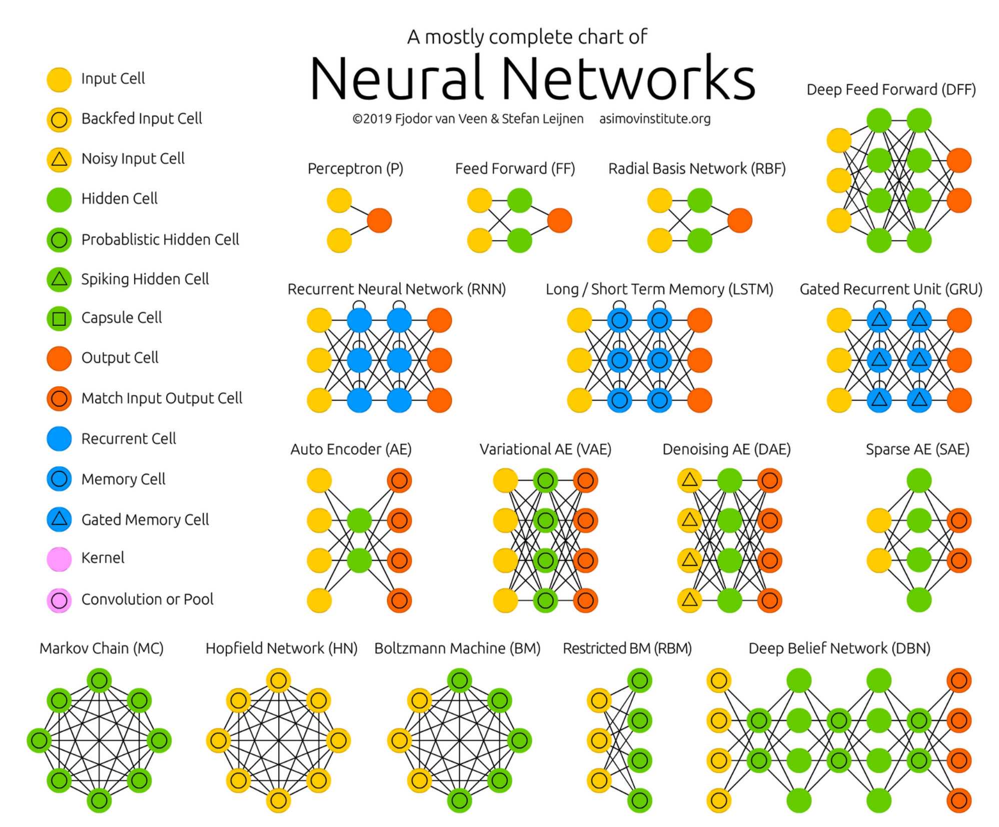
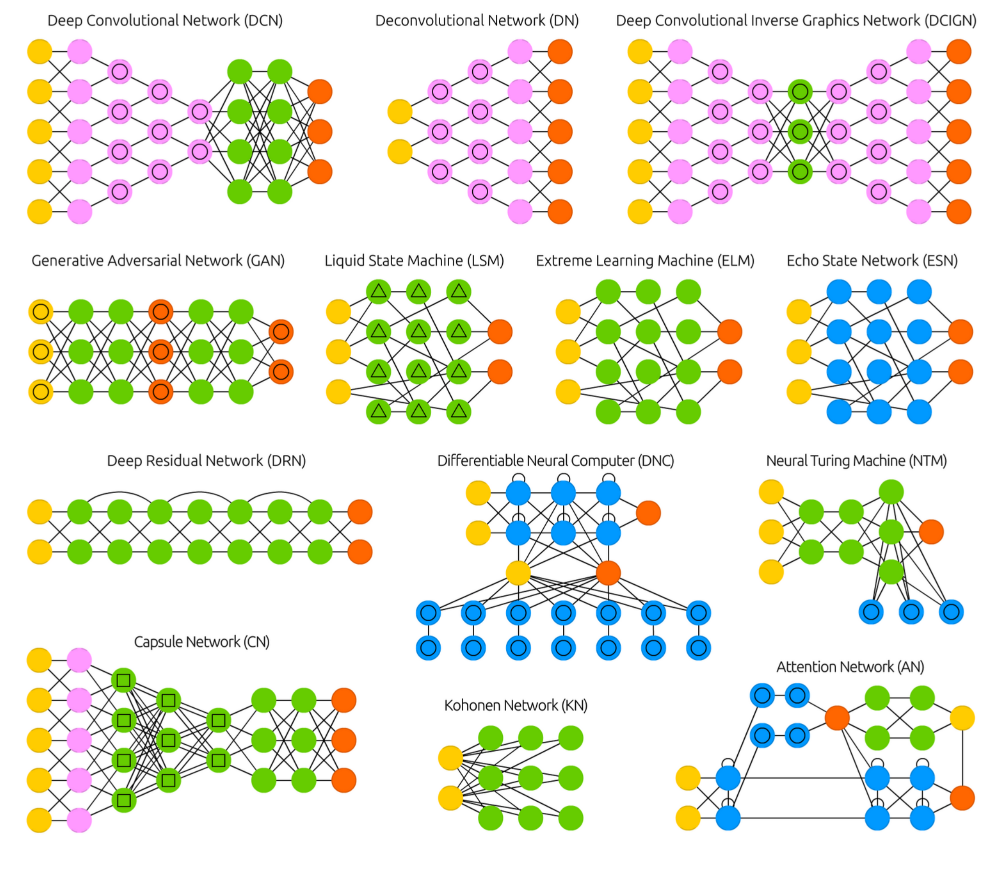
Introduction to Machine Learning for Chemists:
Visualization, Data Processing, Analysis,
Molecular Design

Bibliography: review articles

Version: Sunday, October 1, 2023

The NN zoo

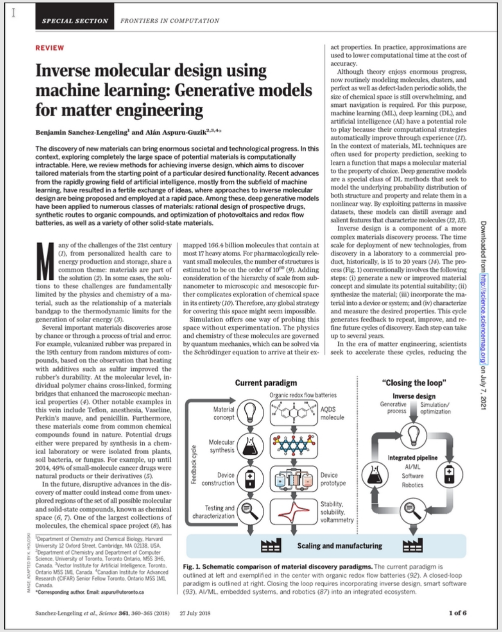

A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)


Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

AI in chemistry (AIChem)

dizzying – and somewhat demoralizing - array of ML applications to chemistry

Review articles



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.

Many of the challenges of the 21st century (*1*), from personalized health care to energy production and storage, share a common theme: materials are part of the solution (*2*). In some cases, the solutions to these challenges are fundamentally limited by the physics and chemistry of a material, such as the relationship of a materials bandgap to the thermodynamic limits for the generation of solar energy (*3*).

Several important materials discoveries arose by chance or through a process of trial and error. For example, vulcanized rubber was prepared in the 19th century from random mixtures of compounds, based on the observation that heating with additives such as sulfur improved the rubber's durability. At the molecular level, individual polymer chains cross-linked, forming bridges that enhanced the macroscopic mechanical properties (*4*). Other notable examples in this vein include Teflon, anesthesia, Vaseline, Perkin's mauve, and penicillin. Furthermore, these materials come from common chemical compounds found in nature. Potential drugs either were prepared by synthesis in a chemical laboratory or were isolated from plants, soil bacteria, or fungus. For example, up until 2014, 49% of small-molecule cancer drugs were natural products or their derivatives (*5*).

In the future, disruptive advances in the discovery of matter could instead come from unexplored regions of the set of all possible molecular and solid-state compounds, known as chemical space (*6, 7*). One of the largest collections of molecules, the chemical space project (*8*), has mapped 166.4 billion molecules that contain at most 17 heavy atoms. For pharmacologically relevant small molecules, the number of structures is estimated to be on the order of 10^{60} (*9*). Adding consideration of the hierarchy of scale from subnanometer to microscopic and mesoscopic further complicates exploration of chemical space in its entirety (*10*). Therefore, any global strategy for covering this space might seem impossible.

Simulation offers one way of probing this space without experimentation. The physics and chemistry of these molecules are governed by quantum mechanics, which can be solved via the Schrödinger equation to arrive at their exact 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 scale for deployment of new technologies, from discovery in a laboratory to a commercial product, historically, is 15 to 20 years (*14*). The process (Fig. 1) conventionally involves the following steps: (i) generate a new or improved material concept and simulate its potential suitability; (ii) synthesize the material; (iii) incorporate the material into a device or system; and (iv) characterize and measure the desired properties. This cycle generates feedback to repeat, improve, and refine future cycles of discovery. Each step can take up to several years.

In the era of matter engineering, scientists seek to accelerate these cycles, reducing the

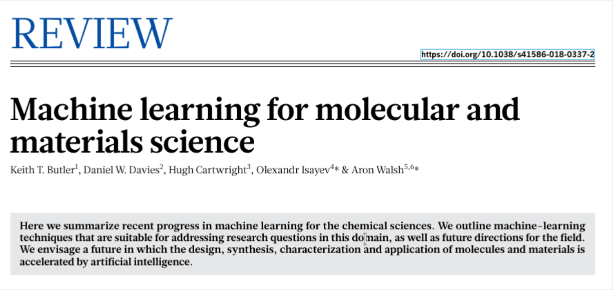
Fig. 1. Schematic comparison of material discovery paradigms. The current paradigm is outlined at left and exemplified in the center with organic redox flow batteries (*92*). A closed-loop paradigm is outlined at right. Closing the loop requires incorporating inverse design, smart software (*93*), AI/ML, embedded systems, and robotics (*87*) into an integrated ecosystem.

IMAGE: ADAPTED BY K. HOLOSKI

[1] Department of Chemistry and Chemical Biology, Harvard University 12 Oxford Street, Cambridge, MA 02138, USA. [2] Department of Chemistry and Department of Computer Science, University of Toronto, Toronto Ontario, M5S 3H6, Canada. [3] Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada. [4] Canadian Institute for Advanced Research (CIFAR) Senior Fellow Toronto, Ontario M5S 1M1, Canada.
*Corresponding author. Email: aspuru@utoronto.ca

Sanchez-Lengeling *et al.*, *Science* **361**, 360–365 (2018) 27 July 2018 **1 of 6**

B. Sanchez-Lengeling & A. Aspuru-Guzik (**2018**)
Inverse molecular design using machine learning: Generative models for matter engineering
Science **361**, 360-365



REVIEW

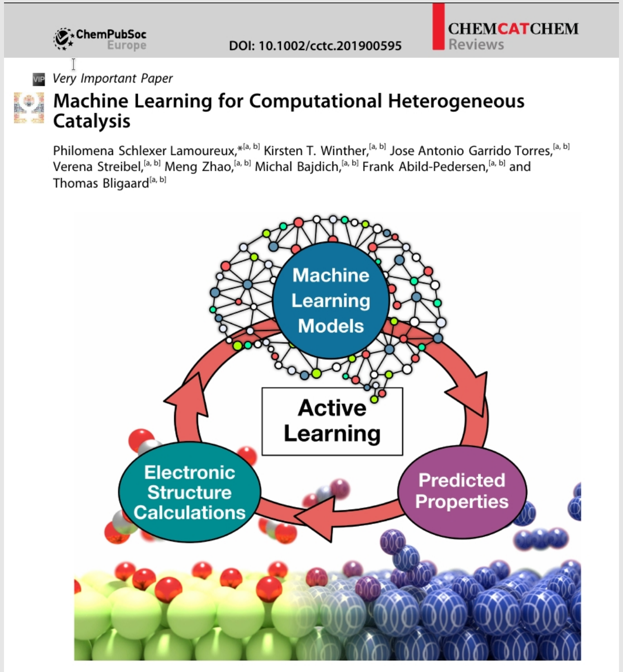
https://doi.org/10.1038/s41586-018-0337-2

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.

