The interplay between data-driven and theory-driven methods for chemical sciences

Ichigaku Takigawa

ichigaku.takigawa@riken.jp

- Medical-risk Avoidance based on iPS Cells Team, **RIKEN Center for Advanced Intelligence Project (AIP)**
- Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), **Hokkaido University**

Research Interests: Ichigaku Takigawa

- Machine learning and data mining technologies
- Data-intensive approaches to natural sciences

10 years (1995-2004)
Hokkaido University (Computer Sciences)
- *Grad School. Engineering*

7 years (2005-2011)
Kyoto University (Bioinformatics, Chemoinformatics)
- *Bioinformatics Center, Inst. Chemical Research*
- *Grad School. Pharmaceutical Sciences*

7 years (2012-2018)
Hokkaido University (Computer Sciences)
- *Grad School. Information Science and Technology*
- *JST PRESTO on Materials Informatics (2015-2018)*

? years (2019-)
RIKEN@**Kyoto** (Machine Learning + Stem Cell Biology)
- *Center for AI Project*

Hokkaido University (Machine Learning + Chemistry)
- *Inst. Chemical Reaction Design and Discovery*

Research Interests: Ichigaku Takigawa

- Machine learning and data mining technologies
- Data-intensive approaches to natural sciences

10 years
(1995-2004)

Hokkaido University (Computer Sciences)
- *Grad School. Engineering*

7 years
(2005-2011)

Kyoto University (Bioinform
- *Bioinformatics Center, In*
- *Grad School. Pharmaceu*

7 years
(2012-2018)

Hokkaido University (Comp
- *Grad School. Information*
- *JST PRESTO on Materia*

? years
(2019-)

RIKEN@**Kyoto** (Machine L
- *Center for AI Project*

Hokkaido University (Mach
- *Inst. Chemical Reaction Design and Discovery*

From AAAI-20 Oxford-Style Debate

Research Interests: Ichigaku Takigawa

- Machine learning and data mining technologies
- Data-intensive approaches to natural sciences

10 years (1995-2004)
Hokkaido University (Computer Sciences)
- *Grad School. Engineering*

7 years (2005-2011)
Kyoto University (Bioinformatics, Chemoinformatics)
- *Bioinformatics Center, Inst. Chemical Research*
- *Grad School. Pharmaceutical Sciences*

7 years (2012-2018)
Hokkaido University (Computer Sciences)
- *Grad School. Information Science and Technology*
- *JST PRESTO on Materials Informatics (2015-2018)*

? years (2019-)
RIKEN@**Kyoto** (Machine Learning + Stem Cell Biology)
- *Center for AI Project*

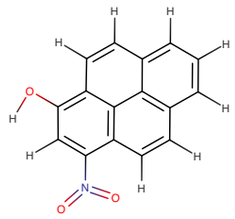
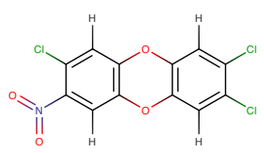
Hokkaido University (Machine Learning + Chemistry)
- *Inst. Chemical Reaction Design and Discovery*

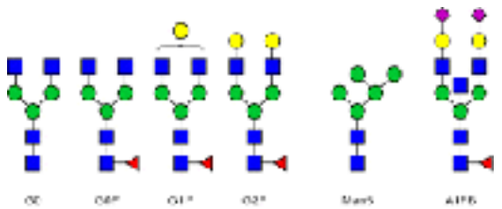
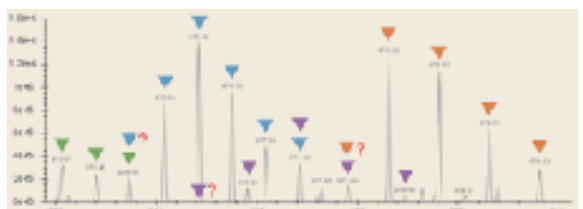
Machine Learning with "Discrete Structures"

Q. How we can make use of "discrete structures" for prediction...?

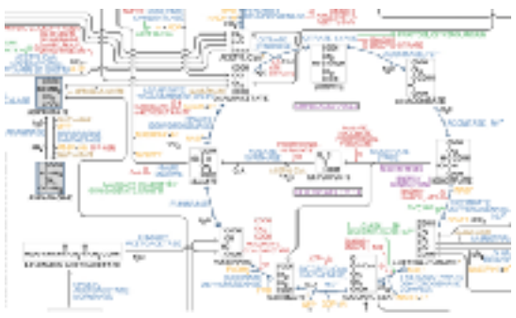
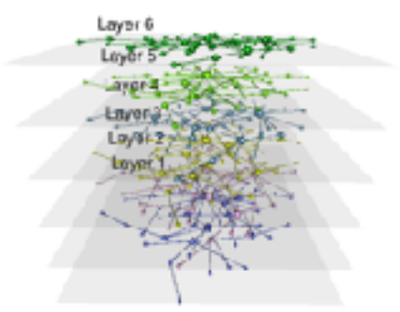
sets, sequences, branchings or hiearchies (trees), networks (graphs), relations, logic, rules, combinations, permutations,, point clouds, algebra, langages, ...

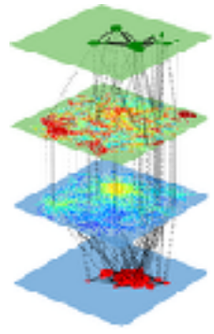
- The target variables can come with "discrete structures"

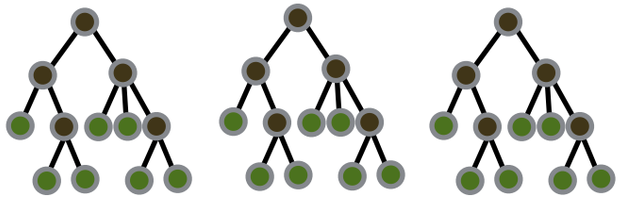
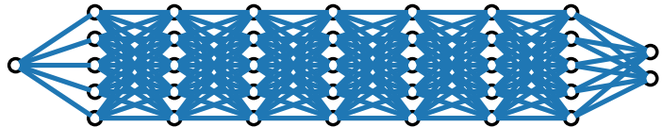
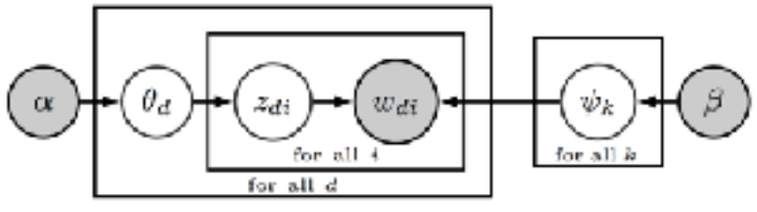



- The relationship between variables can have "discrete structures"

- The ML models themselves can involve "discrete structures"

Past work: Data-intensive approaches to life science

Pathway Commons

Access and discover data integrated from public pathway and interactions databases.

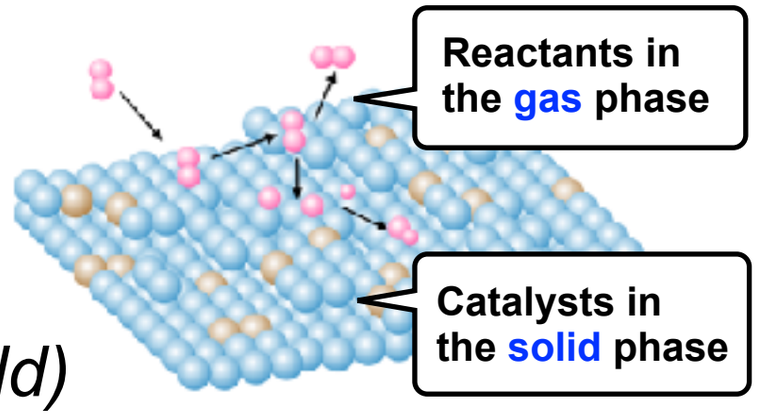
reactome

- **Transcriptional regulation of metabolism**
  Bioinformatics 2007, 2008a, 2008b, 2009, 2010
  Nucleic Acids Res 2011, PLoS One 2012, 2013, KDD'07
- **Transcription regulation by mediator complex**
  Nat Commun 2015, Nat Commun 2020
- **Repetitive sequences in genomes**
  Discrete Appl Math 2013, 2016, AAAI 2020
- **Polypharmacology of drug-target interaction networks**
  PLoS One 2011, Drug Discov Today 2013, Brief Bioinform 2014, BMC Bioinformatics 2020
- **Copy number variations in neurological disorder (MSA, MS)**
  Mol Brain 2017
- **Substrate analysis of modulator protease (Calpain family)**
  Mol Cell Proteom 2016, Genome Informatics 2009, http://calpain.org
- **Cell competition in cancer**
  Cell Reports 2018, Sci Rep 2015
- **Phenotype patterns of somatic mutations in breast cancer**
  Brief Bioinform 2014

Recent work: Data-intensive approaches to chemistry

Machine learning for heterogeneous catalysis

catalysis where the phase of the **catalyst** differs from the phase of the reactants or products.



