orange,blue
R04779 purple,yellow
R01070 magenta

Ex.3 PathComp output can be used.
C00345 <R02740> C05345 <R04779> C05378 <R01070> C00111

Exec Clear

Use auto bgcolor.
Skip repeated reaction.

PathPred: Pathway Prediction server

PathSearch PathComp PathPred KEGG2

About PathPred

PathPred is a web-based server to predict plausible enzyme-catalyzed reaction pathways from a query compound using the information of RDM patterns and chemical structure alignments of substrate-product pairs. This server provides plausible reactions and transformed compounds, and displays all predicted reaction pathways in tree-shaped graph.

- PathPred help

Compute Clear

Reference pathway:

Xenobiotics Biodegradation (Bacteria)

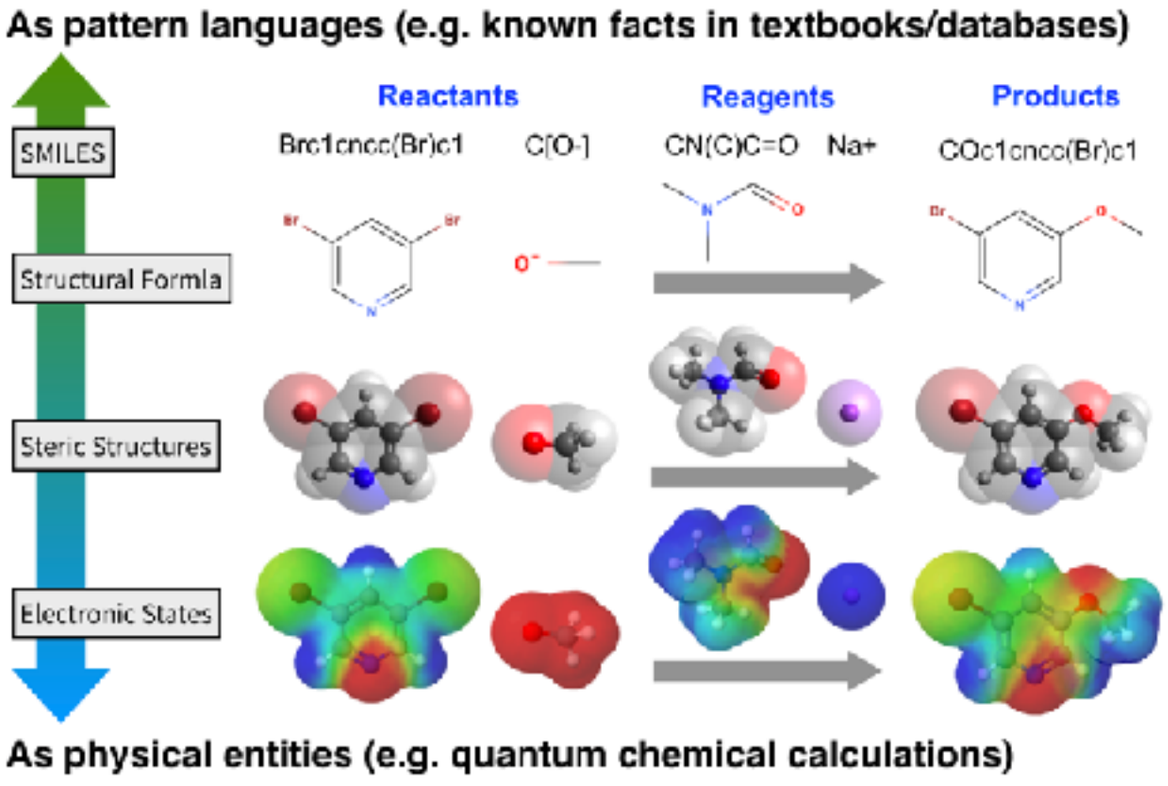
Enter initial compound: (in one of the four forms)

KEGG Compound ID (Ex) C06594 View structure
MOL File Name ファイルを選択 ファイル未選択
MOL File Text

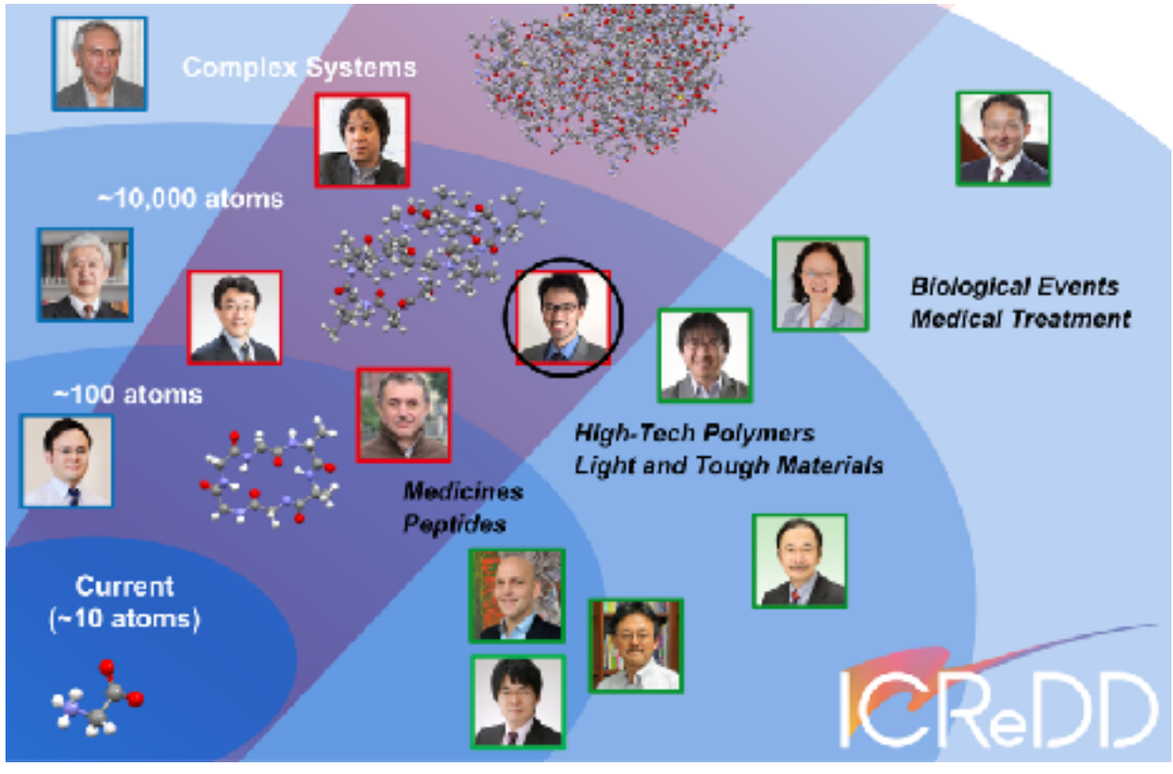
化学反応の予測

化学反応をどのような表現・レベルで捉えるか？

１分子レベルでの表現の多様性



量子・電子系 ～ 複雑系・生体系



① Theory-driven (Quantum Chem)

② Knowledge-driven (Knowledge Bases)

③ Data-driven (Machine Learning)

Automated reaction-path search via GRRM strategy

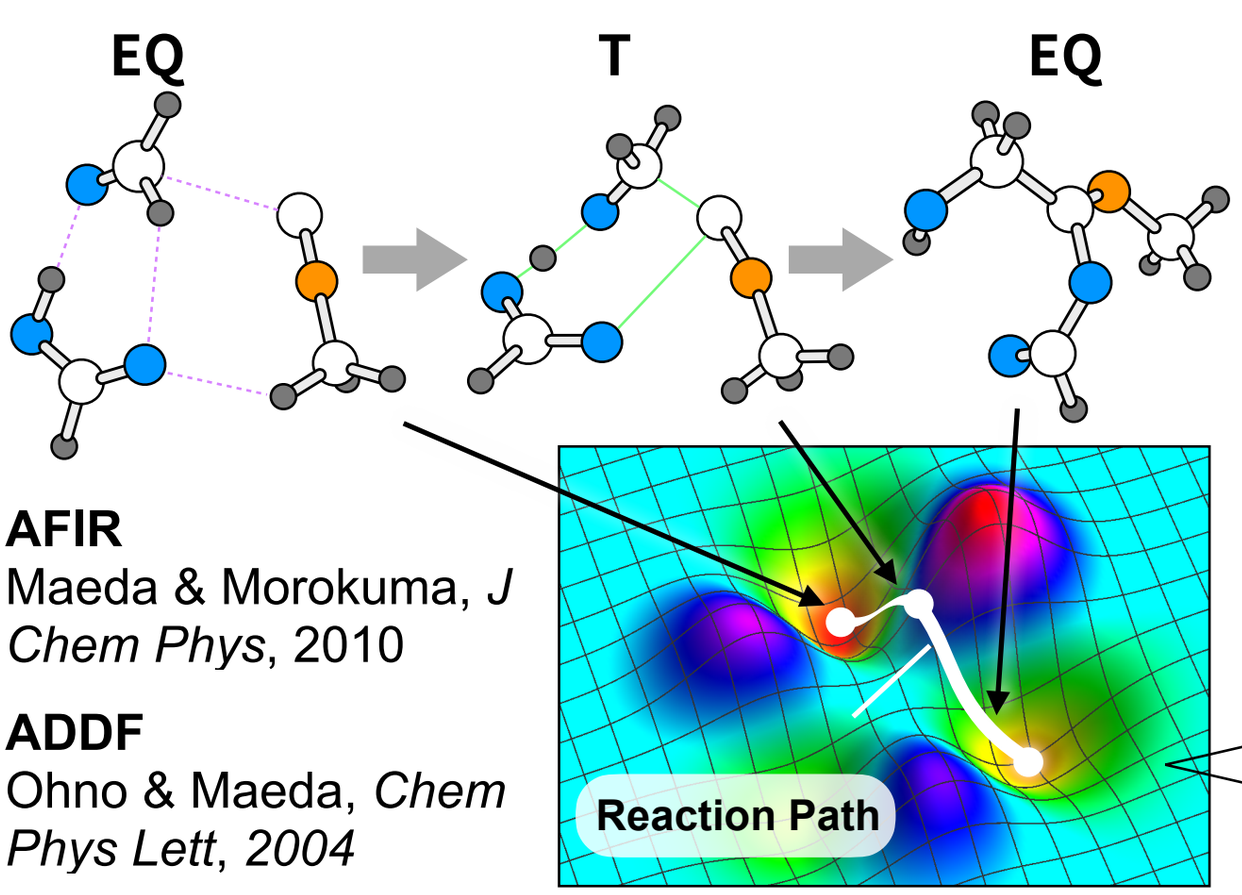
PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683

Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods

Satoshi Maeda,*[a] Koichi Ohno*[b] and Keiji Morokuma*[cd]

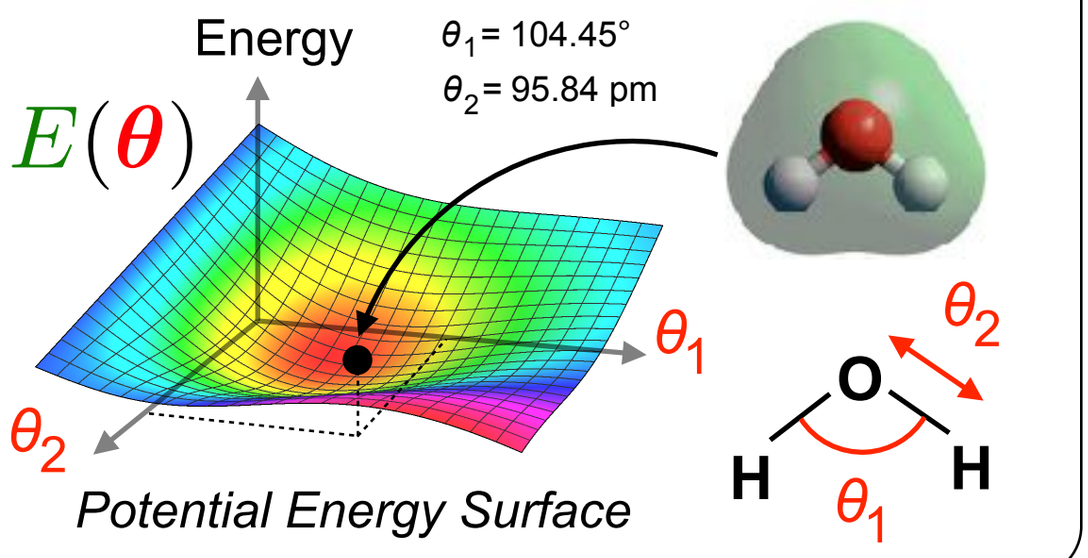


AFIR
Maeda & Morokuma, *J Chem Phys*, 2010

ADDF
Ohno & Maeda, *Chem Phys Lett*, *2004*

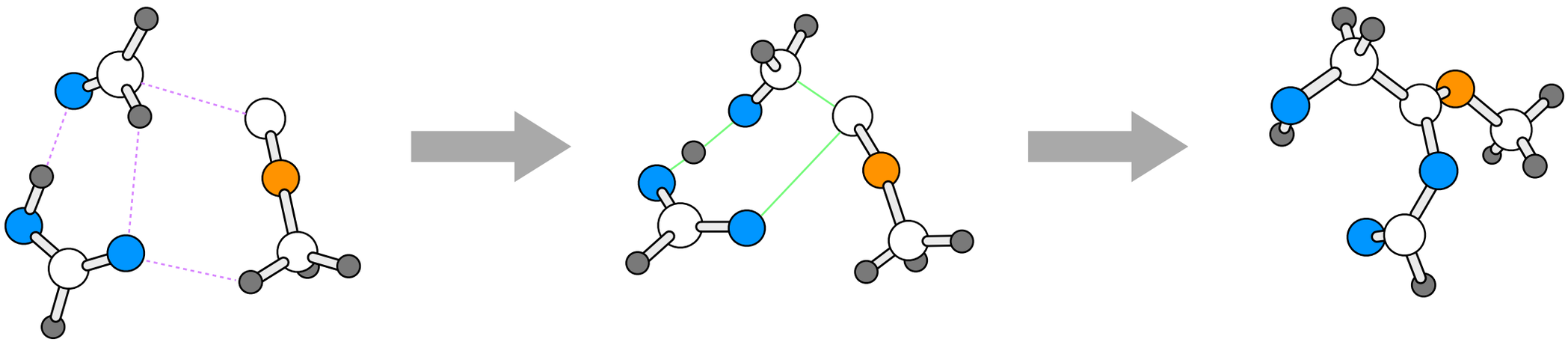
Schrödinger equation

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



But computational chemistry has limitations for now...

Chemical reactions = recombinations of atoms and chemical bonds subjected to *the laws of nature*

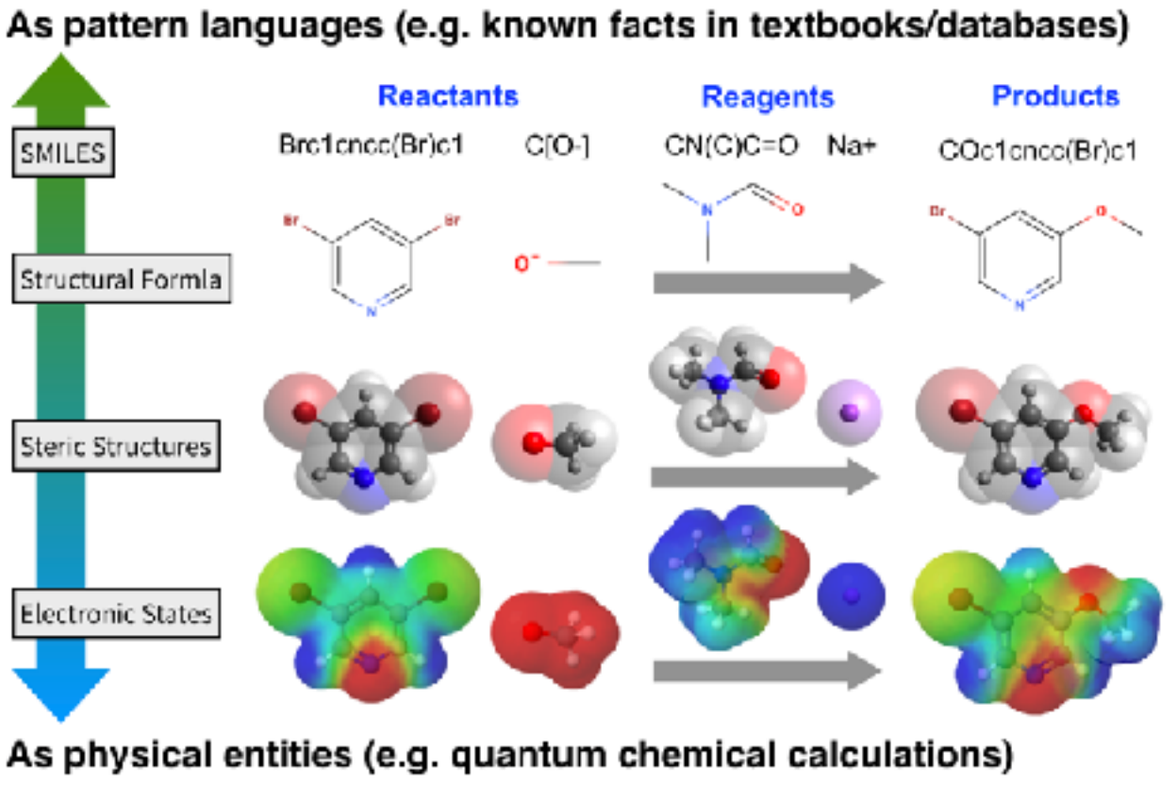


- **Intractably large chemical space**: A intractably large number of "theoretically possible" candidates for reactions and compounds...
- **Scalability issue**: Simulating an Avogadro-constant number of atoms is utterly infeasible... (After all, we need some compromise here)
- **Complexity and uncertainty of real-world systems**:
  Many uncertain factors and arbitrary parameters are involved...
- **Known and unknown imperfections of currently established theories**:
  Current theoretical calculations have many exceptions and limitations...