K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev & A. Walsh (**2018**)
Machine learning for molecular and materials science
Nature **559**, 547-555



ChemPubSoc Europe

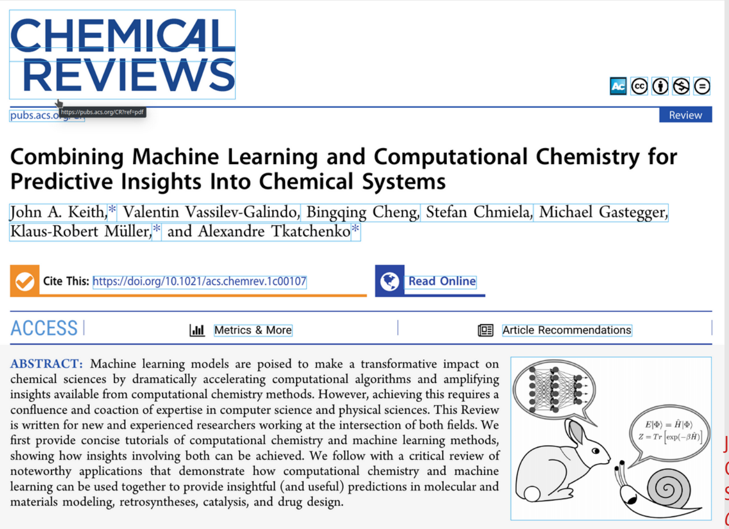
DOI: 10.1002/cctc.201900595

CHEMCATCHEM Reviews

Very Important Paper

Machine Learning for Computational Heterogeneous Catalysis

Philomena Schlexer Lamoureux,*[a, b] Kirsten T. Winther,[a, b] Jose Antonio Garrido Torres,[a, b] Verena Streibel,[a, b] Meng Zhao,[a, b] Michal Bajdich,[a, b] Frank Abild-Pedersen,[a, b] and Thomas Bligaard[a, b]



CHEMICAL REVIEWS

pubs.acs.org/CR

Review

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

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

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

ABSTRACT: Machine learning models are poised to make a transformative impact on chemical sciences by dramatically accelerating computational algorithms and amplifying insights available from computational chemistry methods. However, achieving this requires a confluence and coaction of expertise in computer science and physical sciences. This Review is written for new and experienced researchers working at the intersection of both fields. We first provide concise tutorials of computational chemistry and machine learning methods, showing how insights involving both can be achieved. We follow with a critical review of noteworthy applications that demonstrate how computational chemistry and machine learning can be used together to provide insightful (and useful) predictions in molecular and materials modeling, retrosyntheses, catalysis, and drug design.

$E|\Phi\rangle = \hat{H}|\Phi\rangle$
$Z = Tr\left[\exp(-\beta\hat{H})\right]$

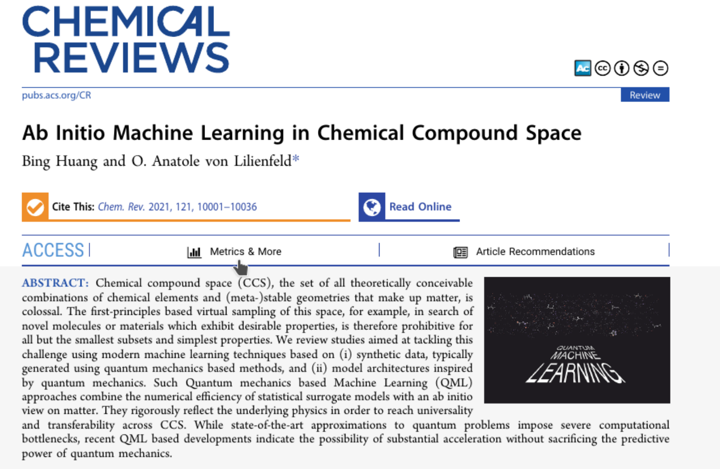
P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen & T. Bligaard (**2019**)
Machine Learning for Computational Heterogeneous Catalysis
ChemCatChem **11**, 3581-3601

J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller & A. Tkatchenko (**2021**)
Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems
Chem. Rev. **121**, 9816–9872

Version: Sunday, October 1, 2023

AI in chemistry (AIChem)

Review articles



CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Ab Initio Machine Learning in Chemical Compound Space

Bing Huang and O. Anatole von Lilienfeld*

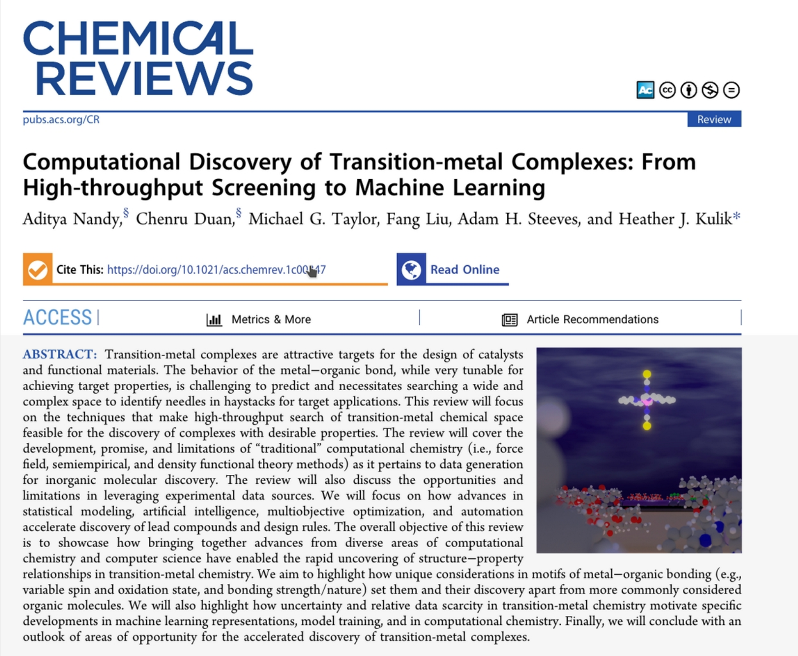
Cite This: *Chem. Rev.* 2021, 121, 10001−10036

Read Online

ACCESS | Metrics & More | Article Recommendations |

ABSTRACT: Chemical compound space (CCS), the set of all theoretically conceivable combinations of chemical elements and (meta-)stable geometries that make up matter, is colossal. The first-principles based virtual sampling of this space, for example, in search of novel molecules or materials which exhibit desirable properties, is therefore prohibitive for all but the smallest subsets and simplest properties. We review studies aimed at tackling this challenge using modern machine learning techniques based on (i) synthetic data, typically generated using quantum mechanics based methods, and (ii) model architectures inspired by quantum mechanics. Such Quantum mechanics based Machine Learning (QML) approaches combine the numerical efficiency of statistical surrogate models with an ab initio view on matter. They rigorously reflect the underlying physics in order to reach universality and transferability across CCS. While state-of-the-art approximations to quantum problems impose severe computational bottlenecks, recent QML based developments indicate the possibility of substantial acceleration without sacrificing the predictive power of quantum mechanics.

B. Huang & O. A. von Lilienfeld (**2021**)
Ab Initio Machine Learning in Chemical Compound Space
Chem. Rev. **121**, 10001-10036



CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning

Aditya Nandy,[§] Chenru Duan,[§] Michael G. Taylor, Fang Liu, Adam H. Steeves, and Heather J. Kulik*

Cite This: https://doi.org/10.1021/acs.chemrev.1c00347

Read Online

ACCESS | Metrics & More | Article Recommendations |

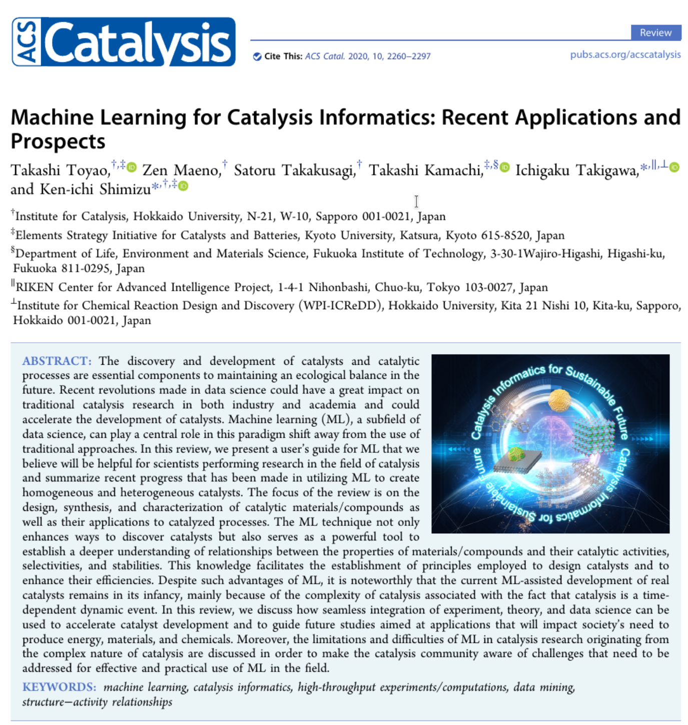
ABSTRACT: Transition-metal complexes are attractive targets for the design of catalysts and functional materials. The behavior of the metal−organic bond, while very tunable for achieving target properties, is challenging to predict and necessitates searching a wide and complex space to identify needles in haystacks for target applications. This review will focus on the techniques that make high-throughput search of transition-metal chemical space feasible for the discovery of complexes with desirable properties. The review will cover the development, promise, and limitations of "traditional" computational chemistry (i.e., force field, semiempirical, and density functional theory methods) as it pertains to data generation for inorganic molecular discovery. The review will also discuss the opportunities and limitations in leveraging experimental data sources. We will focus on how advances in statistical modeling, artificial intelligence, multiobjective optimization, and automation accelerate discovery of lead compounds and design rules. The overall objective of this review is to showcase how bringing together advances from diverse areas of computational chemistry and computer science have enabled the rapid uncovering of structure−property relationships in transition-metal chemistry. We aim to highlight how unique considerations in motifs of metal−organic bonding (e.g., variable spin and oxidation state, and bonding strength/nature) set them and their discovery apart from more commonly considered organic molecules. We will also highlight how uncertainty and relative data scarcity in transition-metal chemistry motivate specific developments in machine learning representations, model training, and in computational chemistry. Finally, we will conclude with an outlook of areas of opportunity for the accelerated discovery of transition-metal complexes.

Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. (**2021**)
Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning.
Chem. Rev. **121**, 9927–10000

Version: Sunday, October 1, 2023

AI in chemistry (AIChem)

Review articles


Catalysis Informatics for Sustainable Future
Catalysis Informatics for Sustainable Future

ACS Catalysis

Review

Cite This: *ACS Catal.* 2020, 10, 2260−2297

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[*,†,‡]

[†] Institute for Catalysis, Hokkaido University, N-21, W-10, Sapporo 001-0021, Japan

[‡] Elements Strategy Initiative for Catalysts and Batteries, Kyoto University, Katsura, Kyoto 615-8520, Japan

[§] Department of Life, Environment and Materials Science, Fukuoka Institute of Technology, 3-30-1Wajiro-Higashi, Higashi-ku, Fukuoka 811-0295, Japan

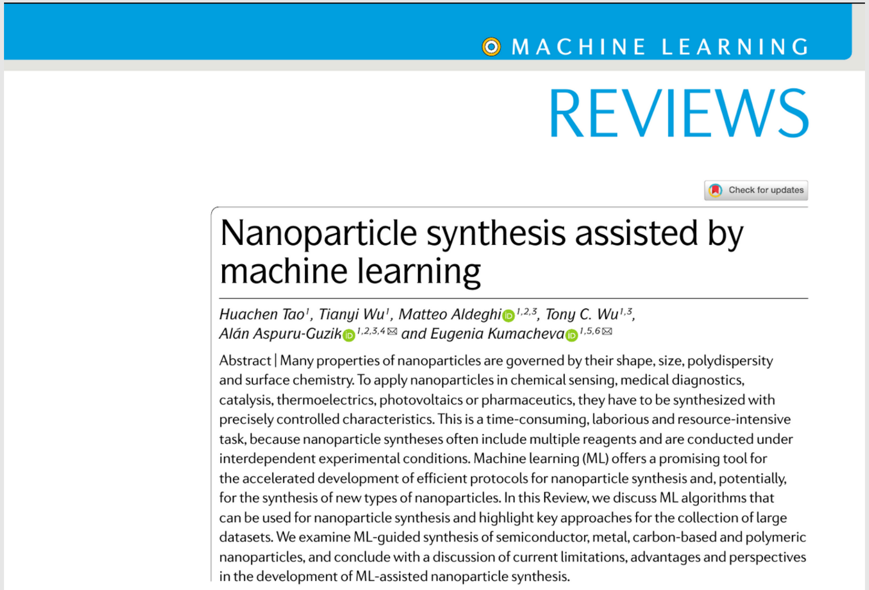
[∥] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[⊥] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan

ABSTRACT: The discovery and development of catalysts and catalytic processes are essential components to maintaining an ecological balance in the future. Recent revolutions made in data science could have a great impact on traditional catalysis research in both industry and academia and could accelerate the development of catalysts. Machine learning (ML), a subfield of data science, can play a central role in this paradigm shift away from the use of traditional approaches. In this review, we present a user's guide for ML that we believe will be helpful for scientists performing research in the field of catalysis and summarize recent progress that has been made in utilizing ML to create homogeneous and heterogeneous catalysts. The focus of the review is on the design, synthesis, and characterization of catalytic materials/compounds as well as their applications to catalyzed processes. The ML technique not only enhances ways to discover catalysts but also serves as a powerful tool to establish a deeper understanding of relationships between the properties of materials/compounds and their catalytic activities, selectivities, and stabilities. This knowledge facilitates the establishment of principles employed to design catalysts and to enhance their efficiencies. Despite such advantages of ML, it is noteworthy that the current ML-assisted development of real catalysts remains in its infancy, mainly because of the complexity of catalysis associated with the fact that catalysis is a time-dependent dynamic event. In this review, we discuss how seamless integration of experiment, theory, and data science can be used to accelerate catalyst development and to guide future studies aimed at applications that will impact society's need to produce energy, materials, and chemicals. Moreover, the limitations and difficulties of ML in catalysis research originating from the complex nature of catalysis are discussed in order to make the catalysis community aware of challenges that need to be addressed for effective and practical use of ML in the field.