- *ACS Catalysis 2020 (review)*
- *ChemCatChem 2019 (front cover)*
- *J Phys Chem C 2018a, 2018b, 2019 (cover)*
- *RSC Advances 2016 (highlighted in Chemical World)*

Haber–Bosch Process

(industrial synthesis of ammonia)

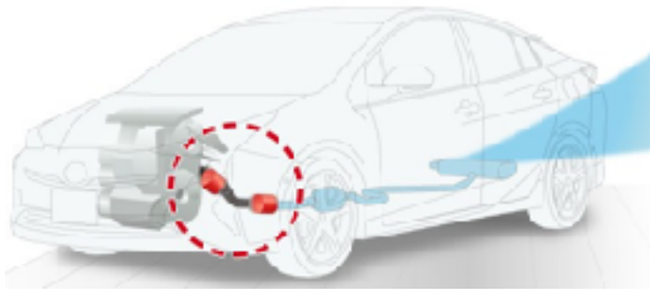
“Fertilizer from Air”

artificial nitrogen fixation

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

Ferrous Metal Catalysis

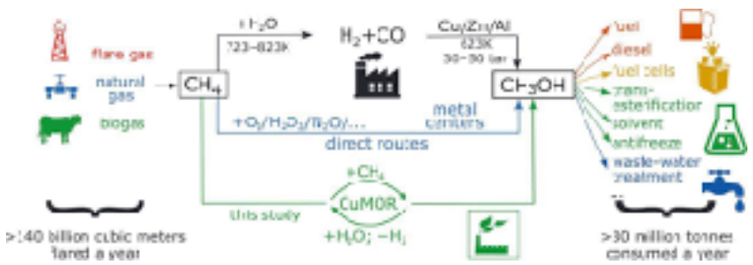
Exhaust Gas Purification



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)

Effective use of data is another key in natural sciences

(alongside experiments and simulations)

And please keep in mind that **unplanned data collection is too risky**.
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

Today's AI has stark limitations, facing big problems

- Deep learning techniques thus far have proven to be data hungry, shallow, brittle, and limited in their ability to generalize (Marcus, 2018)
- Current machine learning techniques are data-hungry and brittle—they can only make sense of patterns they've seen before. (Chollet, 2020)
- A growing body of evidence shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets... instead of learning meaning in the flexible and generalizable way that humans do. (Nie et al., 2019)
- Current machine learning methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice. (Bengio et al., 2019)

Though it's extremely powerful and useful in suitable applications!

Forbes | Technology Council

Feb 19, 2020, 08:00am

In Praise Of Boring AI (A.K.A. Machine Learning)

JC Schutterle Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program

The AI/ML community is now struggling with these...

- **The Winograd Schema Challenge** (Levesque et al., 2011)

  A collection of 273 multiple-choice problems that requires *commonsense reasoning* to solve as an alternative to Turing test (Turing, 1950).

  ↓ Many methods passed this test of "Imitation Games" were actually tricks being fairly adept at fooling humans...

  Unexpectedly, it turned out that several tricks using BERT/Transformer can crack many of these (with >90% accuracy) ...

  - **WinoGrande** (Sakaguchi et al., 2020)

    An improved collection of 44k very hard problems
    (AAAI-20 Outstanding Paper Award)
    Though some top-ranking methods can have 70-80% accuracy ...

- **Abstraction and Reasoning Challenge at Kaggle (**Chollet, 2020)

  At the moment, to what extent we can create an AI capable of solving reasoning tasks it has never seen before?

  "Kaggle's most important AI competition" "Impossible AI Challenge"

Examples

WSC-273

(Q0) The town councillors refused to give the angry demonstrators a permit because **they** feared violence. **Who** feared violence?

- the town councillors
- the angry demonstrators

(due to Terry Winograd)

(Q1) The trophy would not fit in the brown suitcase because **it** was so small. **What** was so small?

- the trophy
- the brown suitcase

WinoGrande-1.1

(Q) He never comes to my home, but I always go to his house because the _____ is smaller.

1. home
2. house

Take home message

The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

Take home message

The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

- Model-based learning, Neuro-symbolic or ML-GOFAI integration.
  We can partly use *explicit models* for well-understood parts for sample efficiency and for filling the gap between correlation and causation.

Take home message

The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

- Model-based learning, Neuro-symbolic or ML-GOFAI integration.
  We can partly use *explicit models* for well-understood parts for sample efficiency and for filling the gap between correlation and causation.
- Needs for modeling unverbalizable comon sense of domain experts
  We need a good strategy for building a *gigantic* data collection for this, as well as *self-supervised learning* and/or *meta learning* algorithms.

  *"Self-supervised is training a model to **fill in the blanks**. This is what is going to allow our AI systems to go to the next level. Some kind of common sense will emerge." (Yann LeCun)*

Take home message

The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

- Model-based learning, Neuro-symbolic or ML-GOFAI integration.
  We can partly use *explicit models* for well-understood parts for sample efficiency and for filling the gap between correlation and causation.
- Needs for modeling unverbalizable comon sense of domain experts
  We need a good strategy for building a *gigantic* data collection for this, as well as *self-supervised learning* and/or *meta learning* algorithms.

  *"Self-supervised is training a model to **fill in the blanks**. This is what is going to allow our AI systems to go to the next level. Some kind of common sense will emerge." (Yann LeCun)*
- Needs for novel techniques for compositionality (combinatorial generalization), out-of-distribution prediction (extrapolation), and their flexible transfer
  We need to somehow combine and flexibly transfer partial knowledge to generate a new thing or deal with completely new situations.

Take home message

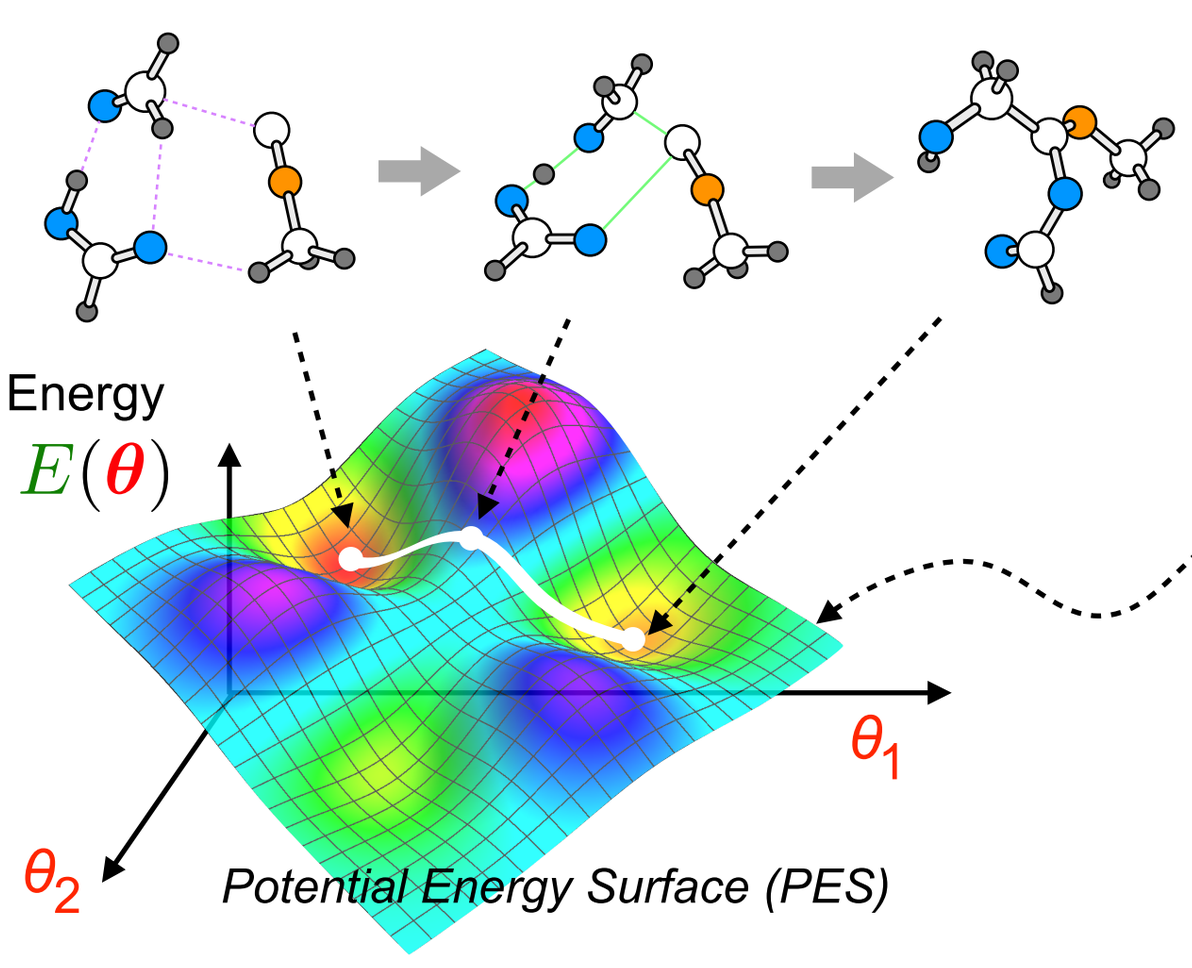
The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

- Model-based learning, Neuro-symbolic or ML-GOFAI integration.
  We can partly use *explicit models* for well-understood parts for sample efficiency and for filling the gap between correlation and causation.
- Needs for modeling unverbalizable comon sense of domain experts
  We need a good strategy for building a *gigantic* data collection for this, as well as *self-supervised learning* and/or *meta learning* algorithms.

  *"Self-supervised is training a model to **fill in the blanks**. This is what is going to allow our AI systems to go to the next level. Some kind of common sense will emerge." (Yann LeCun)*
- Needs for novel techniques for compositionality (combinatorial generalization), out-of-distribution prediction (extrapolation), and their flexible transfer
  We need to somehow combine and flexibly transfer partial knowledge to generate a new thing or deal with completely new situations.