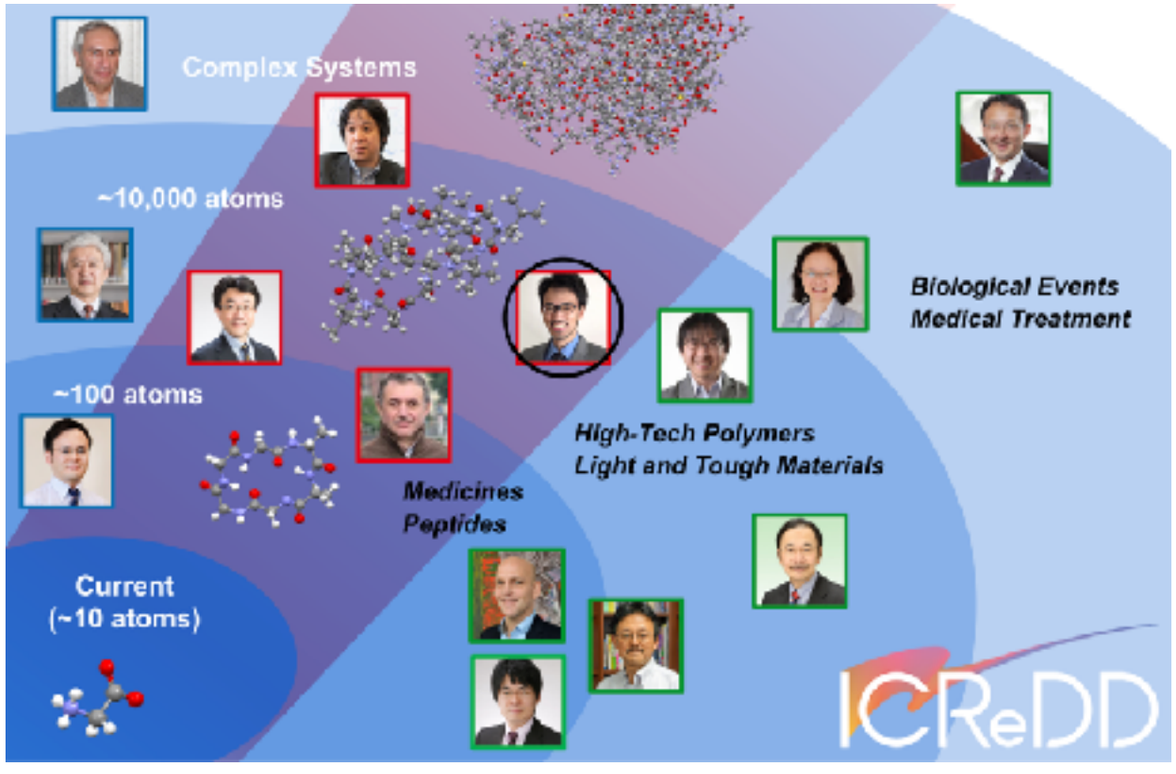
化学反応の予測

化学反応をどのような表現・レベルで捉えるか？

１分子レベルでの表現の多様性



量子・電子系 ～ 複雑系・生体系



① Theory-driven (Quantum Chem)

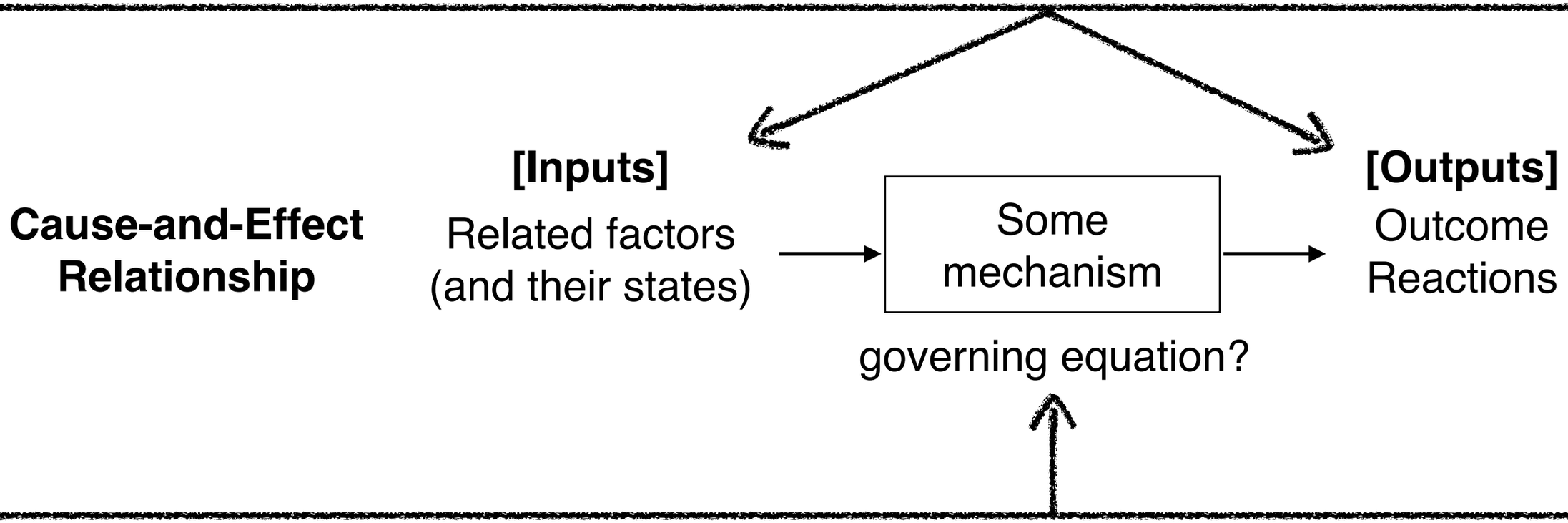
② Knowledge-driven (Knowledge Bases)

③ Data-driven (Machine Learning)

Yet another approach: Data-driven

based on very different principles and quite complementary!

Data-driven methods try to precisely approximate its outer behavior (the input-output relationship) observable as "data".
(e.g. through *machine learning* from a large collection of data)

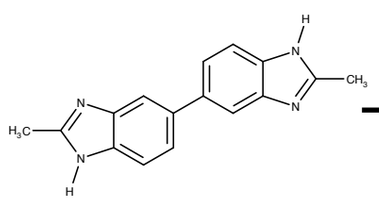


Theory-driven methods try to explicitly model the inner workings of a target phenomenon (e.g. through first-principles simulations)

Machine Learning (ML)

A new style of programming

a technique to reproduce a *transformation process (or function)* where the underlying principle is unclear and hard to be explicitly modelled just by giving a lot of **input-output examples**.



QSAR/QSPR Prediction

Growth inhibition
0.739

Generic Object Recognition

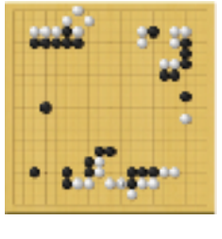
Speech Recognition

“ありがとう”

J’aime la musique

Machine Translation

I love music

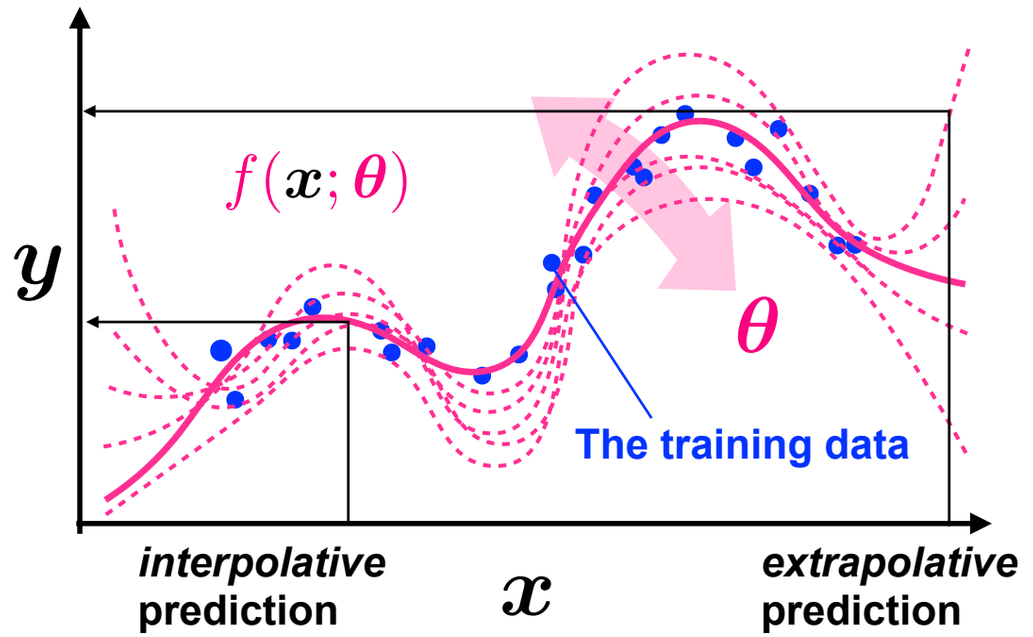


AI Game Players

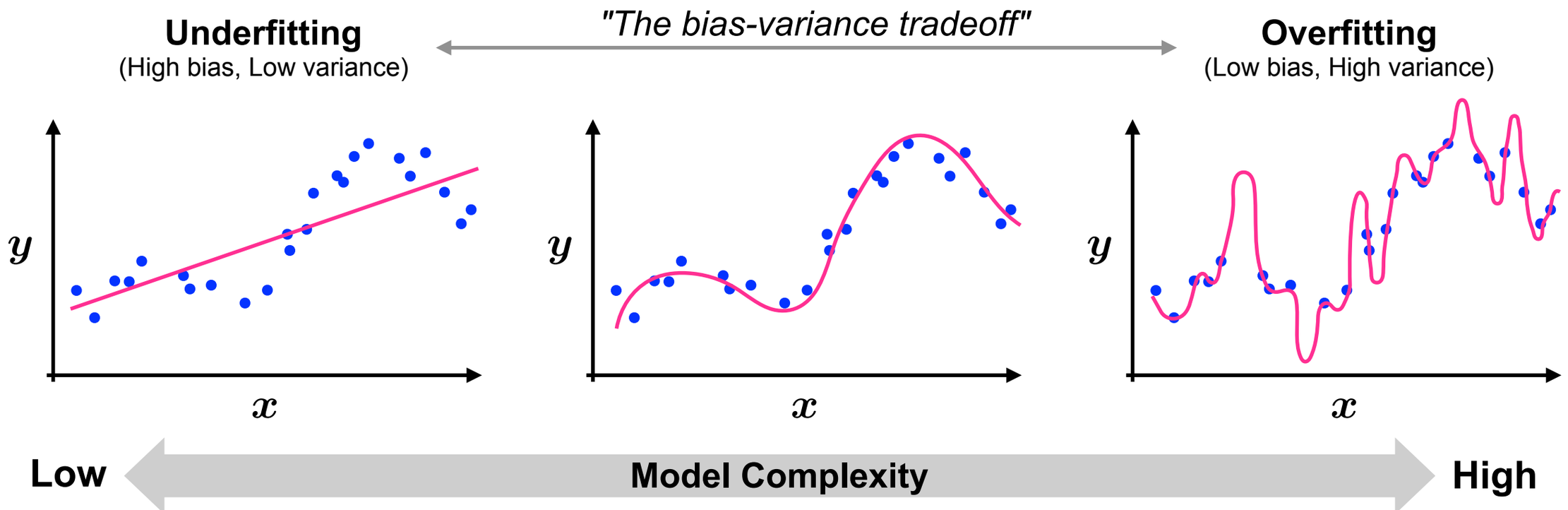
How ML works: fitting a function to data

A function $f(x;\theta)$ best fitted to
a given set of example input-output pairs
(the training data).

$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$



(High-dimensional)



Multilevel representations of chemical reactions

As pattern languages (e.g. known facts in textbooks/databases)

SMILES

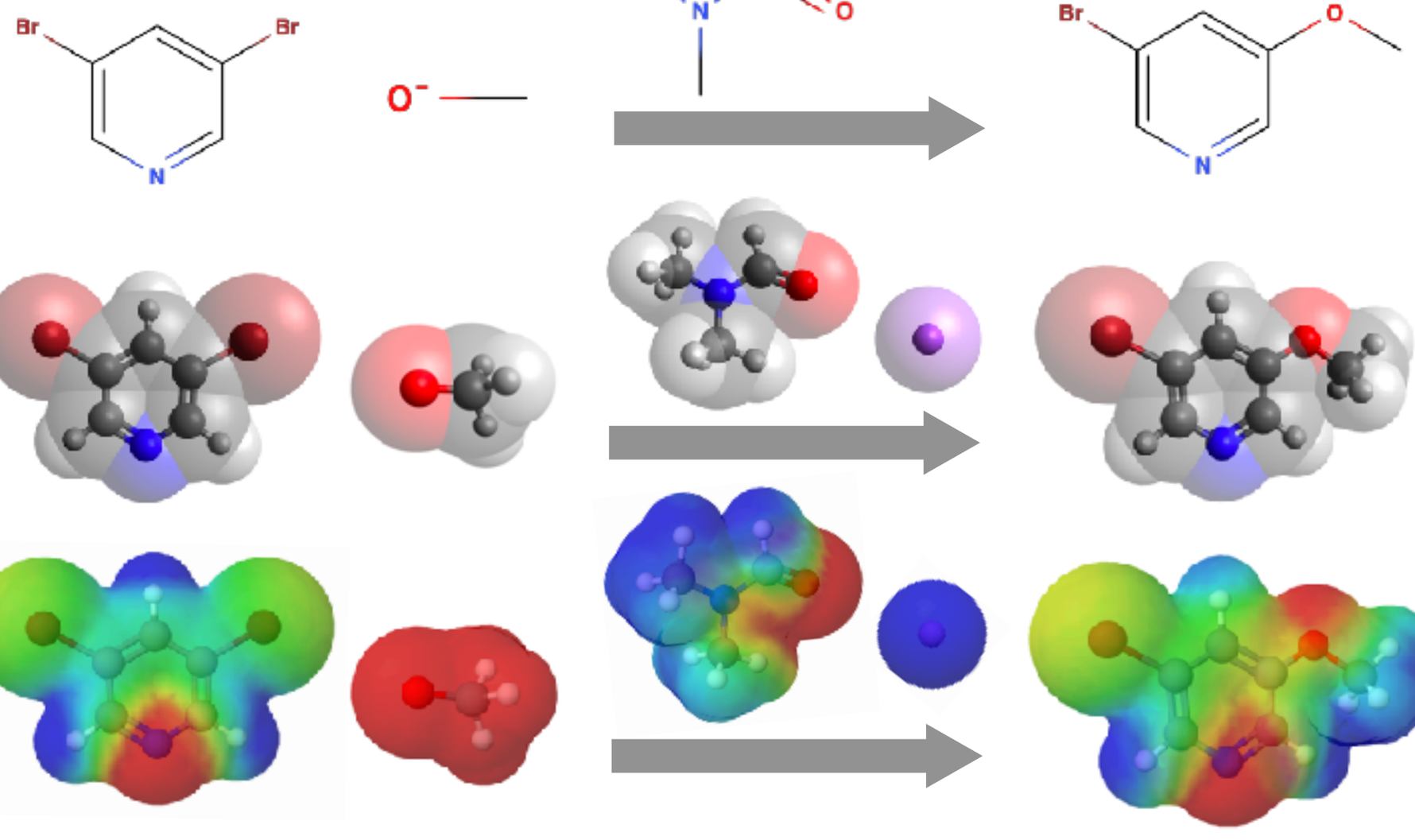
Structural Formla

Steric Structures

Electronic States

Reactants **Reagents** **Products**

Brc1cncc(Br)c1 C[O-] CN(C)C=O Na+ COc1cncc(Br)c1

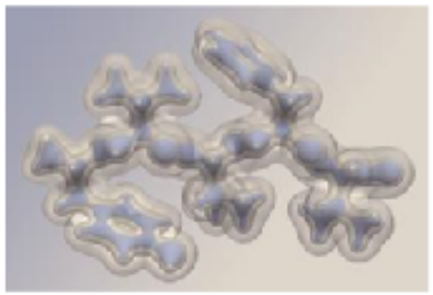


As physical entities (e.g. quantum chemical calculations)

https://www.chemistryworld.com/machine-learning/1616.tag

CHEMISTRY WORLD

All machine learning articles



RESEARCH

Machine learning predicts electron densities with DFT accuracy

2 OCTOBER 2019

Non-covalent interactions and electron densities can be explored quickly without the need for expensive and time-consuming quantum chemical calculations

RESEARCH

Language-based software's accurate predictions translate to benefits for chemists

30 SEPTEMBER 2019

State-of-the-art design for computer language processing results in improved models for predicting chemistry

RESEARCH

Human biases cause problems for machines trying to learn chemistry

13 SEPTEMBER 2019

Including 'unpopular' reagents and reaction conditions into datasets could lead to better machine-learning models

RESEARCH

Retrosynthetic algorithm broadened to design similar, but different, molecules

26 AUGUST 2019

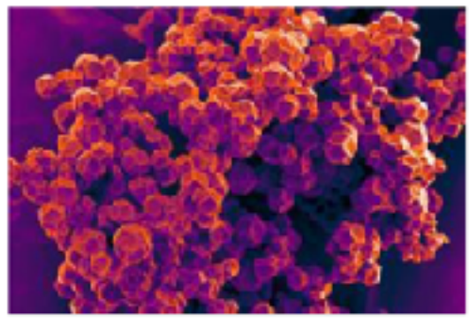
Chematica can now design efficient syntheses for large compound libraries

RESEARCH

Are synthetic chemists out of a job as AI meets automation?

9 AUGUST 2019

Platform can weigh up a synthetic route, plan it and then carry out it

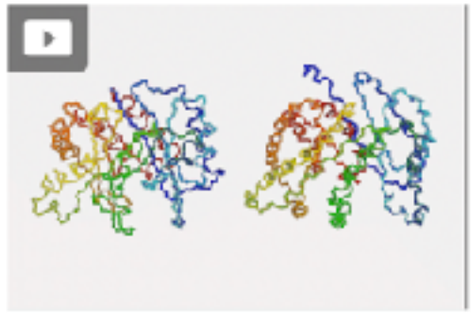


RESEARCH

Algorithm accurately predicts mechanical properties of existing and theoretical MOFs

17 MAY 2019

Machine learning could speed up the production and use of coordination polymers in industry



RESEARCH

Neural network folds proteins a million times faster than its competitors

8 MAY 2019

Machine learning algorithm that predicts protein structures in milliseconds could top next protein folding contest

RESEARCH

Dispute over reaction prediction puts machine learning's pitfalls in spotlight

18 DECEMBER 2018

Two research teams' argument over a reaction-predicting algorithm show that there is still a lot to understand when applying machine learning to chemistry

ML-based chemical reaction predictions

Graph NN	*Sequence NN*	*Combined or Other*
WLDN Jin+ *NeurIPS 2017*	**seq2seq** Liu+ *ACS Cent Sci 2017*	**Neural-Symbolic ML** Segler+ *Chemistry 2017*
ELECTRO Bradshaw+*ICLR 2019*	**IBM RXN** Schwaller+ *Chem Sci 2018*	**Similarity-based** Coley+ *ACS Cent Sci 2017*
GPTN Do+ *KDD 2019*	**Transformer** Karpov+ *ICANN 2019*	**3N-MCTS/AlphaChem** Segler+ *Nature 2018*
WLN Coley+ *Chem Sci 2019*	**Molecular Transformer** Schwaller+ *ACS Cent Sci 2019*	**Molecule Chef** Bradshaw+ *DeepGenStruct (ICLR WS) 2019*
GLN Dai+ *NeurIPS2019*		

ML + First-principle simulations

Both from Google DeepMind

Fermionic Neural Network

Pfau+ Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks. arXiv:1909.02487, Sep 2019.

Hamiltonian Graph Networks with ODE Integrators

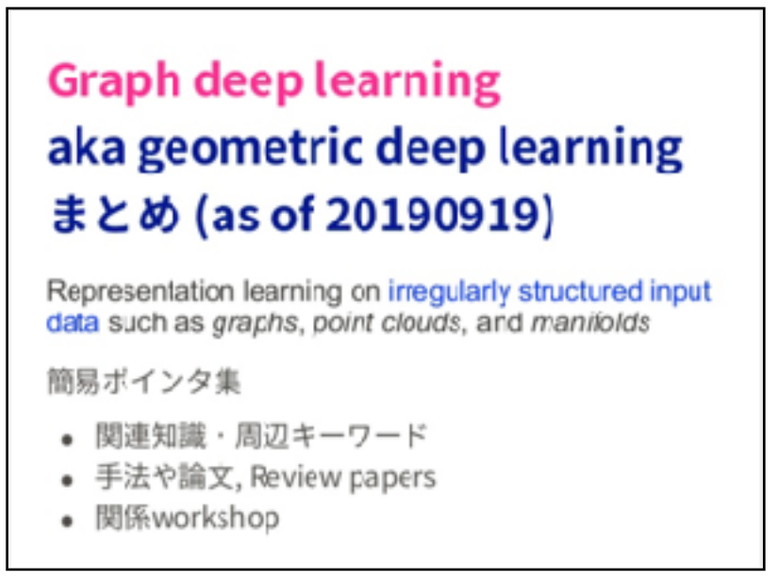
Sanchez-Gonzalez+ Hamiltonian Graph Networks with ODE Integrators. arXiv:1909.12790, Sep 2019.

