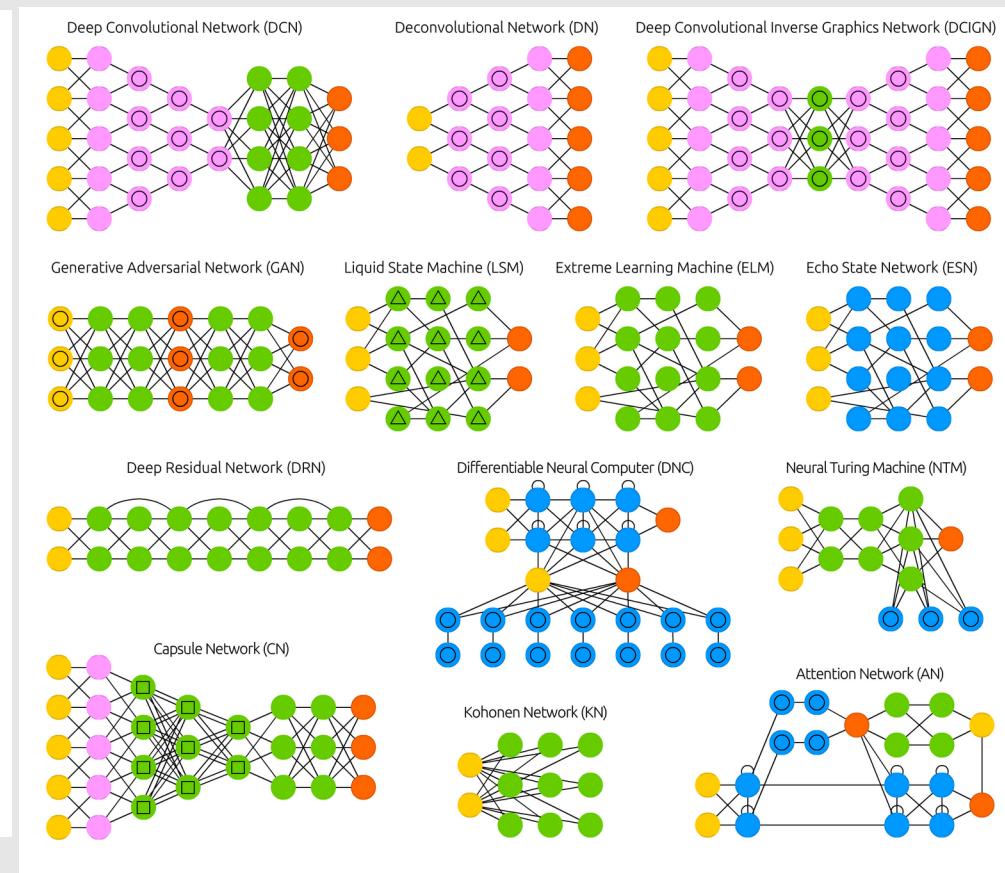
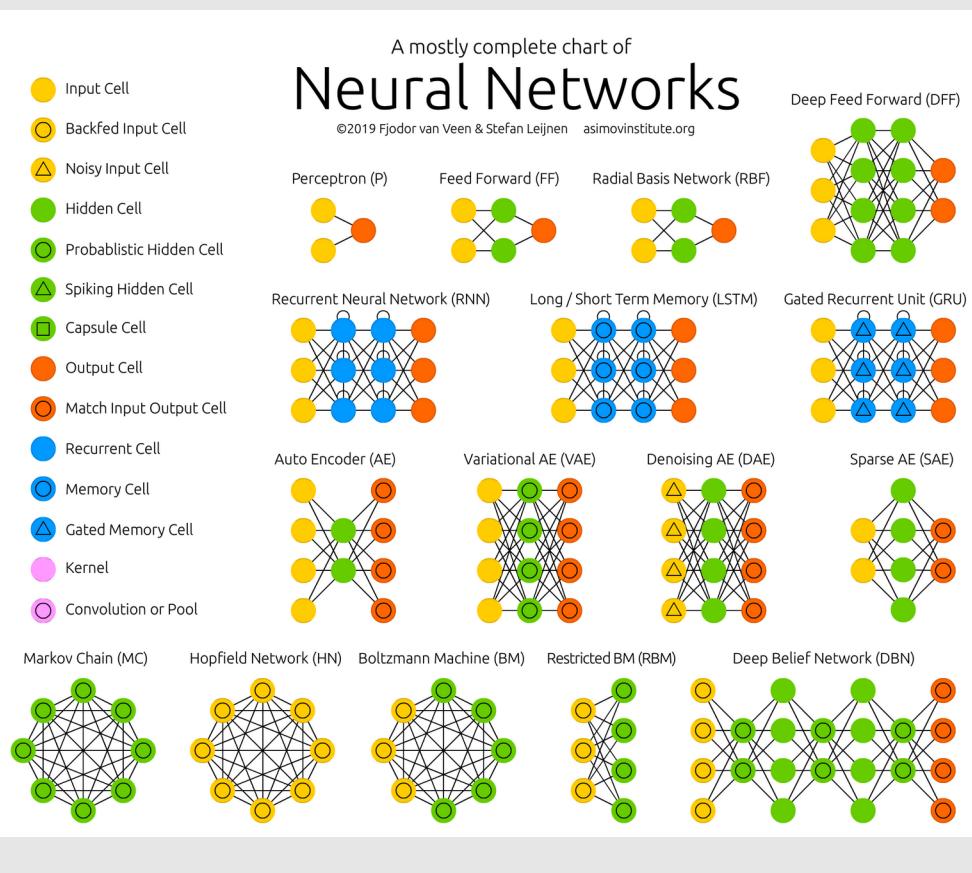


Introduction to Machine Learning for Chemists:
Visualization, Data Processing, Analysis,
Molecular Design

Bibliography: review articles



Review articles

SPECIAL SECTION FRONTIERS IN COMPUTATION

REVIEW

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

Benjamin Sanchez-Lengeling^a and Alán Aspuru-Guzik^{a,b,c,d}

The discovery of new materials can be enormous scientific and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for addressing inverse design, which aims to discover materials with specific properties. We focus on generative models, which have emerged from the rapidly growing field of artificial intelligence, mostly from the outset of machine learning. We describe how generative models have been applied to inverse design. These deep generative models 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, synthesis of organic molecules, retrosynthetic analysis, photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

Many of the challenges of the 21st century (*i.e.*, from personalized health care to energy generation and storage, share a common theme: they are all limited by the physics and chemistry of a material, such as the relativity of materials based on their atomic structure or the generation of solar energy (*i.e.*, devices). Despite these important discoveries area by example, valuable robots were prepared in the past few days. The synthesis of compounds, based on the observation that heating with a flame can increase the durability of rubber's durability. At the molecular level, individual polymer chains cross-linked, forming bridges between the atoms of different properties (*i.e.*, Other notable examples in this regard include the synthesis of penicillin, Perkin's mace, and penicillins. Furthermore, these materials come from common chemical compounds found in nature. Potential drugs either were prepared by synthesis in a chemist laboratory or isolated from natural sources, such as bacteria, fungi. For example, up until 2014, 62% of all molecular drug drugs were natural products or semi-synthetic.

