Machine Learning and Model-Based Optimization for Heterogeneous Catalyst Design and Discovery

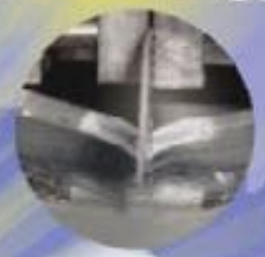
Ichigaku Takigawa

- Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University
- RIKEN Center for Advanced Intelligence Project (AIP)

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

Rationalize & Accelerate Chemical Design and Discovery

This Symposium Topic
Toward Interdisciplinary Research Guided by Theory and Calculation

To reach "**experimental realizations**" through "theory and calculation", supports by *implicit/empirical chemical knowledge* are still required.

Even though current chemical calculations are fairly accurate, chemists would still be driven by ...

1. **empirical findings (patterns)** reported in papers and textbooks
2. **experiences** actually doing a lot of experiments

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!

Automated reaction-path search via GRRM strategy

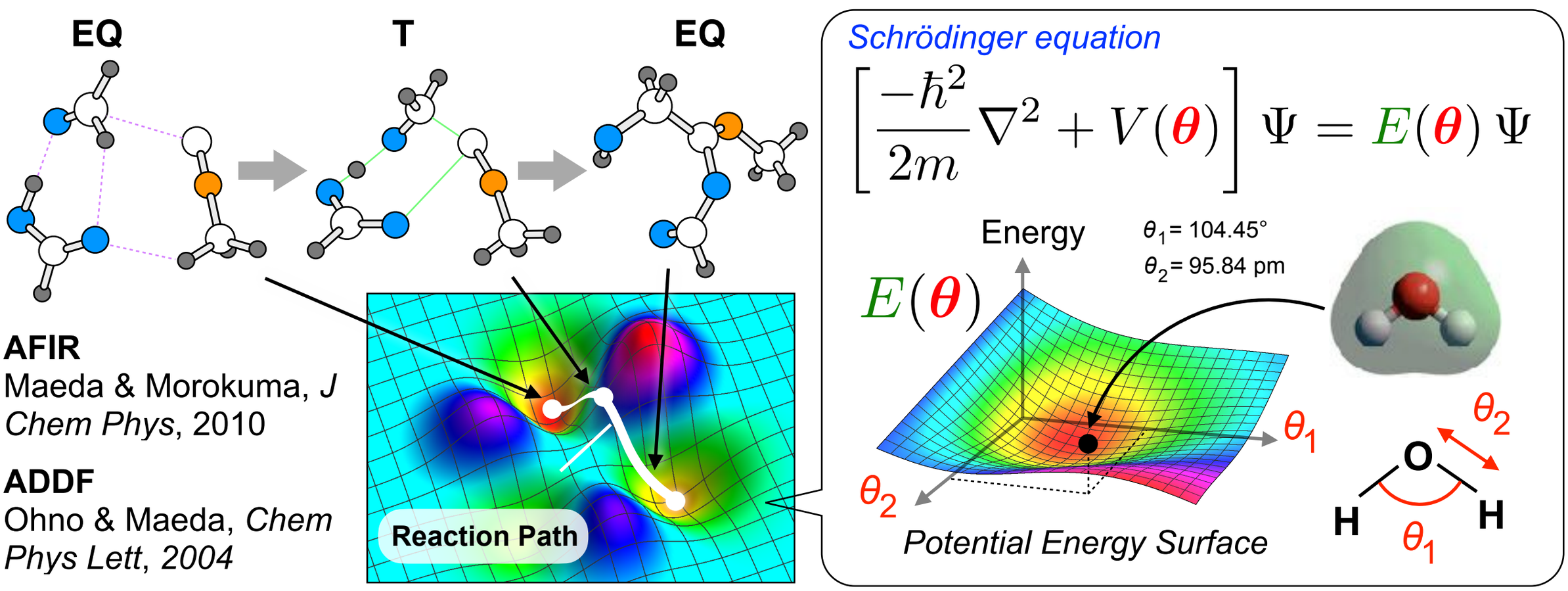
PERSPECTIVE

View Article Online
View Journal | View Issue

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

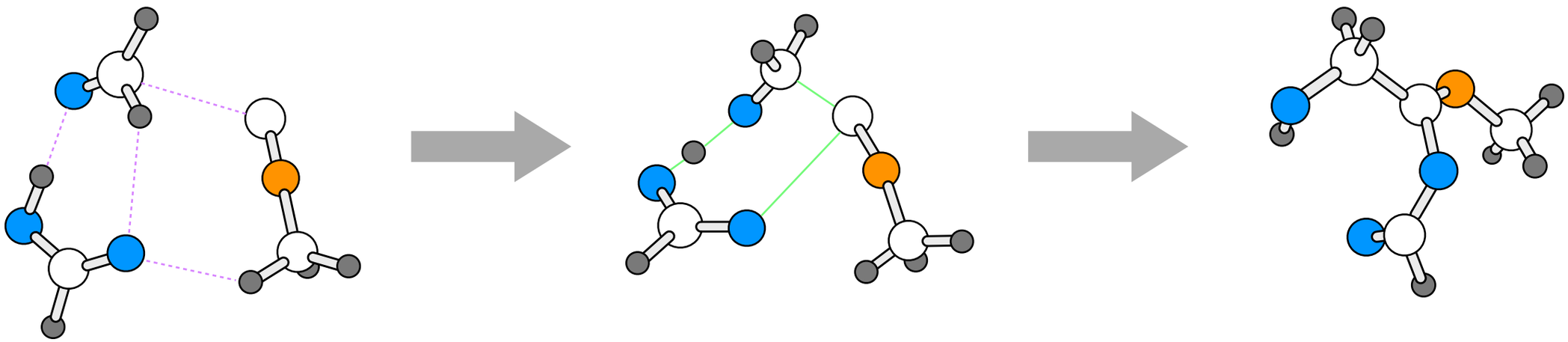
Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods

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



But computational chemistry has limitations for now...

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

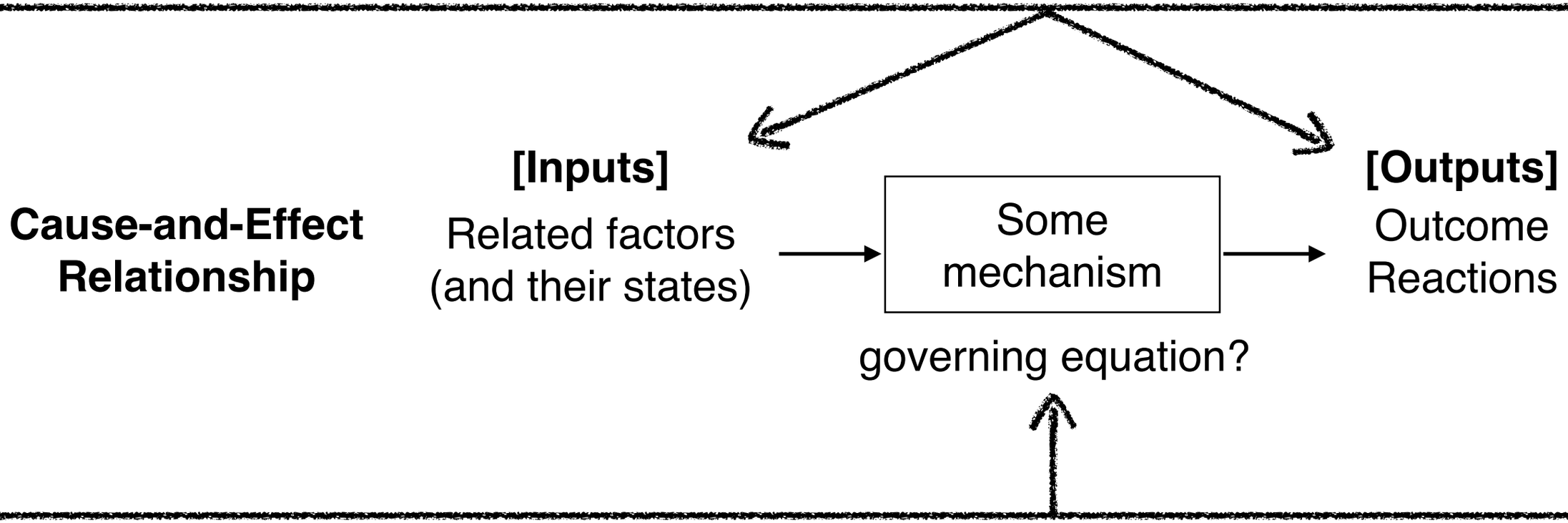


- **Intractably large chemical space**: A intractably large number of "theoretically possible" candidates for reactions and compounds...
- **Scalability issue**: Simulating an Avogadro-constant number of atoms is utterly infeasible... (After all, we need some compromise here)
- **Complexity and uncertainty 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...