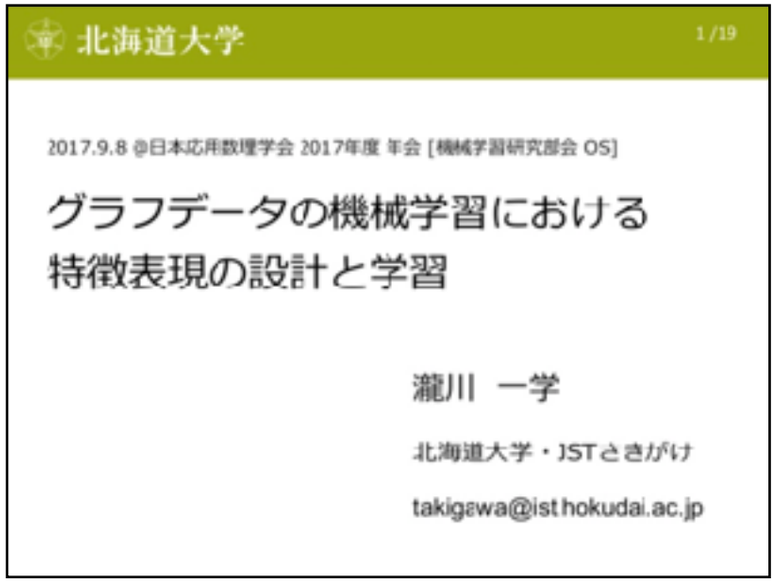
参考１) GNNs (Graph Neural Networks)

- **"グラフ構造"を入力にとるニューラルネットワーク.**
- **主要な機械学習の国際会議ではほぼこのテーマのWorkshopを開催**
- **2019夏以降論文も激増, 様々なタスクで道具に使われている**

  e.g.) Bengio+ Machine Learning for Combinatorial Optimization: a Methodological Tour d'Horizon. **https://arxiv.org/abs/1811.06128** (2018)
- **低分子化合物は分子グラフとして分析されることから、Chemoinformatics/ Materials informatics分野の応用論文も最近たくさん**
- **現行のGNNは技術的限界もよくわかってきた (NeurIPS2019で複数の分析論文)**

  こうしたいくつかの限界の技術的解決が可能なのかが機械学習分野の関心





参考２) Sequence Neural Networks (特にTransformers/Attention)

- 何らかのSequenceを受け取るニューラルネットワーク
- 機械翻訳や文章要約などの自然言語処理(NLP)タスクが主対象
- 古典的にはLSTM/GRUなどをユニットにしたRecurrent NNで行われてきた
- Googleの"Transformers"の発表とそれに基づく言語モデルBERTの大ブレイクで分野が様変わり (言語タスクでも事前学習が極めて有効に)
- RNNは学習が遅いので(Attentionつきの)CNNが使われるようになったあとに登場
- Transformer構造はマルチヘッドのAttentionをstackするとてもシンプルな構造

(かつAttention Weightsを行列積ベースにモデル化することでScalabilityを持つ)

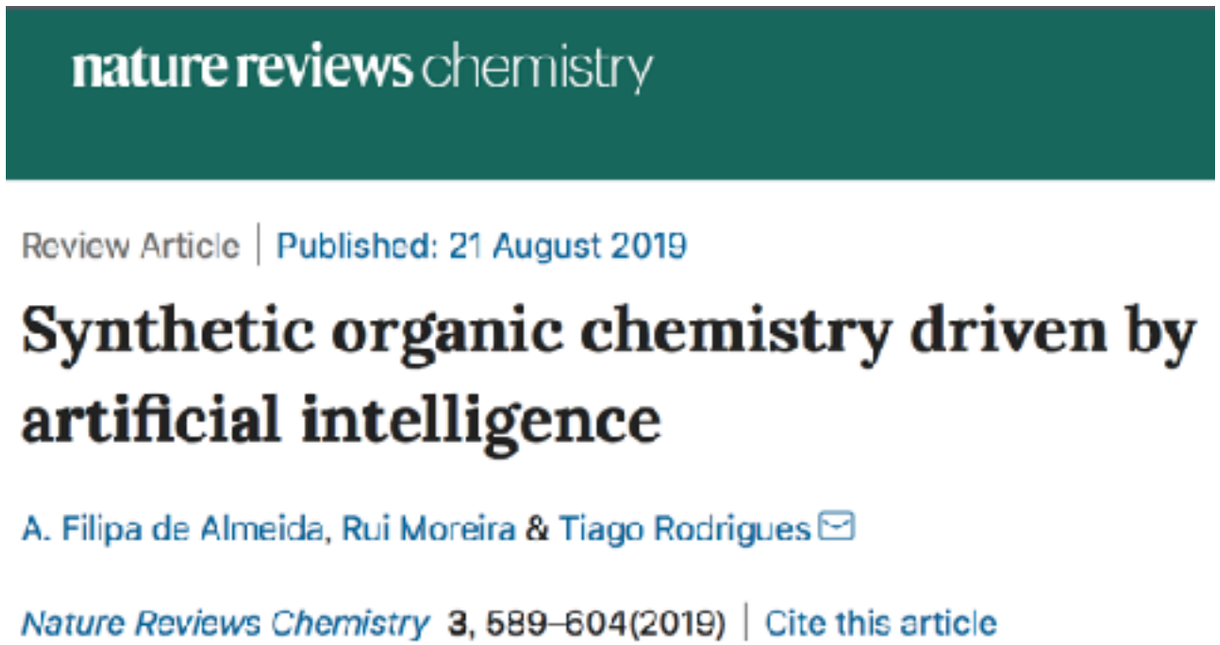
AI翻訳が人間超え、言葉の壁崩壊へ

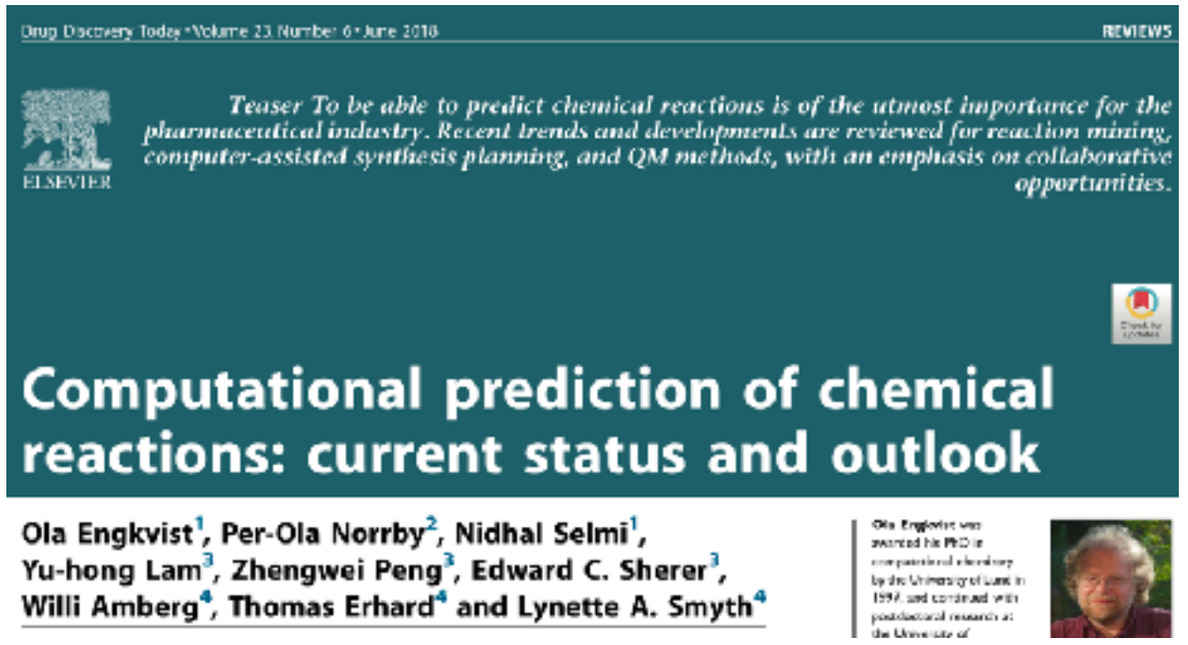
第２部：技術動向

トランスフォーマー時代到来
翻訳技術から汎用言語系AIに

2016年のニューラル機械翻訳（NMT）の実用化は、翻訳業界に衝撃を与え、ポケトークのような自動翻訳端末の市場拡大につながるなど、社会に大きなインパクトを与えた。ただし、翻訳技術や自然言語処理技術（NLP）分野では、その後も革命級のブレークスルーが相次いでいる。翻訳を含む言語系の人工知能（AI）が従来の常識を次々と塗り替え、ありえないペースで発展している。

化学反応/合成経路の予測に関係するReview Papers





- Coley CW, Green WH, Jensen KF.
  Machine Learning in Computer-Aided Synthesis Planning.
  Acc Chem Res. 2018 May 15;51(5):1281-1289.
  doi: 10.1021/acs.accounts.8b00087.

- Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, Bajczyk M, Grzybowski BA.
  Computer-Assisted Synthetic Planning: The End of the Beginning.
  Angew Chem Int Ed Engl. 2016 May 10;55(20):5904-37.
  doi: 10.1002/anie.201506101.

「データ利活用技術」は科学研究の道具の一つに

> **Science is changing, the tools of science are changing. And that requires different approaches.** —— Erich Bloch, 1925-2016

Nature, 559
pp. 547–555 (2018)

Science, 361
pp. 360-365 (2018)

Science, 293
pp. 2051-2055 (2001)

REVIEW

https://doi.org/10.1038/s41586-018-0337-2

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]*

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The

generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

SPECIAL SECTION *FRONTIERS IN COMPUTATION*

REVIEW

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

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

The discovery of new materials can bring enormous societal and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for achieving inverse design, which aims to discover tailored materials from the starting point of a particular desired functionality. Recent advances from the rapidly growing field of artificial intelligence, mostly from the subfield of machine learning, have resulted in a fertile exchange of ideas, where approaches to inverse molecular design are being proposed and employed at a rapid pace. Among these, deep generative models have been applied to numerous classes of materials: rational design of prospective drugs, synthetic routes to organic compounds, and optimization of photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

act properties. In practice, approximations are used to lower computational time at the cost of accuracy.

Although theory enjoys enormous progress, now routinely modeling molecules, clusters, and perfect as well as defect-laden periodic solids, the size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience (*11*). In the context of materials, ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice. Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploiting patterns in massive datasets, these models can distill average and salient features that characterize molecules (*12*, *13*).

Inverse design is a component of a more complex materials discovery process. The time

VIEWPOINT

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

Machine learning (ML) (*1*) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the stan-

correlate surprisingly well with subsequent gene expression analysis (*3*). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (*4*), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to amplify every aspect of a working scientist's

creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern recognition. A recent example is event analysis

教訓 "low input, high throughput, no output science." (Sydney Brenner)

→ 雑な設定・系で網羅的なハイスループット実験をいくらしても何も得られない

素朴な疑問：なぜたくさん手法があるの？どの手法を使えば良い？

Machine Learning Landscape

Supervised Learning

Classification

[Linear classification]
- Logistic / Softmax regregression
- Linear discriminant analysis
- Naive Bayes classifiers
- Perceptron
- Linear Support Vector Machines (SVM)

[Nonlinear classification]
- k-nearest neighbor classifiers
- Decision trees (Classification trees)
- Polynomial classifiers / Factorization machines
- Tree ensemble classifiers
  - Random Forest classifiers
  - Extra Trees classifiers
  - Gradient Boosted Decision Trees (GBDT)
- Kernel method classifiers
  - Support Vector Machines (SVM)
- Gaussian process classifiers
- Neural network (Deep learning) classifiers
  - Multi-layer perceptrons (MLP)
  - Convolutional networks (CNN)
    - VGG (OxfordNet)
    - Inception (GoogLeNet)
    - ResNet / ResNeXt
    - DenseNet
  - Recurrent networks (RNN)

Regression

[Linear regression]
- Least squares regression
- Principal component regression
- Partial Least Squares (PLS) regression
- Penalized linear regression
  - LASSO regression (L1-penalized)
  - Ridge regression (L2-penalized)
  - ElasticNet regression (L1 & L2-penalized)

[Nonlinear regression]
- k-nearest neighbor regressors
- Decision trees (Regression trees, Model trees)
- Polynomial regressors / Factorization machines
- Tree ensemble regressors
  - Random Forest regressors
  - Extra Trees regressors
  - Gradient Boosted Regression Trees (GBRT)
- Kernel method regressors
  - Support Vector Regression (SVR)
  - Kernel Ridge Regression
- Gaussian process regressors
- Neural network (Deep learning) regressors
  - Multi-layer perceptrons (MLP)
  - Convolutional networks (CNN)
    - VGG (OxfordNet)
    - Inception (GoogLeNet)
    - ResNet / ResNeXt
    - DenseNet
  - Recurrent networks (RNN)

Unsupervised Learning

Clustering

- k-means
- Hierarchical clustering
- Gaussian mixtures
- Spectral methods
- DBSCAN

Decomposition

- Principal component analysis (PCA)
- Independent component analoysis (ICA)
- Canonical correlation analysis (CCA)
- Nonnegative matrix factorization (NMF)
- Latent Dirichlet allocation (LDA)

Manifold learning

- Multidimensional scaling (MDS)
- Self-organizing maps (SOM)
- Isomap
- Locally linear embedding (LLE)
- Spectral embedding (Laplacian eigenmaps)
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Autoencoders

Density estimation

Others

Semi supervised learning
Ranking
Transfer learning
K-shot learning
Domain adaptation
Multitask learning
Reinforcement learning
Active learning
Model-based optimization
Time series/Sequence models
Probabilistic inference (Bayesian, Generative, Graphical)
Causal inference
Online/Incremental learning
Anomaly/Outlier detection
Ensemble learning
Relational/Network learning
Representation learning
Structured prediction
Meta Learning
:

観察データだけでは不十分 + データの解析のアヤ

機械学習は与えられた見本データを代表する予測モデルを作る技術で「どのようなデータで学習したか」が肝. 結果の再現性のなさは機械学習分野でも問題に

- Science論文 "Predicting reaction performance in C–N cross-coupling using machine learning"
  - Main paper https://doi.org/10.1126/science.aar5169
  - Erratum https://doi.org/10.1126/science.aat7648
  - Negative comment paper https://doi.org/10.1126/science.aat8603
  - Author's response https://doi.org/10.1126/science.aat8763

https://www.chemistryworld.com/news/dispute-over-reaction-prediction-puts-machine-learnings-pitfalls-in-spotlight/3009912.article

(Review) **Machine learning for catalysis informatics: recent applications and prospects.**

Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I*, Shimizu K*.
ACS Catalysis, Accepted.

Chapter 2が機械学習のユーザガイドになっています！↓査読者からのお薦めの言葉

Reviewer: 1
I don't usually recommend that papers should be accepted "as is", but in this case I don't see the need for changes. This review should be accepted and published in ACS Catalysis. ... I will certainly recommend it to my group and my students when it is published.
Reviewer: 2
The manuscript gives an excellent over the field of machine learning especially with regard to heterogeneous catalysis and I would highly recommend the article for the publication in ACS Catalysis.
Reviewer: 3
This is one of the best reviews for catalyst informatics that the Reviewer has read. In particular, the chapter 2 delivers a very good tutorial, which is concisely and professionally written.

ACS Catalysis

Review

pubs.acs.org/acscatalysis

Machine Learning for Catalysis Informatics: Recent Applications and Prospects

Takashi Toyao,[†,‡] Zen Maeno,[†] Satoru Takakusagi,[†] Takashi Kamachi,[‡,§] Ichigaku Takigawa,[*,∥,⊥] and Ken-ichi Shimizu[*,†,‡]

参考) 関係する講演スライド

https://itakigawa.github.io/news.html

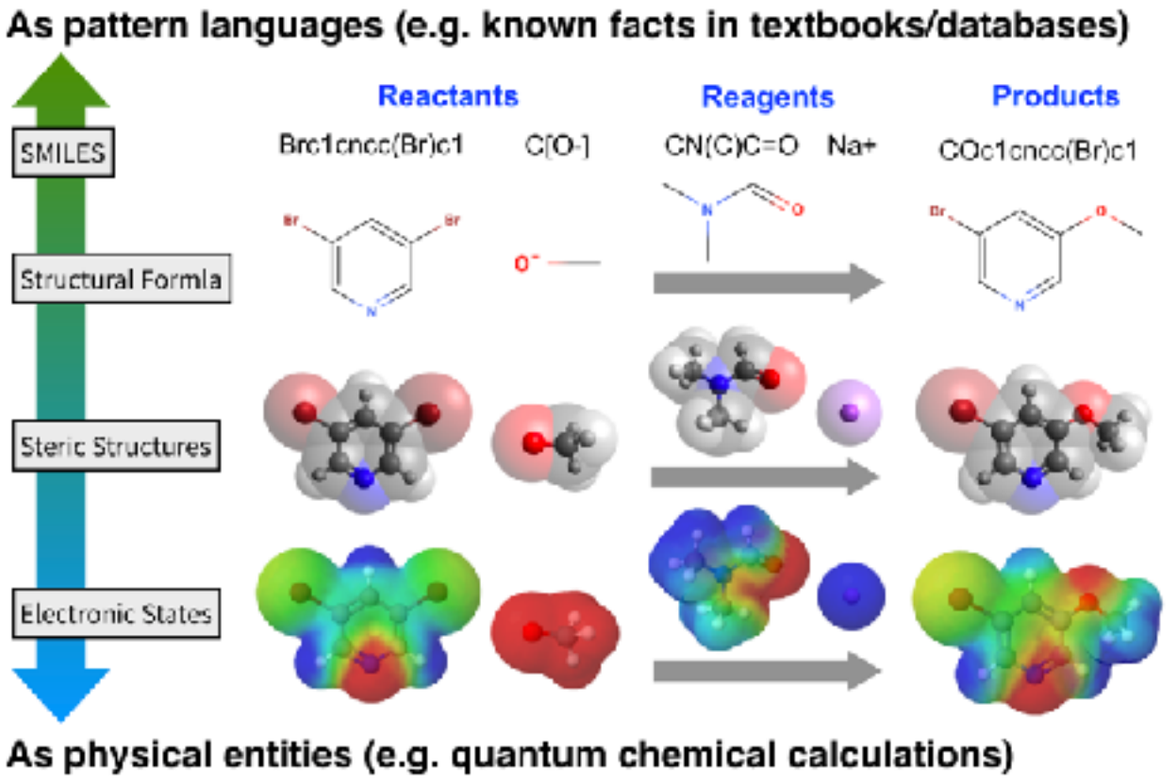
2019.10.4@早稲田大

2019.12.4@JAIST

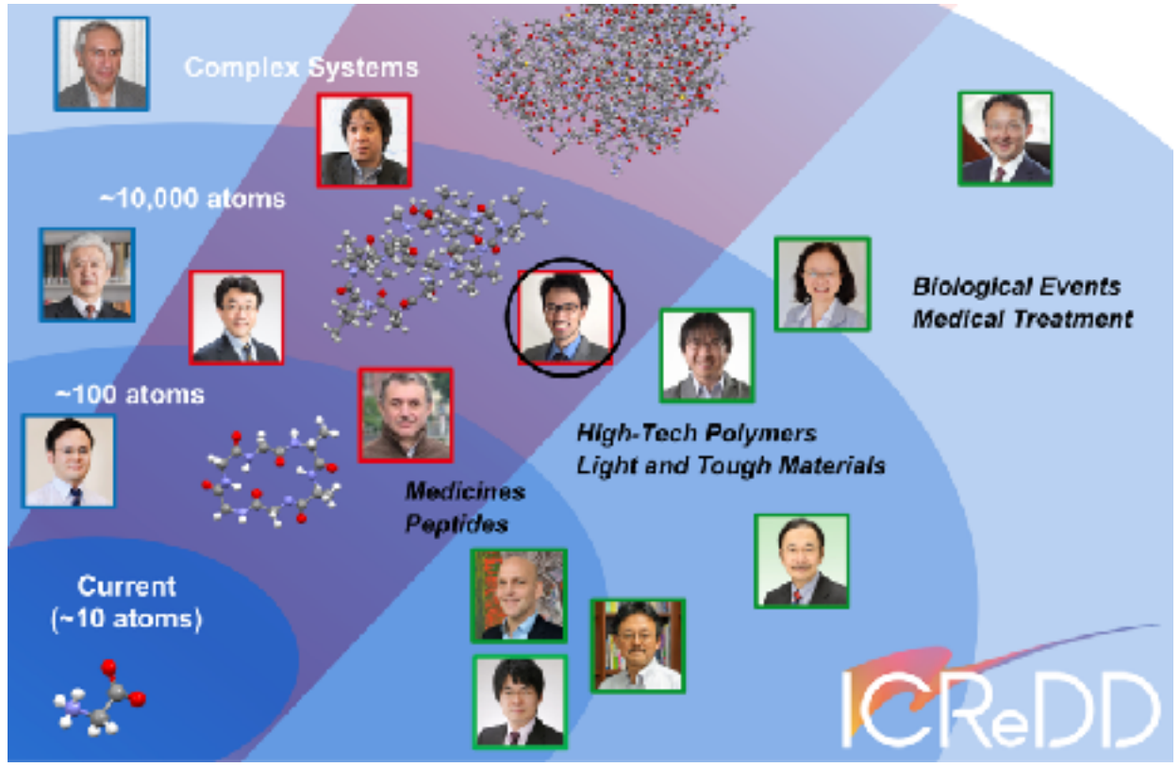
化学反応の予測

化学反応をどのような表現・レベルで捉えるか？

１分子レベルでの表現の多様性



量子・電子系 ～ 複雑系・生体系



技術的融合により化学反応の多角性・多因子性を捉える

① Theory-driven (Quantum Chem)　② Knowledge-driven (Knowledge Bases)　③ Data-driven (Machine Learning)

