Intelligent Pharmacy 4 (2026) 45–66

Contents lists available at ScienceDirect

Intelligent Pharmacy

journal homepage: www.keaipublishing.com/en/journals/intelligent-pharmacy

Review article

Review of the latest progress of AI and machine learning methods in the free energy kinetics estimation and synthesis analysis for organic chemistry applications

Chengze Yang

Department of Molecular Engineering, University of Chicago, Chicago, IL, USA

A B S T R A C T

The accurate prediction of free energy, kinetics, and reaction outcomes in organic chemistry has long presented significant computational challenges, which are often limited by the prohibitive cost of high-precision ab initio methods for complex systems. [1-2] Artificial intelligence and machine learning have transformed computational chemistry through a data-driven approach in property prediction, kinetics, and synthetic design, offering an alternative option for addressing this challenge. This review examines the latest advancements and the potential of artificial intelligence (AI) and machine learning (ML) to overcome these limitations. We highlight groundbreaking progress across several key domains: Dataset-based machine learning and hybrid quantum mechanical/machine learning models that achieve superior accuracy in free energy and kinetics predictions with reduced computational costs. Machine learning models that enable rapid pK_a predictions across a wide range of diverse solvents. The integration of thermodynamic principles into a machine learning model for accurate and consistent macro-micro pK_a prediction. Graph-convolutional neural networks that demonstrate high accuracy in reaction outcome prediction with interpretable mechanisms. Neural-symbolic frameworks and Monte Carlo Tree Search (MCTS) integrated with deep neural networks that revolutionize retrosynthetic planning, generating expert-quality routes at unprecedented speeds. A machine learning model based on molecular orbital reaction theory for organic reaction outcome prediction with remarkable accuracy and generalizability. Finally, hierarchical neural networks that predict comprehensive reaction conditions interdependently with exceptional speed. These advancements in a data-driven approach enhance precision, efficiency, and scalability, but challenges such as data quality, stereochemical prediction, and explicit mechanistic incorporation persist. The convergence of these AI/ML capabilities is paving the way for fully automated chemical discovery, addressing critical global challenges in medicine, materials, and energy.[3,4]

1. Introduction

1.1. The challenge faced by computational chemistry

The accurate prediction of free energy, kinetics, and reaction outcomes stands as a cornerstone in the field of organic chemistry, being indispensable for advancing rational drug design, catalyst development, and the optimization of synthetic routes.[1–6] Computational chemistry has historically provided a robust framework for addressing these challenges. It employs sophisticated computer programs and mathematical descriptions to model intricate chemical systems, with applications spanning reaction kinetics prediction, transition state identification, and mechanistic elucidation. A successful computational approach to these traditionally experimental tasks offers substantial benefits: it can confirm chemists' intuitive understanding of reactivity modes, predict potential side products before extensive structural elucidation, and validate retrosynthetic suggestions derived from computer-aided synthesis planning, which enhances confidence in experimental success, reducing resource expenditure and accelerating discovery.[7]

A critical distinction exists between computational chemistry and theoretical chemistry, primarily in their emphasis on practical applications. While theoretical chemistry focuses on developing the foundational mathematical frameworks, theories, and models that explain chemical phenomena,[8,9] computational chemistry prioritizes the implementation of computer simulations and mathematical approximations to model chemical problems accurately. These methods frequently incorporate *ab initio* quantum mechanical calculations to determine molecular structures, energies, and properties with high precision. However, such calculations can be computationally prohibitive, especially for large molecular systems or when extensive conformational sampling is required.[10,11] This limitation is particularly pronounced in the study of complex organic reactions, which often involve multiple intermediates and transition states. Consequently, computational chemists have developed hierarchical approaches that combine different levels of theory and employ various approximations

E-mail address: chengze@uchicago.edu.