In the future, distinctive advances in the design of materials will be achieved by exploring regions of the set of all possible molecular systems. In this context, known as chemical space (6, 7). One of the largest collections of molecules, the chemical space project (8), has mapped 100+ billion molecules that contain at most 17 heavy atoms. For pharmaceutically relevant small molecules, it is on the order of 10^{10} . Addressing the complexity of the hierarchy of scale subproblems in this space is a major challenge. This work explores the application of machine learning to the discovery of a material that is suitable for a specific purpose. The process of discovery in a laboratory is a commercial problem. The steps of the process are: (i) generate a new or improved material model; (ii) use the model to predict the properties of the material; (iii) synthesize the material; (iv) characterize the synthesized material; (v) use the characterization to repeat, improve, and refine the discovery. Each step can take up to several years.

In the era of matter engineering, scientists seek to accelerate these cycles, reducing the time required for the discovery of a material that is suitable for a specific purpose. The current paradigm is shown in Figure 1A. It consists of a sequence of steps: (i) material concept, (ii) device construction, (iii) testing and characterization, (iv) scaling and manufacturing. The process is iterative, with feedback loops at each stage. The final product is a device that is stable, reliable, and efficient.

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 (9). A closed-loop paradigm is outlined at right. Closing the loop requires incorporating inverse design, smart software (10), AI/machine learning, embedded systems, and robotics (11) into an integrated ecosystem.

^aDepartment of Chemistry and Chemical Biology, Howard University, Washington, DC 20059, USA.
^bDepartment of Chemistry and Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309, USA.
^cVenter Institute for Artificial Intelligence, Toronto, ON M3J 2M1, Canada.
^dQuantum Machine Learning Research Institute (QML), Sunny Hill, Toronto, Ontario M3J 2M2, Canada.

*Corresponding author. Email: asparu@mit.edu

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

REVIEW

<https://doi.org/10.1038/s41566-018-0337-2>

Machine learning for molecular and materials science

Keith T. Butler¹, Daniel W. Davies², Hugh Cartwright³, Olexandr Isayev^{4,*} & Aron Walsh^{5,*}

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

CHEMICAL REVIEWS

<https://pubs.acs.org/doi/pdf/10.1021/cr000059k>

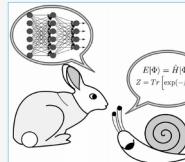
Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems

John A. Keith,¹* Valentin Vassilev-Galindo,² Bingqiang Cheng,³ Stefan Chmiela,⁴ Michael Gastegger,⁵ Klaus-Robert Müller,⁶* and Alexandre Tkatchenko⁷

ACCESS | Metrics & More | Article Recommendations

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.

Cite This: <https://doi.org/10.1021/cr000059k> Read Online





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



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.1c00467>

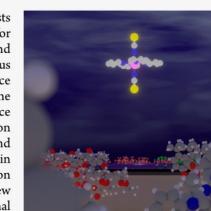
Read Online

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

Review articles

ACS Catalysis Review [Cite This: ACS Catal. 2020, 10, 2260–2297](#) [pubs.acs.org/acscatalysis](#)

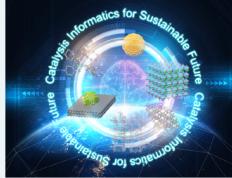
Machine Learning for Catalysis Informatics: Recent Applications and Prospects

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 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 Tao¹, Tianyi Wu¹, Matteo Aldeghi^{1,2,3}, Tony C. Wu^{1,3}, Alán Aspuru-Guzik^{1,2,3,4,5} 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 pharmaceuticals, 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

AI in chemistry (AIChem)

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

Calculation of energies & forces (MD / geometry optimization)