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)

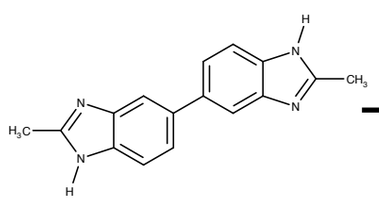


Theory-driven methods try to explicitly model the inner workings of a target phenomenon (e.g. through first-principles simulations)

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

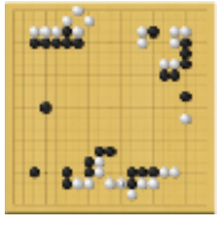
Speech Recognition

“ありがとう”

J’aime la musique

Machine Translation

I love music

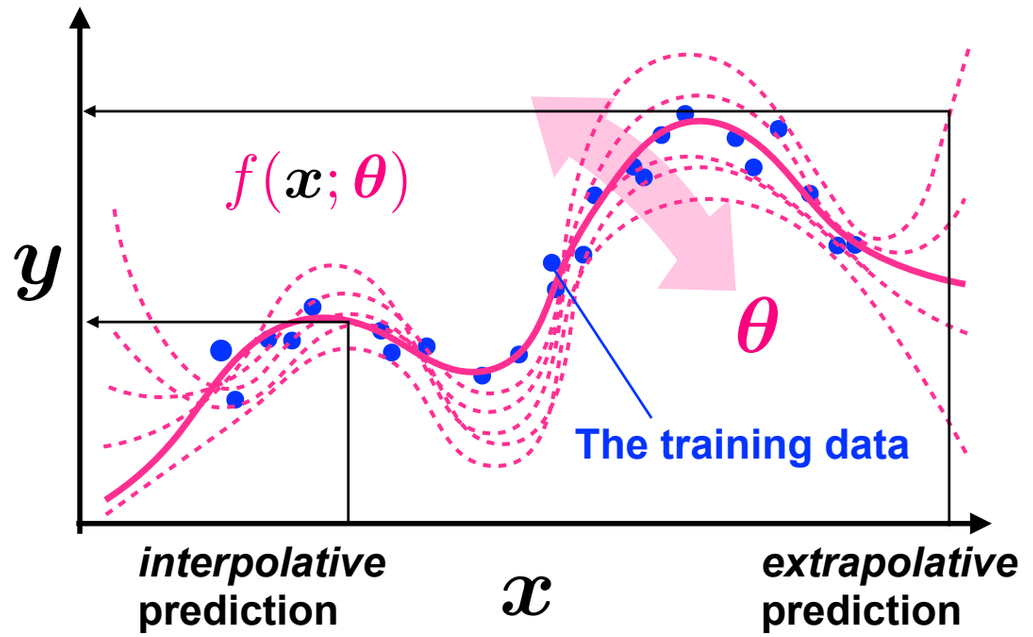


AI Game Players

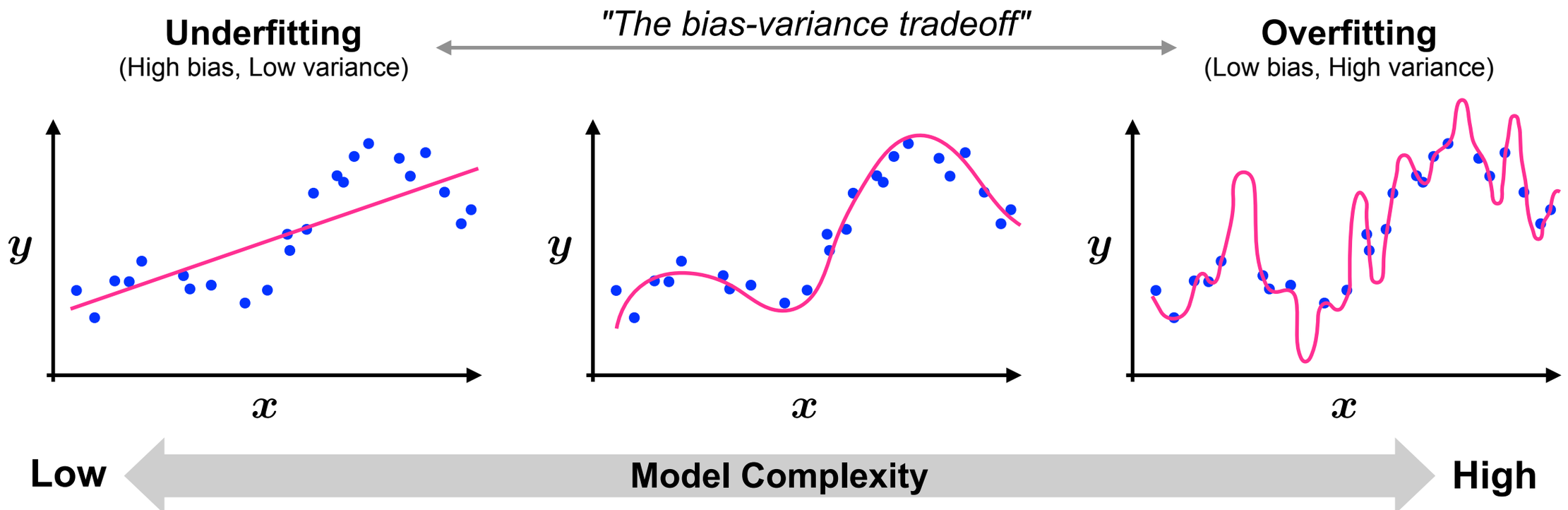
How ML works: fitting a function to data

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

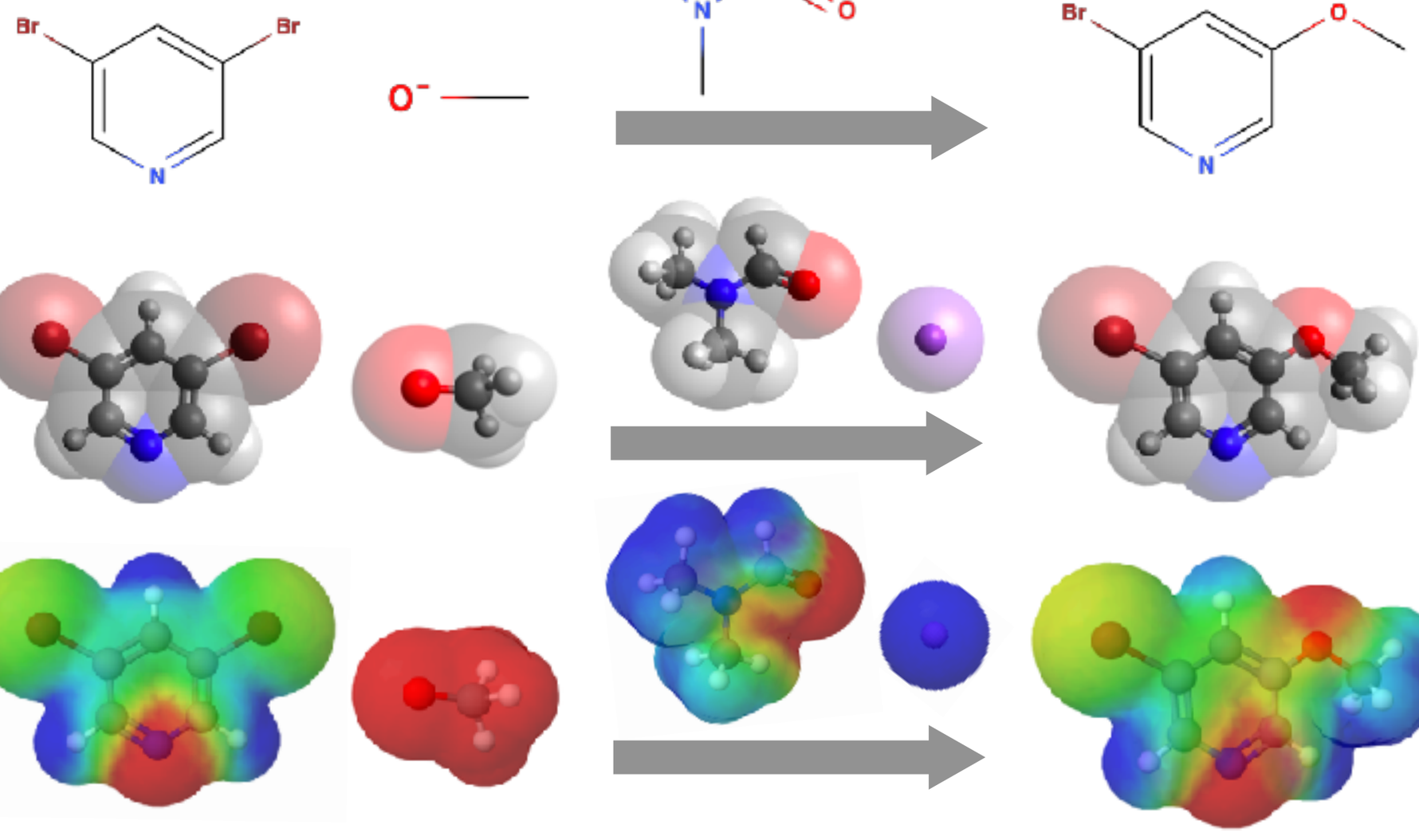
Structural Formla

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)

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

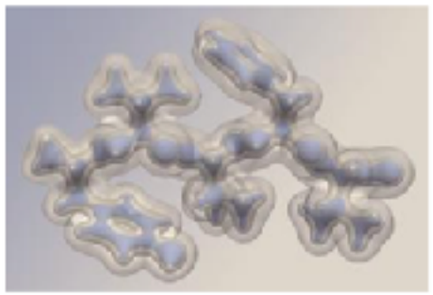
*Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski**

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

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

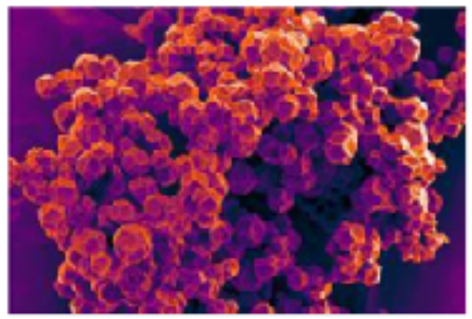
Chematica can now design efficient syntheses for large compound libraries

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

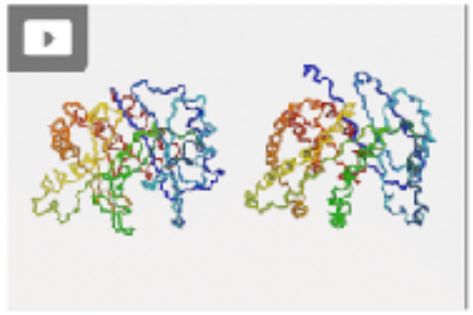


RESEARCH

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

Also hot topics in ML!