KEYWORDS: *machine learning, catalysis informatics, high-throughput experiments/computations, data mining, structure−activity relationships*

Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. (**2020**)
Machine Learning for Catalysis Informatics: Recent Applications and Prospects
ACS Catal. **10** (3), 2260–2297.



MACHINE LEARNING

REVIEWS

Check for updates

Nanoparticle synthesis assisted by machine learning

Huachen Tao[1], *Tianyi Wu*[1], *Matteo Aldeghi*[1,2,3], *Tony C. Wu*[1,3], *Alán Aspuru-Guzik*[1,2,3,4 ✉] *and Eugenia Kumacheva*[1,5,6 ✉]

Abstract | Many properties of nanoparticles are governed by their shape, size, polydispersity and surface chemistry. To apply nanoparticles in chemical sensing, medical diagnostics, catalysis, thermoelectrics, photovoltaics or pharmaceutics, they have to be synthesized with precisely controlled characteristics. This is a time-consuming, laborious and resource-intensive task, because nanoparticle syntheses often include multiple reagents and are conducted under interdependent experimental conditions. Machine learning (ML) offers a promising tool for the accelerated development of efficient protocols for nanoparticle synthesis and, potentially, for the synthesis of new types of nanoparticles. In this Review, we discuss ML algorithms that can be used for nanoparticle synthesis and highlight key approaches for the collection of large datasets. We examine ML-guided synthesis of semiconductor, metal, carbon-based and polymeric nanoparticles, and conclude with a discussion of current limitations, advantages and perspectives in the development of ML-assisted nanoparticle synthesis.

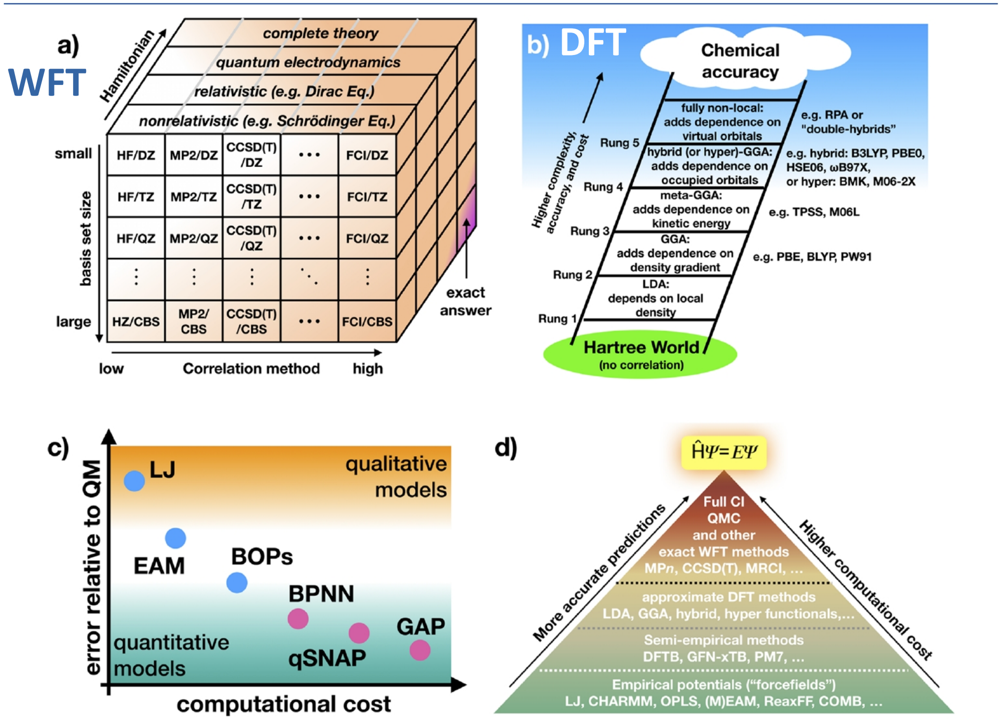
H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik & E. Kumacheva (**2021**)
Nanoparticle synthesis assisted by machine learning
Nat. Rev. Mater. **6**: 701-716

Version: Sunday, October 1, 2023

AI in chemistry (AIChem)

AIChem ← is also currently used to reduce the cost of computational [quantum] chemistry

Calculation of energies & forces (MD / geometry optimization)



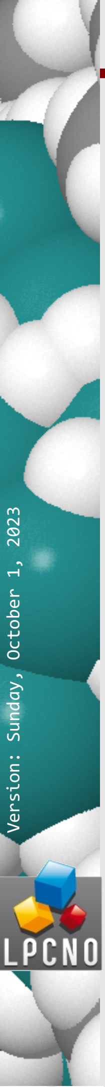
a)
WFT
Hamiltonian
complete theory
quantum electrodynamics
relativistic (e.g. Dirac Eq.)
nonrelativistic (e.g. Schrödinger Eq.)
small
basis set size
large
HF/DZ
MP2/DZ
CCSD(T)/DZ
FCI/DZ
HF/TZ
MP2/TZ
CCSD(T)/TZ
FCI/TZ
HF/QZ
MP2/QZ
CCSD(T)/QZ
FCI/QZ
HZ/CBS
MP2/CBS
CCSD(T)/CBS
FCI/CBS
exact answer
low
Correlation method
high
b) DFT
Chemical accuracy
Higher complexity, accuracy, and cost
Rung 5
fully non-local: adds dependence on virtual orbitals
e.g. RPA or "double-hybrids"
Rung 4
hybrid (or hyper)-GGA: adds dependence on occupied orbitals
e.g. hybrid: B3LYP, PBE0, HSE06, ωB97X, or hyper: BMK, M06-2X
Rung 3
meta-GGA: adds dependence on kinetic energy
e.g. TPSS, M06L
Rung 2
GGA: adds dependence on density gradient
e.g. PBE, BLYP, PW91
Rung 1
LDA: depends on local density
Hartree World
(no correlation)
c)
error relative to QM
LJ
qualitative models
EAM
BOPs
BPNN
GAP
quantitative models
qSNAP
computational cost
d)
ĤΨ=EΨ
More accurate predictions
Higher computational cost
Full CI
QMC
and other
exact WFT methods
MPn, CCSD(T), MRCI, ...
approximate DFT methods
LDA, GGA, hybrid, hyper functionals,...
Semi-empirical methods
DFTB, GFN-xTB, PM7, ...
Empirical potentials ("forcefields")
LJ, CHARMM, OPLS, (M)EAM, ReaxFF, COMB, ...

Figure 5. (a) "Magic cube"[59] depiction of hierarchies of correlated wavefunction approaches. (b) "Jacob's Ladder"[60] depiction of hierarchies of Kohn–Sham density functional theory (DFT) approaches. (c) Hierarchies of atomistic potentials. (d) Overall hierarchies in predictive atomic scale modeling methods.

J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller & A. Tkatchenko (**2021**)
Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **121**, 9816–9872

Version: Sunday, October 1, 2023

LPCNO



Version: Sunday, October 1, 2023

Short selection of [simple] applications of supervised learning to chemistry

All these applications face the same problem

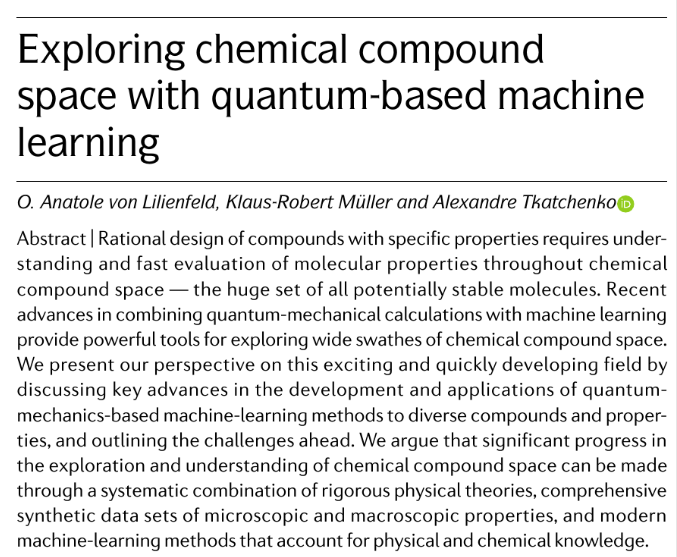
ML is only as good as the data it learned from

quality

diversity

volume

when ML is applied for prediction purpose (ML-QSPR)



Exploring chemical compound space with quantum-based machine learning

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

Abstract | Rational design of compounds with specific properties requires understanding and fast evaluation of molecular properties throughout chemical compound space — the huge set of all potentially stable molecules. Recent advances in combining quantum-mechanical calculations with machine learning provide powerful tools for exploring wide swathes of chemical compound space. We present our perspective on this exciting and quickly developing field by discussing key advances in the development and applications of quantum-mechanics-based machine-learning methods to diverse compounds and properties, and outlining the challenges ahead. We argue that significant progress in the exploration and understanding of chemical compound space can be made through a systematic combination of rigorous physical theories, comprehensive synthetic data sets of microscopic and macroscopic properties, and modern machine-learning methods that account for physical and chemical knowledge.

O. A. von Lilienfeld, K.-R. Müller & A. Tkatchenko (**2020**)
Exploring chemical compound space with quantum-based machine learning
Nat. Rev. Chem. **4**: 347-358

but it is also a remarkable tool to find correlations between data

Version: Sunday, October 1, 2023

AI in chemistry (AIChem)

Control of an organic synthesis robot

space of chemical reactions explored quickly

organic synthesis robot performs chemical reactions and analysis faster than a human

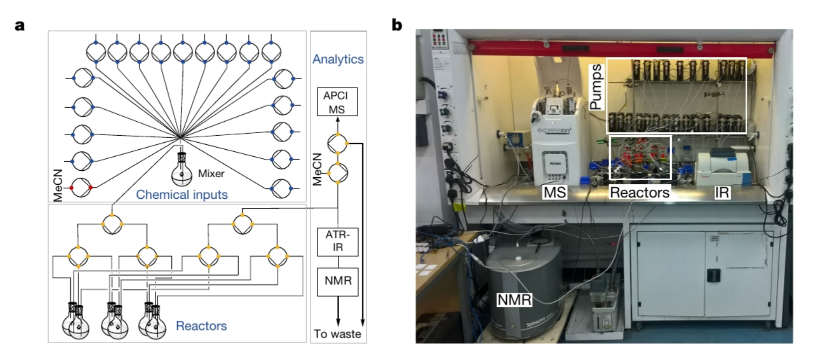
prediction of the reactivity of possible reagent combinations after conducting a small number of experiments

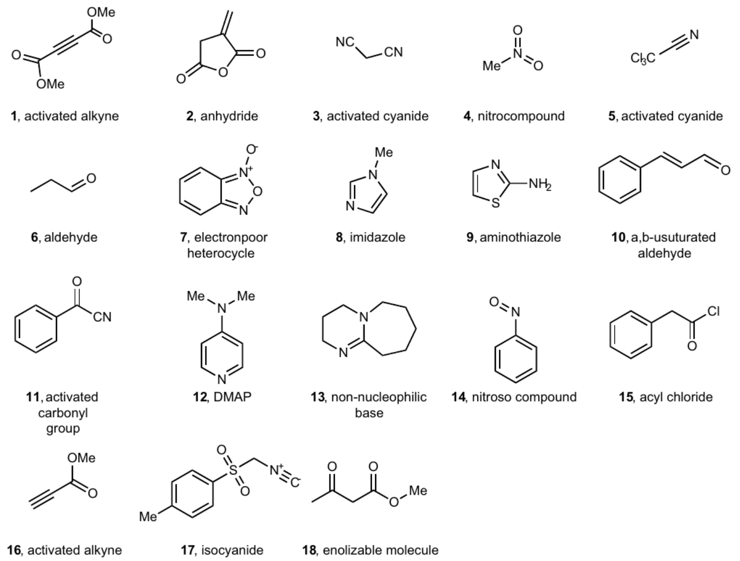
decision making by ML

robot equipped with real-time sensors to record the spectra of the reaction mixtures:

- flow benchtop NMR system
- mass spectrometer
- IR system

reactions mixtures classified as reactive (**1**) or non reactive (**0**) by the robot (uses a **support vector machine (SVM) algorithm** with a **linear kernel model**)