Version: Sunday, October 1, 2023

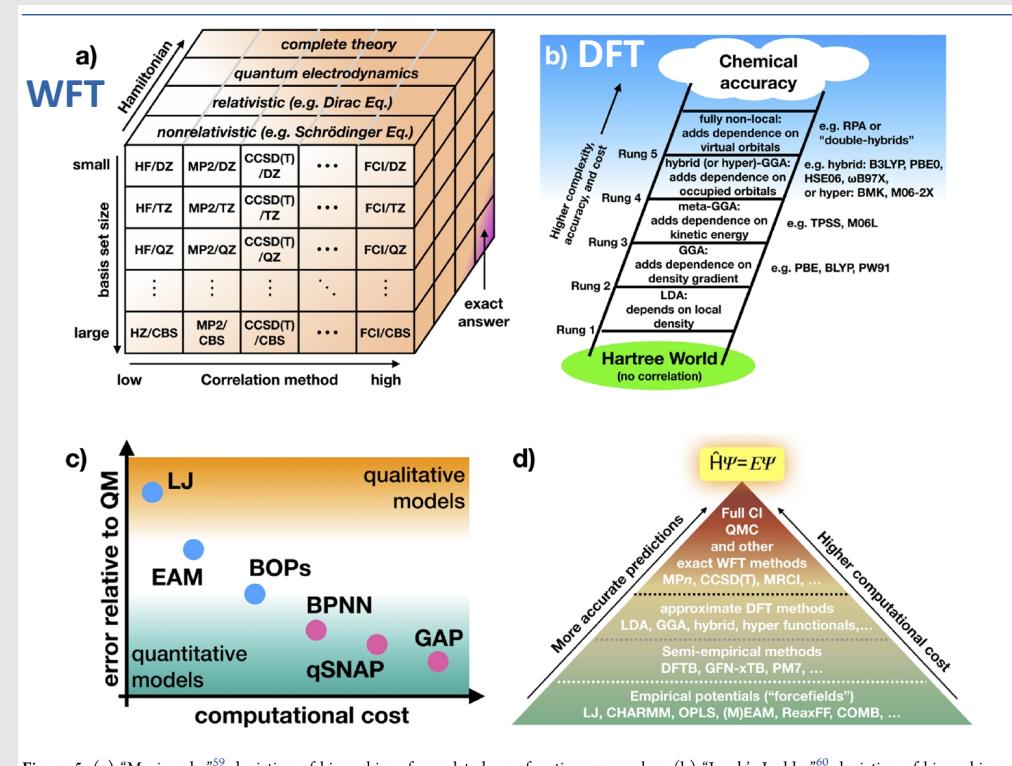


Figure 5. (a) "Magic cube"⁵⁹ depiction of hierarchies of correlated wavefunction approaches. (b) "Jacob's Ladder"⁶⁰ depiction of hierarchies of Kohn–Sham density functional theory (DFT) approaches. (c) Hierarchies of atomistic potentials. (d) Overall hierarchies in predictive atomic scale modeling methods.

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



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

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



quality
diversity
volume

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



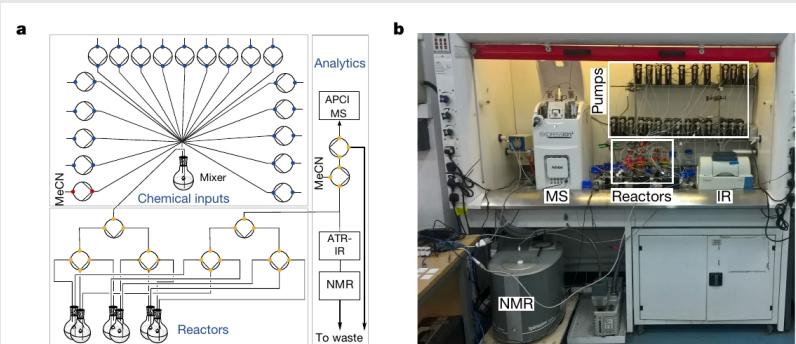
Control of an organic synthesis robot

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

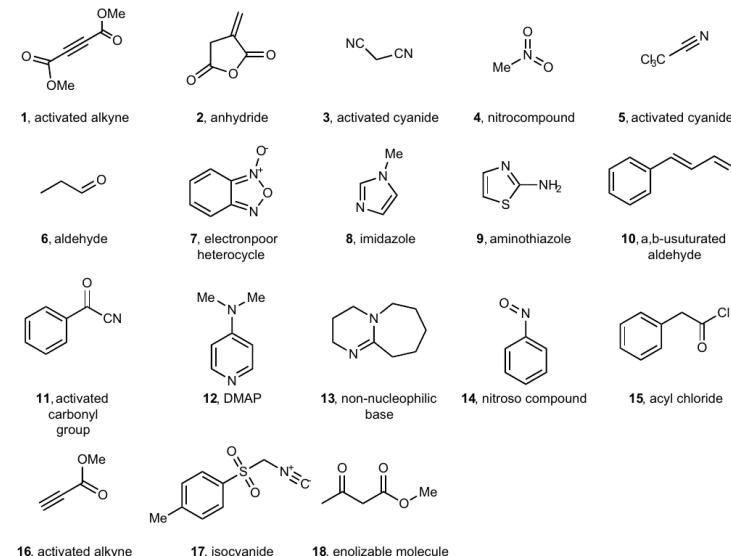
decision making by ML



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.