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*	**Molecular Transformer** Schwaller+ *ACS Cent Sci 2019*	**3N-MCTS/AlphaChem** Segler+ *Nature 2018*
WLN Coley+ *Chem Sci 2019*		**Molecule Chef** Bradshaw+ *DeepGenStruct (ICLR WS) 2019*

ML + First-principle simulations

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.

Both from

Google DeepMind

On-going ICReDD projects

- **Predictive Modeling for Heterogeneous Catalysis**
  with Institute for Catalysis (Hokkaido University)

- **Fast path-ranking algorithms to prioritize candidate reaction paths from automated reaction-path search**
  with Prof. Maeda's group

- **Automatic classification of trajectories from the first-principles molecular dynamics**
  with Prof. Taketsugu's group

- **Pathway analysis from single-cell RNA-seq on DN-gel induced cancer stem cells**
  with Prof. Tanaka's group

- **Output prediction of high-performance liquid chromatography**
  with Prof. Ito's group

Agenda

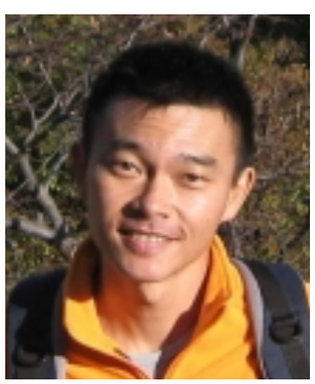
1. **Predicting the d-band centers by ML**
   Takigawa I, Shimizu K, Tsuda K, Takakusagi S
   RSC Advances. 2016; 6: 52587-52595.

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.

Institute for Catalysis
ICAT

CREST



Ken-ichi SHIMIZU (ICAT)

Satoru TAKAKUSAGI (ICAT)

Takashi TOYAO (ICAT)

Keisuke SUZUKI (DENSO)

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)

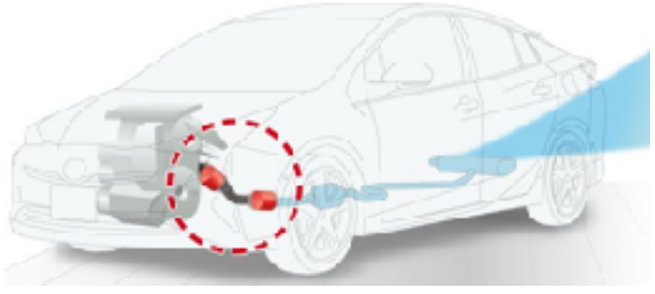
“Fertilizer from Air”

artificial nitrogen fixation

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

Ferrous Metal Catalysis

Exhaust Gas Purification



Exhaust Gas → **Harmless gas**

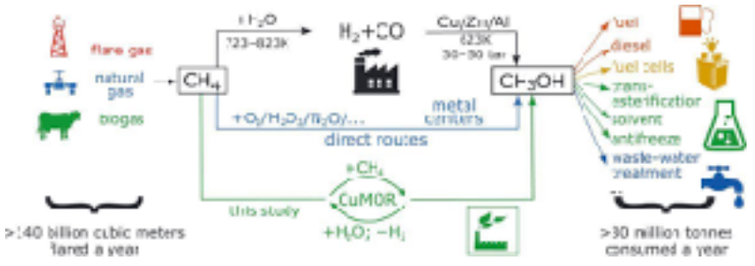
NOx → N_2

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

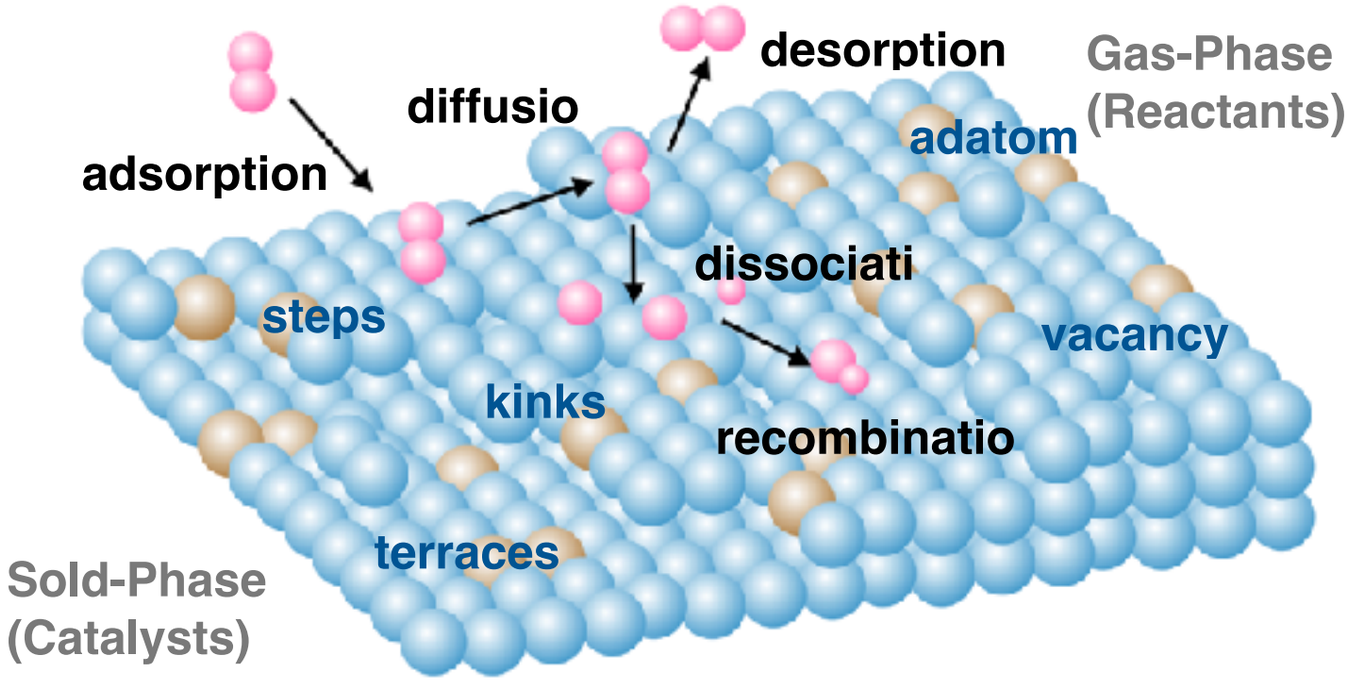
Notoriously complex surface reactions between different phases.

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)

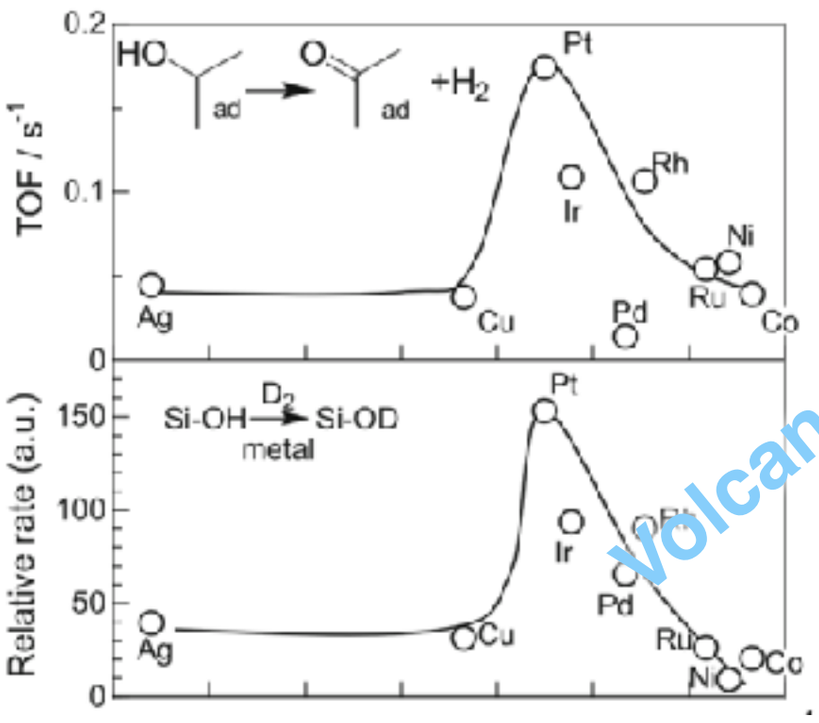
3. Can we predict the catalytic activity?