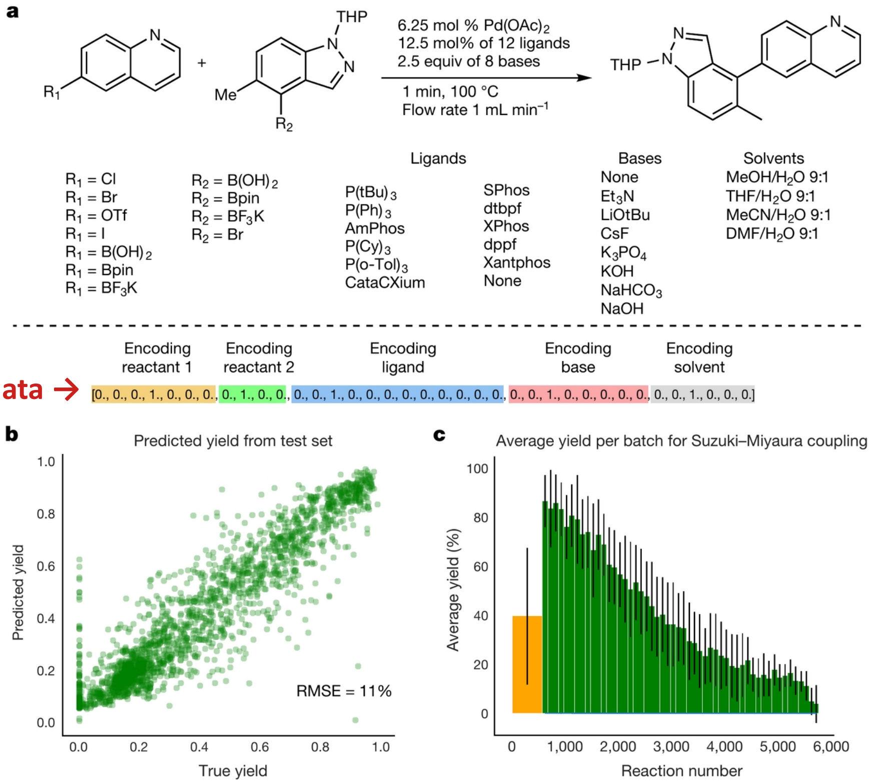
Extended Data Fig. 1 | Reaction space explored. The chemical inputs (**1**–**18**) used in the platform to search for new transformations and to evaluate the performance of the algorithm.

J. M. Granda, L. Donina, V. Dragone, D.-L. Long & L. Cronin (**2018**). Controlling an organic synthesis robot with machine learning to search for new reactivity
Nature **559**, 377-381

Version: Sunday, October 1, 2023

AI in chemistry (AIChem)

Version: Sunday, October 1, 2023



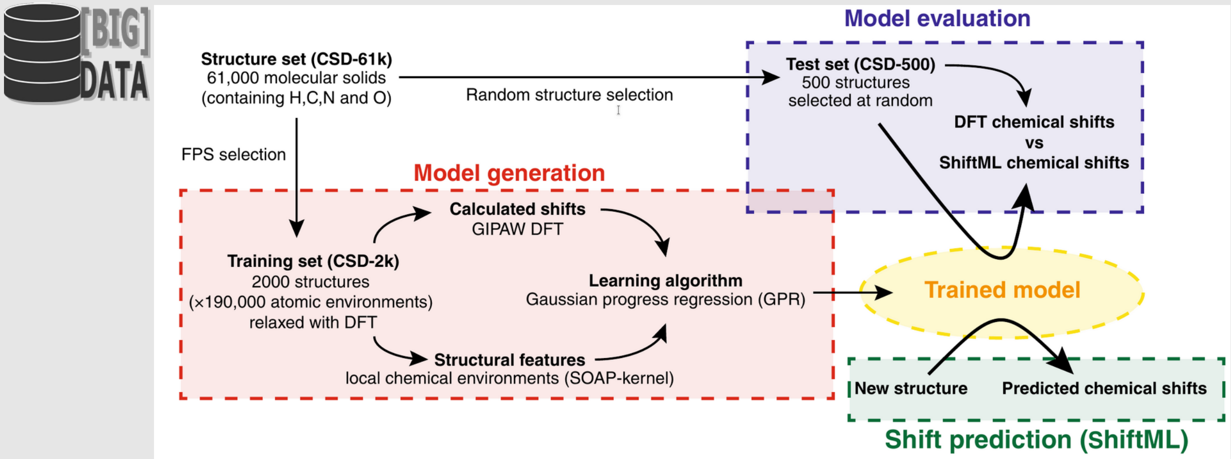
flattening of the data →

Fig. 4 | Exploring the Suzuki–Miyaura reaction using machine learning. a, The reaction space of the Suzuki–Miyaura reaction. Shown are the identity of reactants, ligand, base and solvent, and the vector representation of the reaction for machine learning. **b**, Validation of the predictive power of the model for a test set of 30% of the reactions (1,728 reactions). RMSE, root-mean-square error. **c**, Simulation of the machine-learning-controlled exploration of this reaction space. The yellow bar shows the initial random choice of 10% of reaction space (576 reactions). The green bars show the next batches of 100 reactions chosen by the machine learning algorithm. The error bars represent the standard deviation within individual batches for Suzuki–Miyaura coupling.

J. M. Granda, L. Donina, V. Dragone, D.-L. Long & L. Cronin (**2018**). Controlling an organic synthesis robot with machine learning to search for new reactivity
Nature **559**, 377-381

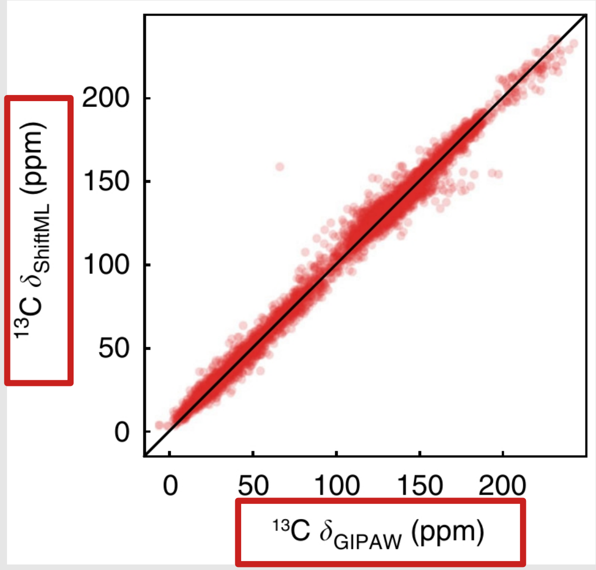
AI in chemistry (AIChem)

Prediction of spectroscopic parameters



F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti & L. Emsley (**2018**)
Chemical shifts in molecular solids by machine learning
Nat. Commun. **9**, 4501

Test



predicted values as a function of the actual values

AI in chemistry (AIChem)

keras software
4 convolutional layers + 2 fully connected layers

Deep Convolutional Network (DCN)

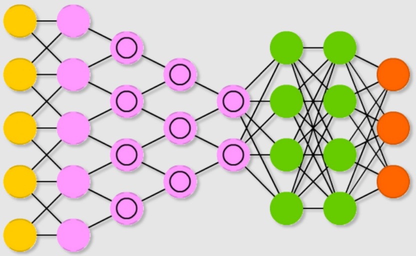


Image recognition

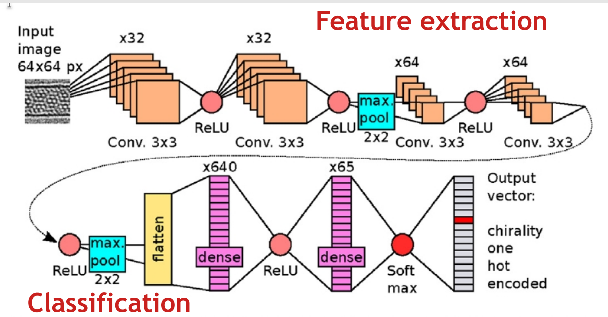


Fig. 3. Schematic detailing the architecture of the CNNs. (A colour version of this figure can be viewed online.)

Molecular Dynamics + STEM image simulation

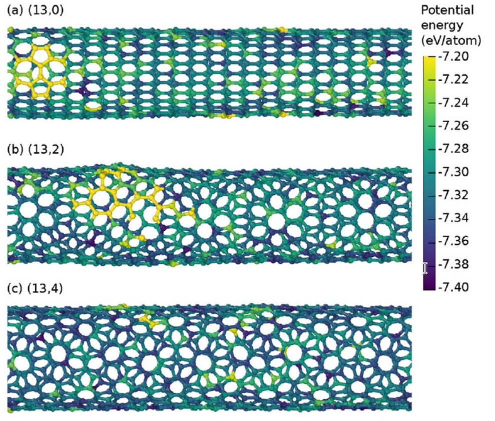
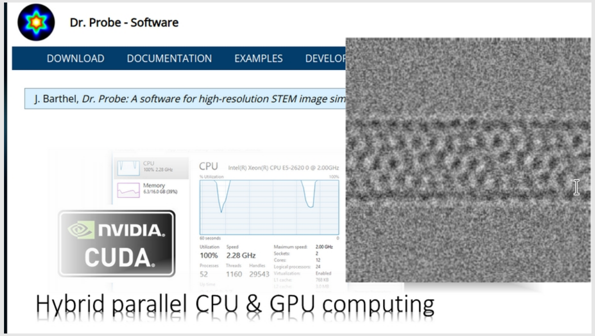


Fig. 1. Some examples of final snapshots from the molecular dynamics simulations. Defects (higher potential energy, appear in yellow) have been introduced by high-temperature annealing. The visualizations are generated using the OVITO software [51]. (A colour version of this figure can be viewed online.)



Dr. Probe - Software

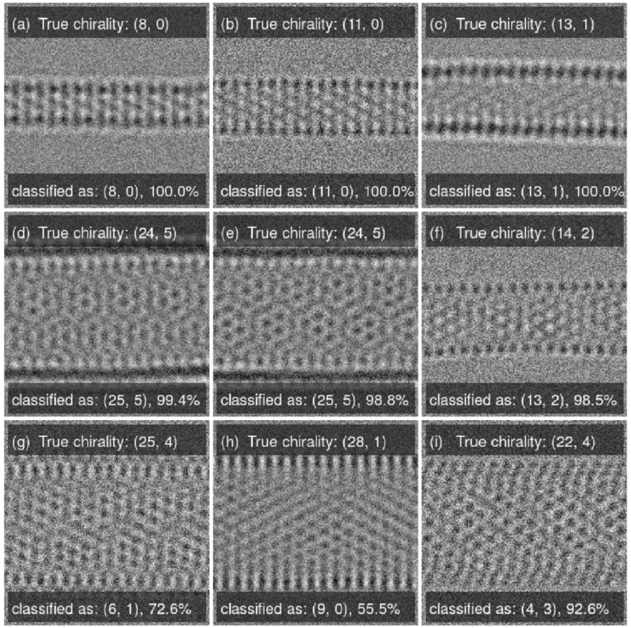
J. Barthel, *Dr. Probe: A software for high-resolution STEM image sim*