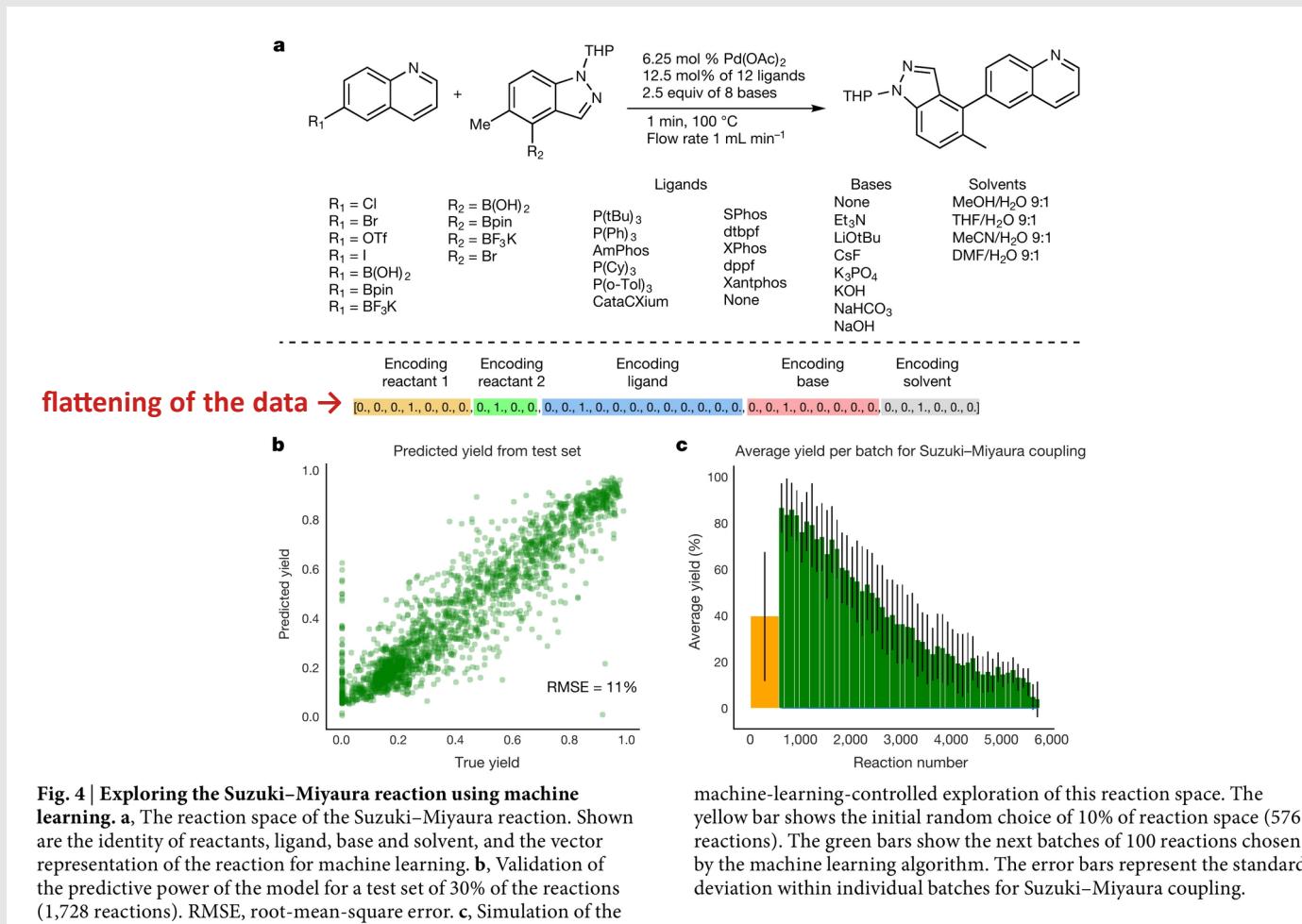
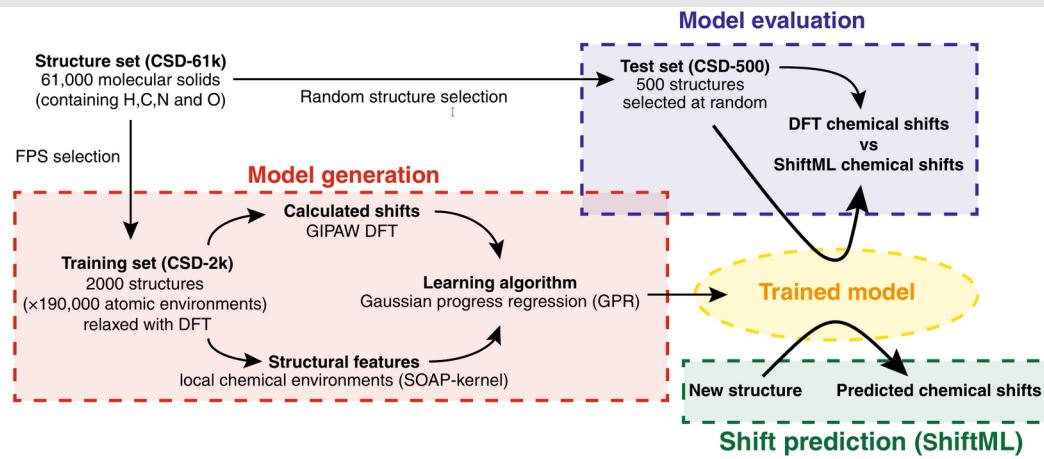


Fig. 4 | Exploring the Suzuki–Miyaura reaction using machine learning. **a**, The reaction space of the Suzuki–Miyaura reaction. Shown are the identity of reactants, ligand, base and solvent, and the vector representation of the reaction for machine learning. **b**, Validation of the predictive power of the model for a test set of 30% of the reactions (1,728 reactions). RMSE, root-mean-square error. **c**, Simulation of the

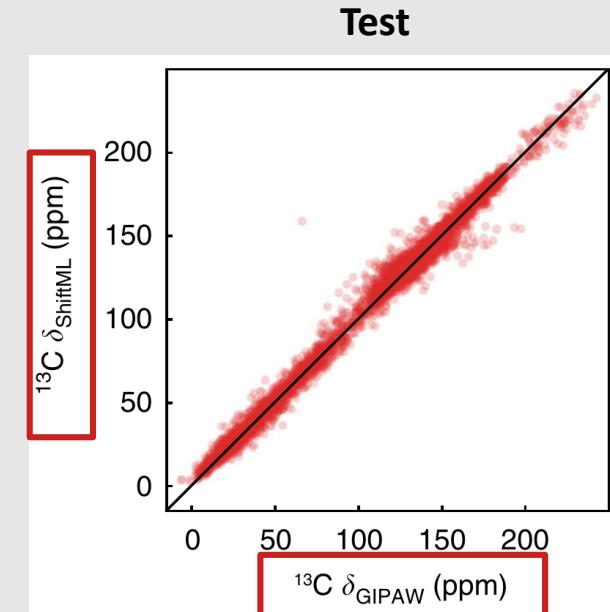
machine-learning-controlled exploration of this reaction space. The yellow bar shows the initial random choice of 10% of reaction space (576 reactions). The green bars show the next batches of 100 reactions chosen by the machine learning algorithm. The error bars represent the standard deviation within individual batches for Suzuki–Miyaura coupling.



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



predicted values as a function of the actual values



Deep Convolutional Network (DCN)

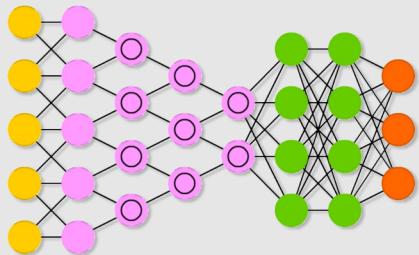


Image recognition

Molecular Dynamics + STEM image simulation

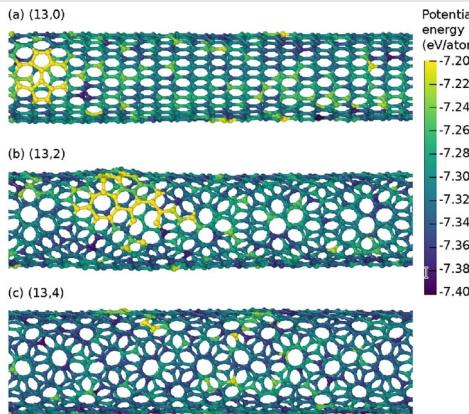


Fig. 1. Some examples of final snapshots from the molecular dynamics simulations. Defects (higher potential energy, appear in yellow) have been introduced by high-temperature annealing. The visualizations are generated using the OVITO software [51]. (A colour version of this figure can be viewed online.)