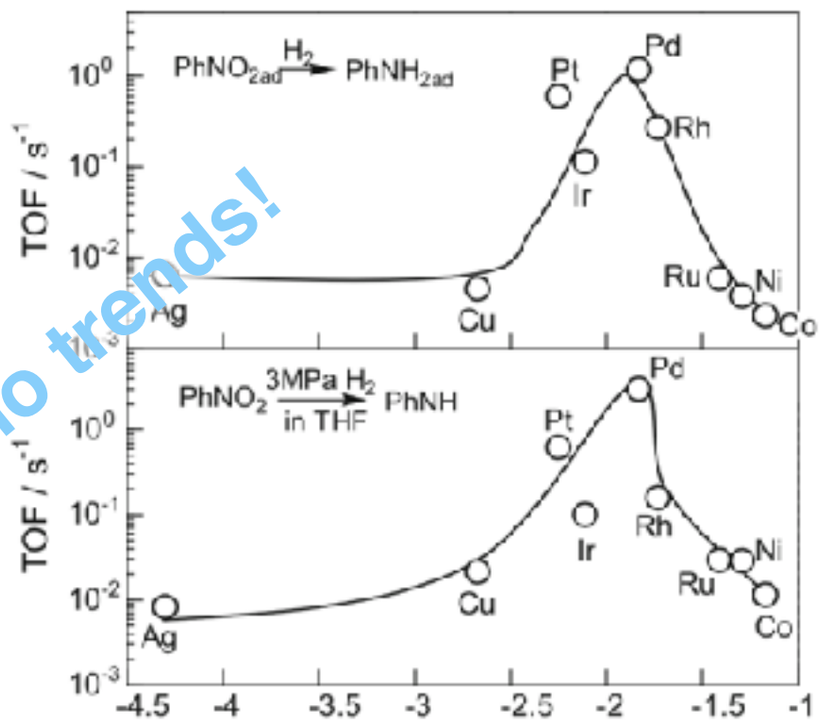
predicting values from experiments reported in the literature by machine learning
(Suzuki et al, ChemCatChem, 2019)

How to understand the catalytic activities?

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

Hammer–Nørskov d-band model





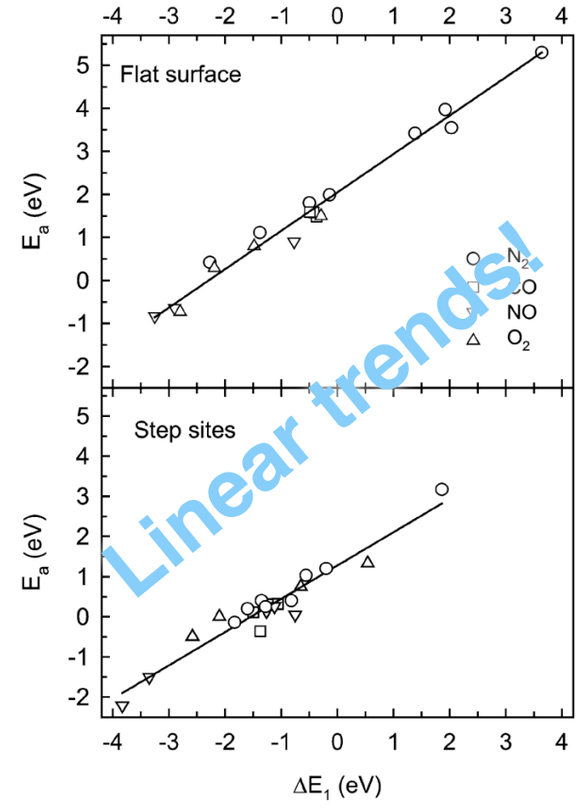
reaction rates

Volcano trends!

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

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

Brønsted-Evans-Polanyi relation



activation energy / eV

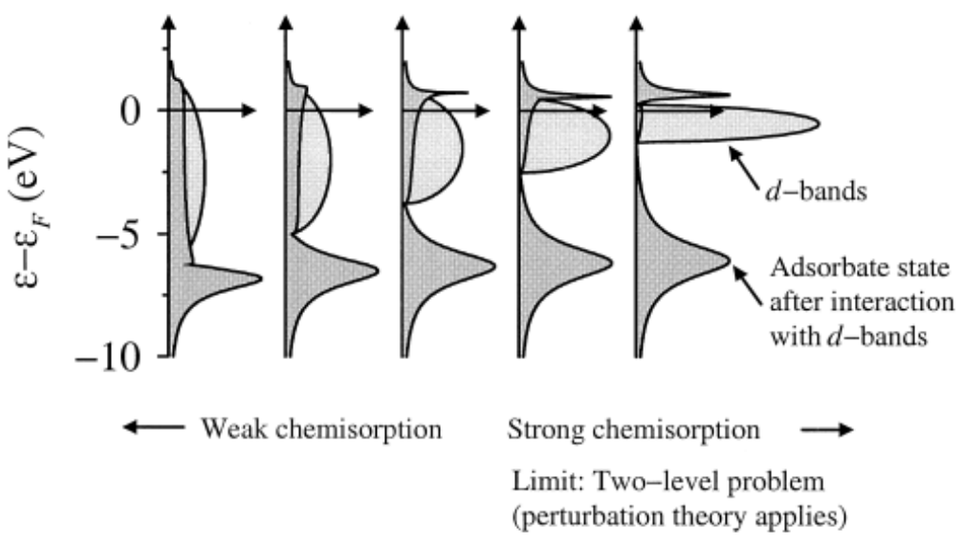
Linear trends!

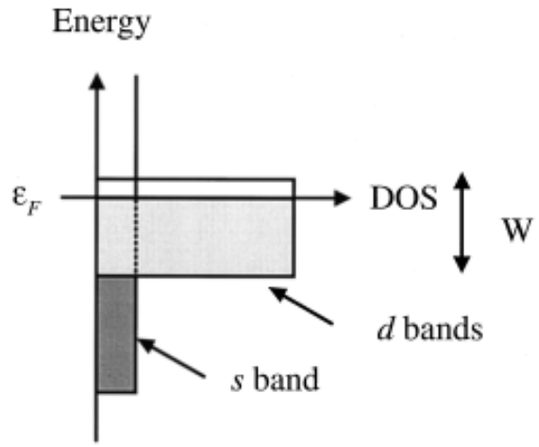
adsorption energy / eV

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).



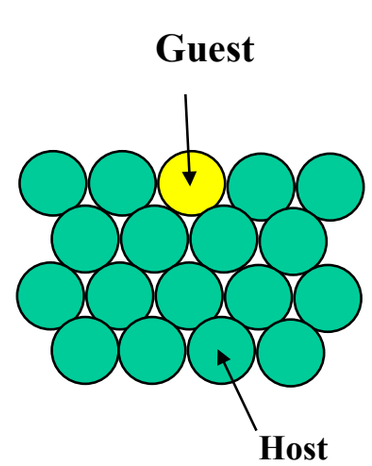


J. K. Nørskov, et al., *Advances in Catalysis*, 2000

[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

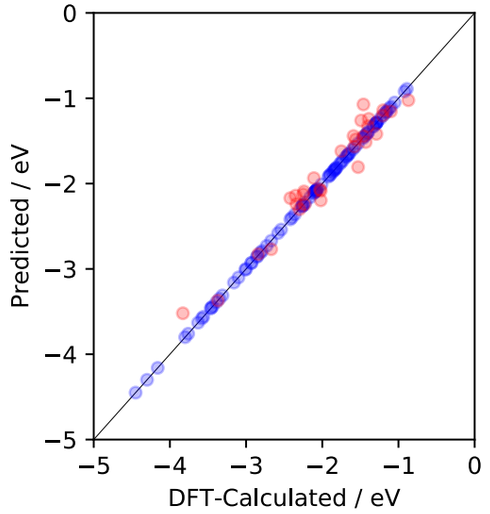
Metal	G	R/Å	AN	AM/ g mol^{-1}	P	EN	IE/ eV	$\Delta_{fus}H$/ J g^{-1}	ρ/ g cm^{-3}
Fe	8	2.66	26	55.85	4	1.83	7.90	247.3	7.87
Co	9	2.62	27	58.93	4	1.88	7.88	272.5	8.86
Ni	10	2.60	28	58.69	4	1.91	7.64	290.3	8.90
Cu	11	2.67	29	63.55	4	1.90	7.73	203.5	8.96
Ru	8	2.79	44	101.07	5	2.20	7.36	381.8	12.10
Rh	9	2.81	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%)

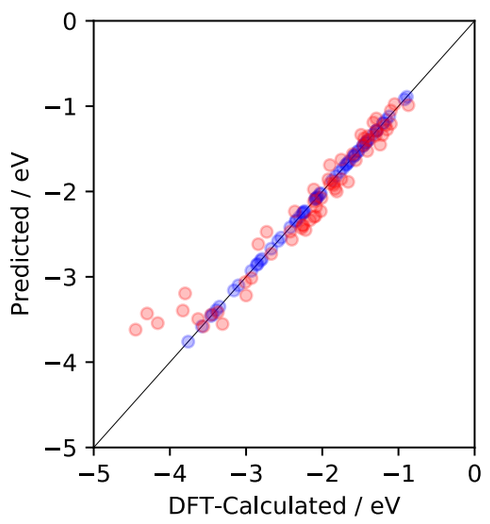


100 times mean RMSE: **0.153** / eV

	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%)