Chemistry has the first principle at the electron level

Our end goal:
take full control over chemical reactions,
the basis for transforming substances to another (energy, materials, foods, colors, ...)



- **Outcomes we see are just discrete and combinatorial**

Every substance is **a combination of only 118 elements** in the periodic table.

- **Recombinations of atoms and chemical bonds are subjected to *the laws of nature***

They are trantitions from a valley to a hill to another valley, but **solving a quantum chemical equation is needed at every point θ** to get the energy surface $E(\theta)$

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

Schrödinger equation

Our core technology for automated reaction-path search

PERSPECTIVE

View Article Online
View Journal | View Issue

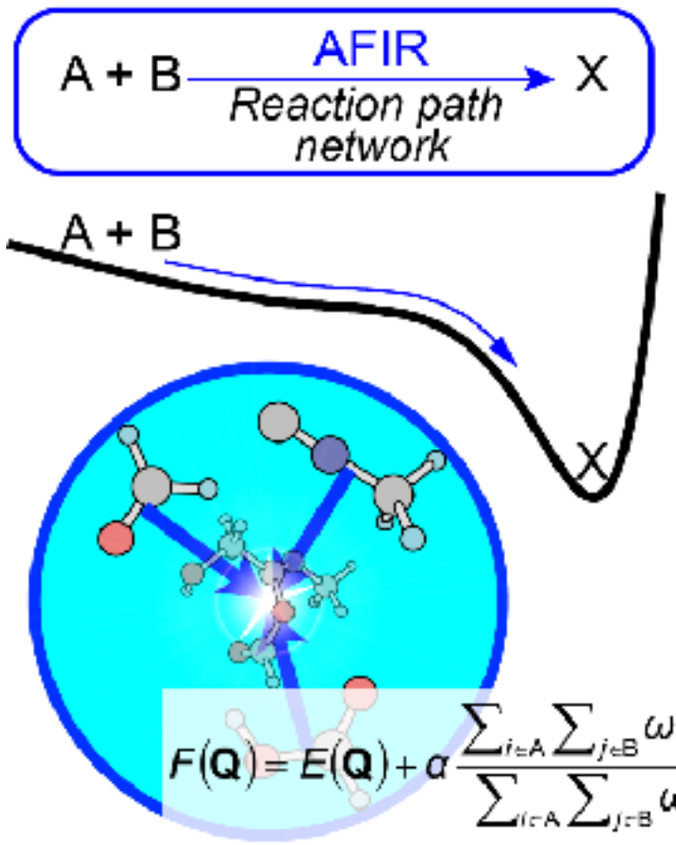
Cite this: *Phys. Chem. Chem. Phys.*, 2013, **15**, 3683

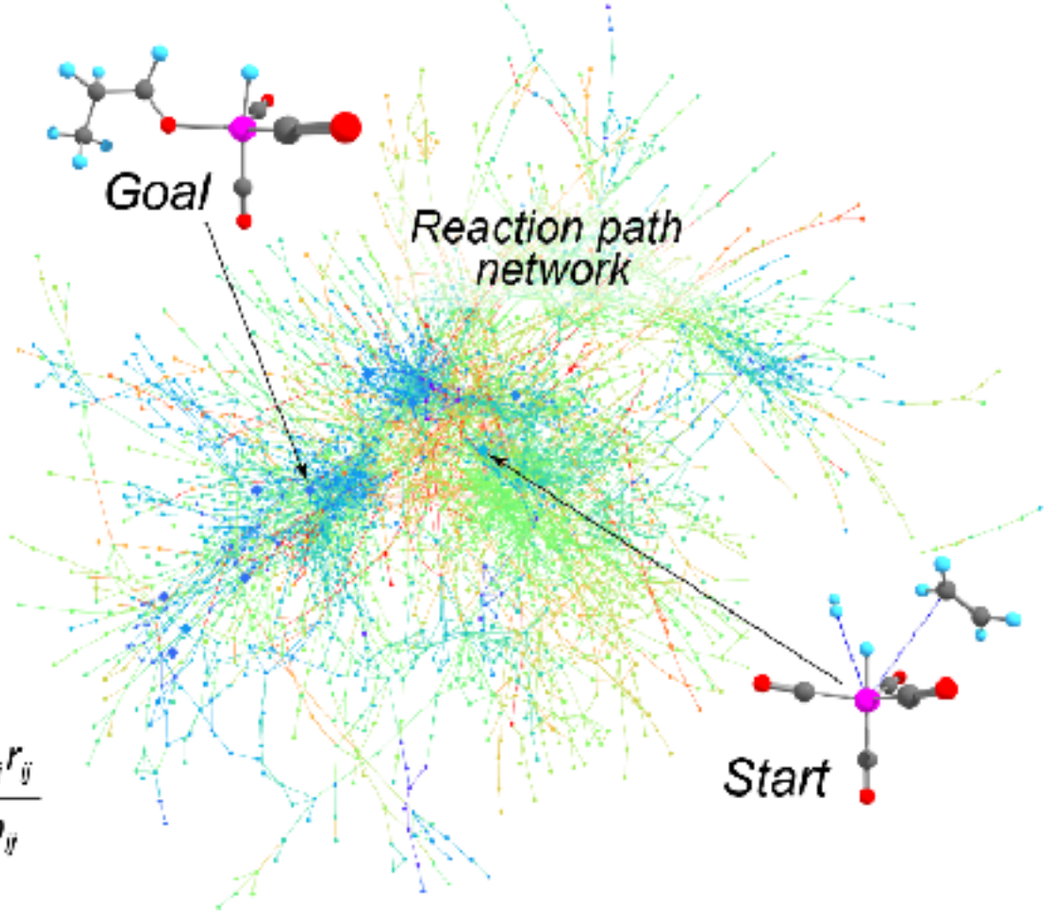
Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods

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

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

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





But chemistry still remains quite empirical, and why?

Even though current chemical calculations are fairly accurate, chemists would still heavily rely on "Edisonian empiricism (trial & error)."

1. empirical knowledge (from the literature) and their flexible transfer
2. intuitions and experiences (unverbalizable *common sense* of experts)

But chemistry still remains quite empirical, and why?

Even though current chemical calculations are fairly accurate, chemists would still heavily rely on "Edisonian empiricism (trial & error)."

1. empirical knowledge (from the literature) and their flexible transfer
2. intuitions and experiences (unverbalizable *common sense* of experts)

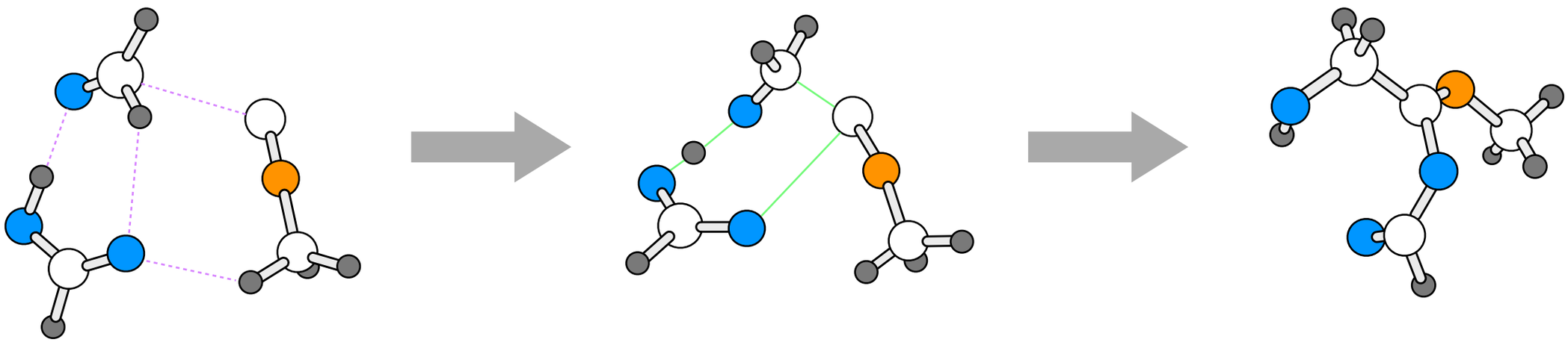
Takeaway: We also need 'data-driven' bridges!
first principles are not enough for us to throw away these *empirical* things; **data-driven approaches (ML)** play a complementary role!

Fusing multi-disciplinary methods...?

1. Theory-driven: (slow but) high-fedelity quantum-chemical simulations
2. Knowledge-driven: explicit knowns and inference
3. Data-driven: *empirical* prediction grounded upon multifaceted data

Indeed computational chemistry has many limitations...