Peer review under the responsibility of Editorial Board of Intelligent Pharmacy.

https://doi.org/10.1016/j.ipha.2025.10.001

Received 27 August 2025; Received in revised form 27 September 2025; Accepted 3 October 2025

Available online 6 October 2025

2949-866X/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of Higher Education Press and KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

to balance accuracy with computational efficiency. Despite these advancements, the accurate prediction of free energy and kinetics for organic reactions remains challenging due to the inherent complexity of chemical systems, the necessity of comprehensive conformational sampling, and the subtle interplay between electronic and steric effects.[12–15] These persistent limitations have created a compelling opportunity for artificial intelligence (AI) approaches to complement and potentially revolutionize traditional computational methodologies.

2. Emergence of AI/ML

Machine learning (ML), a prominent branch of AI research, has experienced unprecedented growth and widespread application across numerous scientific domains in recent years. The foundational concept of machine learning, initially articulated by Arthur Samuel in 1959, centered on enabling computers to learn from experience without explicit programming.[16]

Just as the human brain develops chemical intuition through accumulated experience with diverse molecular structures, intricate reaction mechanisms, and observed experimental outcomes, machine learning algorithms are capable of learning by processing vast amounts of chemical data. By defining appropriate hypothesis spaces and molecular representations, algorithms can discern general patterns from chemical data, construct predictive models for thermodynamic and kinetic properties, selectivity, reaction pathways, and endeavor to approximate the underlying physical relationships that govern chemical reactivity.[17] Integration of machine learning with computational chemistry has opened entirely new avenues for addressing fundamental challenges within the field. Machine learning models possess the capacity to learn from extensive experimental databases, high-level quantum chemical calculations, and molecular dynamics simulations. This learning process allows them to develop sophisticated predictive capabilities that effectively complement existing or traditional computational approaches, enhancing their scope and efficiency.[18] This report aims to examine the latest advancements in applying artificial intelligence and machine learning methods to the prediction of free energy, kinetics, and synthesis analysis in organic chemistry reactions, delineate the challenges and opportunities in developing accurate predictive models, and discuss the profound potential for these hybrid approaches to transform organic chemistry research and synthetic planning.

3. Fundamentals of machine learning for chemical systems

Algorithms constitute the foundational core of artificial intelligence, and a diverse array of machine learning algorithms are currently employed across various scientific domains, including chemistry. Based on their distinct learning strategies, machine learning algorithms can be broadly categorized into supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and deep learning.

4. Supervised learning

Supervised learning is a category of machine learning that utilizes labeled datasets, comprising both features (inputs) and corresponding target outputs, to train algorithms for predicting outcomes and recognizing patterns. Examples of supervised learning methods include the following. Regression: Algorithms like Linear Regression and Random Forest Regressor are used to predict a real or continuous value by detecting relationships between two or more variables. Classification: Algorithms such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Neural Networks are employed to group data by predicting a categorical label or output variable.[19,20]

5. Unsupervised learning

Unsupervised learning operates on unlabeled data, aiming to discover hidden structures, patterns, or intrinsic relationships within the data without any explicit guidance or prior instruction.[21] In this paradigm, self-learning algorithms do not rely on pre-defined labels, but rather infer rules and structural information based solely on similarities, differences, and patterns observed within the raw data, making them exceptionally powerful for exploring vast, uncharted spaces and revealing previously unrecognized relationships between molecules or reactions. Examples include Clustering**:** Techniques like K-Means and DBSCAN group data points into homogeneous subgroups based on inherent similarities. Dimensionality Reduction: Methods such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Autoencoders reduce the number of features or dimensions in a dataset while preserving essential information.[21–24,25,26]

6. Semi-supervised learning

Semi-supervised learning leverages a combination of a small amount of labeled data and a large amount of unlabeled data to train models. This strategy is particularly advantageous when the process of obtaining labeled data is expensive, time-consuming, or otherwise resource-intensive.[21–24] The method often employs techniques like pseudo-labeling, where an initial model trained on the limited labeled data is used to predict outputs (pseudo-labels) for the abundant unlabeled data. These pseudo-labels are then combined with the original labeled data for subsequent retraining, iteratively enhancing the model's accuracy.[27–29]