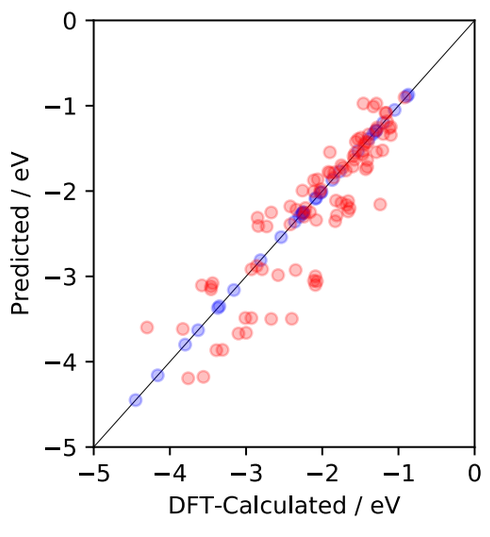


100 times mean RMSE: **0.235** / eV

	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%)

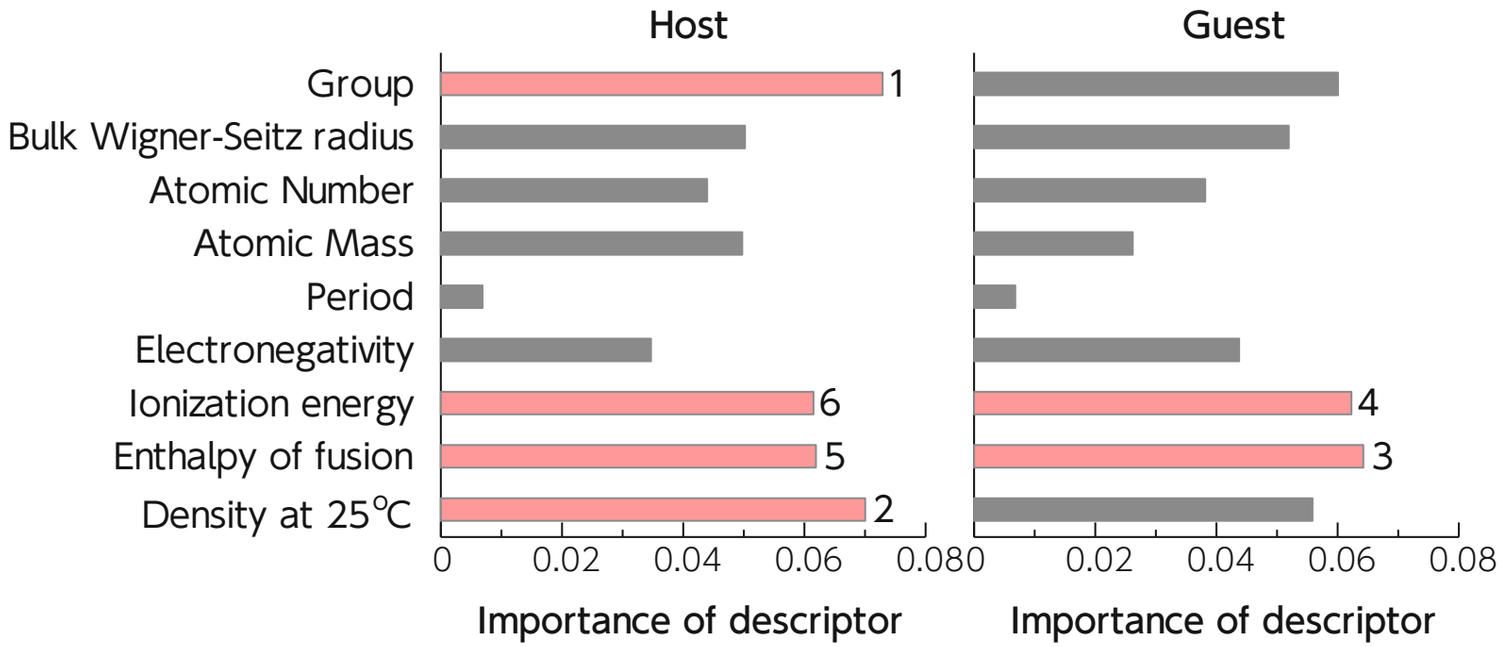


100 times mean RMSE: **0.402** / eV

Descriptor analysis and selection

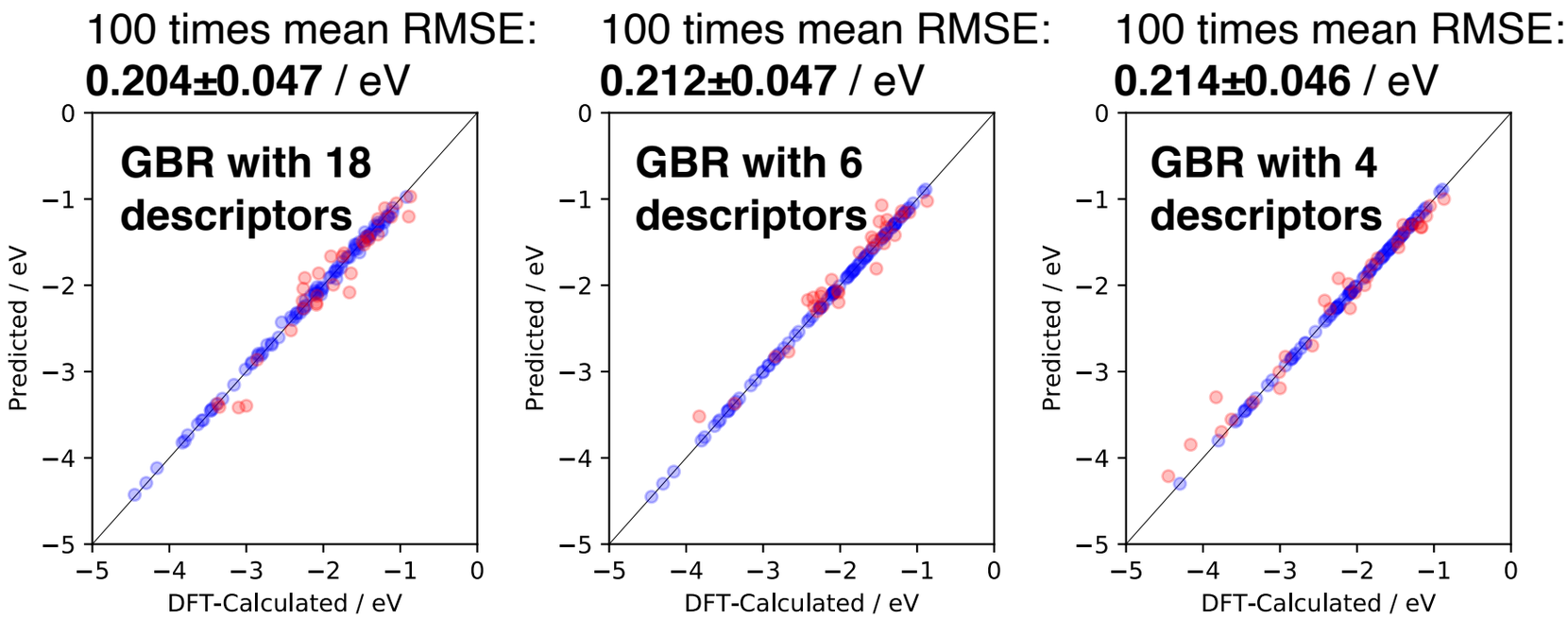
Analyzing the prediction model trained with the current data also provides some insights on contributing factors, and variable selection.