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

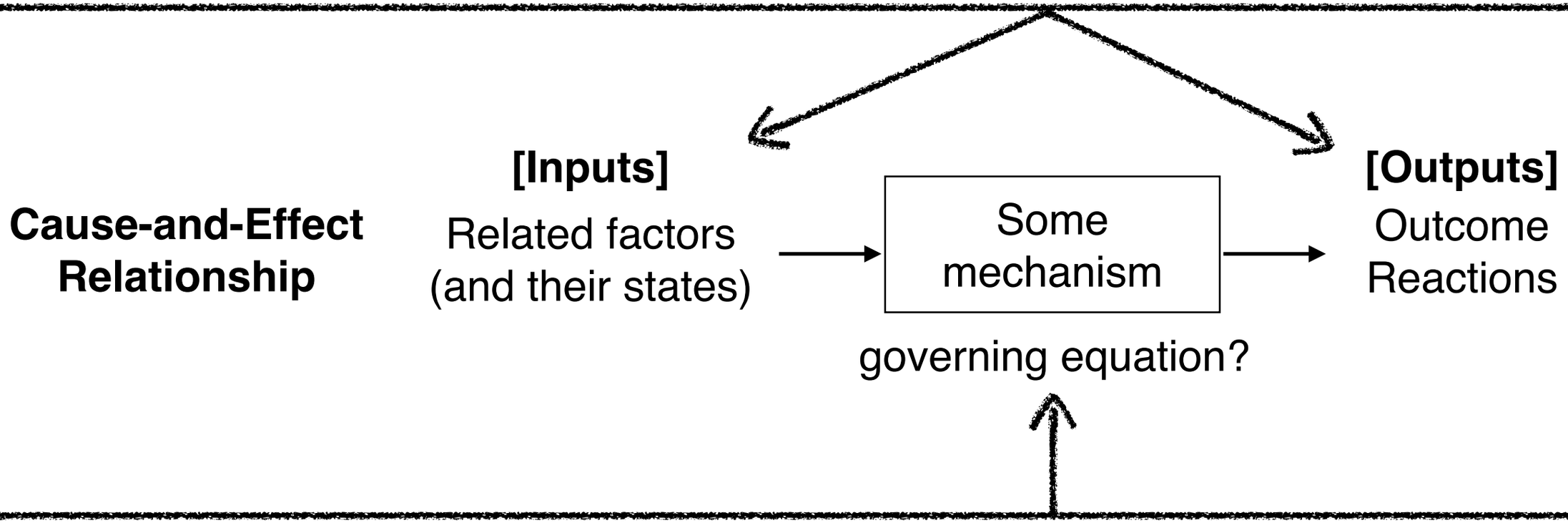


- **Intractably large chemical space**: A intractably large number of "theoretically possible" candidates for reactions and compounds...
- **Computing time and scalability issue**: Simulating an Avogadro-constant number of atoms is utterly infeasible... (we need some compromise here)
- **Complexity and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

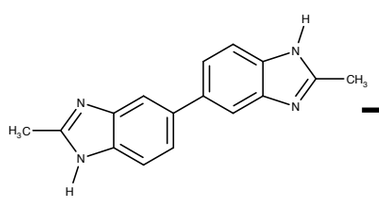


Theory-driven methods try to explicitly model the inner workings of a target phenomenon (e.g. through first-principles simulations)

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

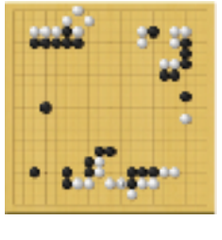
Speech Recognition

“ありがとう”

J’aime la musique

Machine Translation

I love music

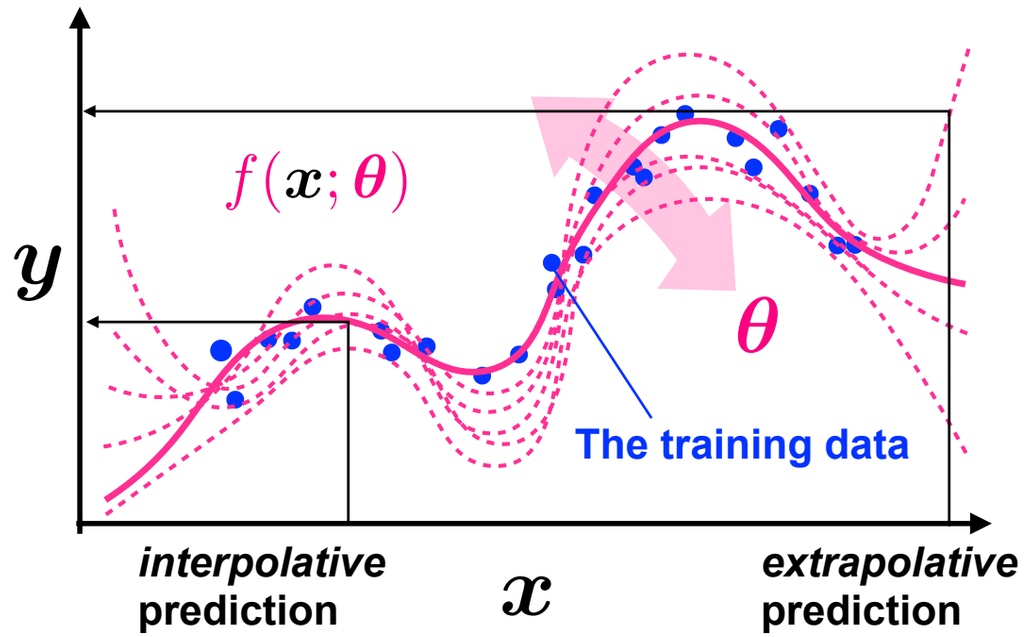


AI Game Players

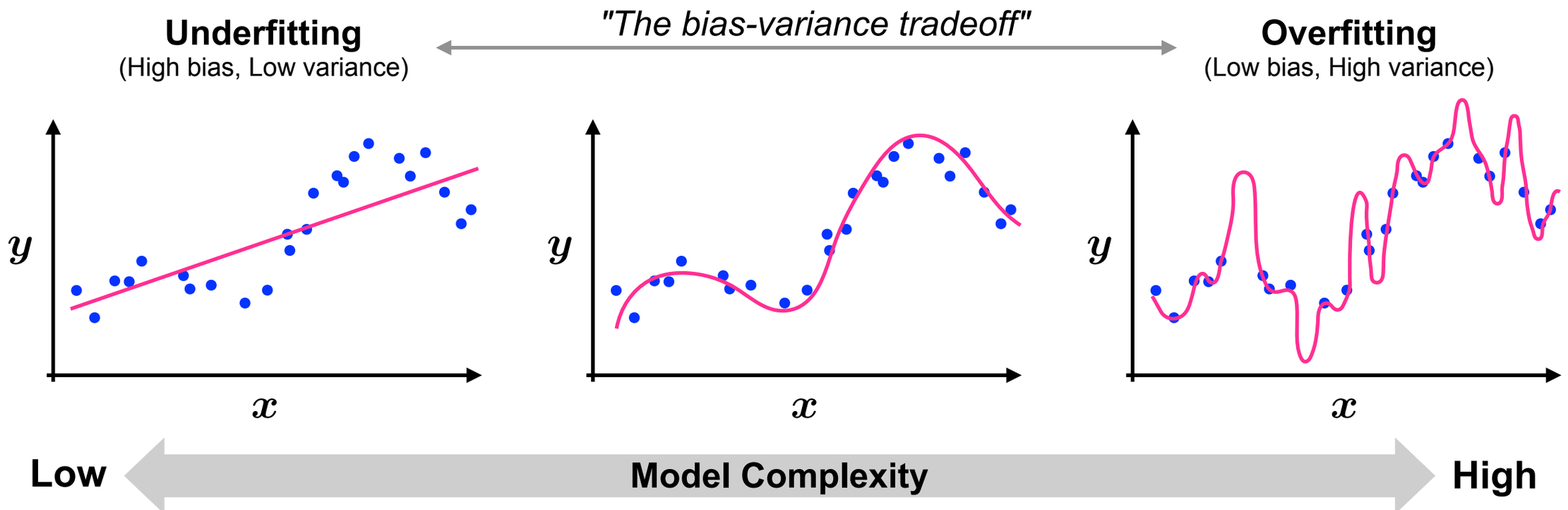
How ML works: fitting a function to data

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

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



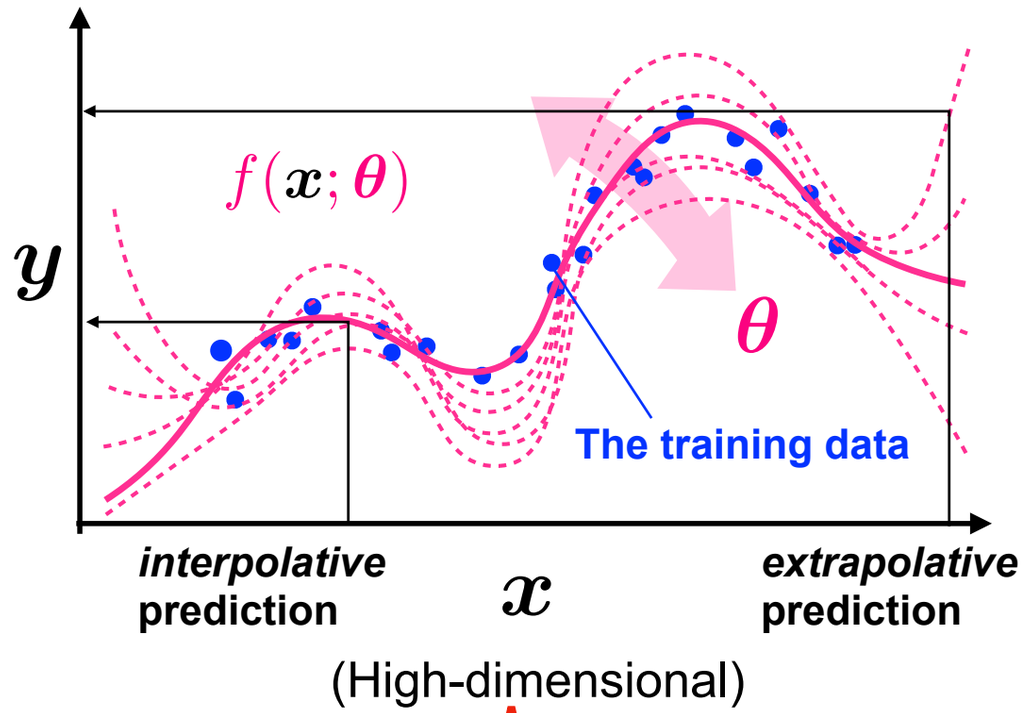
(High-dimensional)