計算科学・実験科学・情報科学による化学反応設計と探索

① Theory-driven (Quantum Chem)　② Knowledge-driven (Knowledge Bases)　③ Data-driven (Machine Learning)

GDCh Communications Angewandte International Edition Chemie

VIP AI-Assisted Synthesis Very Important Paper

International Edition: DOI: 10.1002/anie.201912083
German Edition: DOI: 10.1002/ange.201912083

Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

*Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski**

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Letter

pubs.acs.org/jcim

Synergies Between Quantum Mechanics and Machine Learning in Reaction Prediction

Peter Sadowski,*,† David Fooshee,† Niranjan Subrahmanya,‡ and Pierre Baldi*,†

Heterogeneous Catalysis

計算での扱いがまだまだ難しい例：不均一系触媒

アンモニアの工業的合成 (ハーバー・ボッシュ法)

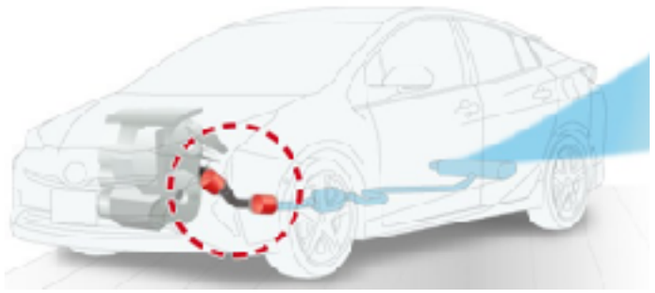
“水と石炭と空気からパンを作る方法”

20世紀の食糧難を解決した人工的窒素固定

$$N_2 + 3\,H_2 \rightarrow 2\,NH_3$$

鉄系触媒など

排気ガスの浄化



排気ガス(有毒) → **無害化**

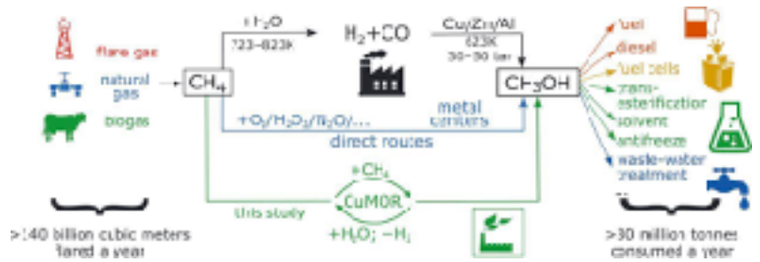
NOx → N_2

CO → CO_2

HC → H_2O

貴金属触媒など (Pt, Pd, Rh…)

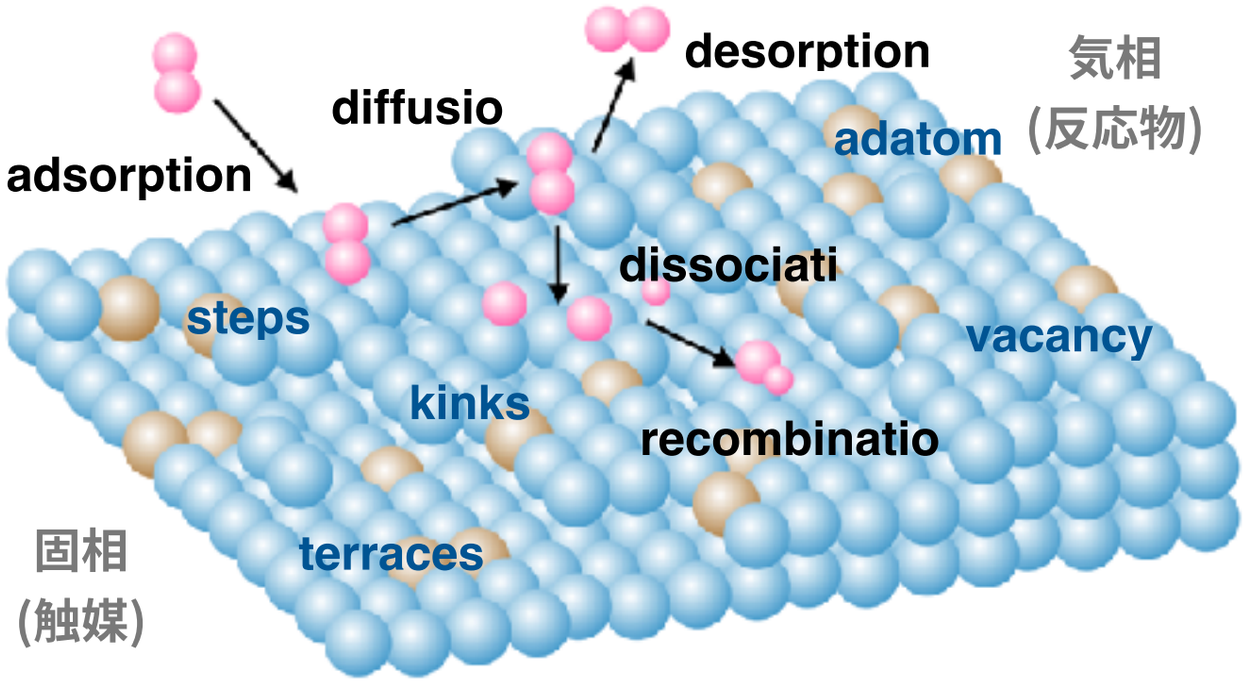
メタン転換



メタン →

- エタン
- エチレン
- メタノール
- ⋮

金属触媒 (Li, 希土類, アルカリ土類)



- **固体表面での気相反応 (複雑系)**
- **吸着・乖離・拡散・反応・脱離など複数の素反応過程の関与**
- **多数の因子に依存:** 触媒組成、担持金属ナノ粒子のサイズや形状、担体表面の終端構造、反応温度やガス圧力等の反応条件、…

機械学習に基づく研究3事例を紹介 (with 触媒科学研究所)


Keisuke Suzuki


Dr. Takashi Toyao


Dr. Zen Maeno


Dr. Satoru Takakusagi


Prof. Ken-ichi Shimizu

1. **Predicting the d-band centers by ML**
   Takigawa I*, Shimizu K, Tsuda K, Takakusagi S
   RSC Advances. 2016; 6: 52587-52595.

2. **Predicting the adsorption energy by ML**
   Toyao T*, Suzuki K, Kikuchi S, Takakusagi S, Shimizu K, Takigawa I*.
   The Journal of Physical Chemistry C. 2018; 122(15): 8315-8326.

3. **Predicting the experimentally-reported catalytic activity by ML**
   Suzuki K, Toyao T, Maeno Z, Takakusagi S, Shimizu K*, Takigawa I*.
   ChemCatChem. 2019; 11(18): 4537-4547. (Front Cover)

- (Review) **Machine learning for catalysis informatics: recent applications and prospects.**
  Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I*, Shimizu K*.
  ACS Catalysis, Accepted.

Heterogeneous Catalysis

- Heterogeneous catalysis is a type of catalysis in which **the catalyst occupies a different phase** from the reactants and products.
- It can be more easily recycled than homogeneous, but characterization of the catalyst and optimization of properties can be **more difficult**.
- It is **of paramount importance** in many areas of the chemical and energy industries.

Haber–Bosch Process

(industrial synthesis of ammonia)

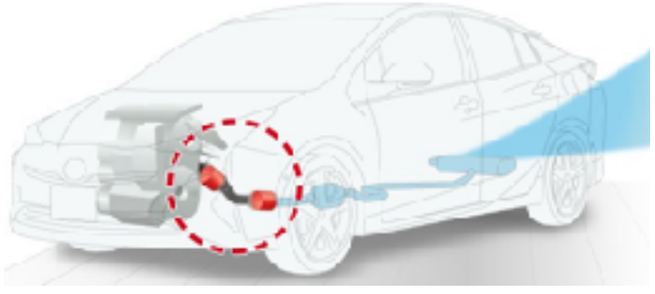
“Fertilizer from Air”

artificial nitrogen fixation

$$N_2 + 3\,H_2 \rightarrow 2\,NH_3$$

Ferrous Metal Catalysis

Exhaust Gas Purification



Exhaust Gas → **Harmless gas**

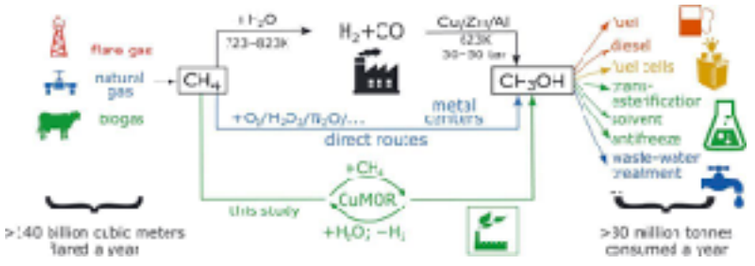
NOx → N_2

CO → CO_2

HC → H_2O

Noble Metal Catalysis (Pt, Pd, Rh…)

Conversion of Methane



Methane →

- **Ethane**
- **Ethylene**
- **Methanol**
- **:**

Various Metallic Catalysts (Li, rare earthes, alkaline earths)

Heterogeneous Catalysis

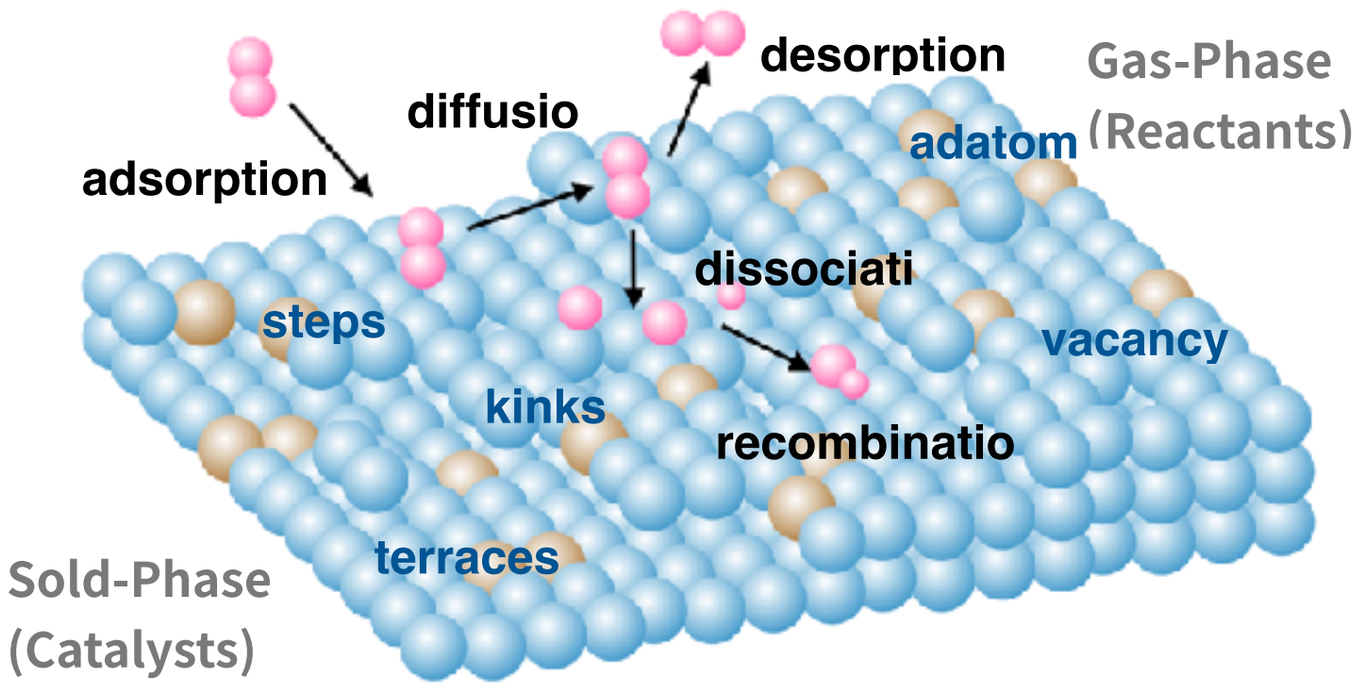
Notoriously complex surface reactions between different phases.

Wolfgang Pauli

**“God made the bulk;
the surface was invented by the devil.”**

Many hard-to-quantify intertwined factors involved.

Too complicated (impossible?) to model everything...



- multiple elementary reaction processes
- composition, support, surface termination, particle size, particle morphology, atomic coordination environment
- reaction conditions

Our ML-based case studies

1. Can we predict the d-band center?

predicting DFT-calculated values by machine learning
(Takigawa et al, RSC Advances, 2016)

2. Can we predict the adsorption energy?

predicting DFT-calculated values by machine learning
(Toyao et al, JPCC, 2018)

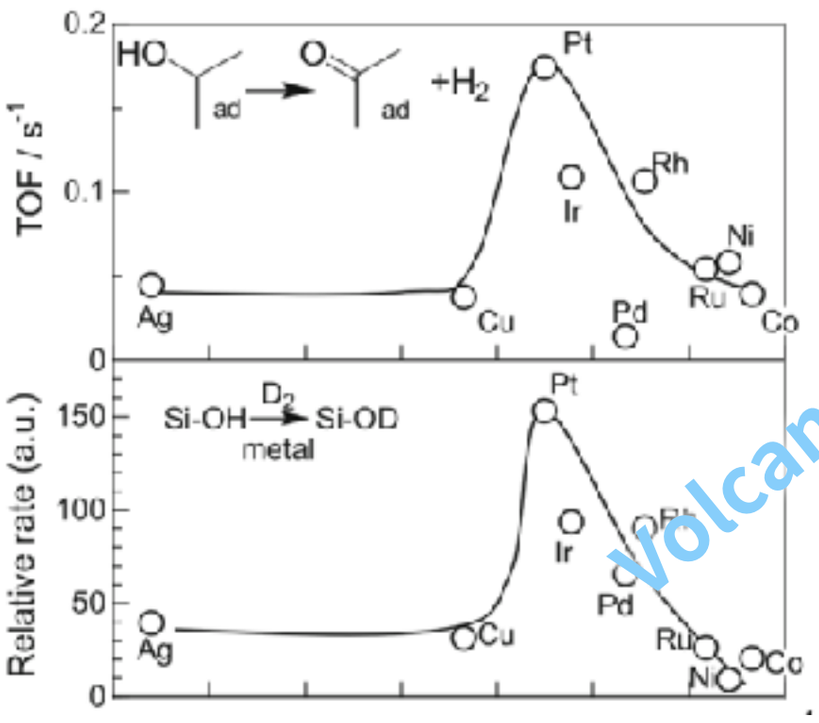
3. Can we predict the catalytic activity?

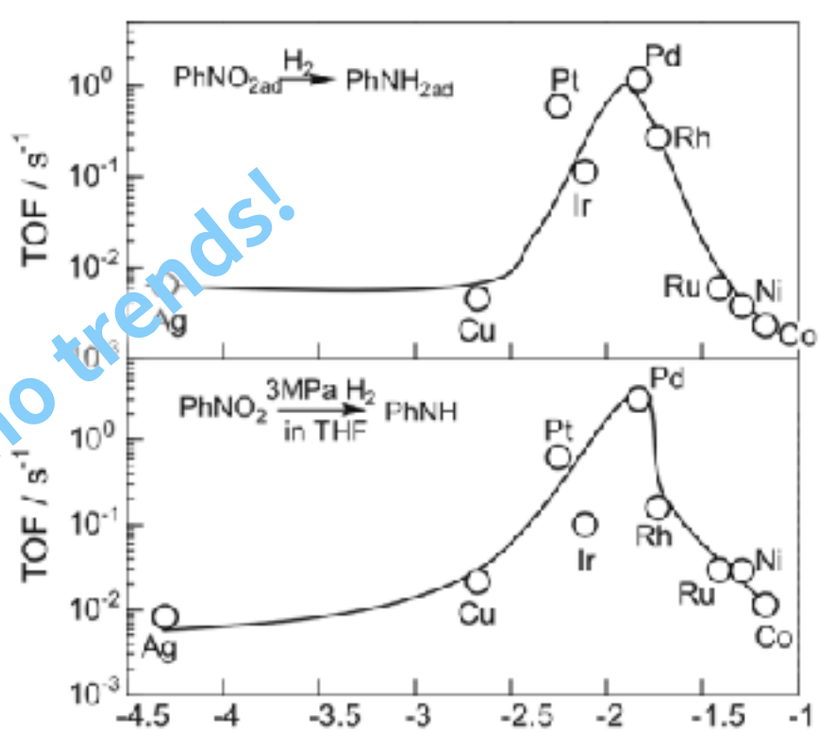
predicting values from experiments reported in the literature by machine learning
(Suzuki et al, ChemCatChem, 2019)

How to understand the catalytic activities?

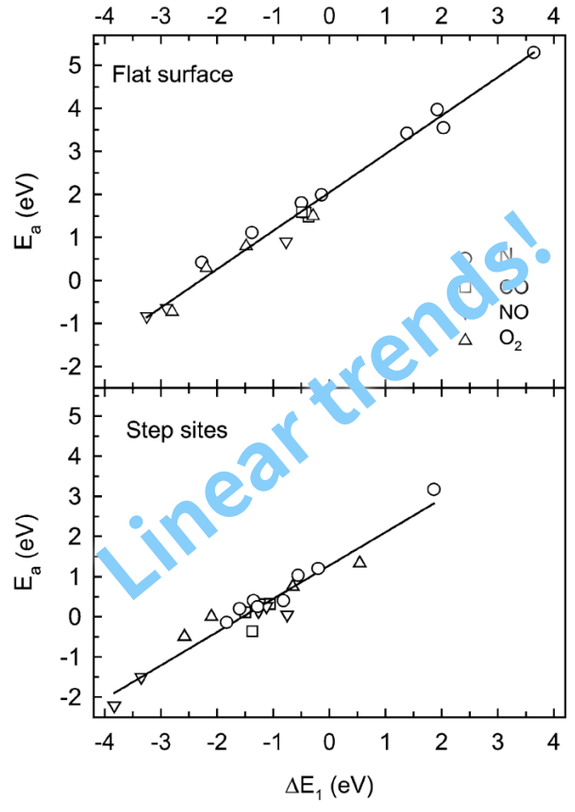
Traditionally, the computable indexes that well *correlate* the catalytic activities have been investigated...