Descriptor Importances



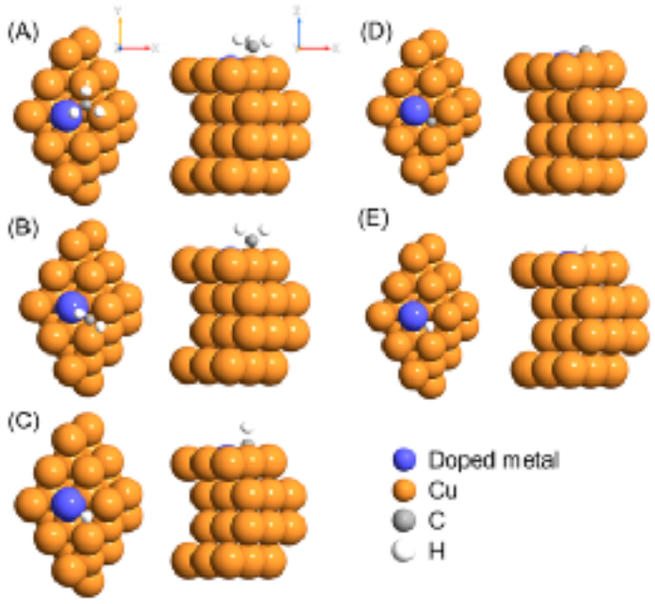
Descriptor Selection (top-k)

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

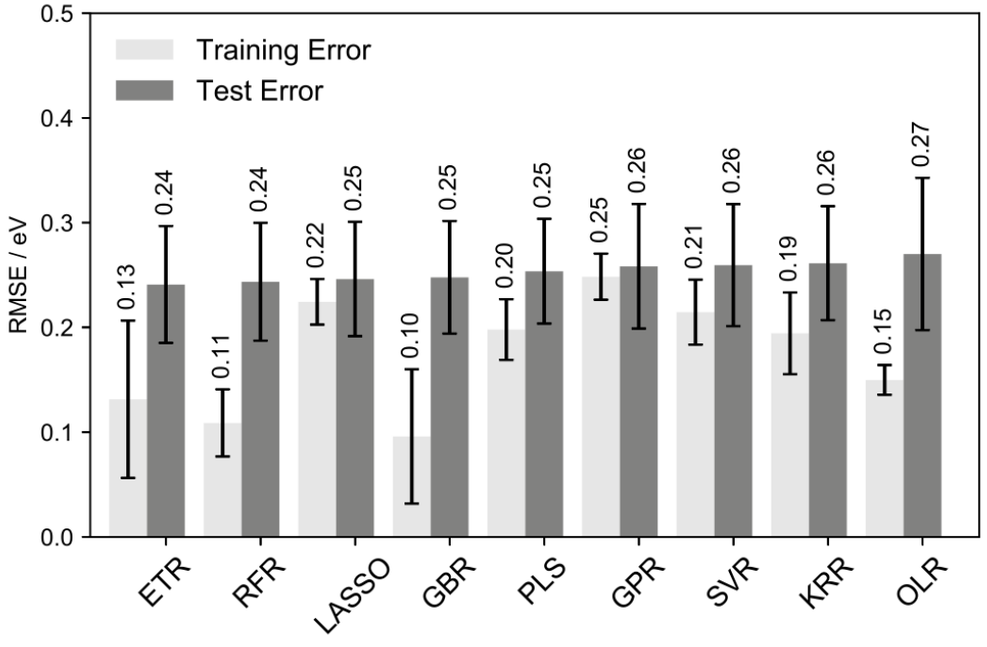


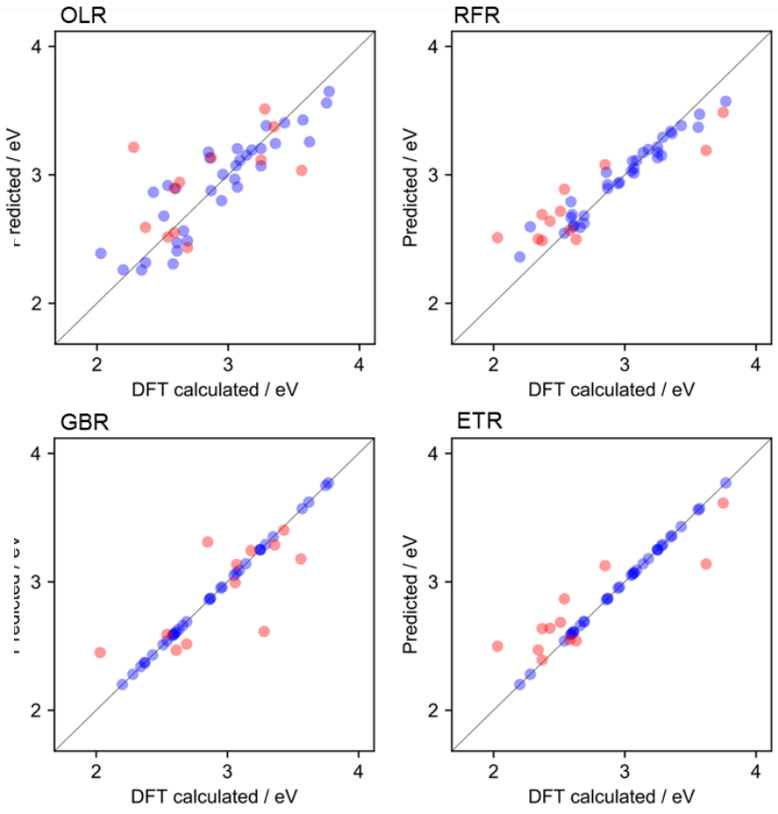
Case 2. Predicting the adsorption energy

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]

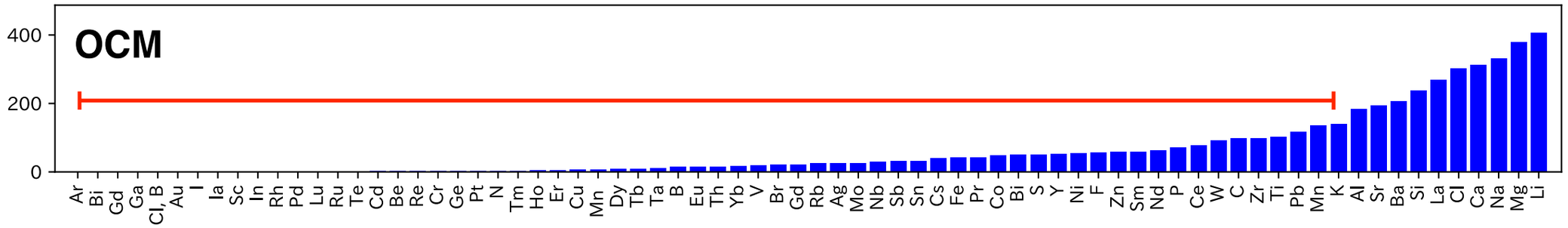
Collections from various papers published in the past including

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

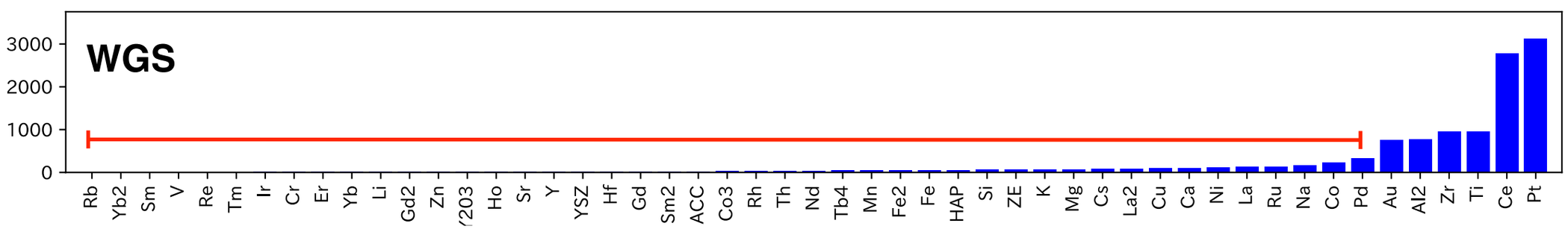
Two big problems we had

- **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**...

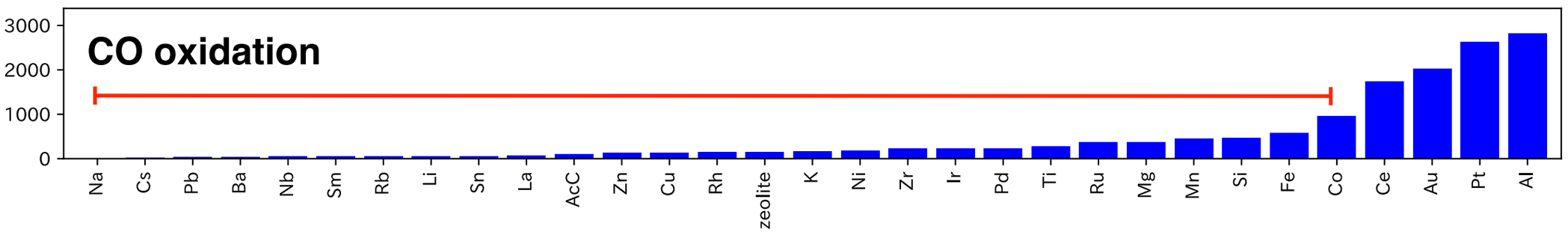
 # Catal: 1866
 # Elem: 74



 # Catal: 4360
 # Elem: 51



 # Catal: 5610
 # Elem: 31



Two big problems we had

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

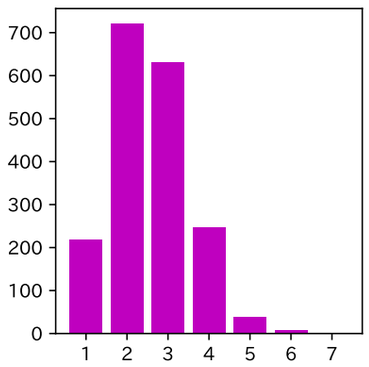
	A	B	C	D	E	
Cat-ABC =	(0.90,	0.06,	0.04,	0.00,	0.00)	catalyst with 90% A, 6% B, and 4% C
Cat-BCD =	(0.00,	0.30,	0.10,	0.60,	0.00)	catalyst with 60% D, 30% B, and 10% C
Cat-BCE =	(0.00,	0.30,	0.10,	0.00,	0.60)	catalyst with 60% E, 30% B, and 10% C

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

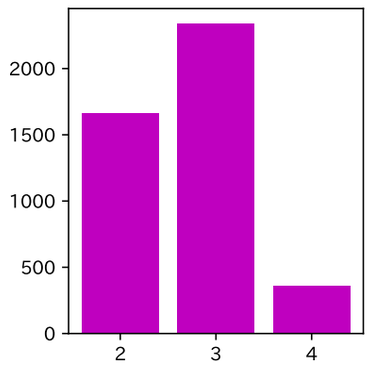
- For large datasets, this composition vectors are very sparse and mostly the overlapped elements are only one or two (or even zero...)