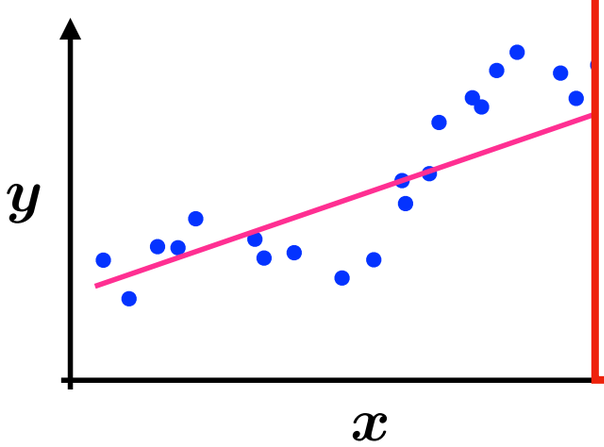


How ML works: fitting a function to data

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

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





x x x

This can be often **ultra high-dimensional** in modern ML

e.g.) Deep learning usually comes with *overparametrization* and the number of parameters can be several tens of millions or even billions... (This is why model-complexity control *or regularization* is so critical and essential in ML)

Low ← **Model Complexity** → **High**

Multilevel representations of chemical reactions

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

SMILES

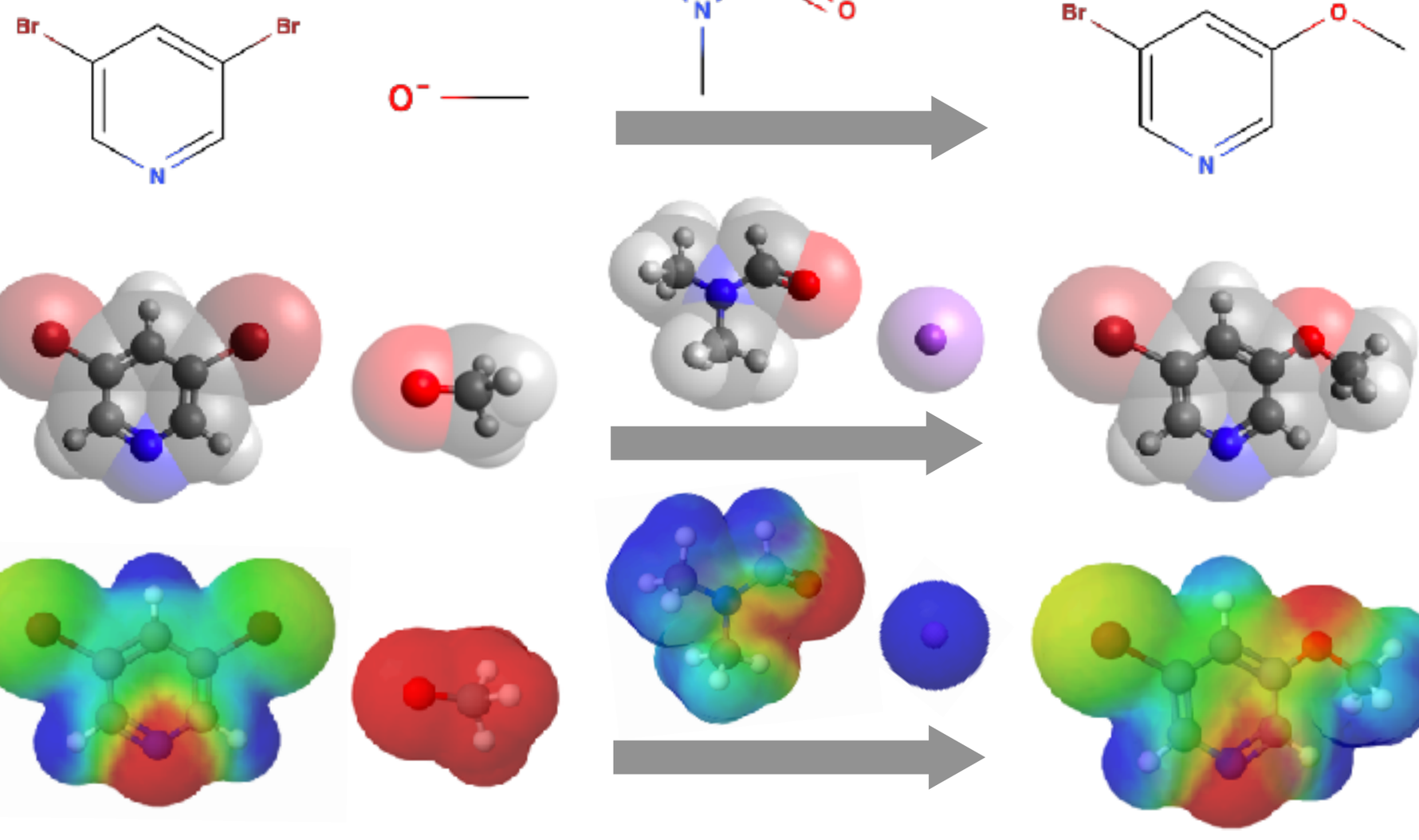
Structural Formla

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)

Computer-assisted synthetic planning

A traditional chemoinformtics topic since 70s:
collect known chemical reactions, and search for reaction paths

e.g.) Representation, classification, exploration of chemical reactions

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

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

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

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,

Knowledge-based approaches

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**

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

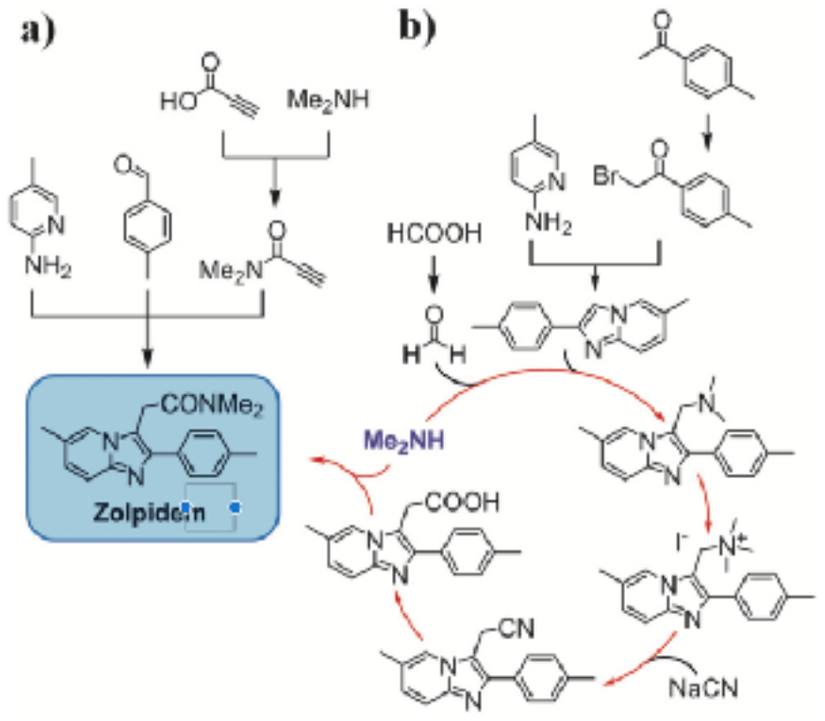


Figure 5. Two different optimal syntheses of zolpidem for relatively low (right, $C_{res} = 0.075$) and high (left, $C_{res} = 7.5$) cost of labor.

Widely-used proprietary systems
(template-based path search and expansions)

Reaxys® SciFinder® A CAS SOLUTION CHEMATICA

Now towards fusing with Machine Learning...?

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

Purely ML-based approaches and more

ML-based chemical reaction predictions

Graph Neural Networks

WLDN
Jin+ *NeurIPS 2017*

ELECTRO
Bradshaw+*ICLR 2019*

GPTN
Do+ *KDD 2019*

WLN
Coley+ *Chem Sci 2019*

GLN
Dai+ *NeurIPS2019*

Sequence Neural Networks

seq2seq
Liu+ *ACS Cent Sci 2017*

IBM RXN
Schwaller+ *Chem Sci 2018*

Transformer
Karpov+ *ICANN 2019*

Molecular Transformer
Schwaller+ *ACS Cent Sci 2019*

Combined or Other

Neural-Symbolic ML
Segler+ *Chemistry 2017*

Similarity-based
Coley+ *ACS Cent Sci 2017*

3N-MCTS/AlphaChem
Segler+ *Nature 2018*

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

Similar technologies go for biology!