Hammer–Nørskov d-band model





Brønsted-Evans-Polanyi relation



reaction rates

activation energy / eV

Volcano trends!

Linear trends!

d-band center ($\varepsilon_d - E_F$) / eV

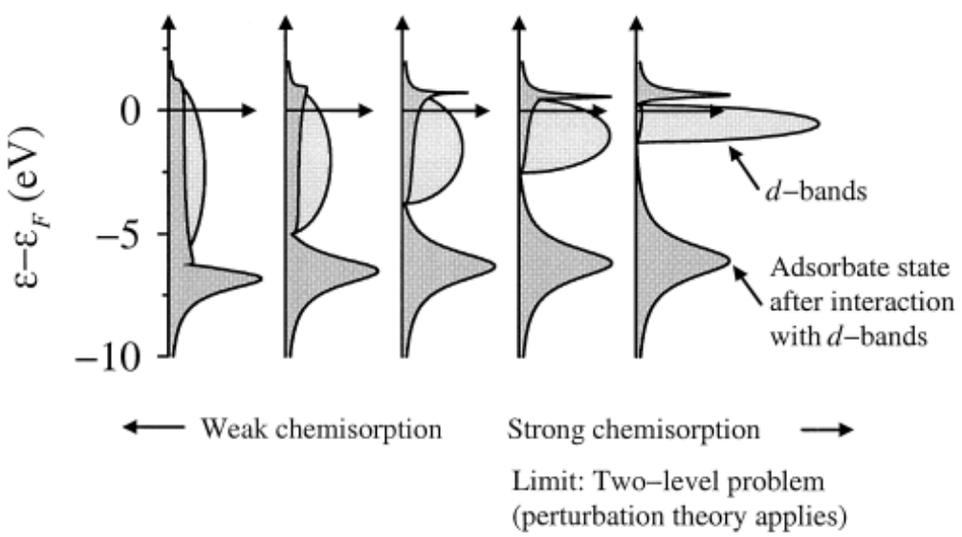
d-band center ($\varepsilon_d - E_F$) / eV

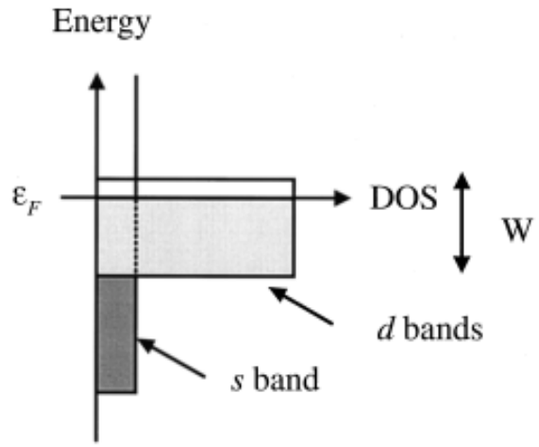
adsorption energy / eV

K. Shimizu et al, *ACS Catal*. 2, 1904 (2012)

Case 1. Predicting the d-band centers

The d-band center is one of the established indexes to understand the trends of heterogeneous catalysts (transition metals based).



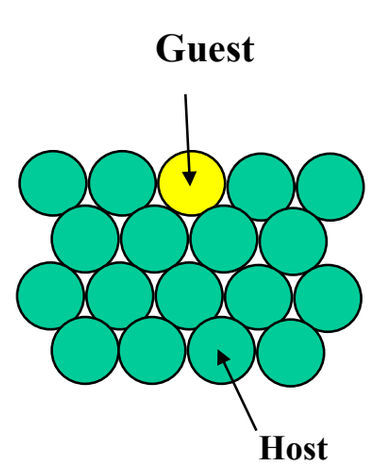


J. K. Nørskov, et al., *Advances in Catalysis*, 2000

[1% doped] **Guest**

Host

M_g / M_h	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	**−0.92**	−0.87	−1.12	−1.05	−1.21	−1.46	−2.16	−1.75	−1.28	−2.01	−2.34
Co	−1.16	**−1.17**	−1.45	−1.33	−1.41	−1.75	−2.54	−2.08	−1.53	−2.36	−2.73
Ni	−1.20	−1.10	**−1.29**	−1.10	−1.43	−1.60	−2.26	−1.82	−1.43	−2.09	−2.42
Cu	−2.11	−2.07	−2.40	**−2.67**	−2.09	−2.35	−3.31	−3.37	−2.09	−3.00	−3.76
Ru	−1.20	−1.15	−1.40	−1.29	**−1.41**	−1.58	−2.23	−1.68	−1.39	−2.03	−2.25
Rh	−1.49	−1.39	−1.57	−1.29	−1.69	**−1.73**	−2.27	−1.66	−1.56	−2.08	−2.22
Pd	−1.46	−1.29	−1.33	−0.89	−1.59	−1.47	**−1.83**	−1.24	−1.30	−1.64	−1.66
Ag	−3.58	−3.46	−3.63	−3.83	−3.46	−3.44	−4.16	**−4.30**	−3.16	−3.80	−4.45
Ir	−1.90	−1.84	−2.06	−1.90	−2.02	−2.26	−2.84	−2.24	**−2.11**	−2.67	−2.85
Pt	−1.92	−1.77	−1.85	−1.53	−2.11	−2.02	−2.42	−1.81	−1.87	**−2.25**	−2.30
Au	−2.93	−2.79	−2.93	−3.01	−2.86	−2.81	−3.39	−3.35	−2.58	−3.10	**−3.56**



Two types of models

- 1% doped
- overlayer

Ruban, Hammer, Stoltze, Skriver, Nørskov, *J Mol Catal A, 115:421-429* (1997)

Simple ML can accurately predict them...

We showed that gradient boosted trees with only 6 descriptors below can predict the d-band centers *without any first-principles calculations*.

(1) Group in the periodic table (host)
(2) Density at 25 °C (host)
(3) Enthalpy of fusion (guest)
(4) Ionization energy (guest)
(5) Enthalpy of fusion (host)
(6) Ionization energy (host)

9 types of readily available values pretested

- Group (G)
- Bulk Wigner–Seitz radius (R) in Å
- Atomic number (AN)
- Atomic mass (AM) in g mol^{-1}
- Period (P)
- Electronegativity (EN)
- Ionization energy (IE) in eV
- Enthalpy of fusion ($\Delta_{fus}H$) in J g^{-1}
- Density at 25 °C (ρ) in g cm^{-3}

Table 3.3 Input features (descriptors) used for the prediction of d-band centers from Ref. [22]. Reproduced from Ref. [19] with permission from the Royal Society of Chemistry

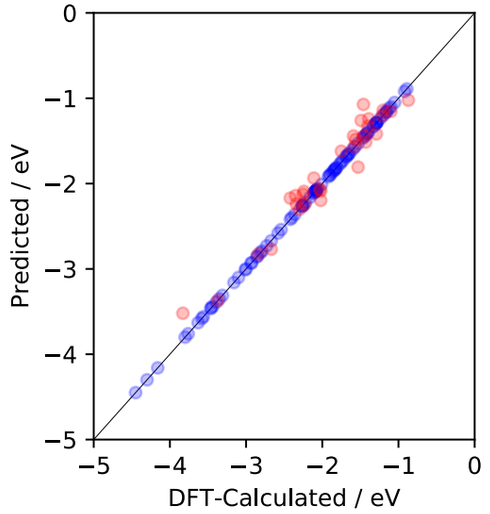
Metal	G	R/Å	AN	AM/ g mol^{-1}	P	EN	IE/ eV	$\Delta_{fus}H$/ J g^{-1}	ρ/ g cm^{-3}
Fe	8	2.66	26	55.85	4	1.83	7.90	247.3	7.87
Co	9	2.62	27	58.93	4	1.88	7.88	272.5	8.86
Ni	10	2.60	28	58.69	4	1.91	7.64	290.3	8.90
Cu	11	2.67	29	63.55	4	1.90	7.73	203.5	8.96
Ru	8	2.79	44	101.07	5	2.20	7.36	381.8	12.10
Rh	9	2.81	45	102.91	5	2.28	7.46	258.4	12.40
Pd	10	2.87	46	106.42	5	2.20	8.34	157.3	12.00
Ag	11	3.01	47	107.87	5	1.93	7.58	104.6	10.50
Ir	9	2.84	77	192.22	6	2.20	8.97	213.9	22.50
Pt	10	2.90	78	195.08	6	2.20	8.96	113.6	21.50
Au	11	3.00	79	196.97	6	2.40	9.23	64.6	19.30

ML Prediction (without any quantum calculations)

	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	-0.92		-0.96	-0.97	-1.65	-1.64	-2.24		-1.87	-2.4	-3.11
Co			-1.37	-1.23		-2.12	-2.82	-2.53	-2.26		-3.56
Ni	-0.33	-1.18			-1.92	-2.03		-2.43	-2.15	-2.82	-3.39
Cu	-2.42		-2.49	-2.67	-2.89	-2.94				-3.82	-4.63
Ru	-1.11	-1.04	-1.12		-1.41		-1.88	-1.81	-1.54		-2.27
Rh	-1.42	-1.32		-1.51	-1.7	-1.73	-2.12	-1.81	-1.7	-2.18	-2.3
Pd	-1.47	-1.29	-1.29	-1.03		-1.58	-1.83	-1.68	-1.52	-1.79	
Ag	-3.75	-3.56	-3.62		-3.8		-4.03		-3.5	-3.93	-4.51
Ir	-1.78	-1.71	-1.78	-1.55		-2.14	-2.53	-2.2	-2.11	-2.6	-2.7
Pt			-1.71	-1.47	-2.13	-2.01	-2.23	-2.06	-1.96		-2.33
Au	-3.03	-2.82	-2.85			-2.89		-3.44			-3.56

gradient boosting w/ 6 descriptors

training sets (75%)
test sets (25%)

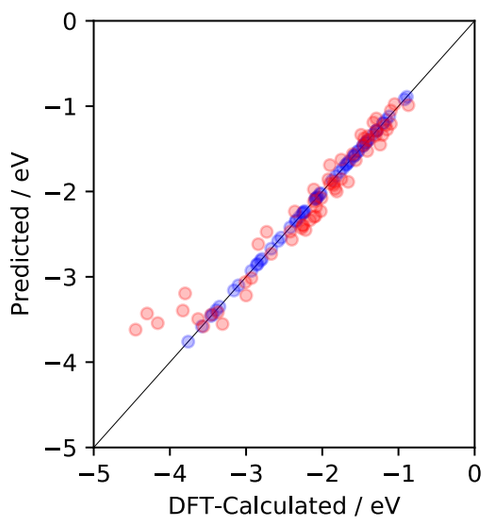


100 times mean RMSE: **0.153** / eV

	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe		-0.78			-1.65	-1.64			-1.87		
Co	-1.18	-1.17	-1.37		-1.87	-2.12	-2.82		-2.26		
Ni	-0.33	-1.18		-1.17			-2.61	-2.43	-2.15	-2.82	
Cu	-2.42				-2.89	-2.94		-3.88			-4.63
Ru	-1.11	-1.04	-1.12	-1.11	-1.41			-1.81			-2.27
Rh	-1.42			-1.51			-2.12	-1.81	-1.7		
Pd		-1.29	-1.29	-1.03		-1.58	-1.83		-1.52	-1.79	
Ag				-3.68	-3.8	-3.63					-4.51
Ir						-2.14			-2.11		-2.7
Pt			-1.71	-1.47	-2.13	-2.01	-2.23	-2.06			
Au				-2.86	-3.09	-2.89		-3.44			-3.56

gradient boosting w/ 6 descriptors

training sets (50%)
test sets (50%)

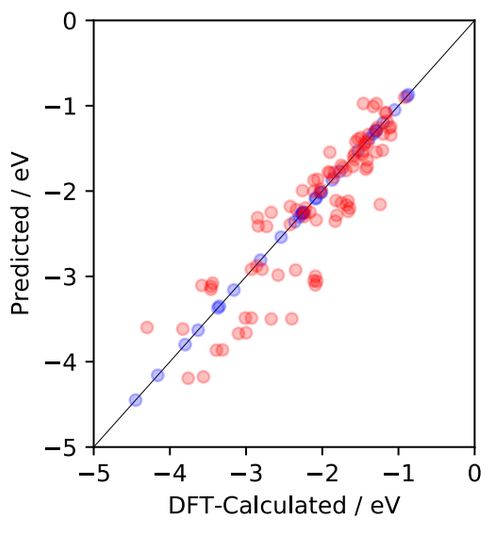


100 times mean RMSE: **0.235** / eV

	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe								-2.17			-3.11
Co		-1.17	-1.37			-2.12					
Ni	-0.33	-1.18					-2.61	-2.43			
Cu	-2.42	-2.29	-2.49				-3.71				-4.63
Ru										-2.02	
Rh		-1.32				-1.73	-2.12				
Pd					-1.94		-1.83				-1.97
Ag	-3.75			-3.68							-4.51
Ir	-1.78	-1.71									-2.7
Pt					-2.13						
Au					-3.09	-2.89					

gradient boosting w/ 6 descriptors

training sets (25%)
test sets (75%)

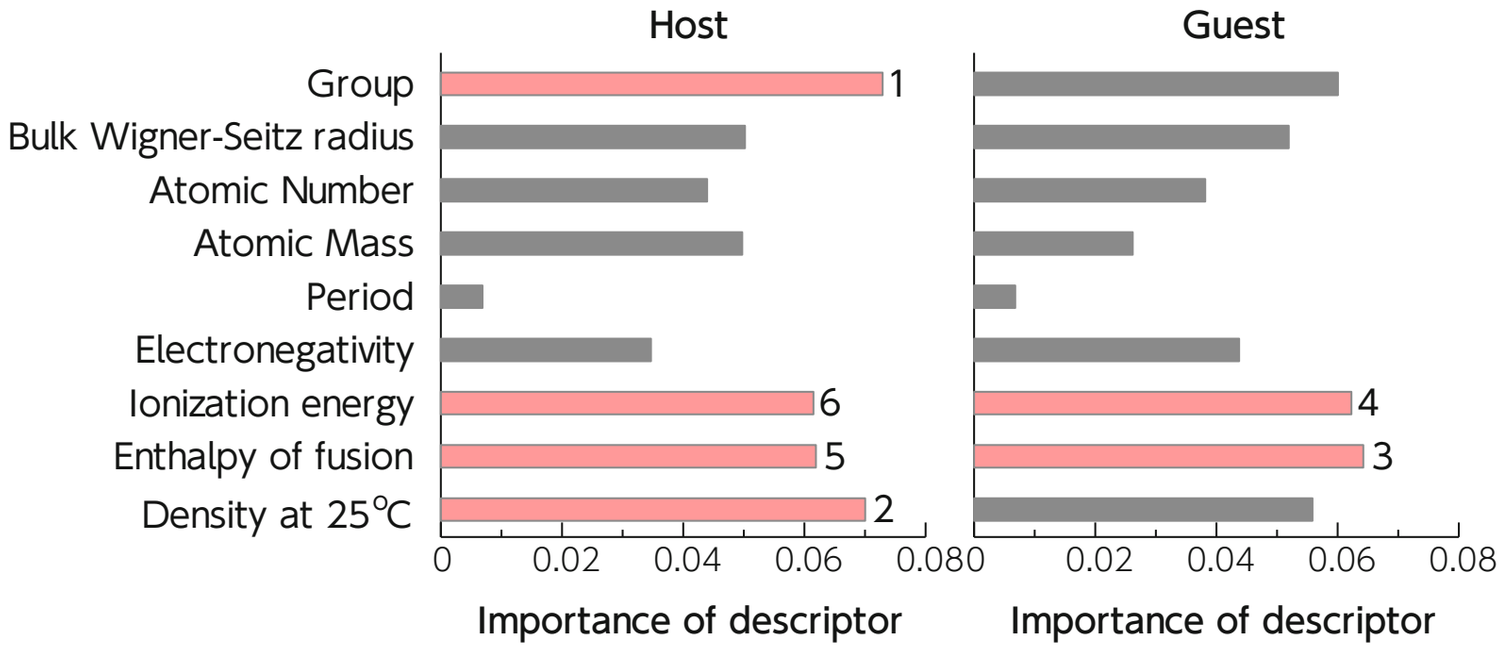


100 times mean RMSE: **0.402** / eV

Descriptor analysis and selection

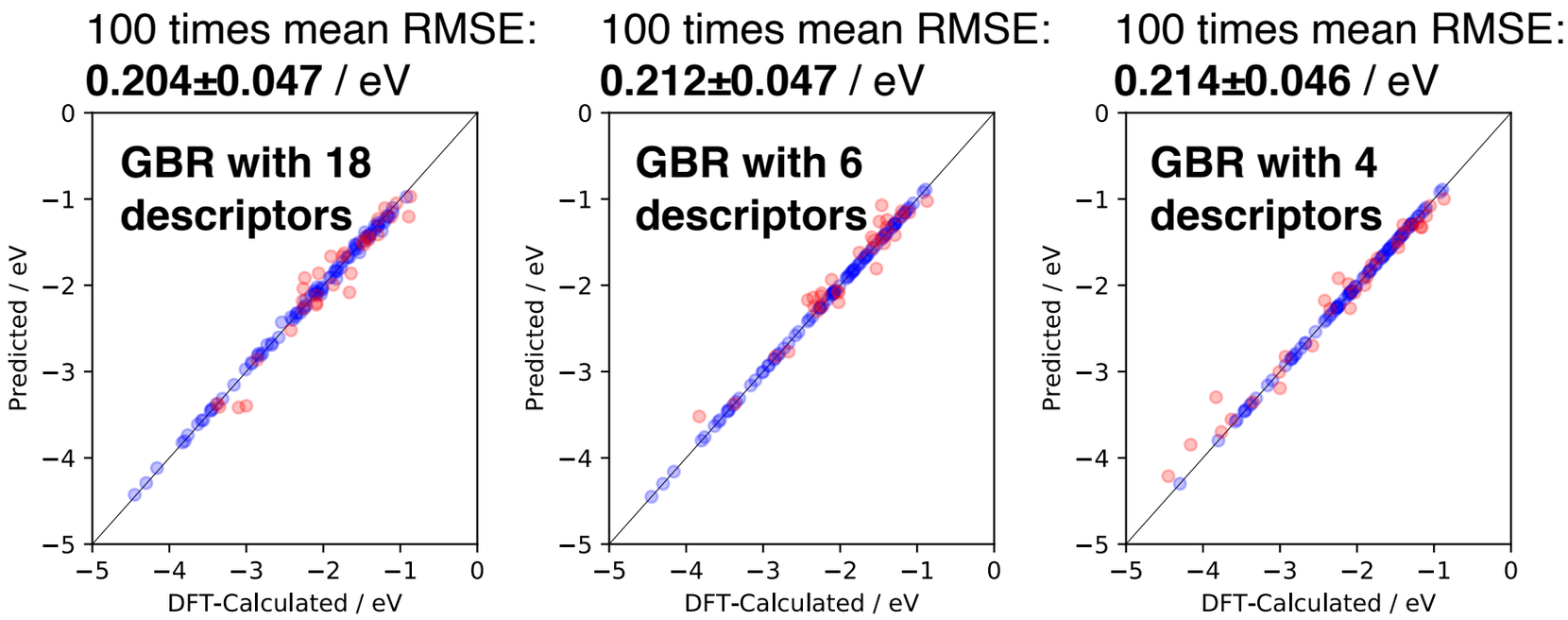
Analyzing the prediction model trained with the current data also provides some insights on contributing factors, and variable selection.