Hybrid parallel CPU & GPU computing

[BIG] DATA

5000 images per chirality

~1.3 10^6 images in total

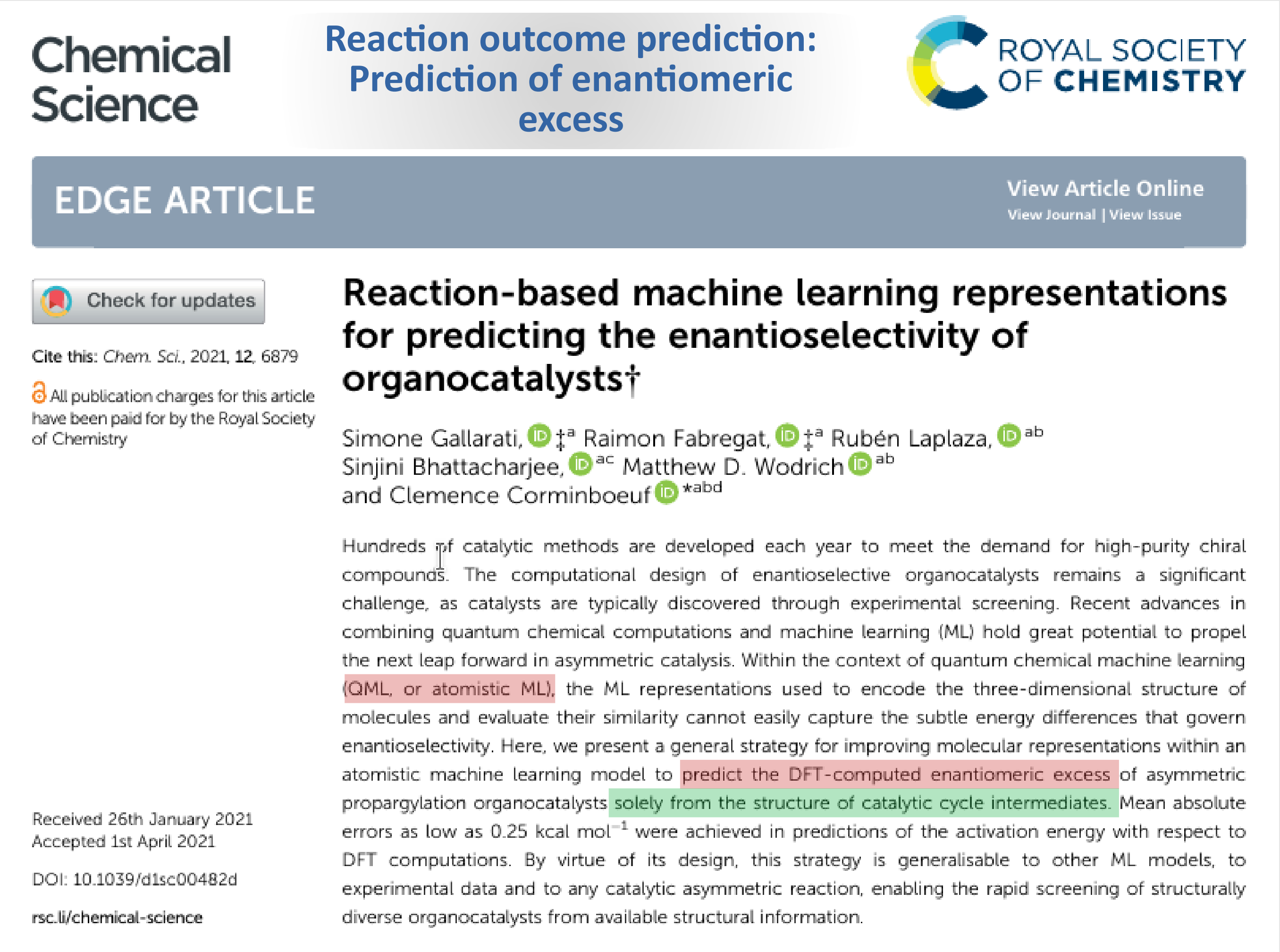

(a) True chirality: (8, 0)
classified as: (8, 0), 100.0%
(b) True chirality: (11, 0)
classified as: (11, 0), 100.0%
(c) True chirality: (13, 1)
classified as: (13, 1), 100.0%
(d) True chirality: (24, 5)
classified as: (25, 5), 99.4%
(e) True chirality: (24, 5)
classified as: (25, 5), 98.8%
(f) True chirality: (14, 2)
classified as: (13, 2), 98.5%
(g) True chirality: (25, 4)
classified as: (6, 1), 72.6%
(h) True chirality: (28, 1)
classified as: (9, 0), 55.5%
(i) True chirality: (22, 4)
classified as: (4, 3), 92.6%

G. D. Förster, A. Castan, A. Loiseau, J. Nelayah, D. Alloyeau, F. Fossard, C. Bichara & H. Amara (**2020**)
A deep learning approach for determining the chiral indices of carbon nanotubes from high-resolution transmission electron microscopy images
Carbon **169**, 465-474

AI in chemistry (AIChem)

Version: Sunday, October 1, 2023

Reaction outcome prediction: Prediction of enantiomeric excess



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2021, 12, 6879

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

Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts†

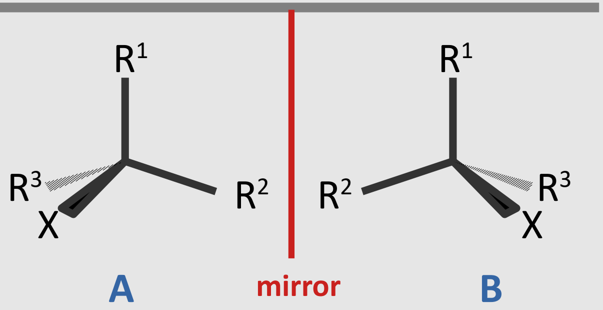
Simone Gallarati,‡[a] Raimon Fabregat,‡[a] Rubén Laplaza,[ab] Sinjini Bhattacharjee,[ac] Matthew D. Wodrich[ab] and Clemence Corminboeuf*[abd]

Hundreds of catalytic methods are developed each year to meet the demand for high-purity chiral compounds. The computational design of enantioselective organocatalysts remains a significant challenge, as catalysts are typically discovered through experimental screening. Recent advances in combining quantum chemical computations and machine learning (ML) hold great potential to propel the next leap forward in asymmetric catalysis. Within the context of quantum chemical machine learning (QML, or atomistic ML), the ML representations used to encode the three-dimensional structure of molecules and evaluate their similarity cannot easily capture the subtle energy differences that govern enantioselectivity. Here, we present a general strategy for improving molecular representations within an atomistic machine learning model to predict the DFT-computed enantiomeric excess of asymmetric propargylation organocatalysts solely from the structure of catalytic cycle intermediates. Mean absolute errors as low as 0.25 kcal mol^{-1} were achieved in predictions of the activation energy with respect to DFT computations. By virtue of its design, this strategy is generalisable to other ML models, to experimental data and to any catalytic asymmetric reaction, enabling the rapid screening of structurally diverse organocatalysts from available structural information.

Received 26th January 2021
Accepted 1st April 2021

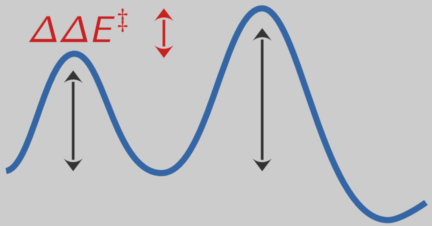
DOI: 10.1039/d1sc00482d

rsc.li/chemical-science



$$\text{e.e.} = \frac{|A - B|}{A + B} \times 100$$

Computational chemistry



$$\text{e.e.} = \frac{1 - e^{\Delta\Delta E^{\ddagger}/RT}}{1 + e^{\Delta\Delta E^{\ddagger}/RT}} \times 100$$

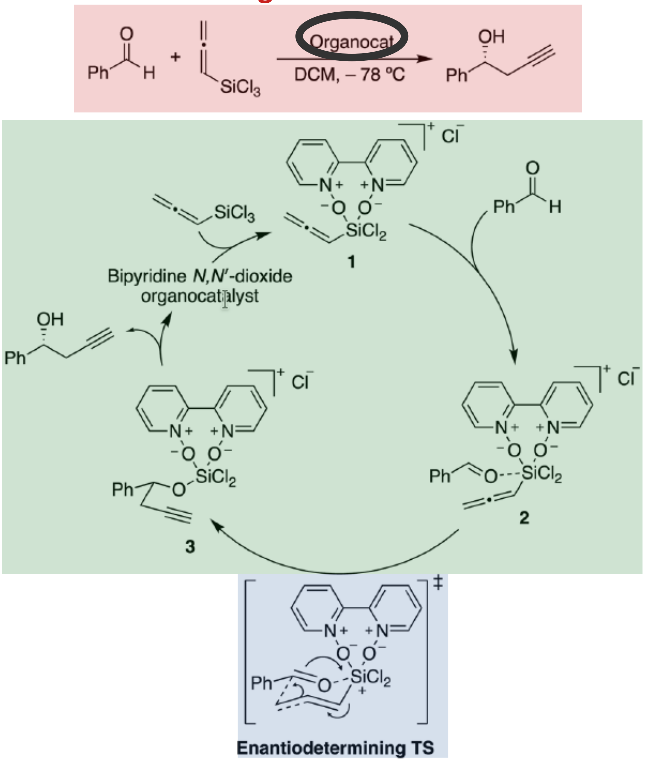
With $\Delta\Delta E^{\ddagger} \lesssim 5$ kcal.mol^{-1}

S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich & C. Corminboeuf (**2021**)
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts, *Chem. Sci.* **12**, 6879-6889

Reaction outcome prediction. Prediction of enantiomeric excess

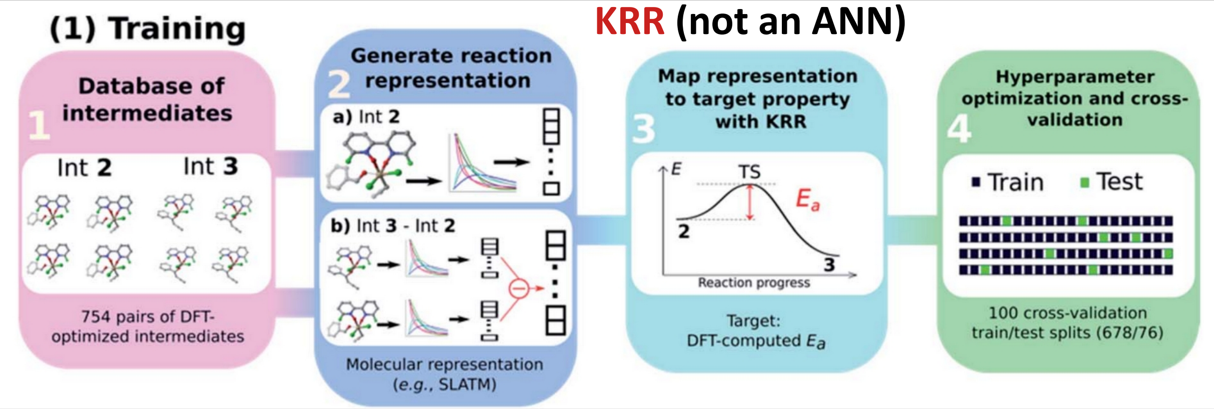
Chemical context

Target reaction



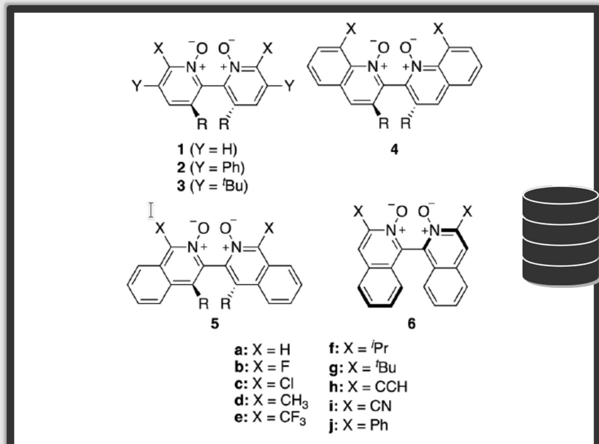
Scheme 2 Catalytic cycle for the propargylation of benzaldehyde with allenyltrichlorosilane, showing the rate-limiting and stereocontrolling transition state. Adapted from ref. 85.

KRR (not an ANN)



(1) Training
1
Database of intermediates
Int 2
Int 3
754 pairs of DFT-optimized intermediates
2
Generate reaction representation
a) Int 2
b) Int 3 - Int 2
Molecular representation (e.g., SLATM)
3
Map representation to target property with KRR
E
TS
E_a
2
3
Reaction progress
Target: DFT-computed E_a
4
Hyperparameter optimization and cross-validation
Train
Test
100 cross-validation train/test splits (678/76)

76 Lewis base organocatalysts



Scheme 1 Library of axially chiral bipyridine *N,N'*-dioxide organocatalysts. R = H or Me. Adapted from ref. 74.

[BIG] DATA

5 possible ligand arrangements around the hexacoordinate silicon in the TS
+
the alkyl nucleophile can add to either face of benzaldehyde

760 DFT TS geometries and activation energies

Version: Sunday, October 1, 2023

S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich & C. Corminboeuf (**2021**)
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts, *Chem. Sci.* **12**, 6879-6889

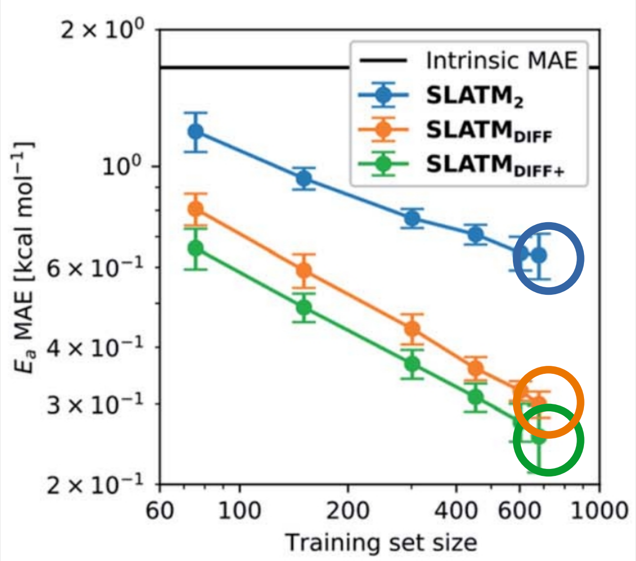
Reaction outcome prediction. Prediction of enantiomeric excess

Main results

Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|E_a^{(\mathrm{ref.})} - \hat{E}_a\right|$$

SLATM = structural descriptors



Learning curves:
MAE in test sets predictions of E_a

0.54 ± 0.06 kcal mol^{-1} (insufficient accuracy)

0.31 ± 0.20 kcal mol^{-1}

0.25 ± 0.40 kcal mol^{-1}

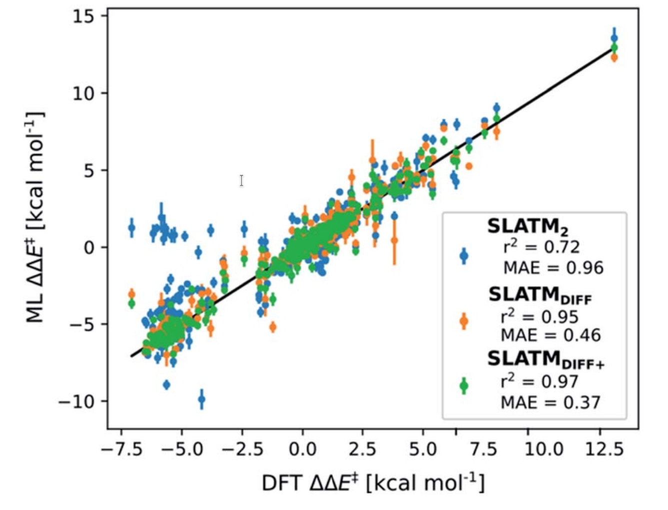


Fig. 2 Predictions of ΔΔ$E^‡$ *vs.* DFT reference for the three approaches discussed. Mean Absolute Errors (MAE) are reported in kcal mol^{-1}. These predictions are obtained by averaging the predictions obtained from the cross-validation scheme with 100 different random train/test splits. The error bars indicate the standard deviation of ML ΔΔ$E^‡$, derived from the standard deviations in the E_a prediction of the 100 different random train/test splits.