706 | Nature | Vol 577 | 30 January 2020

DeepMind's AlphaFold paper

Article

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7

Received: 2 April 2019

Accepted: 10 December 2019

Published online: 15 January 2020

Andrew W. Senior[1,4*], Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]

688 Cell *180*, 688–702, February 20, 2020

A successful example from MIT MLPDS

Cell

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes,[1,2,3] Kevin Yang,[3,4,10] Kyle Swanson,[3,4,10] Wengong Jin,[3,4] Andres Cubillos-Ruiz,[1,2,5] Nina M. Donghia,[1,5] Craig R. MacNair,[6] Shawn French,[6] Lindsey A. Carfrae,[6] Zohar Bloom-Ackerman,[2,7] Victoria M. Tran,[2] Anush Chiappino-Pepe,[5,7] Ahmed H. Badran,[2] Ian W. Andrews,[1,2,5] Emma J. Chory,[1,2] George M. Church,[5,7,8] Eric D. Brown,[6] Tommi S. Jaakkola,[3,4] Regina Barzilay,[3,4,9,*] and James J. Collins[1,2,5,8,9,11,*]

MIT MLPDS

Machine Learning for Pharmaceutical Discovery and Synthesis Consortium

How to integrate data-driven and theory-driven?

- Theory-driven: first-principles simulations, logical reasoning, mathematical models, etc.
- Data-driven: machine learning

Still emerging topics,
but examples of already available approaches are

1. ***Wrapper***: Use ML to control or plan simulations
   Basically we solve the problem by simulations, but use ML models to ask "what if?" questions to guide what simulations to run.
   (Model-based optimization, Sequential design of experiments, Reinforcement learning, Generative methods, etc)

2. ***Hybrid***: Use ML as approximator for unsure parts of simulations
   Plugging ML models into the unsure part of simulations or calling simulations when ML predictions are less confident
   (Data assimilation, domain adaptation, semi-empirical methods, etc)

Back to our example: Heterogeneous Catalysis

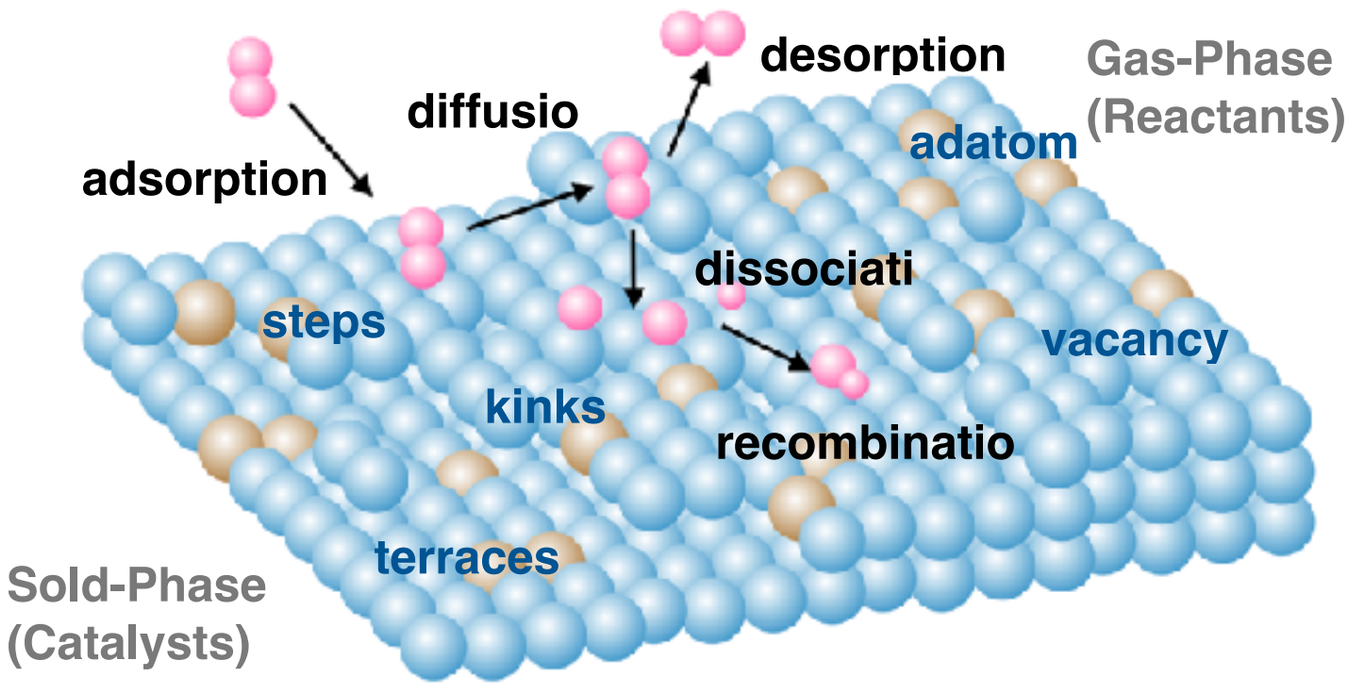
Notoriously complex surface reactions between different phases.

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)

One of big problems we had

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

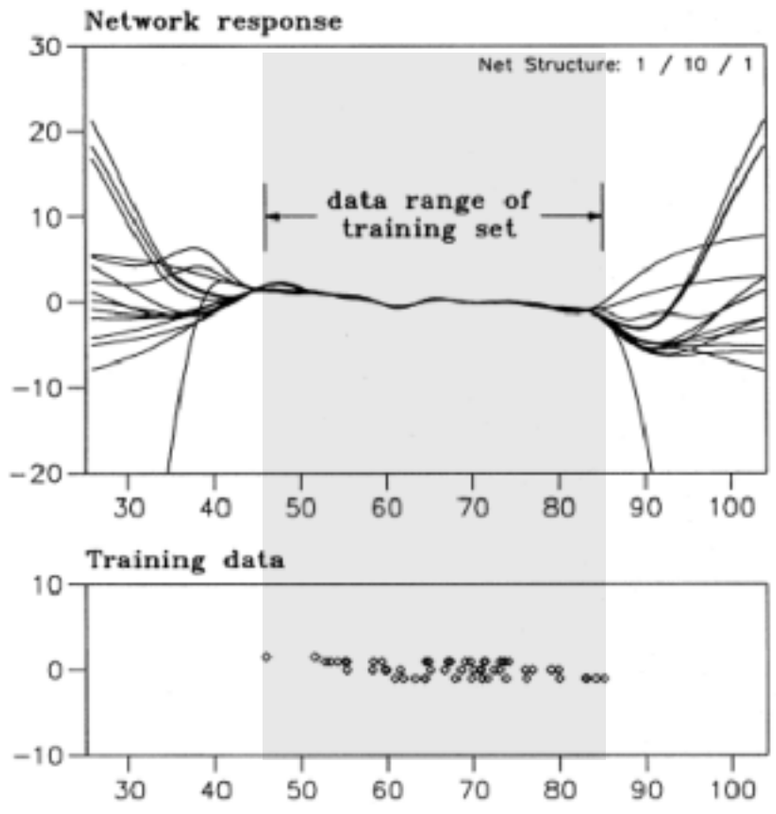
  Catalyst research has **relied heavily on prior published data**, strongly biased toward catalyst composition that were successful

  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
  - 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)



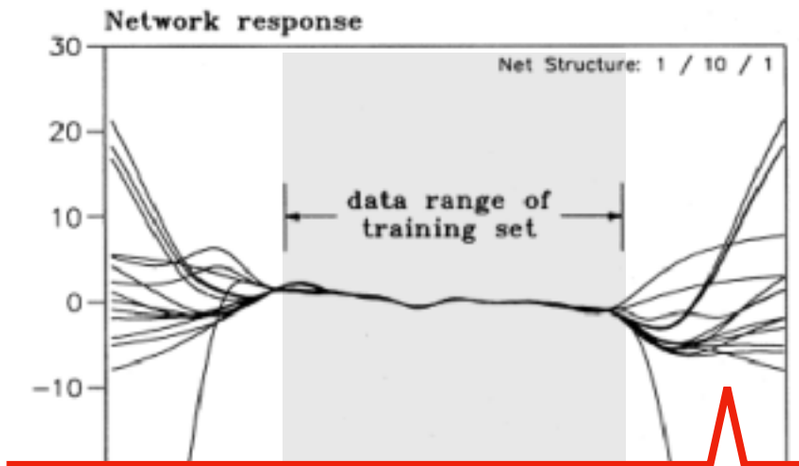
"Beware of the perils of extrapolation, and understand that ML algorithms build models that are representative of the available training samples."

Simply focusing on targets predicted as high-performance ones by ML built on currently avialable data is clearly not a good idea...

Like too much scrutinizing any so-so candidates at the moment that we stumbled upon at the very early stage of research...

An ML model is just representative of the training data

Highly Inaccurate Model Predictions from Extrapolation (Lohninger 1999)



In reality, distinction between interpolation and extrapolation are not that clear due to the **high-dimensionality**.

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

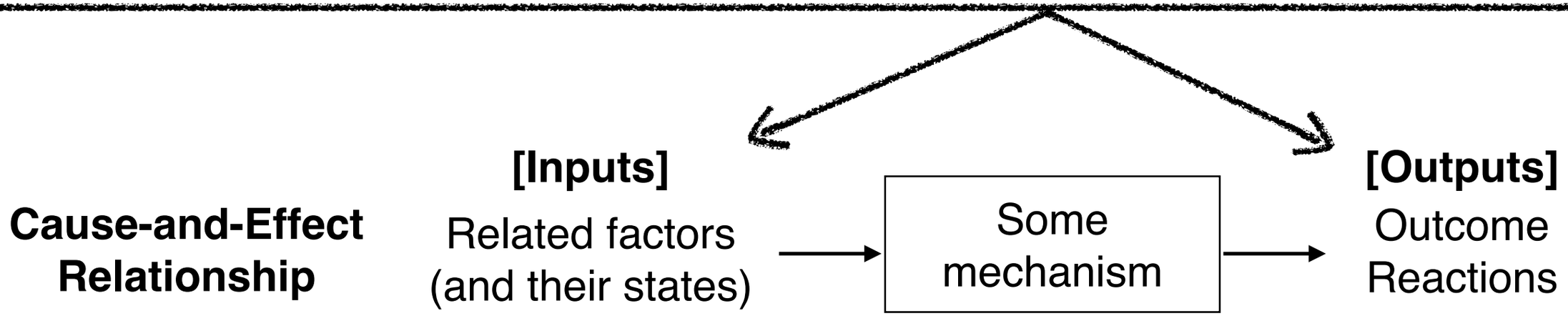
Simply focusing on targets predicted as high-performance ones by ML built on currently avialable data is clearly not a good idea...