Descriptor Importances



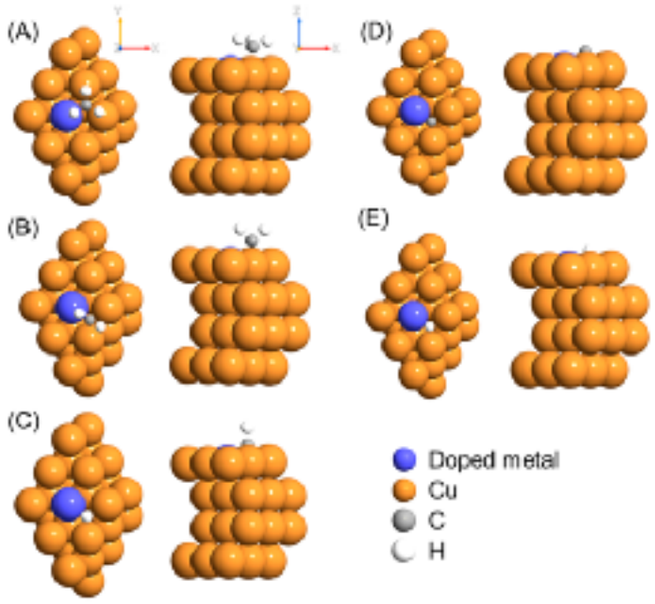
Descriptor Selection (top-k)

training sets (75%)
test sets (25%)

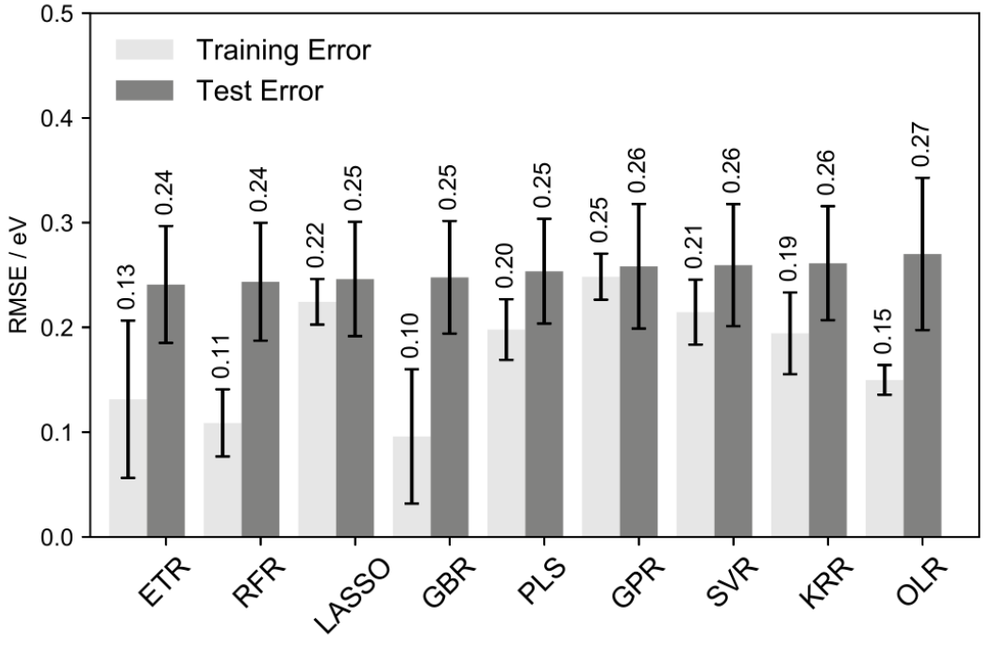


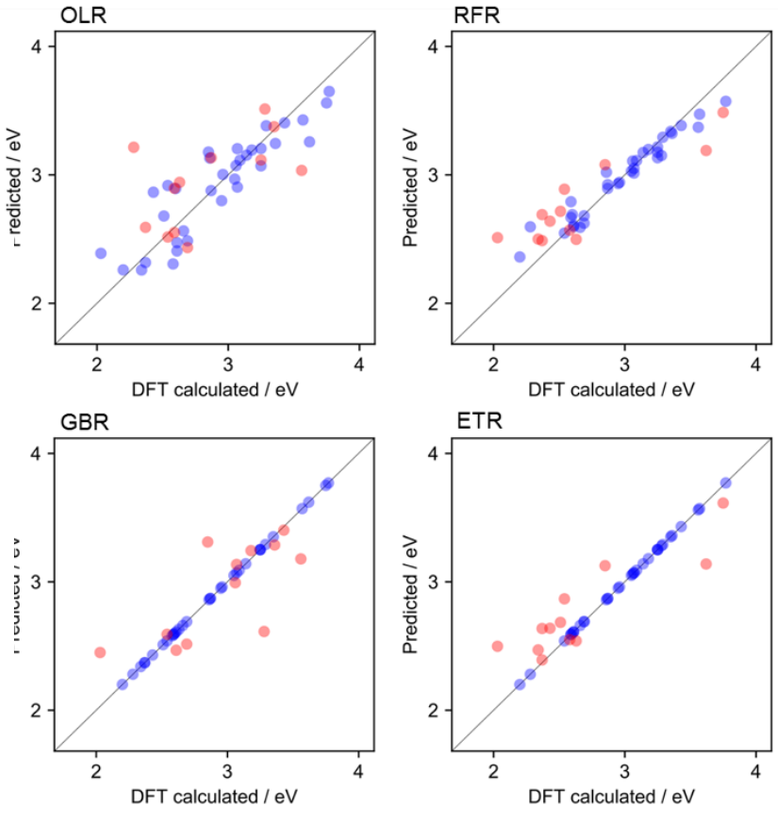
Case 2. Predicting the adsorption energy

Adsorbates: CH_3, CH_2, CH, C, H



Predicting Adsorption energy of CH_3 (on 46 Cu-based alloys)





training sets (75%)
test sets (25%)

DFT calculation of adsorption energy

- **10 hours with our 32 cores workstation** (CH_3 on the Cu monometallic surface)
- even longer time (about 34 hours) for the system containing another metal such as Pb

ML prediction

- **< 1 sec with our 1 core laptop**
- not dependent on target systems, but methods we choose

Case 3. Predicting the experimental catalyst activities

For some reactions, large datasets from already published results are available. **Why not just directly applying ML to them!**

- **Oxidative coupling of methane (OCM)**
  1866 catalysts [Zavyalova+ 2011]
- **Water gas shift reaction (WGS)**
  4360 catalysts [Odabaşi+ 2014]
- **CO oxidation**
  5610 catalysts [Günay+ 2013]

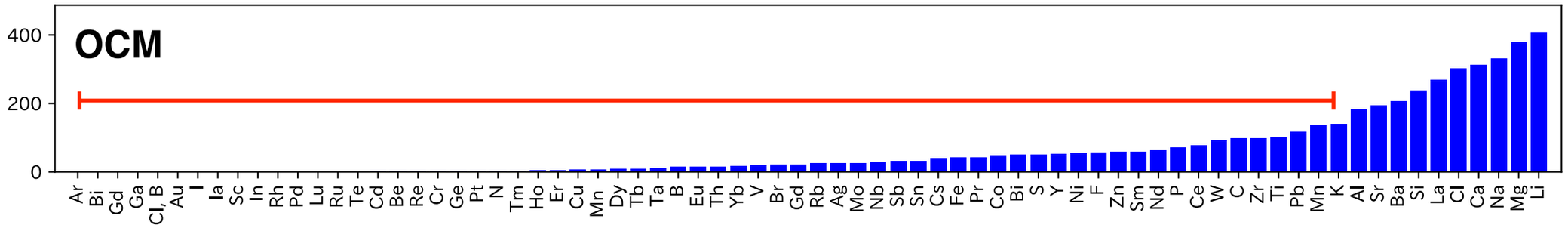
Collections from various papers published in the past including

- catalyst compositions, support types, promotor types
- catalyst performance (C_2 yields, CO conversion)
- experimental conditions (pressure, temperature, etc)

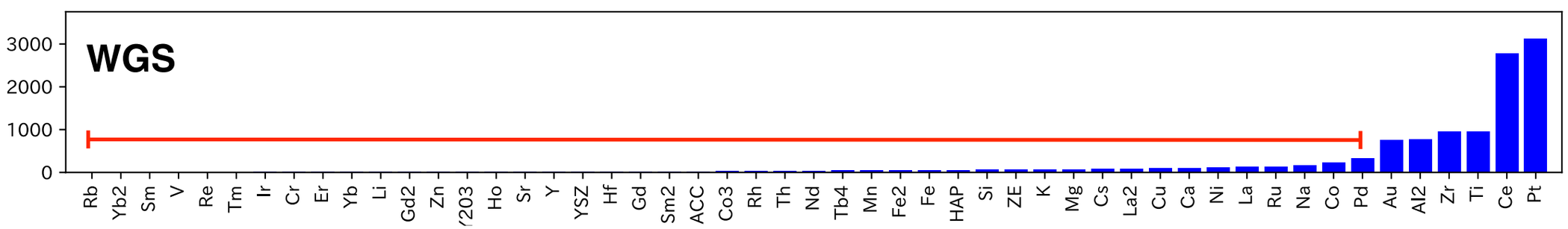
Two big problems we had

- **Problem 1: Data sparsity (Low sample counts for many elements)**
  - For compositions, **only a few are non-zero**. (very sparse table)
  - Non-zero elements are very biased, many elements have only a few nonzero samples (low sample counts), and **statistically negligible**...

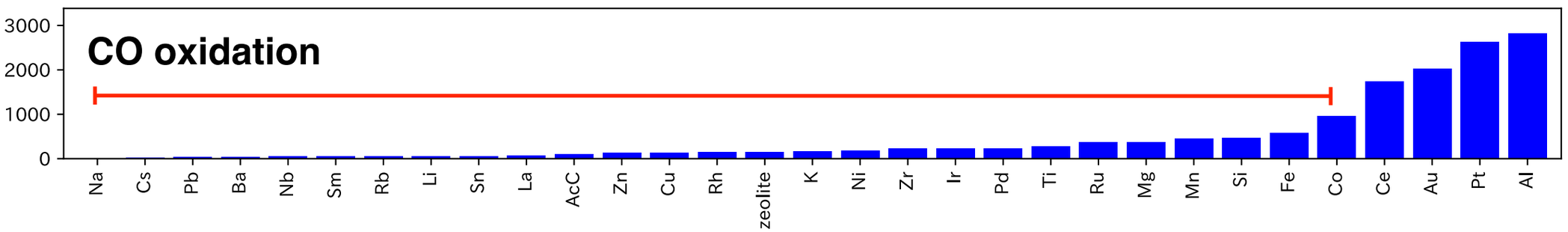
 # Catal: 1866
 # Elem: 74



 # Catal: 4360
 # Elem: 51



 # Catal: 5610
 # Elem: 31



Two big problems we had

- **Problem 1: Data sparsity (Less compositional overlaps)**

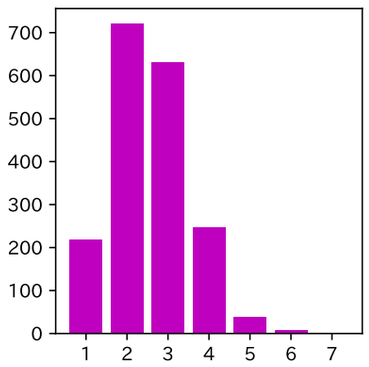
	A	B	C	D	E	
Cat-ABC =	(0.90,	0.06,	0.04,	0.00,	0.00)	catalyst with 90% A, 6% B, and 4% C
Cat-BCD =	(0.00,	0.30,	0.10,	0.60,	0.00)	catalyst with 60% D, 30% B, and 10% C
Cat-BCE =	(0.00,	0.30,	0.10,	0.00,	0.60)	catalyst with 60% E, 30% B, and 10% C

  - The similarities for ABC-BCD and ABC-BCE becomes the same...

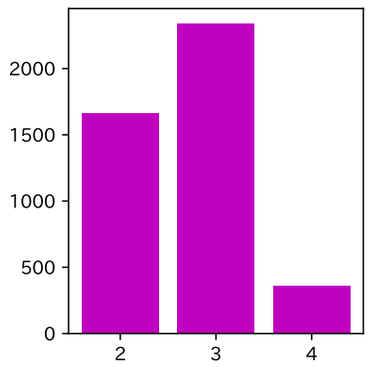
- For large datasets, this composition vectors are very sparse and mostly the overlapped elements are only one or two (or even zero...)

 # elems in a catalyst

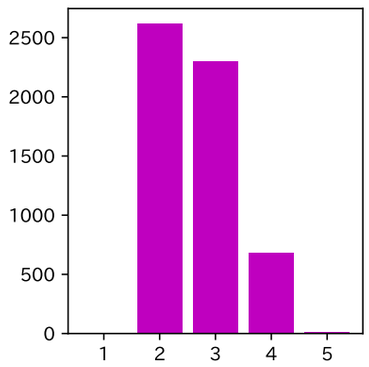
OCM



WGS

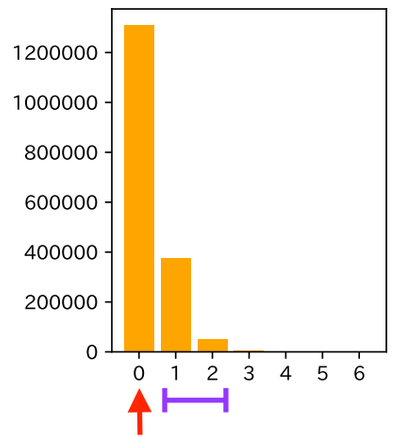


CO oxidation

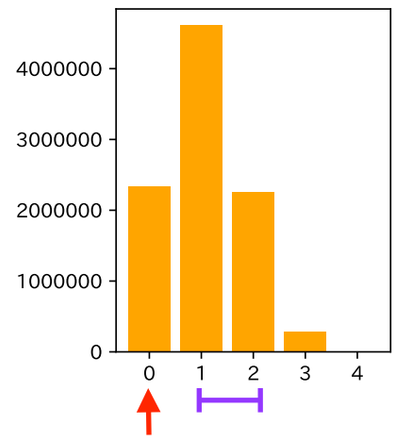


 # overlapped elems (for a pair)

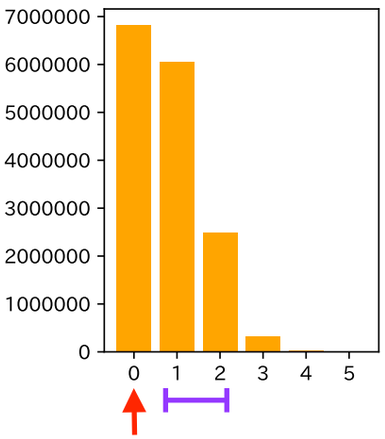
OCM



WGS



CO oxidation



Our solution: Integrating elemental discriptors

All elements appeared in the data (many and diverse elements)

- Sparse (many zeros), and less compositional overlaps
- low sample count for many elements except a few very widely used elements, and hard to directly compare

Literature data

Catalyst	A	B	C	D	Exp. conditions
Catal-1	0.90	0.06	0.04	0.00	
Catal-2	0.00	0.10	0.00	0.90	
⋮					

ML

Prediction

Catalyst	Performance
Catal-1	20.4
Catal-2	6.8
⋮	

Elemental descriptors (features)

Element	Desc-1	Desc-2	…	Desc-p
A	2	1.67		103.1
B	21	0.37		0.9
C	16	2.30		40.1
D	69	6.20		23.6

- Periodic table
- Handbook
- Textbook

Proposed (High-dimensionality is addressed by ML methods)

Catalyst	A	B	C	D	Primary feat.	Secondary feat.	Tertiary feat.	Exp. conditions
Catal-1	0.90	0.06	0.04	0.00	0.90 × Desc(A)	0.06 × Desc(B)	0.04 × Desc(C)	
Catal-2	0.00	0.10	0.00	0.90	0.90 × Desc(D)	0.10 × Desc(B)	0.00	
⋮								

Compositional information

Elemental features are considered for catalyst characterization

Features from not contained elements are zero out

Two big problems we had

- **Problem 2: Very strong "selection bias" in existing datasets**

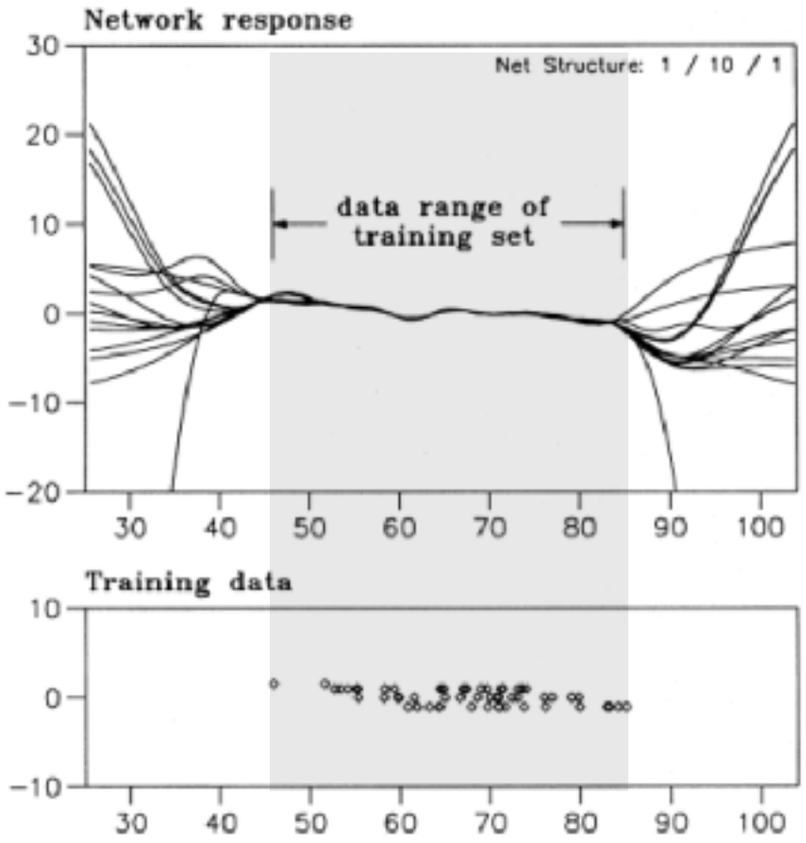
  Catalyst research has relied heavily on prior published data, tends to be biased toward catalyst composition that were successful

  Example) Oxidative coupling of methane (OCM)

  - 1868 catalysts in the original dataset [Zavyalova+ 2011]
  - Composed of 68 different elements: 61 cations and 7 anions (Cl, F, Br, B, S, C, and P) excluding oxygen
  - only 317 catalysts performed well with C_2 yields 15% and C_2 selectivity 50%; Occurrences of only a few elements such as La, Ba, Sr, Cl, Mn, and F are very high.
  - Widely used elements such as Li, Mg, Na, Ca, and La also frequent in the data

An ML model is just representative of the training data

Highly Inaccurate Model Predictions from Extrapolation (Lohninger 1999)



"Beware of the perils of extrapolation, and understand that ML algorithms build models that are representative of the available training samples."