S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich & C. Corminboeuf (**2021**)
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts, *Chem. Sci.* **12**, 6879-6889

Version: Sunday, October 1, 2023

Phase behavior of ionic liquids

Prediction of CO_2 solubility in ionic liquids

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Prediction of CO_2 solubility in ionic liquids using machine learning methods

Zhen Song[a], Huaiwei Shi[a,b], Xiang Zhang[c], Teng Zhou[a,b,*]

[a] Process Systems Engineering, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany
[b] Process Systems Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, D-39106 Magdeburg, Germany
[c] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Database = 10 116 experimental data points

124 different ILs

***T*: [243.2 K – 453.15 K]**

***p*: [0.00798 bar – 499.9 bar]**

training set, 80% of the data = 8093 values

test set, 20% of the data = 2023 values

***cations*: imidazolium, pyrrolidinium, pyridinium, piperidinium, ammonium, phosphonium, sulfonium**

***anions*: tetrafluoroborate [BF4], chloride [Cl], dicyanamide [DCA], nitrate [NO3], hexafluorophosphate [PF6], thiocyanate [SCN], tri-cyanomethanide [C(CN)3], hydrogen sulfate [HSO4], bis(trifluoromethylsulfonyl)amide [Tf2N], methylsulfate [MeSO4], etc**

possible modelling approaches

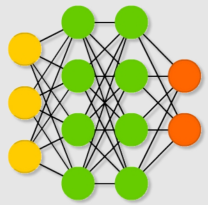
Thermodynamic models
DFT studies in solvent
} **not accurate enough**

other possibility
quantitative structure-property relationship (QSPR)

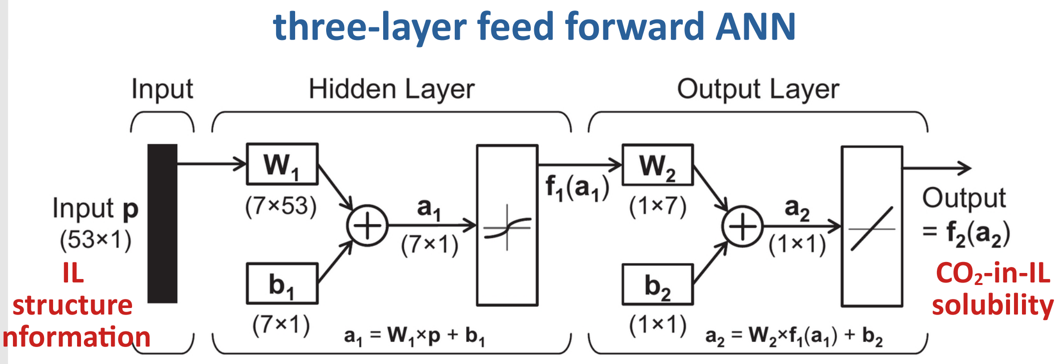
Machine Learning
(surrogate modeling)

Z. Song, H. Shi, X. Zhang & T. Zhou (**2020**)
Prediction of CO2 solubility in ionic liquids using machine learning methods, *Chem. Eng. Sci.* **223**, 115752

Phase behavior of ionic liquids

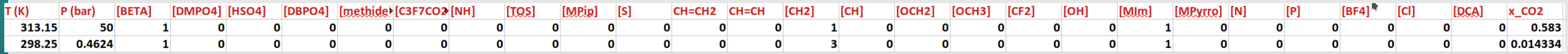


Deep Feed Forward (DFF)



three-layer feed forward ANN

Fig. 1. Schematic structure of the employed three-layer ANN (the dimensions of W_1, W_2, b_1, and b_2 are given in the brackets).



T (K)	P (bar)	[BETA]	[DMPO4]	[HSO4]	[DBPO4]	[methide	[C3F7CO2	[NH]	[TOS]	[MPip]	[S]	CH=CH2	CH=CH	[CH2]	[CH]	[OCH2]	[OCH3]	[CF2]	[OH]	[MIm]	[MPyrro]	[N]	[P]	[BF4]	[Cl]	[DCA]	x_CO2
313.15	50	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0.583
298.25	0.4624	1	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0.014334

data organized as vectors
IL structure information = occurrences of functional groups in the IL

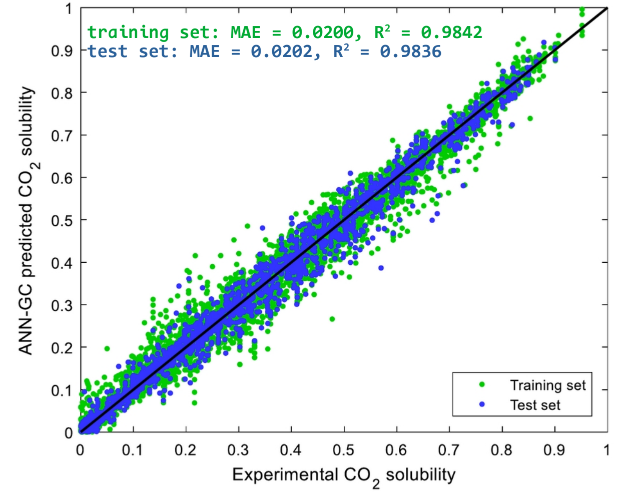


Fig. 2. Comparison between the experimental and ANN-GC predicted CO_2 solubility.

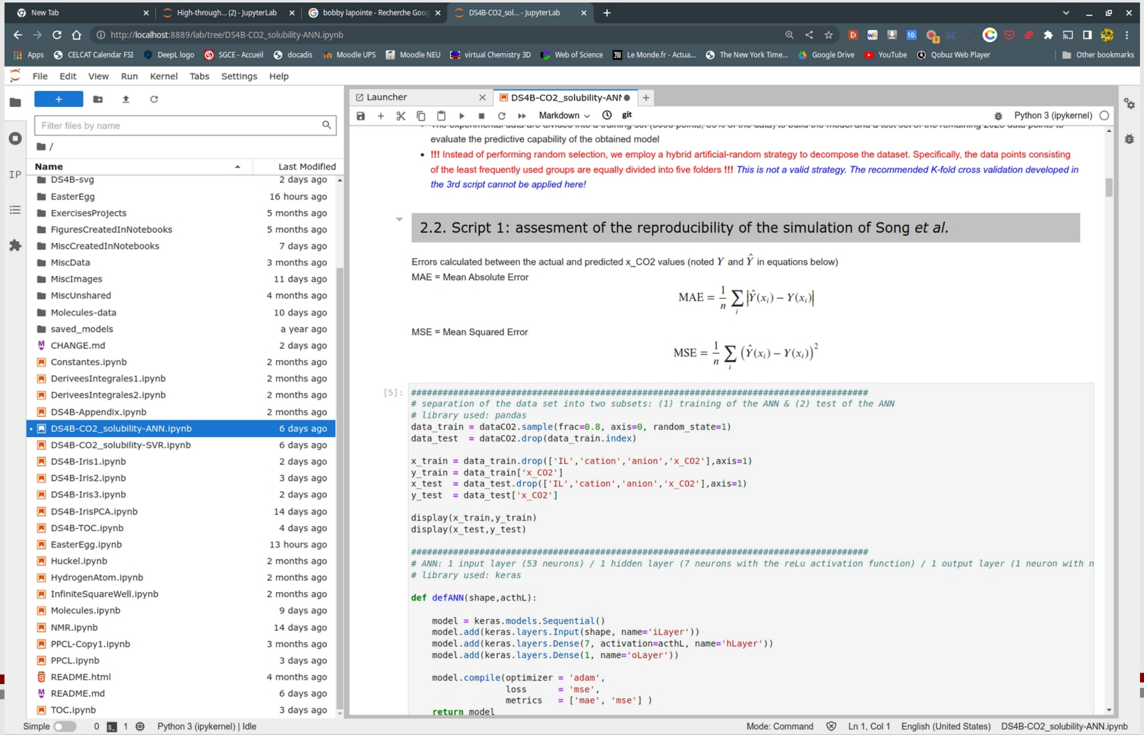
Version: Sunday, October 1, 2023

Two notebooks

Python in the [Physical] Chemistry Lab

[PytChem]

DS4B-CO2_solubility-ANN.ipynb
DS4B-CO2_solubility-SVR.ipynb



Name	Last Modified
DS4B-svg	2 days ago
EasterEgg	16 hours ago
ExercisesProjects	5 months ago
FiguresCreatedInNotebooks	5 months ago
MiscCreatedInNotebooks	7 days ago
MiscData	3 months ago
MiscImages	11 days ago
MiscUnshared	4 months ago
Molecules-data	10 days ago
saved_models	a year ago
CHANGE.md	2 days ago
Constantes.ipynb	2 months ago
DeriveesIntegrales1.ipynb	2 months ago
DeriveesIntegrales2.ipynb	2 months ago
DS4B-Appendix.ipynb	2 months ago
DS4B-CO2_solubility-ANN.ipynb	6 days ago
DS4B-CO2_solubility-SVR.ipynb	6 days ago
DS4B-Iris1.ipynb	2 days ago
DS4B-Iris2.ipynb	3 days ago
DS4B-Iris3.ipynb	2 days ago
DS4B-IrisPCA.ipynb	14 days ago
DS4B-TOC.ipynb	4 days ago
EasterEgg.ipynb	13 hours ago
Huckel.ipynb	2 months ago
HydrogenAtom.ipynb	2 months ago
InfiniteSquareWell.ipynb	2 months ago
Molecules.ipynb	9 days ago
NMR.ipynb	14 days ago
PPCL-Copy1.ipynb	3 months ago
PPCL.ipynb	3 days ago
README.html	4 months ago
README.md	6 days ago
TOC.ipynb	3 days ago

evaluate the predictive capability of the obtained model

- !!! Instead of performing random selection, we employ a hybrid artificial-random strategy to decompose the dataset. Specifically, the data points consisting of the least frequently used groups are equally divided into five folders !!! *This is not a valid strategy. The recommended K-fold cross validation developed in the 3rd script cannot be applied here!*

2.2. Script 1: assesment of the reproducibility of the simulation of Song *et al.*

Errors calculated between the actual and predicted x_CO2 values (noted Y and $\hat{Y}$ in equations below)
MAE = Mean Absolute Error

$$\text{MAE} = \frac{1}{n}\sum_i \left|\hat{Y}(x_i) - Y(x_i)\right|$$

MSE = Mean Squared Error

$$\text{MSE} = \frac{1}{n}\sum_i \left(\hat{Y}(x_i) - Y(x_i)\right)^2$$

[5]:

```
#######################################################################################
# separation of the data set into two subsets: (1) training of the ANN & (2) test of the ANN
# library used: pandas
data_train = dataCO2.sample(frac=0.8, axis=0, random_state=1)
data_test  = dataCO2.drop(data_train.index)

x_train = data_train.drop(['IL','cation','anion','x_CO2'],axis=1)
y_train = data_train['x_CO2']
x_test  = data_test.drop(['IL','cation','anion','x_CO2'],axis=1)
y_test  = data_test['x_CO2']

display(x_train,y_train)
display(x_test,y_test)

#######################################################################################
# ANN: 1 input layer (53 neurons) / 1 hidden layer (7 neurons with the reLu activation function) / 1 output layer (1 neuron with n
# library used: keras

def defANN(shape,acthL):

    model = keras.models.Sequential()
    model.add(keras.layers.Input(shape, name='iLayer'))
    model.add(keras.layers.Dense(7, activation=acthL, name='hLayer'))
    model.add(keras.layers.Dense(1, name='oLayer'))

    model.compile(optimizer = 'adam',
                  loss      = 'mse',
                  metrics   = ['mae', 'mse'] )
    return model
```