Reinforcement Learning (RL):

Reinforcement learning is a machine learning technique in which an agent learns to make decisions by interacting with an environment through trial and error.[30,31] The overarching goal of the agent is to achieve optimal results by maximizing cumulative rewards over time.[32] The strength of RL lies in its capacity to learn optimal sequences of actions within dynamic environments. This approach extends beyond static prediction to active decision-making. Consequently, RL holds the profound promise of achieving truly autonomous discovery, wherein AI systems can iteratively design, test (either virtually or through robotic experimentation), and refine processes.[33] Examples of RL algorithms include Q-learning, Deep *Q* Networks (DQN), and Policy Gradient methods.

7. Deep learning architectures for chemical data

Deep learning (DL) represents a powerful subset of machine learning that employs neural networks with multiple layers to learn hierarchical representations of data. Its inherent ability to process large-scale datasets and identify intricate patterns has led to its transformative impact across numerous fields, including chemistry.[34]

7.1. Feedforward Neural Networks (FNN)

Feedforward Neural Networks are the simplest form of neural network architecture, characterized by information flowing in a single direction, from the input layer through one or more hidden layers to the output layer. Each neuron in a given layer is typically connected to every neuron in the subsequent layer. [35,36–38]

7.2. Convolutional neural networks (CNN)

While primarily known for their success in image analysis, Convolutional Neural Networks are designed to leverage convolutional layers to learn spatial hierarchies of features from input data automatically. They are highly effective at capturing localized patterns such as edges, textures, and complex shapes within data.[35,39] They are particularly useful when spatial relationships or local patterns within a molecule are critical for the predictive task.

7.3. Recurrent Neural Networks (RNN), LSTM, GRU

Recurrent Neural Networks are specifically designed to process sequential data, including text and time series. Unlike feedforward networks, RNNs feature connections that form directed cycles, enabling them to maintain an internal state or "memory" of previous inputs. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) are advanced variants of RNNs that effectively address the vanishing gradient problem, allowing them to learn and retain long-term dependencies in sequential data.[35,40]

7.4. Transformers

Transformers represent state-of-the-art architectures in Natural Language Processing (NLP). Their effectiveness stems from an attention mechanism that dynamically weighs the importance of different parts of the input sequence, allowing them to capture long-range dependencies more effectively than traditional RNNs.[41,42]

7.5. Graph Neural Networks (GNNs)

Graph Neural Networks are designed explicitly for learning over graph-structured data. This architecture naturally aligns with the fundamental graph structure of molecules. Prominent GNN variants include Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Message-Passing Neural Networks (MPNNs).[43] GNNs are arguably the most intuitive and powerful deep learning architecture for molecular data because they directly operate on the intrinsic graph structure of molecules. This approach circumvents the need for manual feature engineering of substructures, as is common with fingerprints, or the complex sequential parsing required for SMILES strings.[44] GNNs can learn intricate relationships between atoms and bonds directly, capturing both local and global molecular features without explicit human intervention. This mode of operation closely mirrors how chemists intuitively understand molecules—as interconnected networks of atoms and bonds. The ability of GNNs to integrate both molecular graph topology and auxiliary molecular descriptors further enhances their representational capacity, allowing for the simultaneous learning of structural and physicochemical features. This dual representation enables GNNs to capture complex intra- and intermolecular interactions, making them particularly powerful for predicting a wide range of molecular and reaction properties. GNN-based models represent a substantial advancement in the field of machine learning for chemistry, moving toward architectures that are not only data-driven but also more aligned with chemical intuition—effectively making them more "chemically aware"[45] (Table 1).

Table 1
Overview of machine learning paradigms and chemical applications.