Like too much scrutinizing any so-so candidates at the moment that we stumbled upon at the very early stage of research...

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)



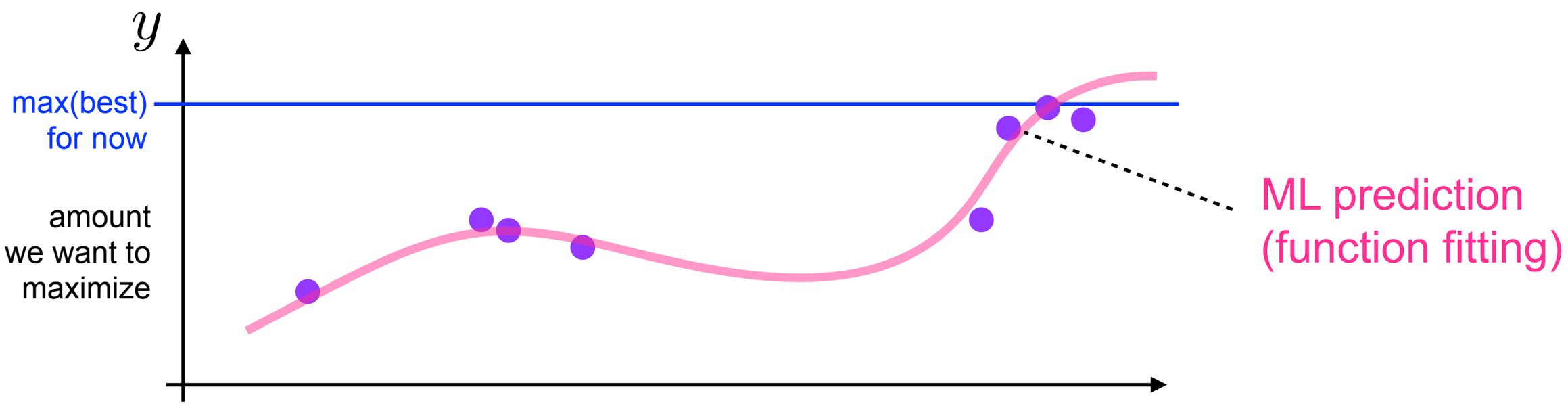
"Current machine learning techniques are data-hungry and brittle —they can only make sense of patterns they've seen before."
(Chollet, 2020)

Our Solution: Model-based optimization

Either through experiments or simulations,
we face a dilemma in choosing between options to learn new things.

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

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

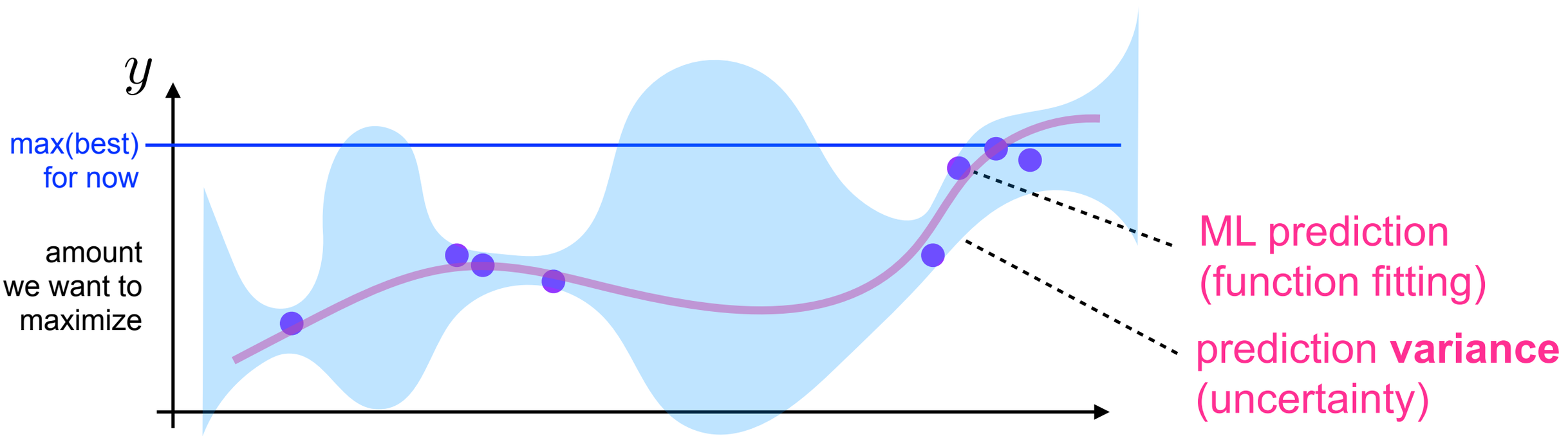


Our Solution: Model-based optimization

Either through experiments or simulations,
we face a dilemma in choosing between options to learn new things.

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

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

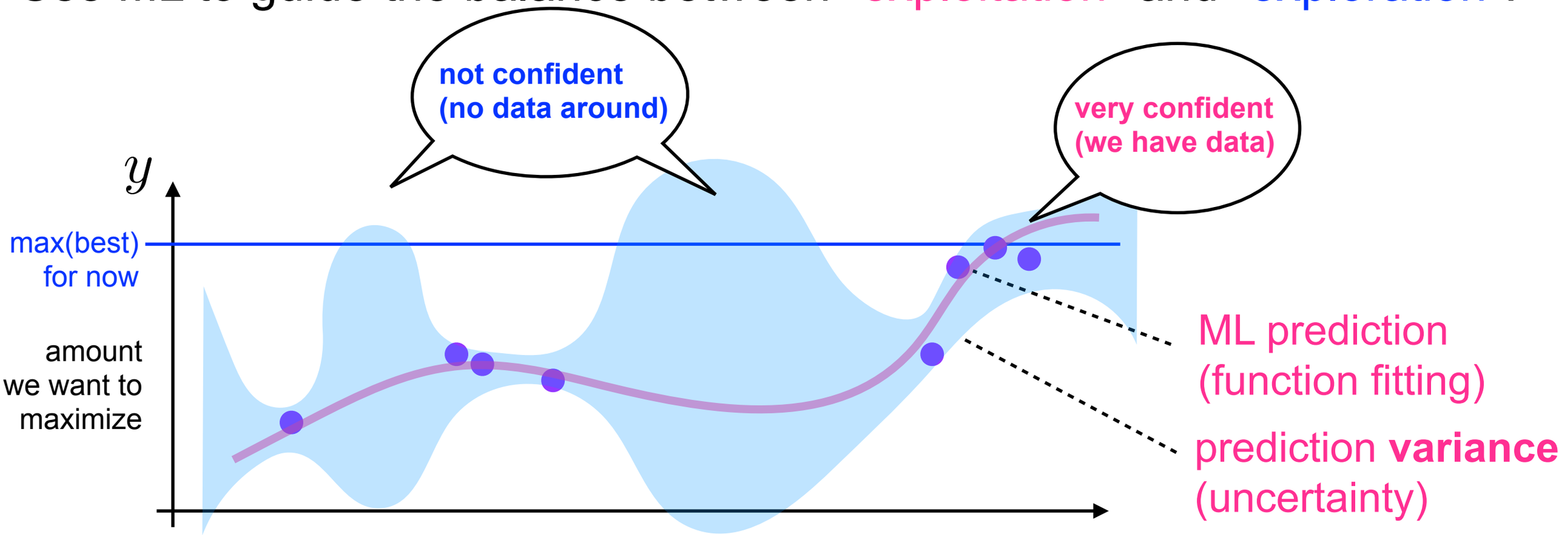


Our Solution: Model-based optimization

Either through experiments or simulations,
we face a dilemma in choosing between options to learn new things.

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

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

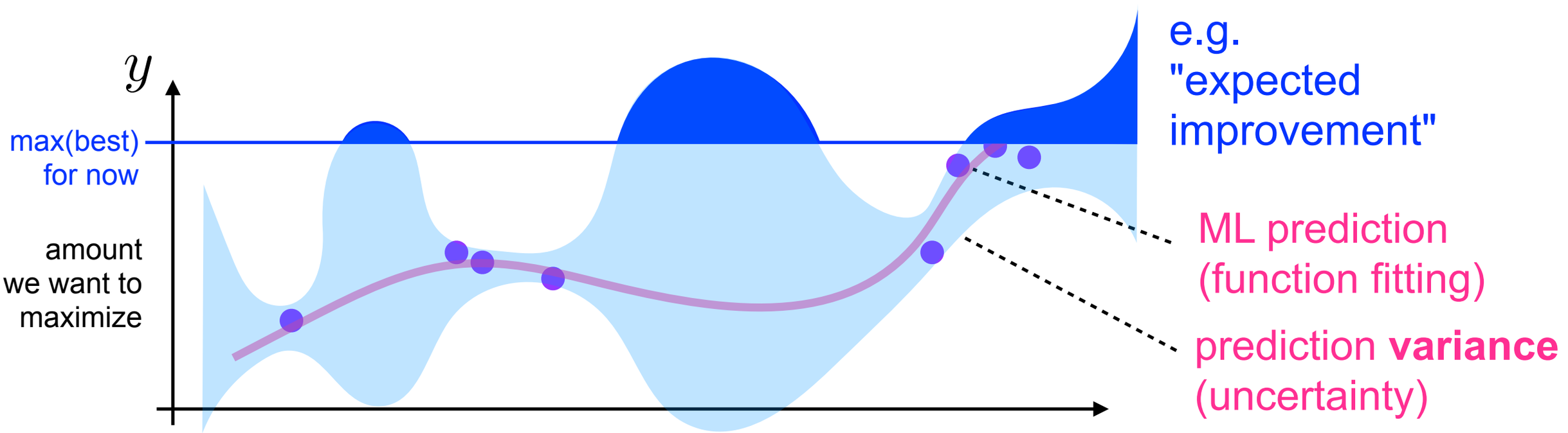


Our Solution: Model-based optimization

Either through experiments or simulations,
we face a dilemma in choosing between options to learn new things.