We also need this

"exploration"

to obtain new knowledge/data

ML basically for this

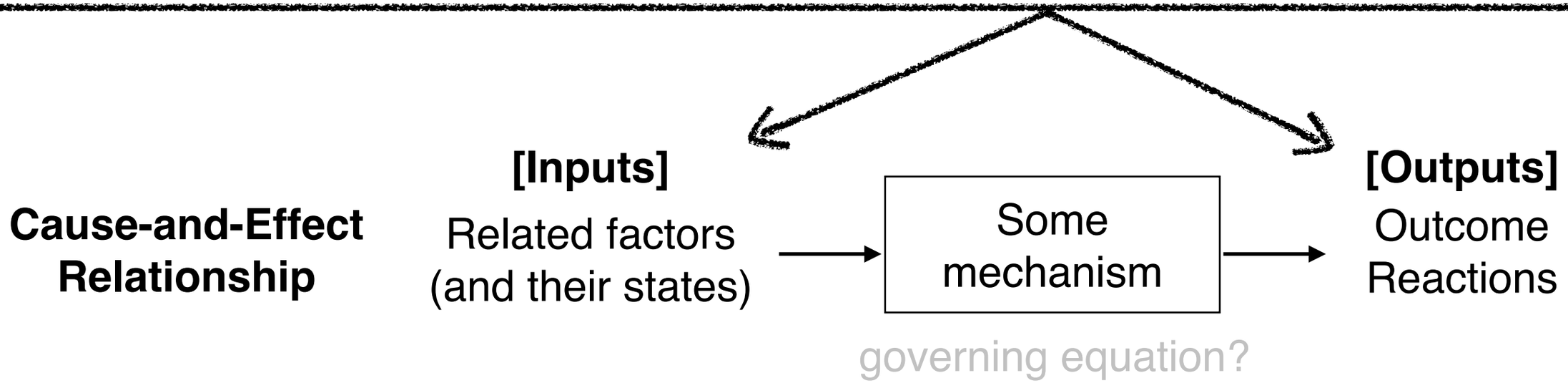
"exploitation"

to use the knowledge/data to improve the performane

No guarantee of data-driven for the outside of given data

Keep in mind: Given data DEFINES the data-driven prediction!

Data-driven methods try to precisely approximate its outer behavior (the input-output relationship) observable as "data".
(e.g. through *machine learning* from a large collection of data)



Theory-driven methods try to explicitly model the inner workings of a target phenomenon (e.g. through first-principles simulations)

Empirical optimization: "Edisonian empiricism"

問題：時間とコストは有限！！

理論的に可能なあらゆる候補を
この方式で検証することは不可能

次の実験計画へ
feedback

既知の知見・
観測(データ)

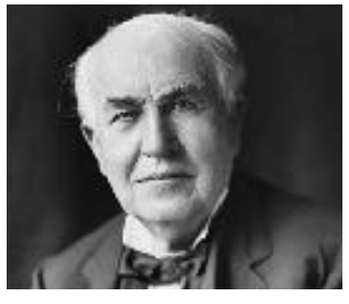
仮説形成

- 実験
- シミュレーション

仮説検証

結果の確認と
検証

"観察と帰納 (empirical/inductive)"　　"論理と演繹 (rational/deductive)"



Thomas Edison先生

- Genius is 1% inspiration and 99% perspiration.
- There is no substitute for hard work.
- I have not failed. I've just found 10,000 ways that won't work.
- :

よく考えるとブラックなことしか言ってない！

科学的発見とセレンディピティ

問題：時間とコストは有限！！
理論的に可能なあらゆる候補を
この方式で検証することは不可能

次の実験計画へ
feedback

既知の知見・観測(データ)

仮説形成

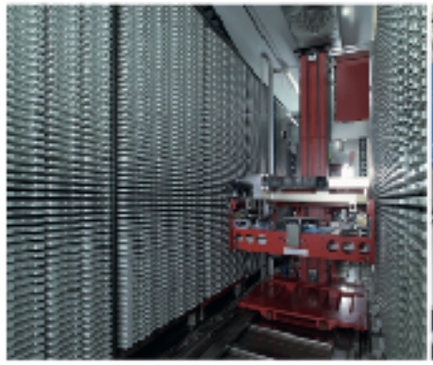
- 実験
- シミュレーション

仮説検証

結果の確認と検証

- **それゆえ「研究者のセンス・腕の見せ所」＋「幸運(セレンディピティ)」に依存する**
  筋の良さそうな候補を選ぶ、今まで試されてない全く新しいやり方を思いつく、etc
- **候補があまりに膨大(実質ほぼ無限)なので(数多く試すのは有利だとは言え…)**
  必ずしも「力技とお金と人海戦術で数多く試した者が勝つ」とは限らない
- **仮説形成はふつう完全に行き当たりばったりではない。「勘と経験」が非常に大切。**
  すばらしい発見は…「完全に運」＜「幸運は準備された者に降りる」

機械学習/データ科学の利用に基づく最適実験計画？



次の実験計画、
feedback

既知の知見・
観測(データ)

高速・高精度な
Data-Driven予測

結果の確認と
検証

仮説形成

(機械学習・データマイニング)

- どういう実験・シミュレーションを次に行うかの計画立案
- 時間のかかる計算の高精度高速近似
- 曖昧な因子や実験条件の最適化
- Multilevelの情報統合

仮説検証

(シミュレーション+実験)

- 再現性を担保する高精度・高速実験系
- 仮想化検証が可能な因子のシミュレーション(計算科学)による探索
  → 望ましい対象のさらなる絞り込み

Model-based optimization

The "Surrogate" model for
- Demanding experiments
- Time-consuming hi-fidelity simulations

Available data and findings

Generate hypotheses

Machine Learning or any "data-driven" predictions

predictions

Check and validate results

Grrr, give me any serendipity...

Model-based optimization

The "Surrogate" model for
- Demanding experiments
- Time-consuming hi-fidelity simulations

Available data and findings

Generate hypotheses

Machine Learning or any "data-driven" predictions

predictions

Check and validate results

CAUTION

We need a mechanism balancing exploitation and exploration

Grrr, give me any serendipity...

Machine Learningは与えたデータを代表するだけ

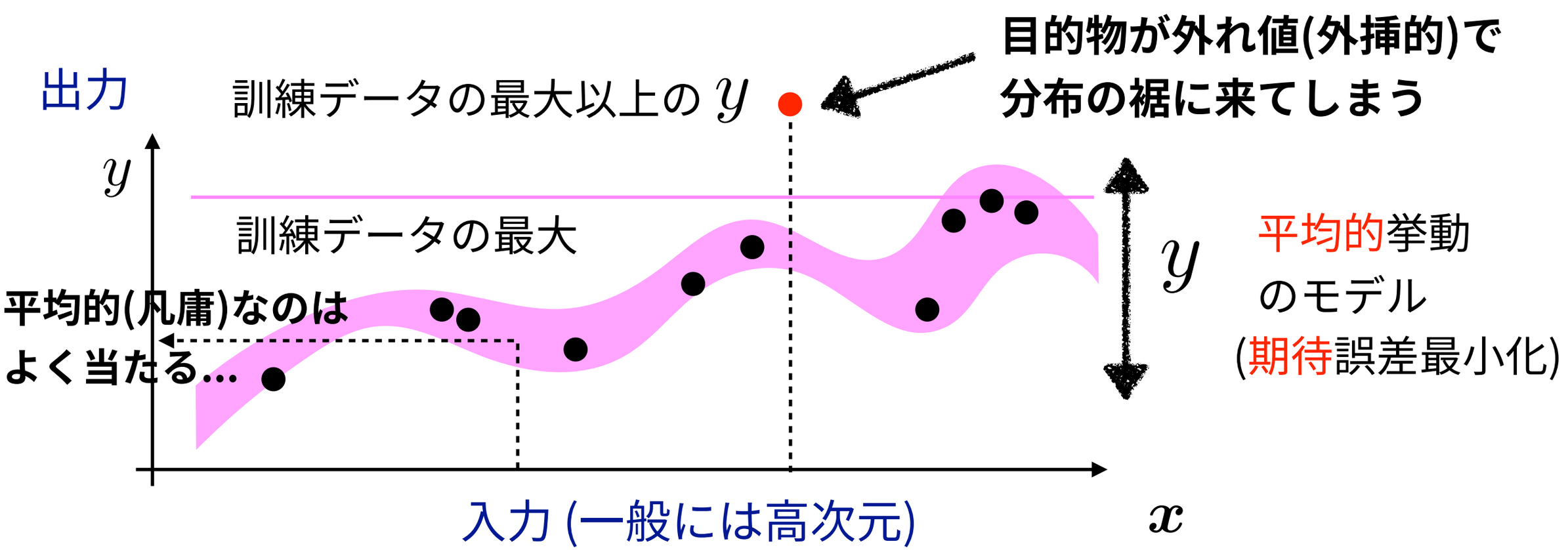
機械学習 = 訓練データの平均的法則性をとらえる

目的が不整合

予測モデルとの誤差の「期待値」を最小化 = 汎化

発見 = 訓練データの中にないものを見つけたい

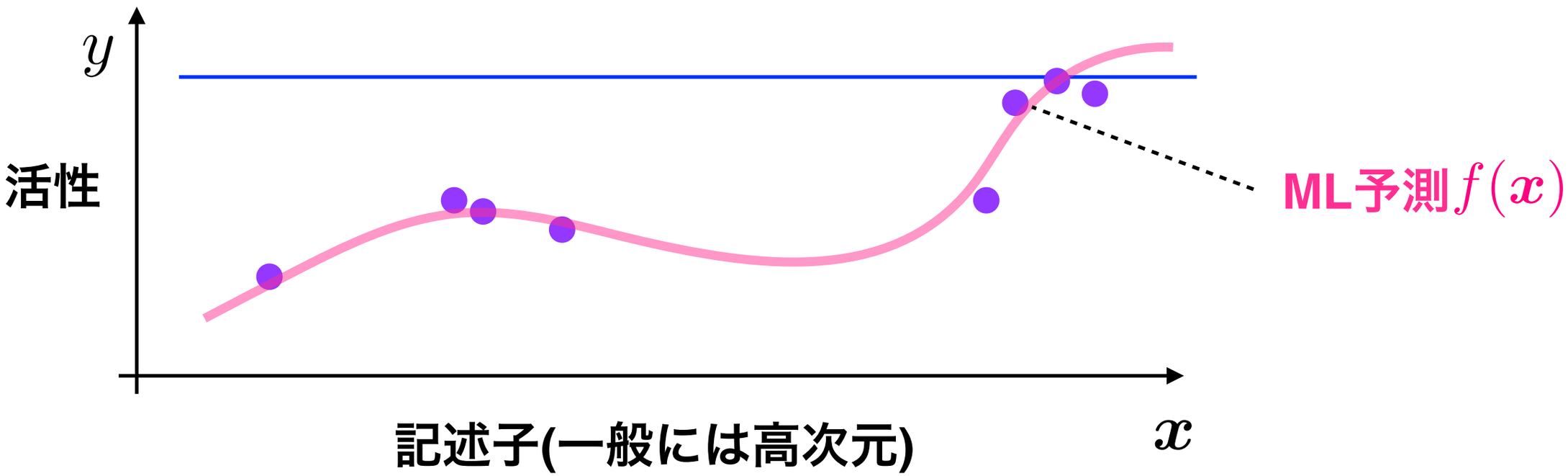
「外れ値」



Optimal design of experiments / Active learning

For next experiments, we face a dilemma in choosing between options.

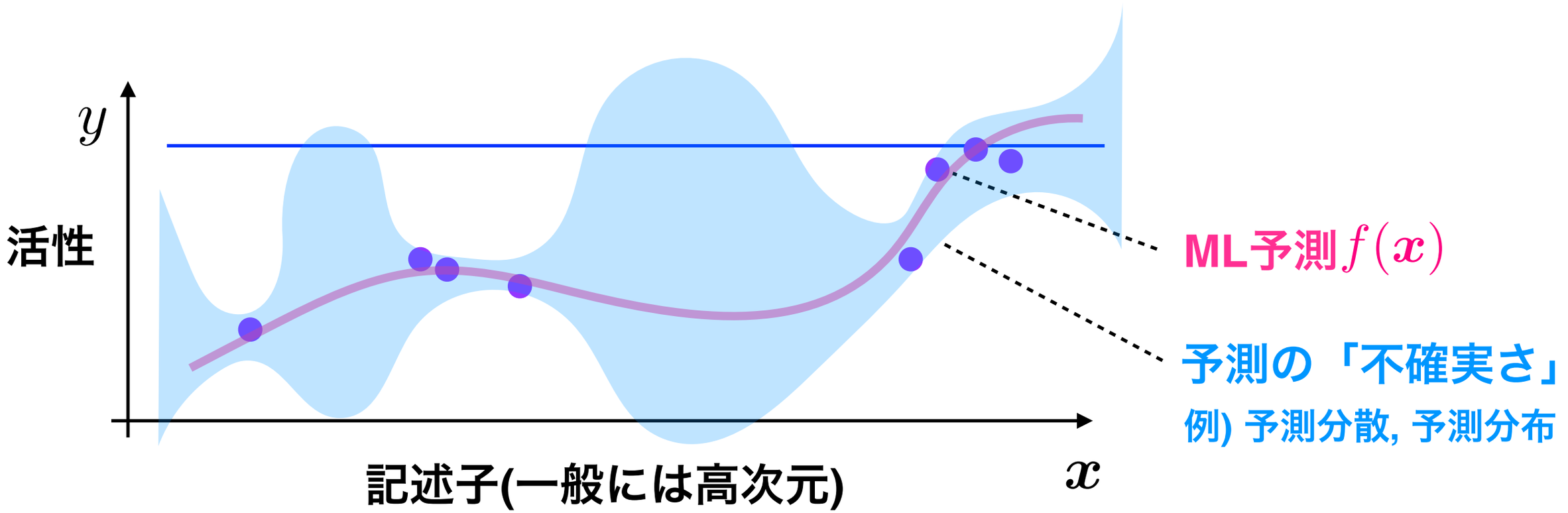
- Exploitation:
  what we already know and get something close to what we expect
- Exploration:
  something we aren't sure about and possibly learn more



Optimal design of experiments / Active learning

For next experiments, we face a dilemma in choosing between options.

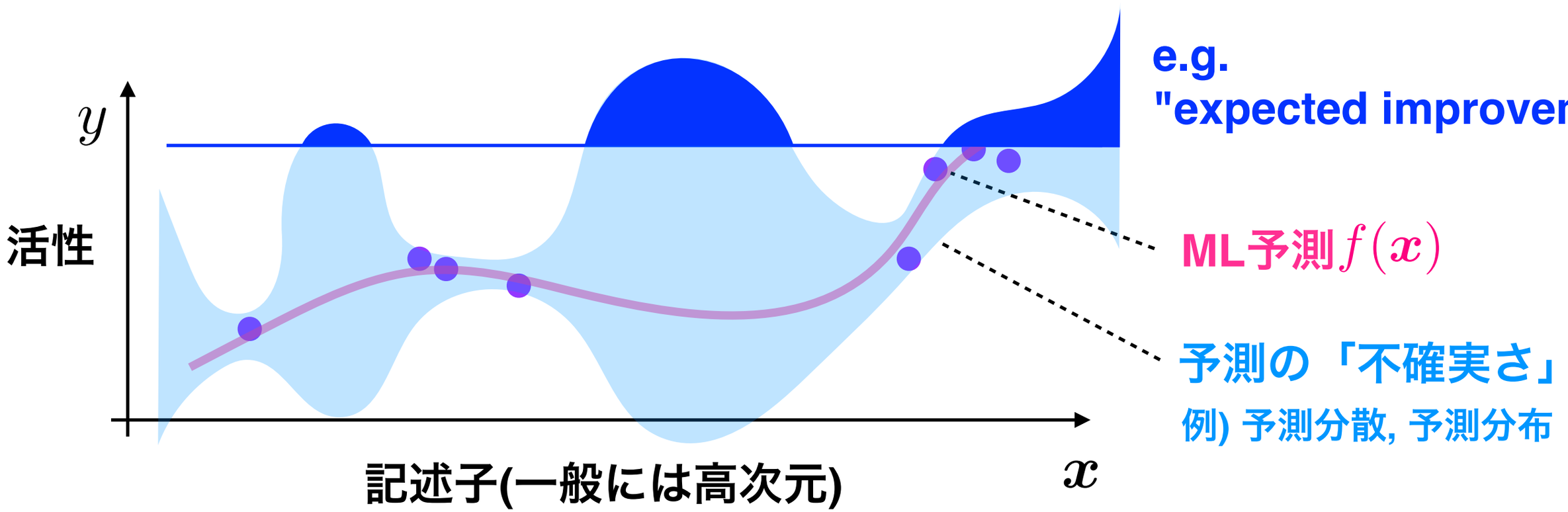
- Exploitation:
  what we already know and get something close to what we expect
- Exploration:
  something we aren't sure about and possibly learn more



Optimal design of experiments / Active learning

For next experiments, we face a dilemma in choosing between options.

- Exploitation:
  what we already know and get something close to what we expect
- Exploration:
  something we aren't sure about and possibly learn more



Model-based optimization

Use ML to guide the balance between "exploitation" and "exploration"!

Model-based optimization

1. Initial Sampling (DoE)
2. Loop:
    1. Construct a Surrogate Model.
    2. Search the Infill Criterion.
    3. Add new samples.

An Open Research Topic in ML

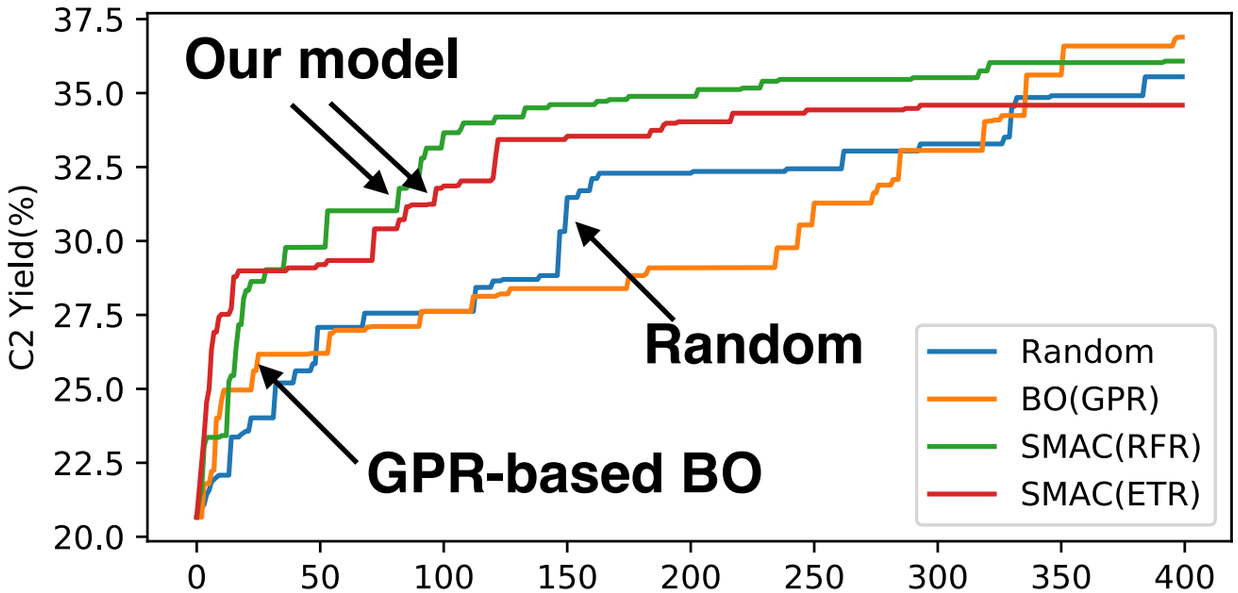
- Optimal design of experiments (DoE)
- Active learning
- Bayesian optimization
- Blackbox optimization
- Reinforcement learning
- Multi-armed bandit
- Evolutional computation
- Game-theoretic approaches

Our solution:

- **Representation**: Our elemental-descriptor based vectors
- **Surrogate**: Tree ensembles with prediction variance
- **Optimization**: Extending SMAC algorithm (Hutter+ 2011) to
    - Constrained search (e.g. sum to 1, [0,1]-valued, one-hot encodings)
    - Sparsity constraint (otherwise it always suggests dense vectors..)
    - Discrete local search for nominal value updating
    - Multiple suggestion at one time (batched optimization)
- **Infill criterion**: Expected improvement (EI) + small explicit exploration

Results

- **Our model finds high-performance catalysts more quickly than other alternatives**