 # elems in a catalyst

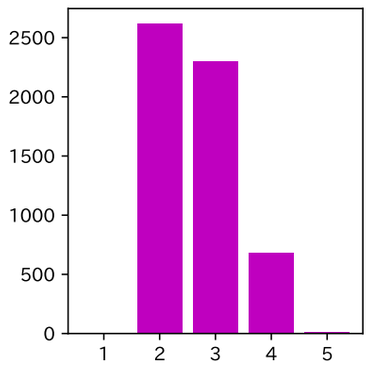
OCM



WGS

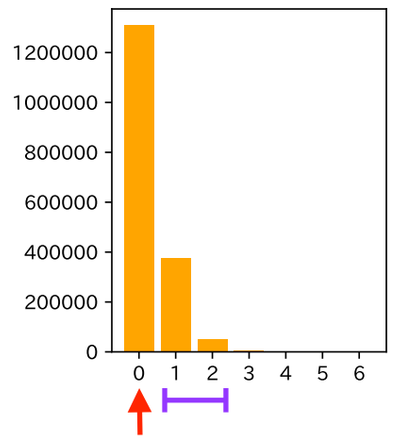


CO oxidation

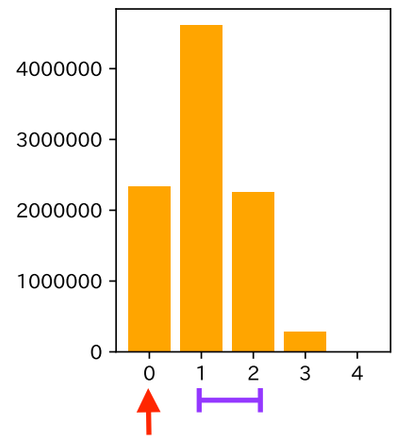


 # overlapped elems (for a pair)

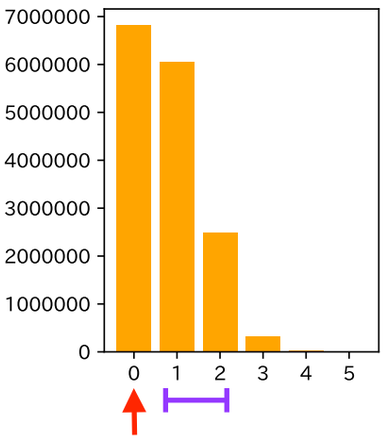
OCM



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**

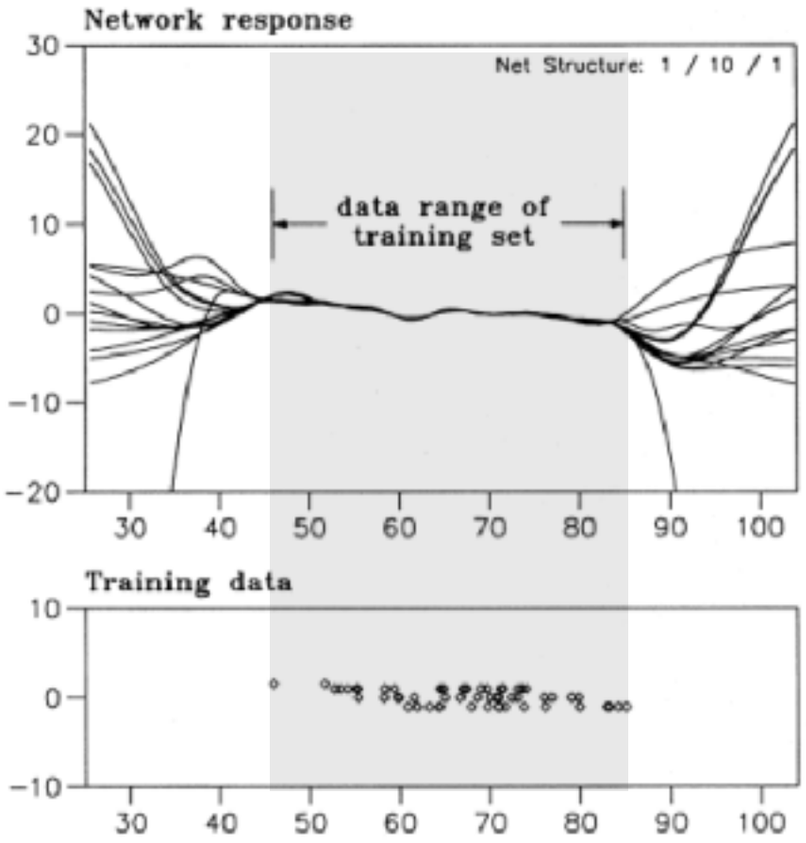
  Catalyst research has relied heavily on prior published data, tends to be biased toward catalyst composition that were successful

  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

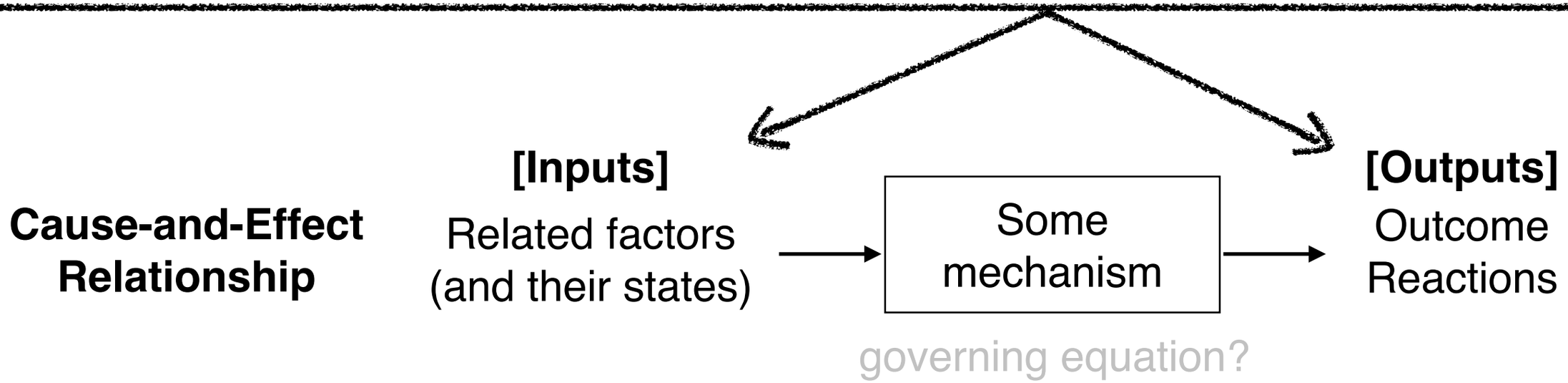
"exploitation"

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)

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

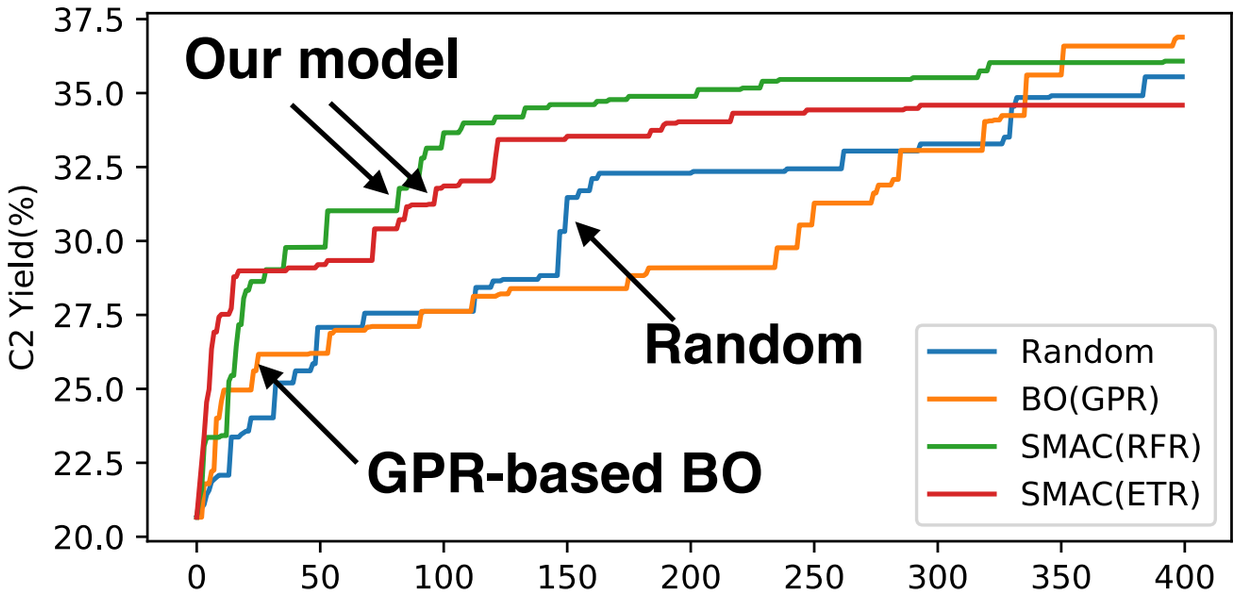
- Optimal design of experiments (DoE)
- Active learning
- Bayesian optimization
- Blackbox optimization
- Reinforcement learning
- Multi-armed bandit
- Evolutional computation
- Game-theoretic approaches