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

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

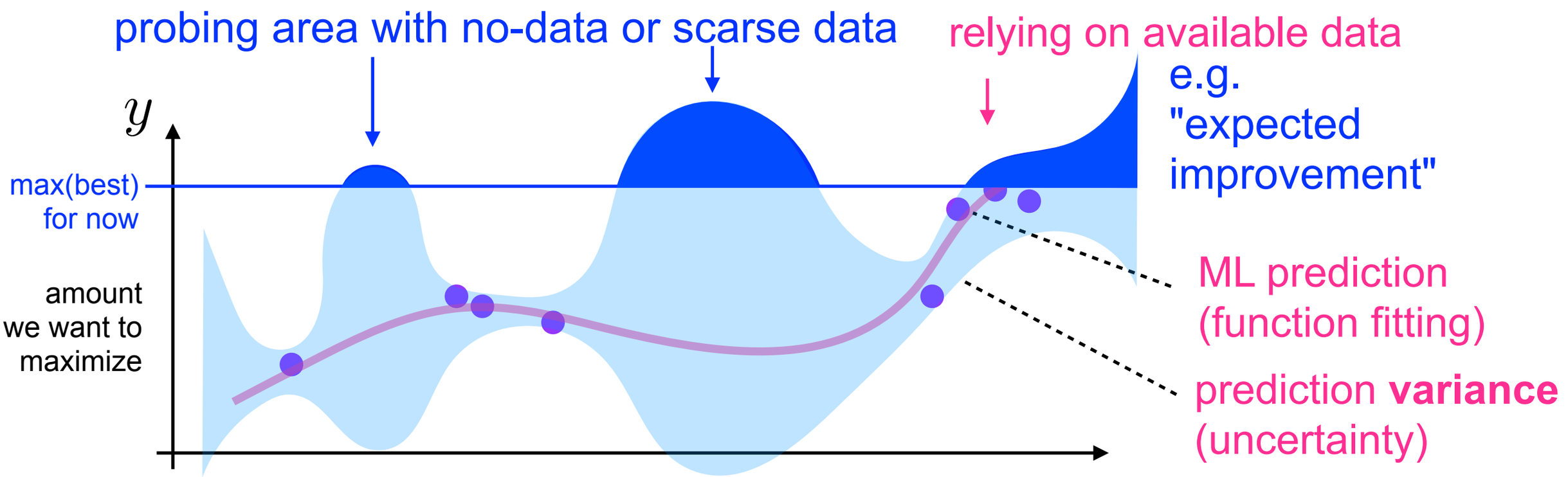


Our Solution: Model-based optimization

Either through experiments or simulations,
we face a dilemma in choosing between options to learn new things.

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

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



Key: model-based and exploitation-exploration balance

Model-based reinforcement learning

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 (Use ML for tuning ML)

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

Cloud AutoML

AutoML

OPTUNA

AutoDL 2019

Amazon SageMaker

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/or 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

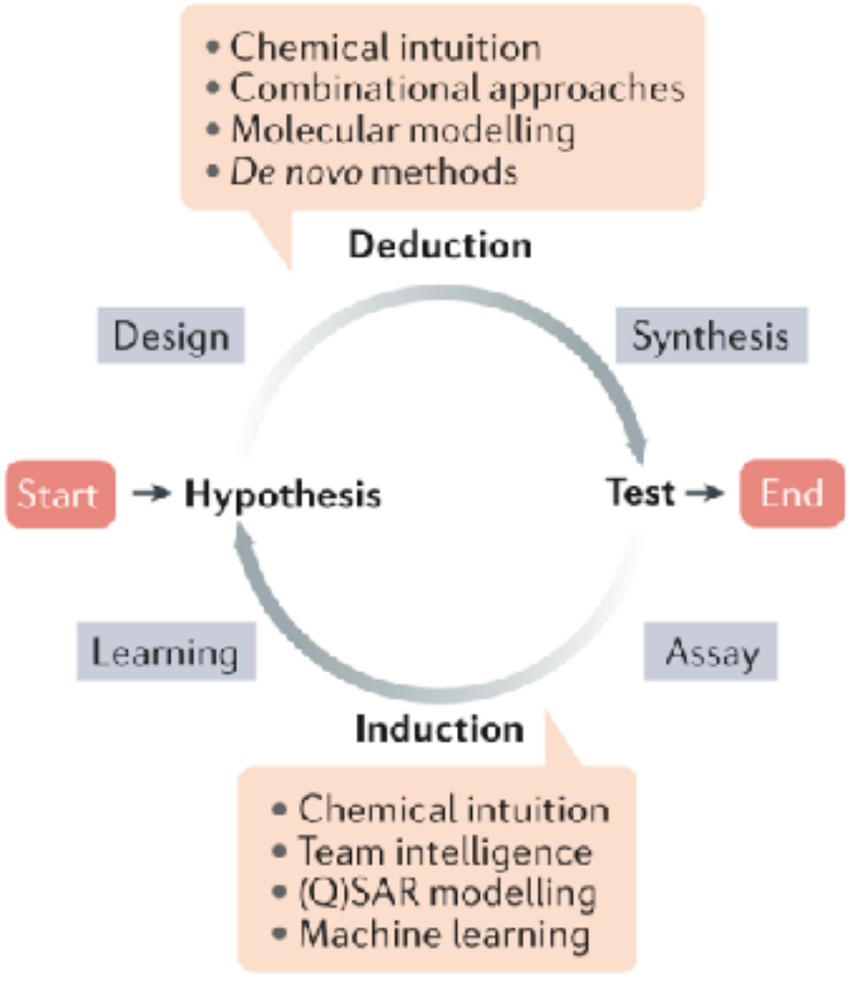
PERSPECTIVES

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)

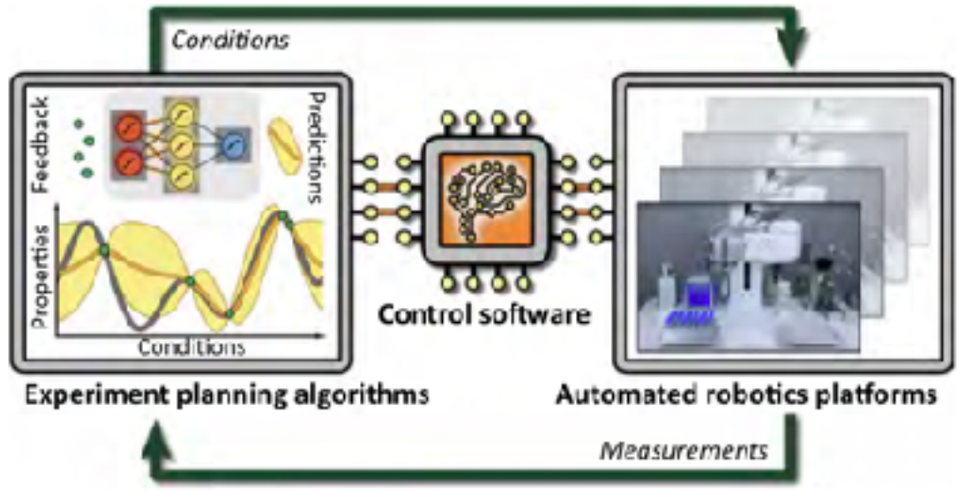
... also to chemistry!

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

Summary

The current "end-to-end" or "fully data-driven" strategy of ML is **too data-hungry**. But in many cases, we **cannot have enough data** for various practical restrictions (cost, time, ethics, privacy, etc).

- Model-based learning, Neuro-symbolic or ML-GOFAI integration.
  We can partly use *explicit models* for well-understood parts for sample efficiency and for filling the gap between correlation and causation
- Needs for modeling unverbalizable comon sense of domain experts
  We need a good strategy for building a *gigantic* data collection for this, as well as *self-supervised learning* and/or *meta learning* algorithms.

  *"Self-supervised is training a model to **fill in the blanks**. This is what is going to allow our AI systems to go to the next level. Some kind of common sense will emerge." (Yann LeCun)*
- Needs for novel techniques for compositionality (combinatorial generalization), out-of-distribution prediction (extrapolation), and their flexible transfer
  We need to somehow combine and flexibly transfer partial knowledge to generate a new thing or deal with completely new situations.

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[*,†,‡]

ACS Catalysis, 2020; 10: 2260-2297.

Chapter 2 is the general user's guide of ML for natural sciences.

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.