Paradigm	Input Data Type	Goal	Common Algorithms/ Architectures	Applications
Supervised Learning	Labeled data	Prediction	Linear Regression, Random Forest, SVM, Neural Networks	Property Prediction
Unsupervised Learning	Unlabeled data	Discover hidden structure or patterns	K-Means, PCA, t-SNE, Autoencoders	Molecular Clustering, Chemical Space Exploration
Semi-Supervised Learning	Labeled and unlabeled data	Cost-effective labeling	Pseudo-labeling	Leveraging unlabeled datasets when labeled data is scarce
Reinforcement Learning	Interacts with environment	Optimal decision making	Q-learning, Deep Q Networks (DQN), Policy Gradient methods	Optimization
Deep Learning	Various	Complex pattern	FNN, CNN, RNN, LSTM, GRU, Transformers, GNNs	Broad applications across all areas

8. Machine learning pipeline in organic chemistry and synthesis

The application of machine learning (ML) to organic chemistry typically adheres to a standardized workflow. The standard procedure usually includes data collection, data representation, data processing, model training, and lastly model validation and testing.

9. Data collection

Data serves as the fundamental bedrock upon which all machine learning models are trained. The quality, quantity, and diversity of the collected data directly influence the model's performance. Data for chemical ML applications are compiled from a variety of sources, including published scientific literature, specialized chemical databases, patents, and meticulously maintained laboratory notebooks. To summarize, data can be generated from experimental measurements or from highly accurate theoretical calculations.

9.1. Databases

Leading commercial databases offer extensive collections of chemical information.

- SciFinder: This is a comprehensive platform for chemical substance information, encompassing structures, names, regulatory data, and various properties. SciFinder also provides advanced tools for optimizing synthesis plans and patent analysis.[46]
- Reaxys: An innovative chemistry database designed to optimize small molecule discovery, Reaxys integrates billions of chemistry data points from chemical patents and literature with powerful analytical and predictive tools. Furthermore, Reaxys offers predictive retrosynthesis capabilities and information on commercial product availability.[47,48]

While these commercial databases offer unparalleled breadth and quality of curated data, their proprietary nature limits widespread access for training and validating advanced ML models.

9.2. Open-access datasets

- GDB-13: GDB-13 enumerates small organic molecules up to 13 atoms of C, N, O, S, and Cl following simple chemical stability and synthetic feasibility rules. With 977 468 314 structures, GDB-13 is the largest publicly available small organic molecule database to date.[49,50]
- QM7/QM9: The QM7 dataset comprises 7,165 organic molecules with up to 7 heavy atoms (C, N, O, S, H) and primarily focuses on atomization energies. The QM9 dataset, a larger collection of approximately 134,000 small organic molecules with up to 9 heavy atoms (C, N, O, F, H), provides 19 electronic properties and includes constitutional isomers, serving as a key resource for de novo molecular design and benchmarking in quantum chemistry. [51]
- ESOL: The ESOL dataset consists of measured aqueous solubilities, which were used to develop a straightforward linear regression model for solubility estimation. [52]

- FreeSolv: This is a meticulously curated database containing experimental and calculated hydration free energies for small neutral molecules in water, along with their molecular structures and input files. [53]
- ZINC: ZINC is a freely accessible database of commercially available compounds primarily used for virtual screening. [54]
- IBOND: The iBond database is a curated collection of bond dissociation energies and related molecular properties. [55]
- USPTO: This widely utilized dataset is crucial for modeling reaction outcomes and retrosynthesis. It was derived from text-mining of United States patents published.

A potential direction lies in effectively combining these data sources to train models with both high accuracy and broader generalizability.

10. Molecular representation (descriptors)

A pivotal step in applying machine learning to chemistry involves converting molecular structures into numerical or symbolic formats that algorithms can effectively interpret.