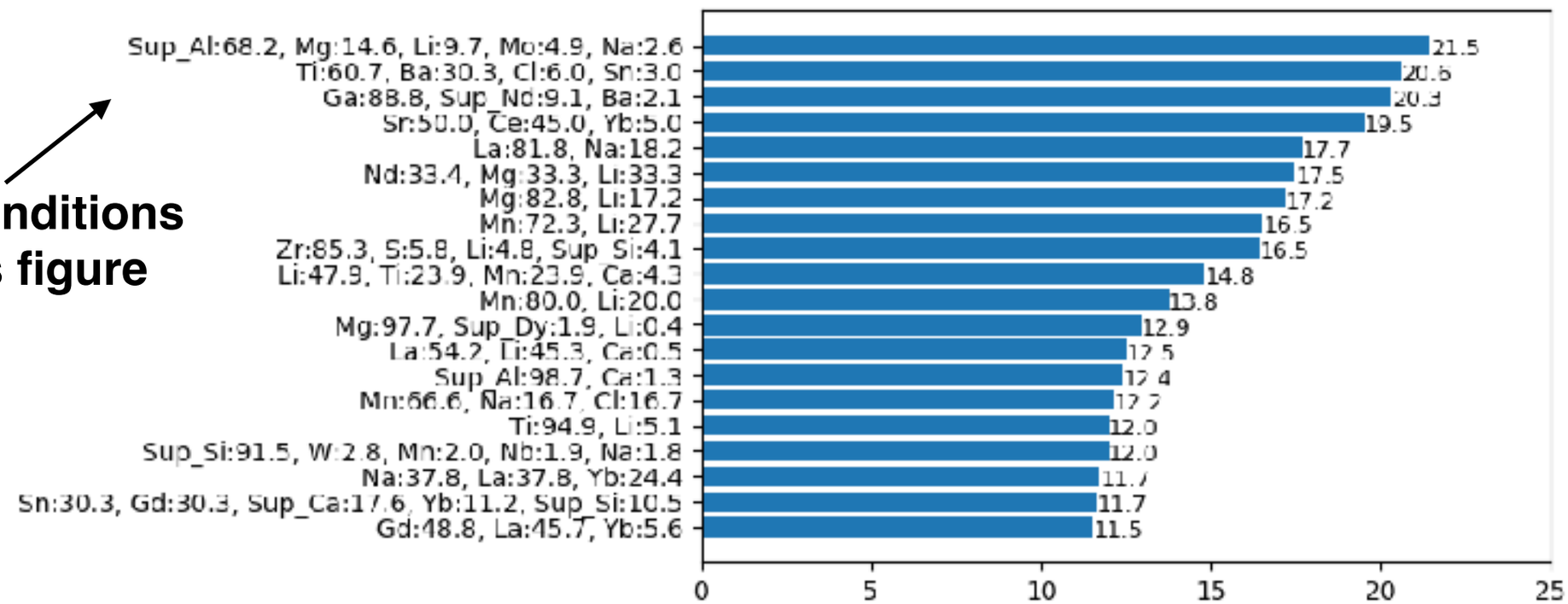


Test on 3 Datasets

- Oxidative coupling of methane (OCM) [Zavyalova+ 2011]
- Water gas shift (WGS) [Odabaşi+ 2014]
- CO oxidation [Günay+ 2013]

- **Our models can suggest a list of promising candidate catalysts with experimental conditions from the entire available data**

Suggested exp conditions are omitted in this figure



機械学習分野の参考事例

モデルベースの強化学習

AlphaGo
(*Nature*, Jan 2016)


ARTICLE

Mastering the game of Go without human knowledge

AlphaGo Zero
(*Nature*, Oct 2017)


Silver *et al.*, *Science* **362**, 1140–1144 (2018) 7 December 2018

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

AlphaZero
(*Science*, Dec 2018)


Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model

MuZero
(*arXiv*, Nov 2019)

AutoML (全自動機械学習)

- Algorithm Configuration
- Hyperparameter Optimization (HPO)
- Neural Architecture Search (NAS)
- Meta Learning / Learning to Learn

Cloud AutoML

AutoML

OPTUNA

AutoDL 2019

Amazon SageMaker

蓄積された「計算・実験・知識データ」の利用

実験データ・計算データ・ファクトの蓄積

In-Houseデータ + Publicデータ + 知識ベース
+ そのQuality Control / Annotations)

仮説形成

(機械学習・データマイニング)

- どういう実験・シミュレーションを次に行うかの計画立案
- 時間のかかる計算の高精度高速近似
- 曖昧な因子や実験条件の最適化
- Multilevelの情報統合

検証

(シミュレーション+実験)

- 再現性を担保する高精度・高速実験系
- 仮想化検証が可能な因子のシミュレーション(計算科学)による探索
  → 望ましい対象のさらなる絞り込み

This trend emerged first in life sciences (drug discovery)

NATURE REVIEWS | DRUG DISCOVERY
VOLUME 17 | FEBRUARY 2018 | 97

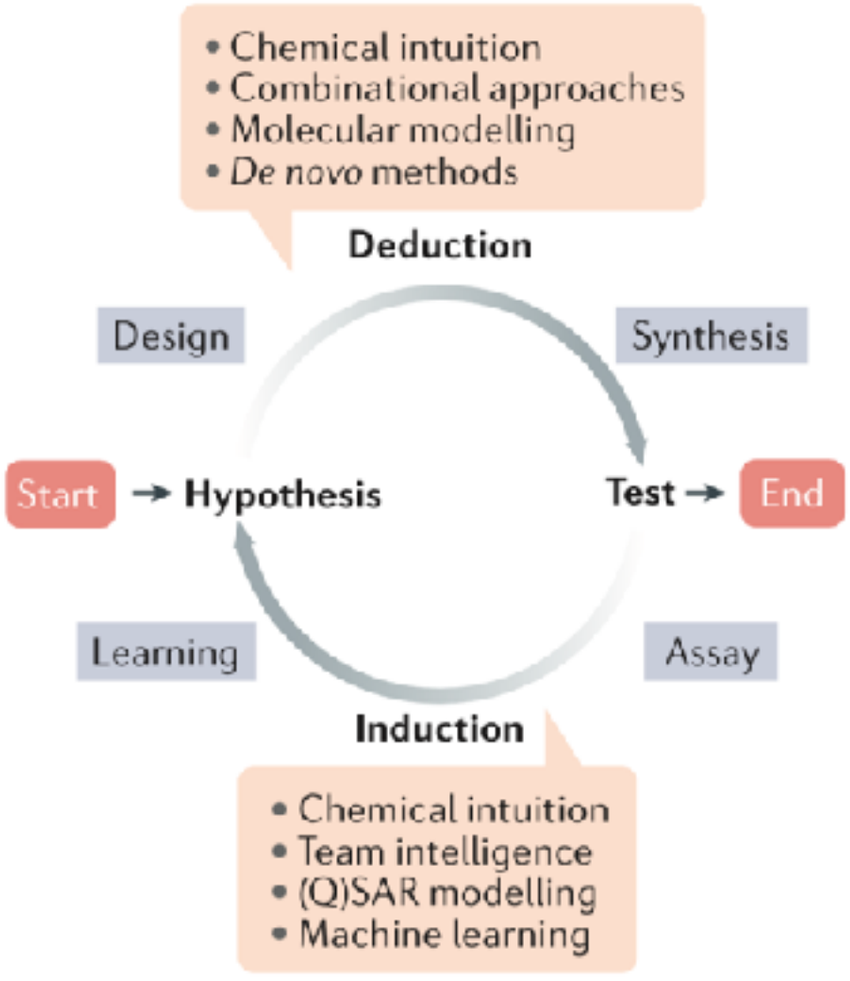
PERSPECTIVES

INNOVATION

Automating drug discovery

Gisbert Schneider

Figure 2 | **Automated drug discovery facilities.** **a** | Millions of compound samples are stored in compact high-capacity facilities and handled by robots. **b** | Robot systems perform both high-throughput and medium-throughput screening of up to ten thousand samples per day to determine the activity against the biological target of interest. Multiple arms and flexible workstations enable fully automated liquid dispensing, compound preparation and testing. These storage and screening systems have become cornerstones of contemporary drug discovery. **c** | A prototype of a novel miniaturized design–synthesize–test–analyse facility for rapid automated drug discovery at AstraZeneca is shown. Images **a** and **b** courtesy of Jan Kriegl, Boehringer–Ingelheim Pharma; image **c** courtesy of Michael Kossenjans, AstraZeneca.



Next expanded to materials science

Toyota teams with China's CATL and BYD to power electric ambitions

Automaker diversifies battery source and moves up electrification goal by 5 years

YUKIHIRO OMOTO, Nikkei staff writer
JUNE 07, 2019 02:00 JST ● UPDATED ON JUNE 07, 2019 14:39 JST

Little human intervention for highly reproducible large-scale production lines

CATL

Automation, monitoring with IoT, and big-data management are also the key to manufacturing.

Now these focuses shifted to the R & D phases.
(very experimental and empirical traditionally)

Next expanded to materials science

Toyota teams with China's CATL and BYD to power electric ambitions

Automaker diversifies battery source and moves up electrification goal by 5 years

YUKIHIRO OMOTO, Nikkei staff writer
JUNE 07, 2019 02:00 JST ● UPDATED ON JUNE 07, 2019 14:39 JST

Little human intervention for highly reproducible large-scale production lines

CATL

Automation, monitoring with IoT, and big-data management are also the key to manufacturing.

Now these focuses shifted to the R & D phases.
(very experimental and empirical traditionally)

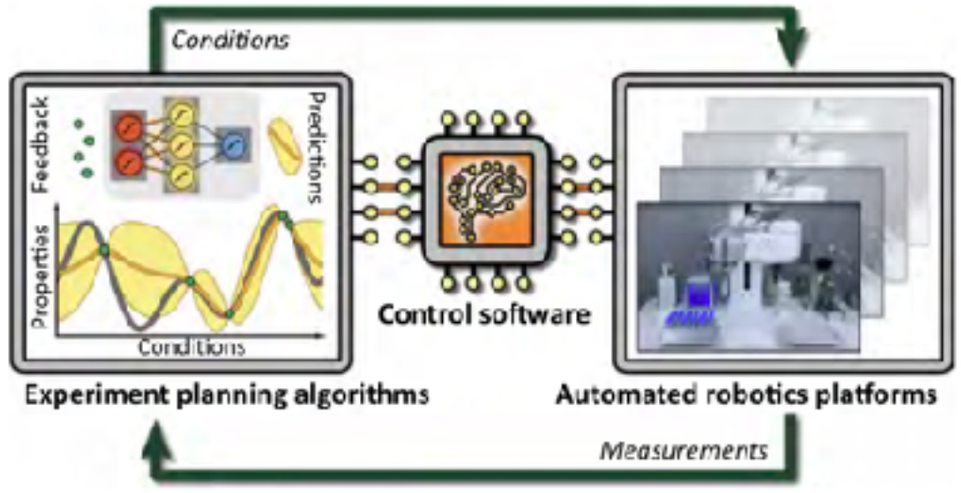
... also to chemistry!

Trends in Chemistry, June 2019, Vol. 1, No. 3 10.1016/j.trechm.2019.02.007

Opinion

Next-Generation Experimentation with Self-Driving Laboratories

Florian Häse,[1,2,3,4] Loïc M. Roch,[1,2,3,4] and Alán Aspuru-Guzik[1,2,3,4,5,*]



How to explore chemical space using algorithms and automation

Piotr S. Gromski, Alon B. Henson, Jarosław M. Granda and Leroy Cronin

PERSPECTIVES
NATURE REVIEWS | CHEMISTRY

Machine-Assisted Chemistry | Special Issue 150 Years of BASF

DOI: 10.1002/anie.201410744

Organic Synthesis: March of the Machines

Steven V. Ley, Daniel E. Fitzpatrick, Richard J. Ingham, and Rebecca M. Myers*

Angew. Chem. Int. Ed. **2015**, *54*, 3449 – 3464

Angewandte International Edition Chemie

Effective use of data is another key in natural sciences

(In addition to experiments and simulations)

And please keep in mind that **unplanned data collection is dangerous**.
We need **right designs for data collection and right tools to analyze.**

A bitter lesson: "low input, high throughput, no output science." (Sydney Brenner)

Science is changing, the tools of science are changing. And that requires different approaches. —— Erich Bloch, 1925-2016

Nature, 559
pp. 547–555 (2018)

REVIEW

https://doi.org/10.1038/s41586-018-0337-2

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]*

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The

generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

Science, 361
pp. 360-365 (2018)

SPECIAL SECTION *FRONTIERS IN COMPUTATION*

REVIEW

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

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

The discovery of new materials can bring enormous societal and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for achieving inverse design, which aims to discover tailored materials from the starting point of a particular desired functionality. Recent advances from the rapidly growing field of artificial intelligence, mostly from the subfield of machine learning, have resulted in a fertile exchange of ideas, where approaches to inverse molecular design are being proposed and employed at a rapid pace. Among these, deep generative models have been applied to numerous classes of materials: rational design of prospective drugs, synthetic routes to organic compounds, and optimization of photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

act properties. In practice, approximations are used to lower computational time at the cost of accuracy.

Although theory enjoys enormous progress, now routinely modeling molecules, clusters, and perfect as well as defect-laden periodic solids, the size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience (*11*). In the context of materials, ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice. Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploiting patterns in massive datasets, these models can distill average and salient features that characterize molecules (*12*, *13*).

Inverse design is a component of a more complex materials discovery process. The time

Science, 293
pp. 2051-2055 (2001)

VIEWPOINT

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

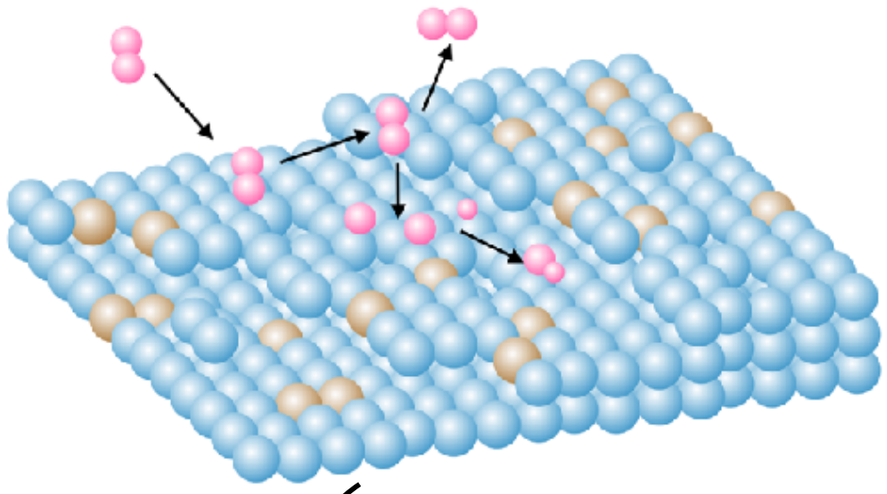
Machine learning (ML) (*1*) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the stan-

correlate surprisingly well with subsequent gene expression analysis (*3*). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (*4*), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to amplify every aspect of a working scientist's

creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern

Theory-driven vs Data-drivenの融合にむけて



伝統的な方法

Data-driven

現象の観察・ファクト

(実験科学)

説明

予測

仮説形成

仮説形成

Theory-driven

現象の理解・原理

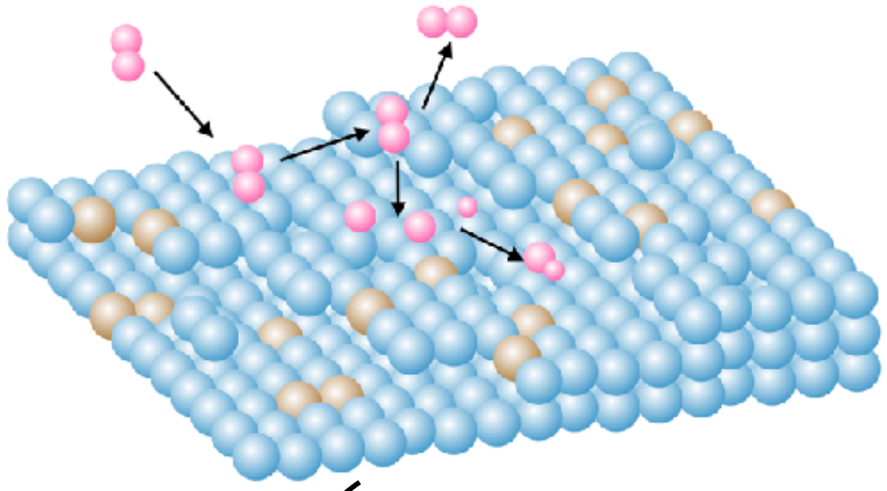
(理論科学・計算科学)

ICReDD
Institute for Chemical Reaction Design and Discovery
HOKKAIDO UNIVERSITY

より合理的・効率的な探索の確立
(新規な触媒・反応の発見へ)

ボトルネックの一つは「(筋の良い)仮説形成」
どのターゲット・条件・パラメタを試すか？

Theory-driven vs Data-drivenの融合にむけて



伝統的な方法

Data-driven

現象の観察・ファクト

(実験科学)

説明

予測

仮説形成

仮説形成

Theory-driven

現象の理解・原理

(理論科学・計算科学)

ICReDD Institute for Chemical Reaction Design and Discovery HOKKAIDO UNIVERSITY

より合理的・効率的な探索の確立
(新規な触媒・反応の発見へ)

ボトルネックの一つは「(筋の良い)仮説形成」
どのターゲット・条件・パラメタを試すか？

根拠・データ・
経験的ファクト

- 予測モデリング
- 不確定因子の最適化
- マルチレベル情報の統合

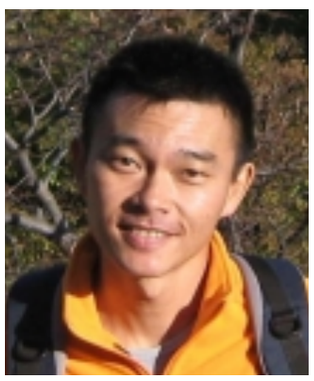
蓄積される知見(データ)の利活用 (情報科学)

- 候補集合を第一原理で提示/制約
- 仮想的検証・計算データの提供
- 明示的ドメイン知識

Summary

Takeaways:
first principles are not enough for us to throw away *empirical* things;
data-driven approaches (such as ML) play a complementary role!

1. Takigawa I, Shimizu K, Tsuda K, Takakusagi S
   RSC Advances. 2016; 6: 52587-52595.
2. Toyao T, Suzuki K, Kikuchi S, Takakusagi S, Shimizu K, Takigawa I.
   The Journal of Physical Chemistry C. 2018; 122(15): 8315-8326.
3. Suzuki K, Toyao T, Maeno Z, Takakusagi S, Shimizu K, Takigawa I.
   ChemCatChem. 2019; 11(18): 4537-4547.



Ken-ichi SHIMIZU (ICAT)

Satoru TAKAKUSAGI (ICAT)

Takashi TOYAO (ICAT)

Keisuke SUZUKI (DENSO)

https://itakigawa.github.io/news.html 瀧川一学 検索

少 年 易 老 學 難 成

Ars longa, vita brevis, occasio praeceps, experimentum periculosum, iudicium difficile.

ニュース・活動

ハイライト

- 最近の発表スライド slideshare
- 人工知能の基本問題：これまでとこれから 総説記事(Open Access) スライド
- 富山県寄附講義 データ社会を生きる技術〜人工知能のHypeとHope〜 スライド
- 統計数理研究所・リーディングDAT講座
  - 2019 L-S 決定木とアンサンブル学習の基礎と実践 Link
  - 2019 L-B2 機械学習とデータサイエンスの現代的手法 Link
  - 2018 L-B2 機械学習とデータサイエンスの現代的手法 Link
- 勾配法と機械学習 link