Dr. Probe - Software

DOWNLOAD DOCUMENTATION EXAMPLES DEVELOP

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

CPU Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.0GHz

NVIDIA CUDA

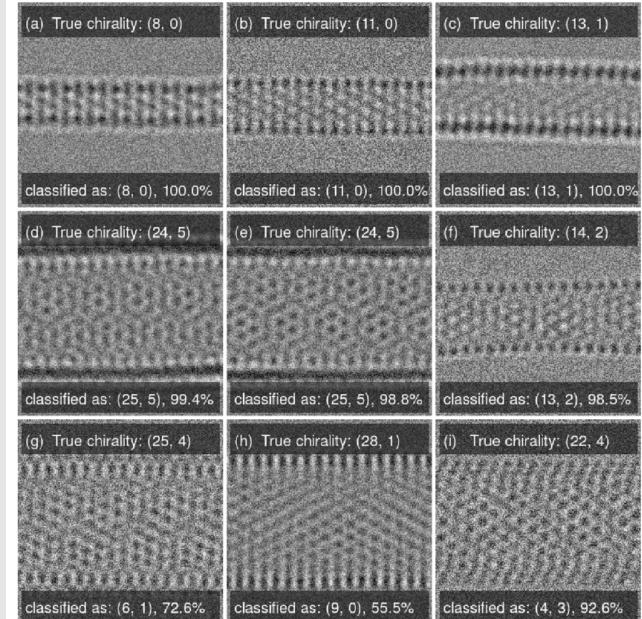
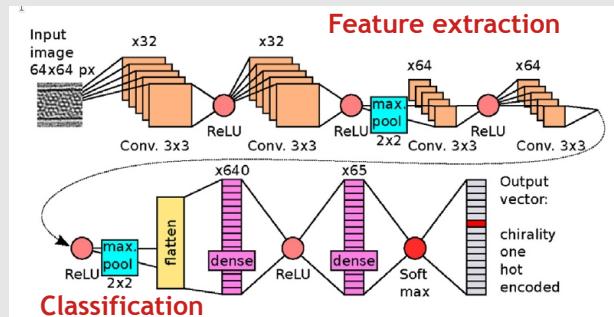
Hybrid parallel CPU & GPU computing



5000 images per chirality
~ $1.3 \cdot 10^6$ images in total

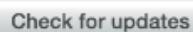
G. D. Förster, A. Castan, A. Loiseau, J. Nelayah, D. Alloyeau, F. Fossard, C. Bichara & H. Amara (2020)

A deep learning approach for determining the chiral indices of carbon nanotubes from high-resolution transmission electron microscopy images
Carbon **169**, 465-474



Chemical
ScienceReaction outcome prediction:
Prediction of enantiomeric
excess

EDGE ARTICLE

[View Article Online](#)[View Journal](#) | [View Issue](#)Cite this: *Chem. Sci.*, 2021, 12, 6879

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

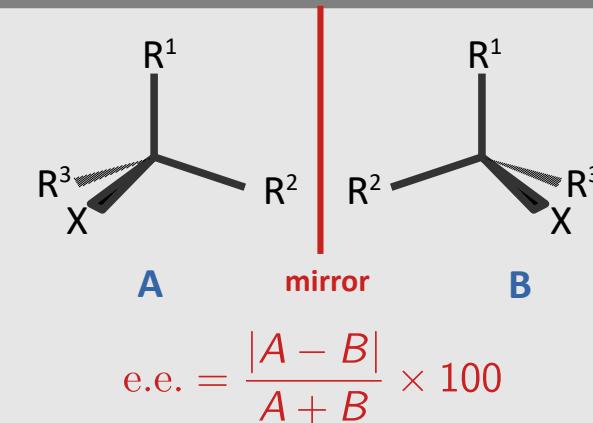
Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts†

Simone Gallarati, ^a Raimon Fabregat, ^a Rubén Laplaza, ^{ab} Sinjini Bhattacharjee, ^{ac} Matthew D. Wodrich ^{ab} and Clemence Corminboeuf ^{*abd}

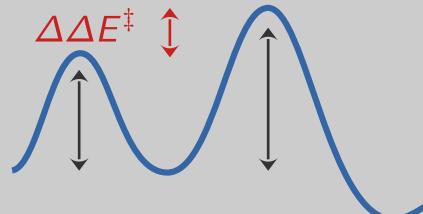
Hundreds of catalytic methods are developed each year to meet the demand for high-purity chiral compounds. The computational design of enantioselective organocatalysts remains a significant challenge, as catalysts are typically discovered through experimental screening. Recent advances in combining quantum 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/disc00482d

rsc.li/chemical-science

Computational chemistry

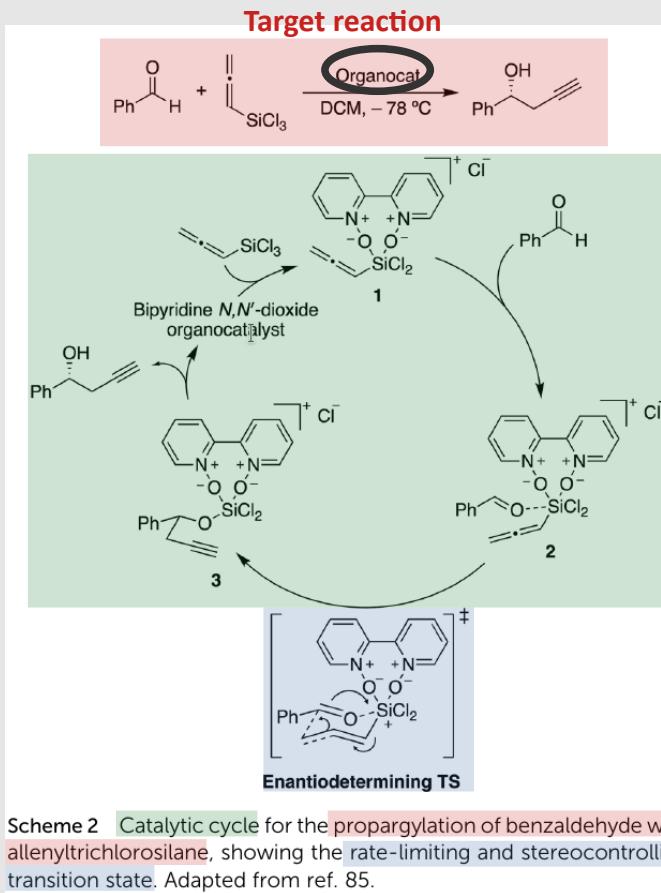


$$\text{e.e.} = \frac{1 - e^{\Delta\Delta E^\ddagger / RT}}{1 + e^{\Delta\Delta E^\ddagger / RT}} \times 100$$