Version: Sunday, October 1, 2023

Catalyst selection for hydrogen production through WGS Reaction

General context: steam methane reforming

WGSR: $CO + H_2O \longrightarrow CO_2 + H_2$

steam methane reforming

+206,2 kJ·mol^{-1}

−41,1 kJ·mol^{-1}

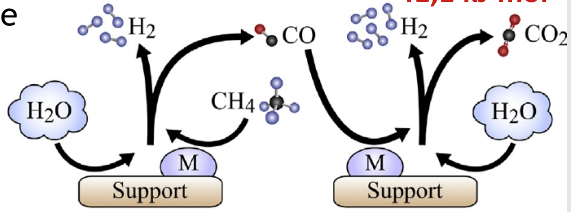
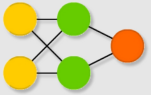
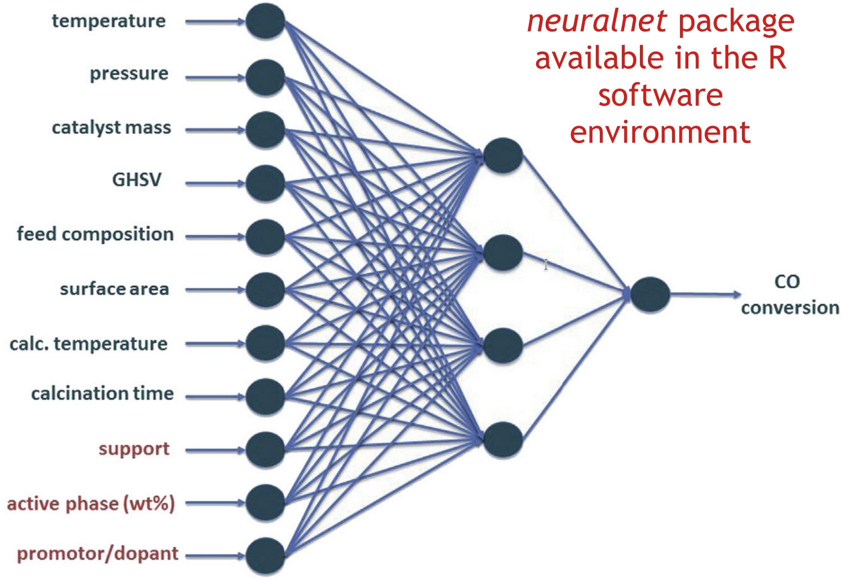


Fig. 1 — **Illustration showing the overall process of methane being converted to CO_2 and H_2 on a supported metal catalyst. Many catalysts follow this same general mechanism.**

Feed Forward (FF)





neuralnet package available in the R software environment

Fig. 2. Three-Layer Feedforward Neural Network employed in this work.

Version: Sunday, October 1, 2023

F. M. Cavalcanti, M. Schmal, R. Giudici & R. M. B. Alves (**2019**) A catalyst selection method for hydrogen production through Water-Gas Shift Reaction using artificial neural networks *J. Environ. Manage*. **237**, 585-594

Catalyst selection for hydrogen production through WGS Reaction

Data and main outcome of the trained ANN

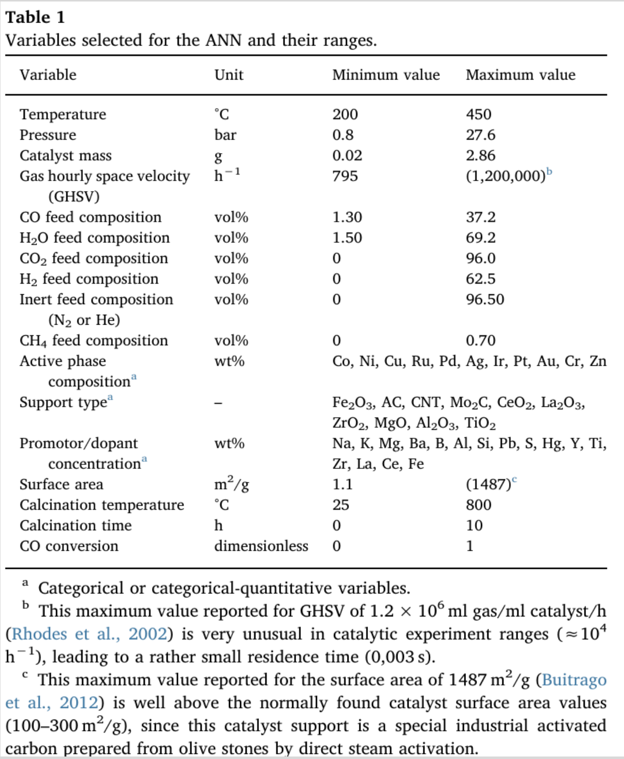


Table 1
Variables selected for the ANN and their ranges.

Variable	Unit	Minimum value	Maximum value
Temperature	°C	200	450
Pressure	bar	0.8	27.6
Catalyst mass	g	0.02	2.86
Gas hourly space velocity (GHSV)	h^{-1}	795	(1,200,000)[b]
CO feed composition	vol%	1.30	37.2
H_2O feed composition	vol%	1.50	69.2
CO_2 feed composition	vol%	0	96.0
H_2 feed composition	vol%	0	62.5
Inert feed composition (N_2 or He)	vol%	0	96.50
CH_4 feed composition	vol%	0	0.70
Active phase composition[a]	wt%	Co, Ni, Cu, Ru, Pd, Ag, Ir, Pt, Au, Cr, Zn	
Support type[a]	–	Fe_2O_3, AC, CNT, Mo_2C, CeO_2, La_2O_3, ZrO_2, MgO, Al_2O_3, TiO_2	
Promotor/dopant concentration[a]	wt%	Na, K, Mg, Ba, B, Al, Si, Pb, S, Hg, Y, Ti, Zr, La, Ce, Fe	
Surface area	m^2/g	1.1	(1487)[c]
Calcination temperature	°C	25	800
Calcination time	h	0	10
CO conversion	dimensionless	0	1

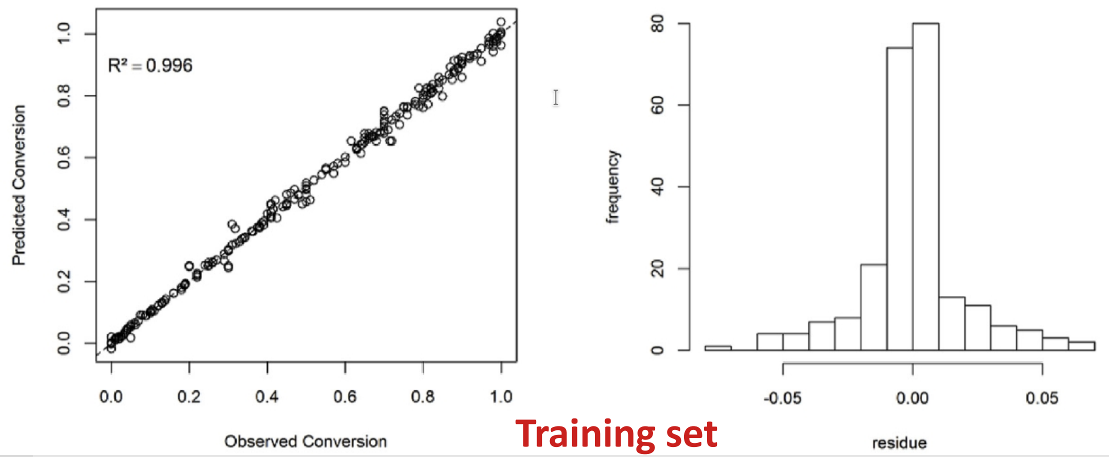
[a] Categorical or categorical-quantitative variables.

[b] This maximum value reported for GHSV of 1.2 × 10^6 ml gas/ml catalyst/h (Rhodes et al., 2002) is very unusual in catalytic experiment ranges (≈10^4 h^{-1}), leading to a rather small residence time (0,003 s).

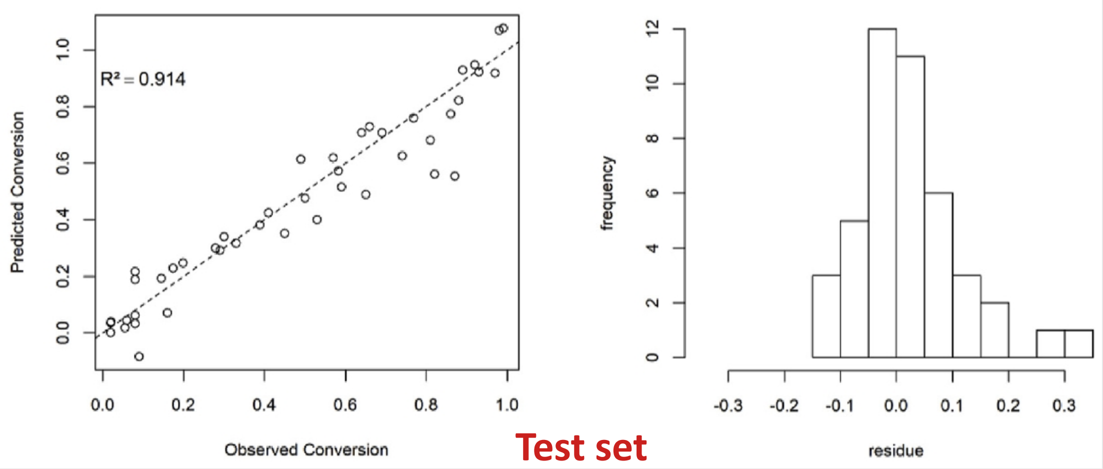
[c] This maximum value reported for the surface area of 1487 m^2/g (Buitrago et al., 2012) is well above the normally found catalyst surface area values (100–300 m^2/g), since this catalyst support is a special industrial activated carbon prepared from olive stones by direct steam activation.

Taken from: T. L. LeValley, A. R. Richard & M. Fan (**2014**)
The progress in water gas shift and steam reforming hydrogen production technologies – A review
Int. J. Hydrogen Energy **39**, 16983-17000

283 experimental data points



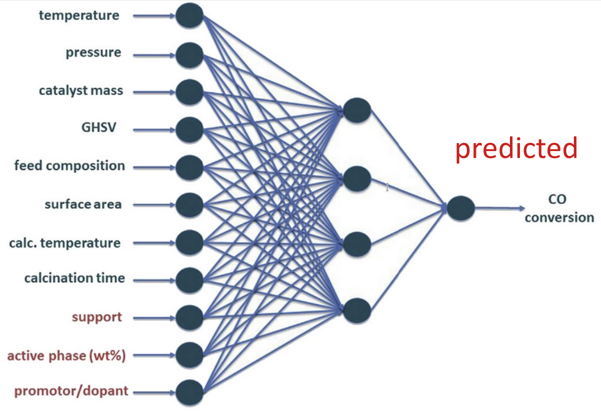
Training set



Test set

Catalyst selection for hydrogen production through WGS Reaction

Best conditions? 1. ANN at work



predicted

Fig. 2. Three-Layer Feedforward Neural Network employed in this work.

Pd and Co: inconsistencies in the ANN model

- lack of data for these conditions?
- or inconsistent kinetic data?

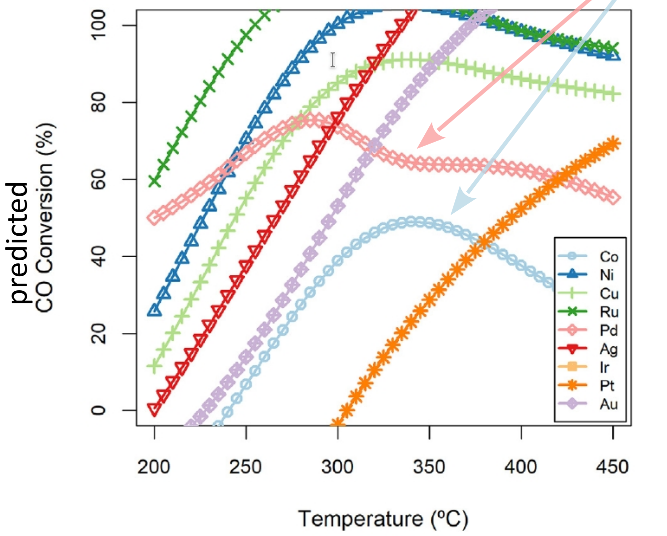


Fig. 8. CO conversion *versus* temperature for different active phases (NH = 12, Metal/CeO_2, Metal = 2 wt%, P = 1 bar, m_{cat} = 0.1 g, GHSV = 1000 h^{-1}, surface area = 100 m^2/g, T_{calc} = 300 °C, t_{calc} = 4 h, feed composition: 2% CO, 10% H_2O, 88% N_2).