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**

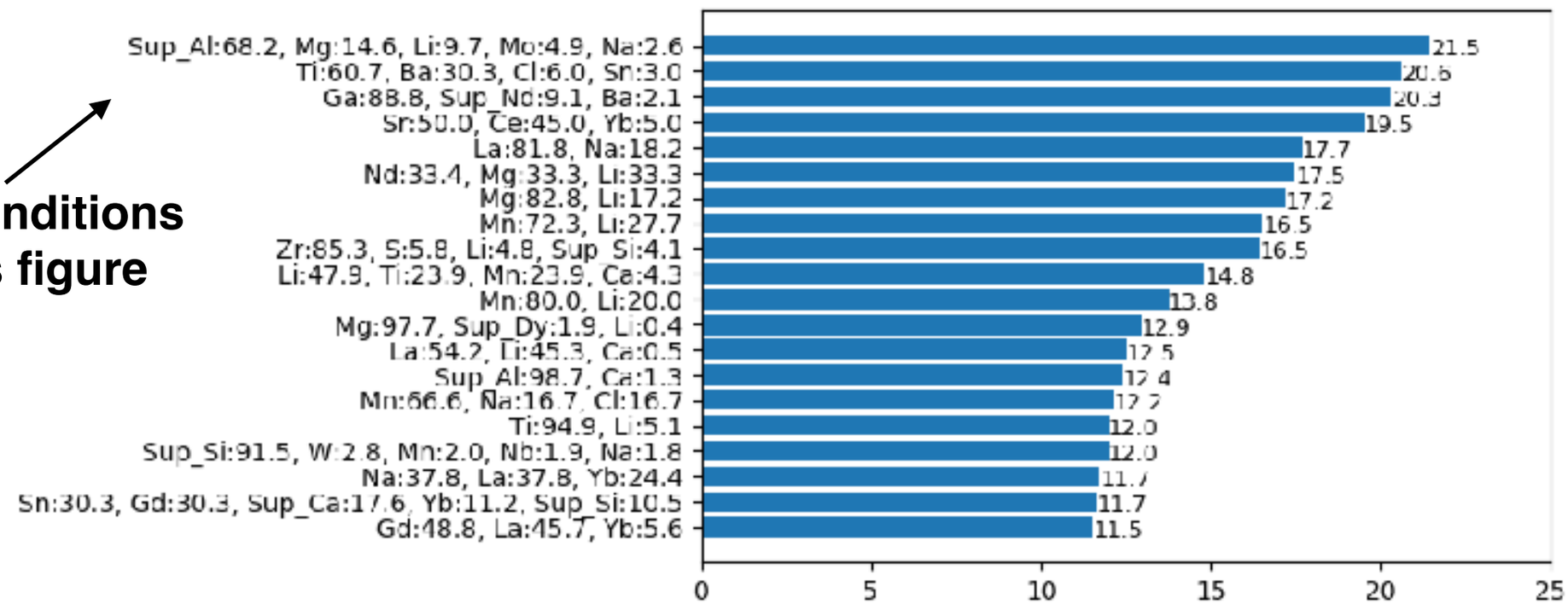


Test on 3 Datasets

- 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



Rationalize & Accelerate Chemical Design and Discovery

Key: Effective use of data with a help with data-driven techniques

Facts from experiments and calculations

In-House data + Public data + Knowledge base
(and their quality control & annotations)

Hypothesis generation
(Machine learning, Data mining)

- Planning what to test in the next experiments or simulations
- Surrogates to expensive or time-consuming experiments or simulations
- Optimize uncertain factors or conditions
- Multilevel information fusion

Validation
(Experiments and Simulations)

- Highly Reproducible experiments with high accuracies and speeds
- Acceleration with ML-based surrogates for time-consuming subproblems
- Simulating many 'what-if' situations

This trend emerged first in life sciences (drug discovery)

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

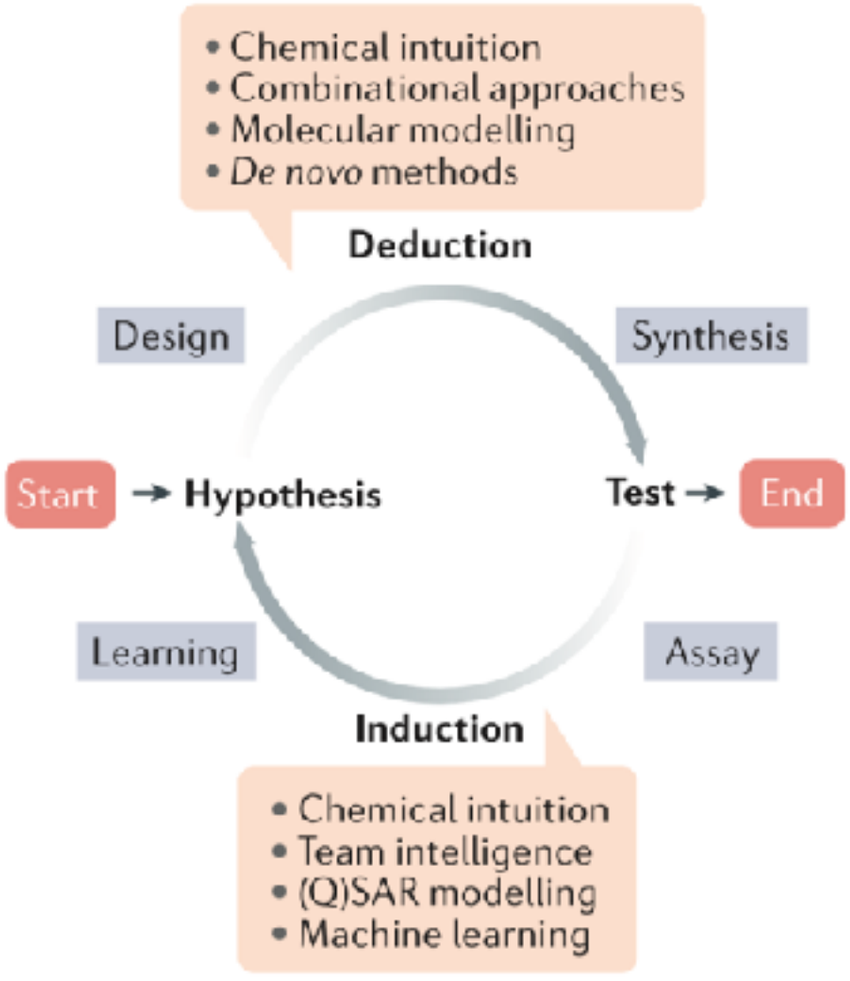
PERSPECTIVES

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)

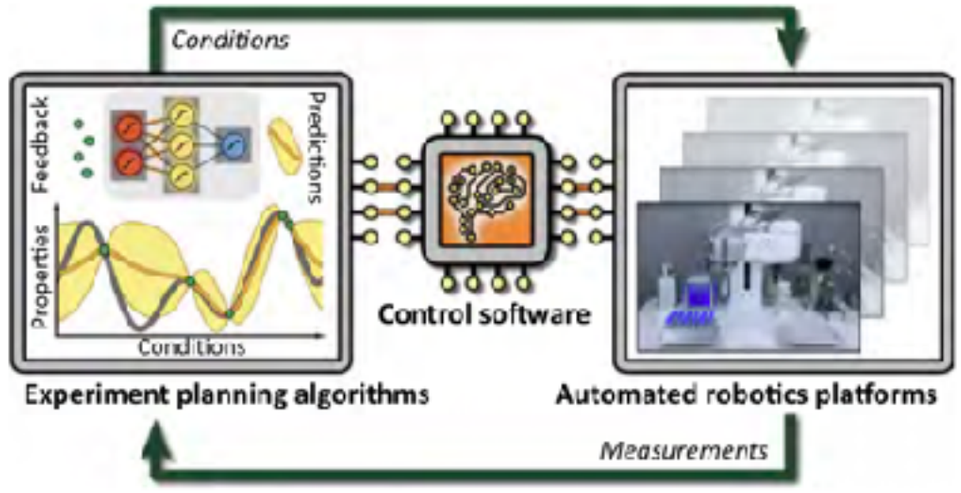
... also to chemistry!

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

Downloa

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

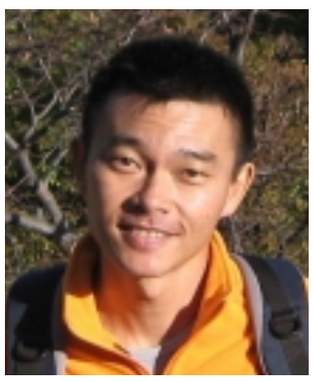
creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition 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

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)