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

Reaction outcome prediction. Prediction of enantiomeric excess

Version : Sunday, October 1, 2023



Chemical context

(1) Training

Database of intermediates

Int 2 Int 3

754 pairs of DFT-optimized intermediates

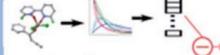
Generate reaction representation

a) Int 2



Molecular representation (e.g., SLATM)

b) Int 3 - Int 2

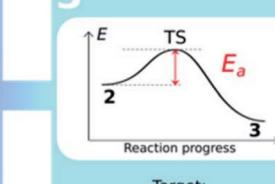


Molecular representation (e.g., SLATM)

KRR (not an ANN)

Map representation to target property with KRR

2 3



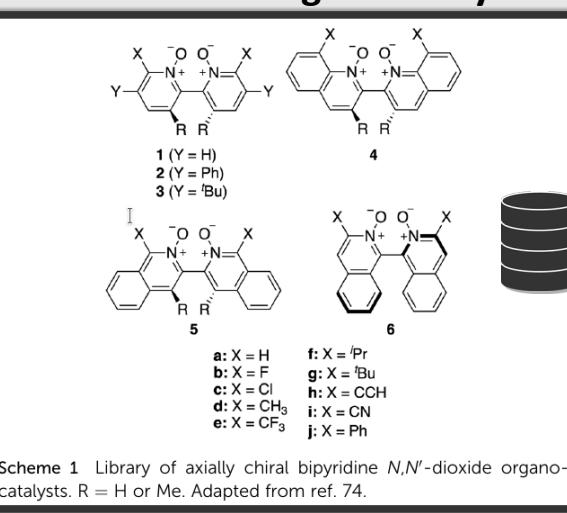
Hyperparameter optimization and cross-validation

■ Train ■ Test

100 cross-validation train/test splits (678/76)

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

76 Lewis base organocatalysts



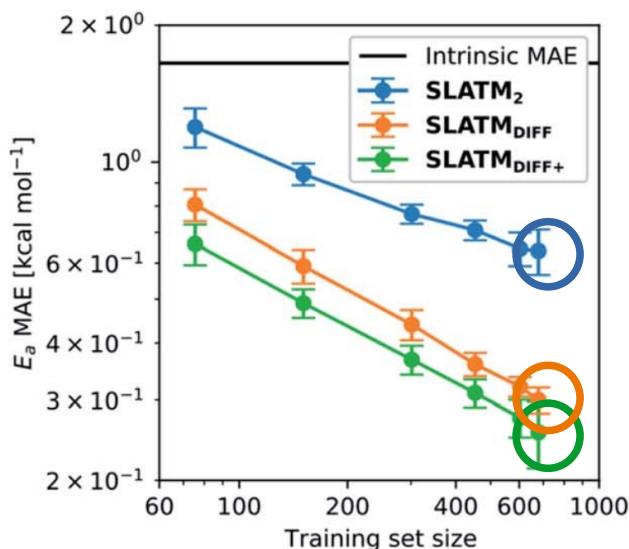
760 DFT TS geometries and activation energies

Main results

Mean Absolute Error

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

SLATM = structural descriptors



0.54 ± 0.06 kcal mol⁻¹
(insufficient accuracy)

0.31 ± 0.20 kcal mol⁻¹

0.25 ± 0.40 kcal mol⁻¹

Learning curves:
MAE in test sets predictions of E_a

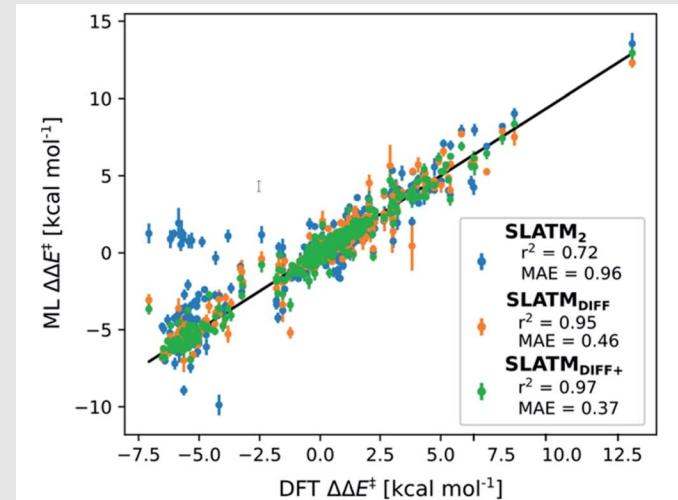


Fig. 2 Predictions of $\Delta\Delta E^\ddagger$ vs. DFT reference for the three approaches discussed. Mean Absolute Errors (MAE) are reported in kcal mol^{-1} . These predictions are obtained by averaging the predictions obtained from the cross-validation scheme with 100 different random train/test splits.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

Phase behavior of ionic liquids

Prediction of CO₂ solubility in ionic liquids



Prediction of CO₂ solubility in ionic liquids using machine learning methods

Zhen Song^a, Huawei 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-cyanomethane [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)**

Phase behavior of ionic liquids

Version: Sunday, October 1, 2023

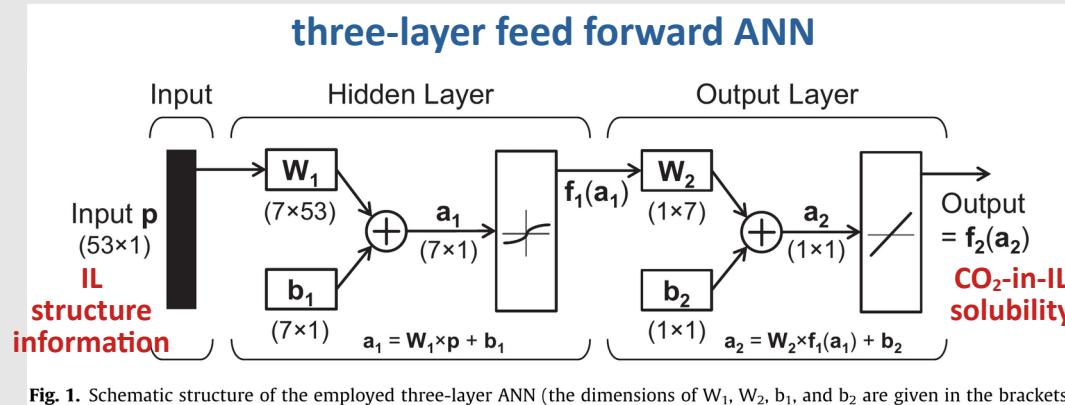
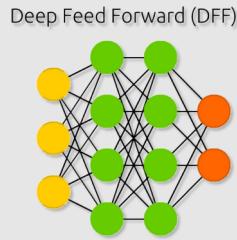


Fig. 1. Schematic structure of the employed three-layer ANN (the dimensions of W_1 , W_2 , b_1 , and b_2 are given in the brackets).