Best metals = Ru, Ni, and Cu

Catalyst selection for hydrogen production through WGS Reaction

Best conditions? 2. Sensitivity analysis using ANN predictions

if…
- relationships between inputs and outputs are poorly understood
- poor or partial understanding of the driving forces and mechanisms

⇒ limits our confidence in the output of the model

⇒ sensitivity analysis provides a kind of “quality assurance”

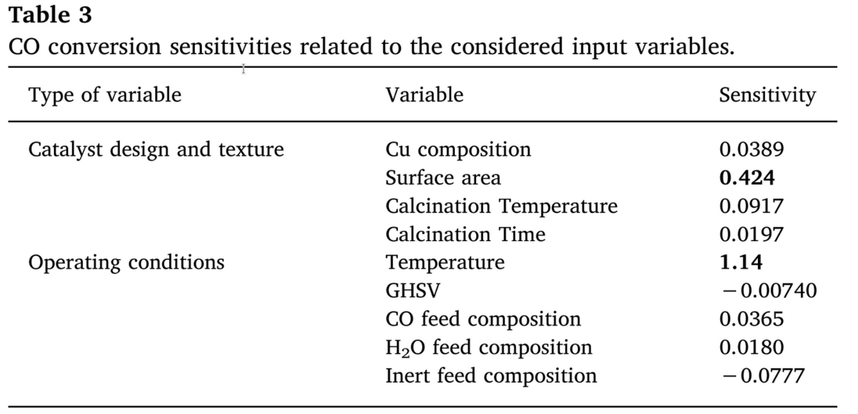


Table 3
CO conversion sensitivities related to the considered input variables.

Type of variable	Variable	Sensitivity
Catalyst design and texture	Cu composition	0.0389
	Surface area	**0.424**
	Calcination Temperature	0.0917
	Calcination Time	0.0197
Operating conditions	Temperature	**1.14**
	GHSV	−0.00740
	CO feed composition	0.0365
	H_2O feed composition	0.0180
	Inert feed composition	−0.0777

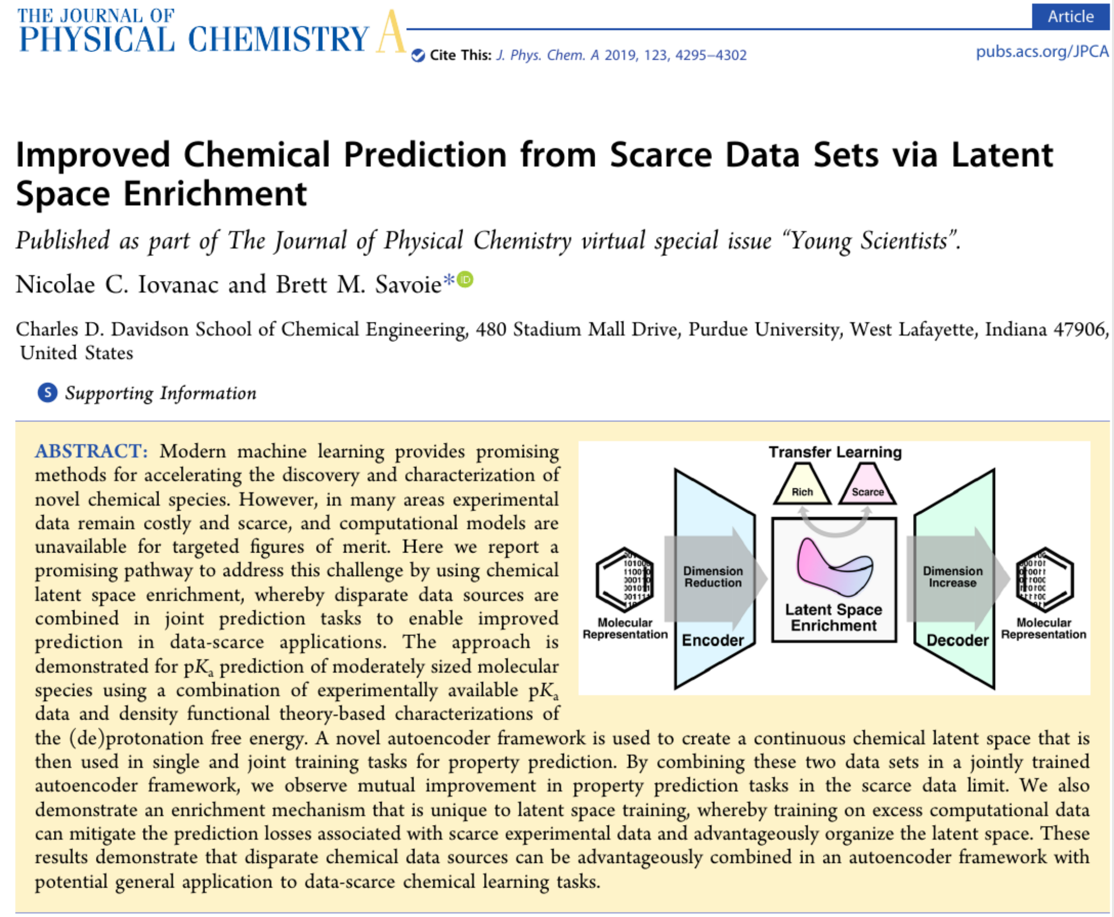
Is such analysis <u>really</u> necessary to demonstrate that surface area and temperature are important to the development of industrial catalysts for the WGS reaction?

→ actually, sensitivity analysis rather orders by importance the strength and relevance of the inputs in determining the variation in the output

Best conditions and evaluation of the most relevant variables? → “standard” statistics is needed

Version: Sunday, October 1, 2023

Improvement of chemical predictions in data-scarce applications



THE JOURNAL OF PHYSICAL CHEMISTRY A

Article

Cite This: *J. Phys. Chem. A* 2019, 123, 4295−4302

pubs.acs.org/JPCA

Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment

Published as part of The Journal of Physical Chemistry virtual special issue "Young Scientists".

Nicolae C. Iovanac and Brett M. Savoie*

Charles D. Davidson School of Chemical Engineering, 480 Stadium Mall Drive, Purdue University, West Lafayette, Indiana 47906, United States

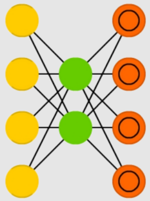
Supporting Information

ABSTRACT: Modern machine learning provides promising methods for accelerating the discovery and characterization of novel chemical species. However, in many areas experimental data remain costly and scarce, and computational models are unavailable for targeted figures of merit. Here we report a promising pathway to address this challenge by using chemical latent space enrichment, whereby disparate data sources are combined in joint prediction tasks to enable improved prediction in data-scarce applications. The approach is demonstrated for pK_a prediction of moderately sized molecular species using a combination of experimentally available pK_a data and density functional theory-based characterizations of the (de)protonation free energy. A novel autoencoder framework is used to create a continuous chemical latent space that is then used in single and joint training tasks for property prediction. By combining these two data sets in a jointly trained autoencoder framework, we observe mutual improvement in property prediction tasks in the scarce data limit. We also demonstrate an enrichment mechanism that is unique to latent space training, whereby training on excess computational data can mitigate the prediction losses associated with scarce experimental data and advantageously organize the latent space. These results demonstrate that disparate chemical data sources can be advantageously combined in an autoencoder framework with potential general application to data-scarce chemical learning tasks.

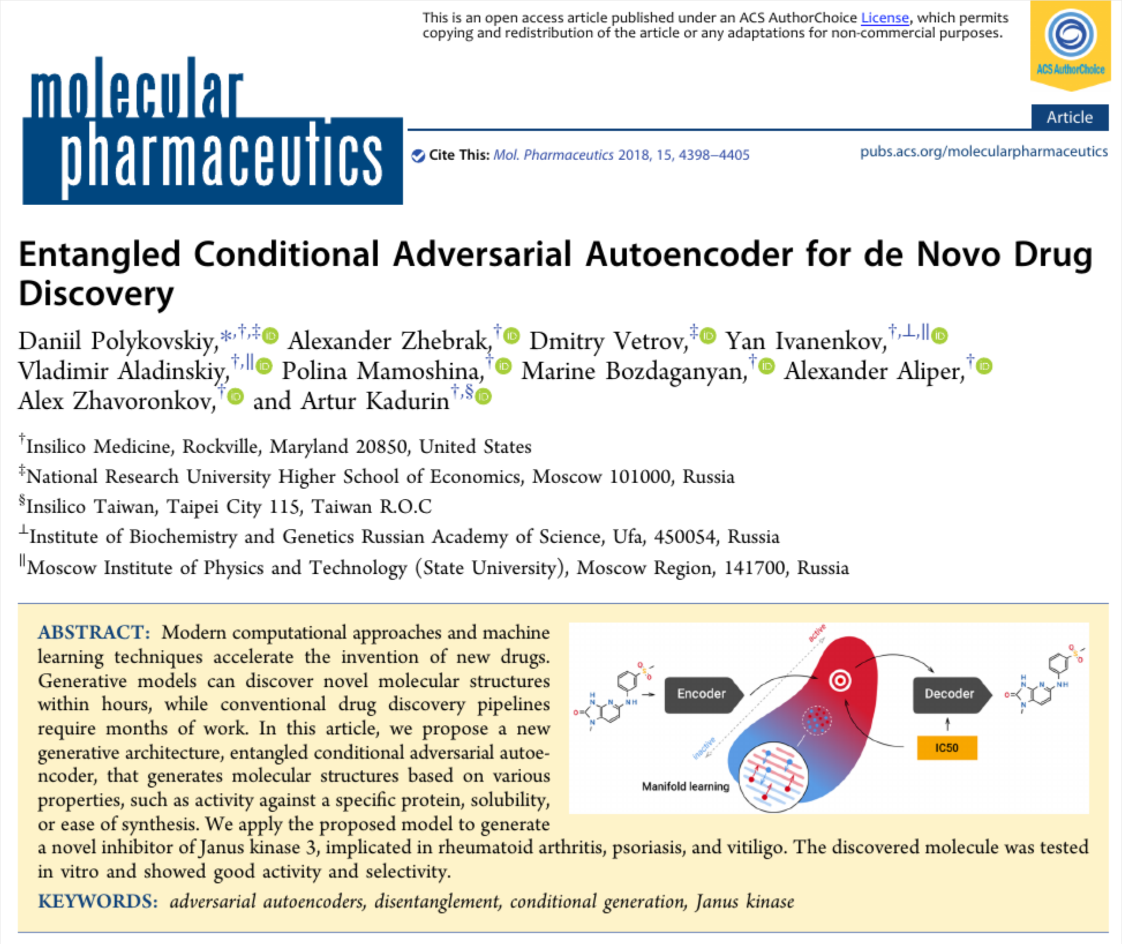
scarce experimental data are supplemented in learning tasks with more abundant computational properties

exp pK_a ↔ DFT deprotonation energies

Auto Encoder (AE)



Version: Sunday, October 1, 2023



molecular pharmaceutics

Article

Cite This: Mol. Pharmaceutics 2018, 15, 4398–4405

pubs.acs.org/molecularpharmaceutics

Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery

Daniil Polykovskiy,*[†,‡] Alexander Zhebrak,[†] Dmitry Vetrov,[‡] Yan Ivanenkov,[†,⊥,∥] Vladimir Aladinskiy,[†,∥] Polina Mamoshina,[†] Marine Bozdaganyan,[†] Alexander Aliper,[†] Alex Zhavoronkov,[†] and Artur Kadurin[†,§]

[†]Insilico Medicine, Rockville, Maryland 20850, United States

[‡]National Research University Higher School of Economics, Moscow 101000, Russia

[§]Insilico Taiwan, Taipei City 115, Taiwan R.O.C

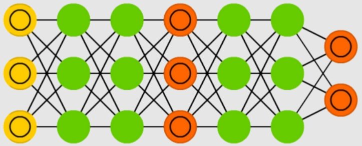
[⊥]Institute of Biochemistry and Genetics Russian Academy of Science, Ufa, 450054, Russia

[∥]Moscow Institute of Physics and Technology (State University), Moscow Region, 141700, Russia

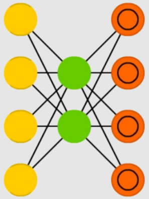
ABSTRACT: Modern computational approaches and machine learning techniques accelerate the invention of new drugs. Generative models can discover novel molecular structures within hours, while conventional drug discovery pipelines require months of work. In this article, we propose a new generative architecture, entangled conditional adversarial autoencoder, that generates molecular structures based on various properties, such as activity against a specific protein, solubility, or ease of synthesis. We apply the proposed model to generate a novel inhibitor of Janus kinase 3, implicated in rheumatoid arthritis, psoriasis, and vitiligo. The discovered molecule was tested in vitro and showed good activity and selectivity.

KEYWORDS: *adversarial autoencoders, disentanglement, conditional generation, Janus kinase*

Generative Adversarial Network (GAN)



Auto Encoder (AE)



ML in chemistry (MLChem)

ML methods are becoming less understood while they are also more regularly used as black box tools.

Many publications show inadequate technical expertise in ML (e.g. inappropriate splitting of training, testing, and validation sets)

It can be difficult to compare different ML methods and know which is the best for a particular application or whether ML should even be used at all

Data quality and context are often missing from ML modeling, and data sets need to be made freely available and clearly explained

J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller & A. Tkatchenko (2021)
Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev*. 121, 9816–9872