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	0	1	0	0	0	0	1	0	0	0	0	0	0.583	
298.25	0.4624	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0.014334

data organized as vectors

IL structure information = occurrences of functional groups in the IL

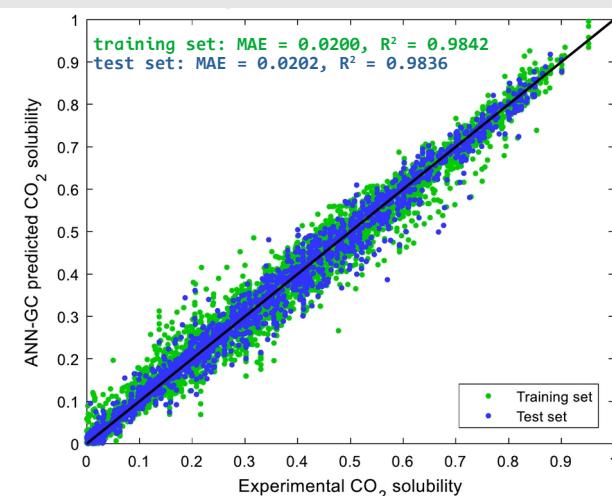


Fig. 2. Comparison between the experimental and ANN-GC predicted CO₂ solubility.



Phase behavior of ionic liquids

Version : Sunday, October 1, 2023



Two notebooks



Python in the [Physical] Chemistry Lab

[PytChem]

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

New Tab | High-through... | Q+ JupyterLab | bobby laponine - Recherche Google | DS4B-CO₂_solubility-ANN.ipynb | http://localhost:8889/lab/tree/DS4B-CO₂_solubility-ANN.ipynb

File Edit View Run Kernel Tabs Settings Help

Launcher DS4B-CO₂_solubility-ANN Python 3 (ipykernel)

evaluate the predictability 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_{CO_2} values (noted Y and \hat{Y} in equations below)

MAE = Mean Absolute Error

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

MSE = Mean Squared Error

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

[5]:

```
#####
# separation of the data set into two subsets: (1) training of the ANN & (2) test of the ANN
# library used: keras
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 (55 neurons) / 1 hidden layer (7 neurons with the relu activation function) / 1 output layer (1 neuron with no activation)
# library used: keras

def dANN(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
```

Mode: Command | Ln 1, Col 1 | English (United States) | DS4B-CO₂_solubility-ANN.ipynb

Catalyst selection for hydrogen production through WGS Reaction

General context: steam methane reforming



steam methane
reforming
 $+206,2 \text{ kJ}\cdot\text{mol}^{-1}$

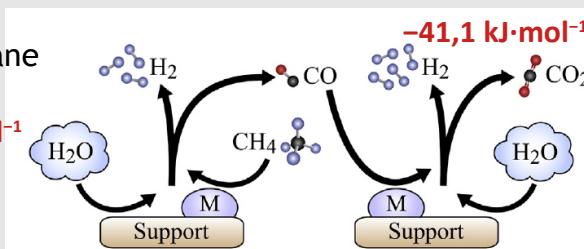


Fig. 1 – Illustration showing the overall process of methane being converted to CO₂ and H₂ on a supported metal catalyst. Many catalysts follow this same general mechanism.

Feed Forward (FF)

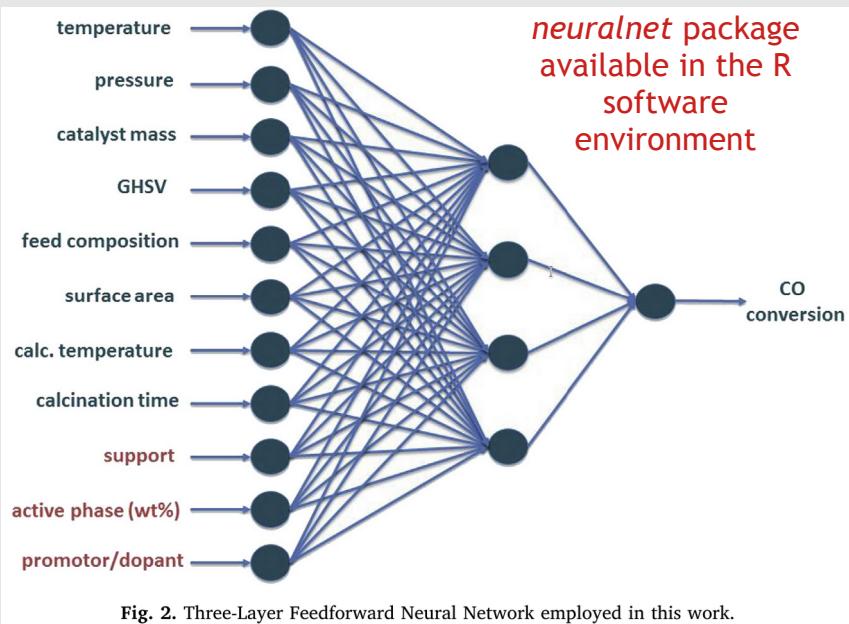


Fig. 2. Three-Layer Feedforward Neural Network employed in this work.

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 ₂ O 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 ₂ O ₃ , AC, CNT, Mo ₂ C, CeO ₂ , La ₂ O ₃ , ZrO ₂ , MgO, Al ₂ O 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 ($\approx 10^4$ h⁻¹), leading to a rather small residence time (0.003 s).

^c This maximum value reported for the surface area of 1487 m²/g (Buitrago et al., 2012) is well above the normally found catalyst surface area values (100–300 m²/g), since this catalyst support is a special industrial activated carbon prepared from olive stones by direct steam activation.

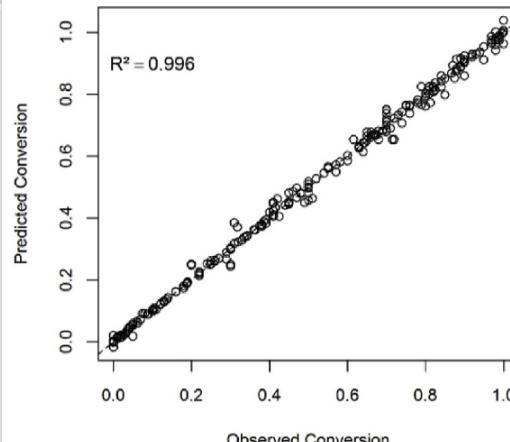
Taken from: T. L. LeValley, A. R. Richard & M. Fan (2014)

The progress in water gas shift and steam reforming hydrogen production technologies – A review

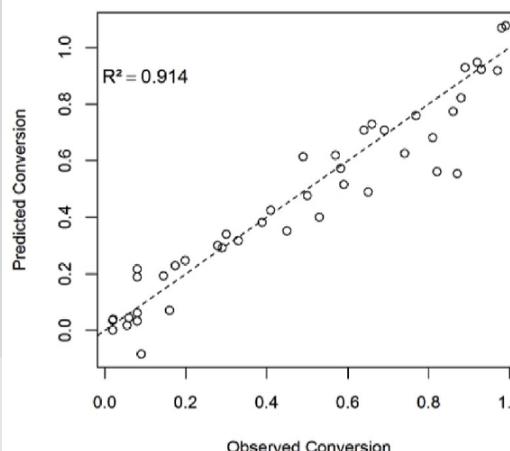
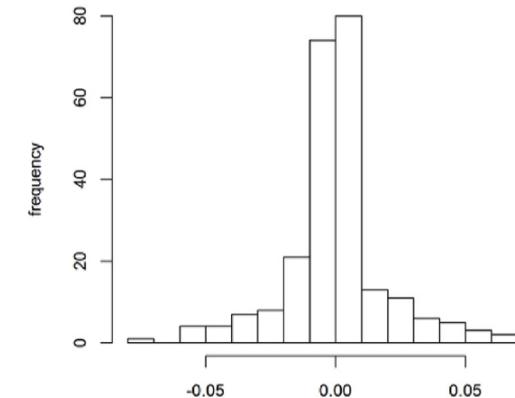
Int. J. Hydrogen Energy 39, 16983-17000



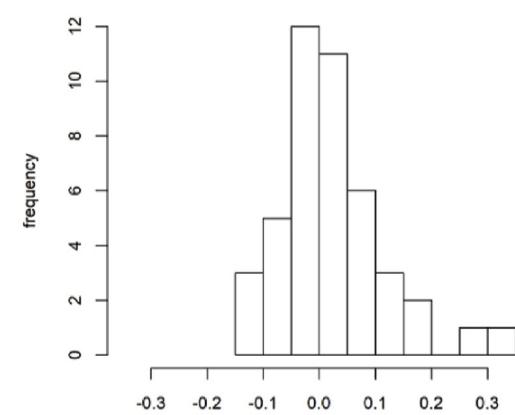
283
experimental
data points



Training set

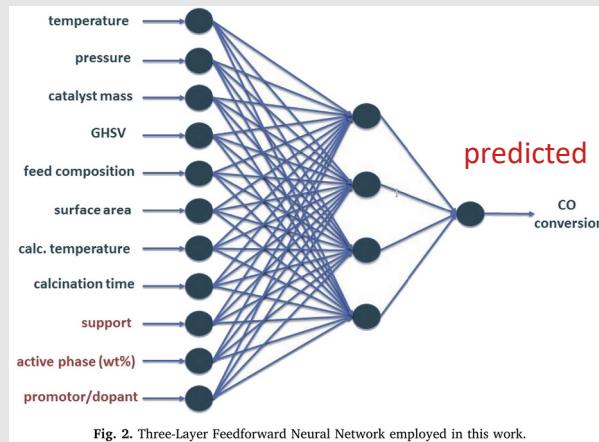


Test set



Catalyst selection for hydrogen production through WGS Reaction

Best conditions? 1. ANN at work



Pd and Co: inconsistencies in the ANN model
- lack of data for these conditions?
- or inconsistent kinetic data?

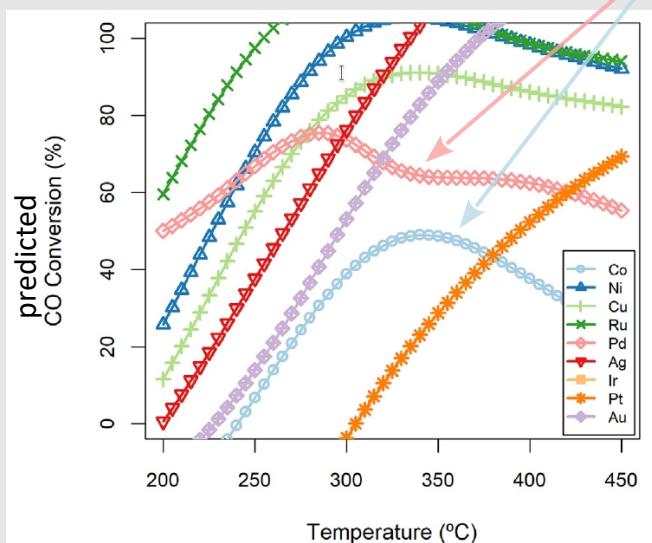


Fig. 8. CO conversion versus temperature for different active phases ($\text{NH}_3 = 12$, Metal/ CeO_2 , Metal = 2 wt%, $P = 1 \text{ bar}$, $m_{\text{cat}} = 0.1 \text{ g}$, GHSV = 1000 h^{-1} , surface area = $100 \text{ m}^2/\text{g}$, $T_{\text{calc}} = 300^\circ\text{C}$, $t_{\text{calc}} = 4 \text{ h}$, feed composition: 2% CO, 10% H_2O , 88% N_2).

Best metals = Ru, Ni, and Cu

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 ₂ O feed composition	0.0180
	Inert feed composition	-0.0777

Is such analysis really 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

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