10.1. Physicochemical descriptors

Physicochemical descriptors encompass calculated molecular properties such as molecular weight, logP (octanol–water partition coefficient), pKa, polar surface area, and the number of rotatable bonds. These properties can be computed using established cheminformatics toolkits, including ChemAxon, RDKit, and CDK.[56–58]

10.2. Molecular fingerprints

While valuable for their interpretability and for initial model development, relying solely on physicochemical descriptors might limit a model's ability to discover novel, non-obvious patterns. A robust approach often combines these chemically intuitive descriptors with more abstract, data-driven representations. Molecular fingerprints encode molecules as binary vectors, where each bit signifies the presence or absence of specific substructures or patterns within the molecule.[59]

- MACCS (Molecular ACCess System) Keys: This is a 166-bit fingerprint where each bit corresponds to a pre-defined structural feature.[60]
- Morgan/ECFP (Extended-Connectivity Fingerprints): These are circular fingerprints that encode substructures around a central atom, sharing conceptual similarities with Daylight fingerprints. ECFP (and its functional class variant, FCFP) are based on the Morgan algorithm and are extensively used for molecular similarity and classification tasks. [61]
- PubChem Fingerprints: This is an 881-bit structural key type fingerprint utilized by the PubChem database for similarity searching. Each bit in this fingerprint tests for the presence of specific structural features, element counts, ring systems, atom pairings, and atom environments. [62]

10.3. SMILES and InChI

These are standardized line notations for representing molecular structures.

- SMILES (Simplified Molecular-Input Line-Entry System): This system represents molecular structures as compact, human-readable ASCII strings. Arthur Weininger and David Weininger developed it in the late 1980s.[63,64]
- InChI (International Chemical Identifier): InChI is a non-proprietary, unique textual identifier for chemical substances. Its design aims to encode molecular information in a standardized manner, thereby facilitating efficient searching in databases and on the web.[65]
- SMARTS (SMILES ARbitrary Target Specification): This is an extension of SMILES specifically designed for substructure matching. SMARTS allows for the specification of wildcard atoms and bonds, making it a powerful tool for defining chemical patterns in database search applications.[66]

SMILES and InChI provide standardized methods for representing chemical structures as strings. However, converting these linear strings into a format suitable for machine learning often involves tokenization, which may not fully capture the inherent three-dimensional or graph-like nature of molecules. Therefore, these string representations might not be the most information-rich input for machine learning models, particularly when complex spatial or topological features are critical.

10.4. Structural representations

Molecular fingerprints provide a fixed-length, compact representation that enables efficient processing of large datasets and rapid similarity searches. However, they lose explicit information about molecular connectivity and spatial arrangement. Their suitability thus depends on the task. Fingerprints perform well for broad similarity searches or classification in established chemical spaces, but the loss of fine structural detail may limit their use in tasks requiring subtle distinctions or the discovery of novel reactivity, where more explicit representations such as graphs may be preferable.[67]

10.5. Coulomb matrix

A notable alternative is the Coulomb matrix, which encodes interatomic electrostatic interactions. The Coulomb matrix mimics the electrostatic interactions between atomic nuclei within a molecule. Its diagonal elements represent a polynomial fit of the potential energies to the nuclear charge of an atom (self-interaction), while the off-diagonal elements quantify the Coulomb repulsion between different pairs of nuclei.[68]

10.6. Molecular graphs

In this representation, molecules are depicted as undirected graphs, where individual atoms are represented as nodes and chemical bonds are represented as edges connecting these nodes. This method naturally aligns with the intrinsic structure of molecules, enabling the direct application of graph-based neural networks, including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Message-Passing Neural Networks (MPNNs).[69]

10.7. Grid-based representations

Molecules can also be embedded into 2D or 3D grids, allowing for the explicit encoding of spatial features. This representation is particularly advantageous in applications such as protein–ligand binding modeling or stereoselectivity prediction, where the precise 3D arrangement of atoms and their interactions with a receptor or catalyst are critical. Grid-based representations enable the use of 3D convolutional neural networks (3D CNNs), which can effectively capture spatial and chemical interaction patterns.[70]

11. Model building

Following the gathering of information from the dataset and encoding of molecular structures into appropriate representations, the next critical phase involves selecting and constructing suitable machine learning models. The choice of model is dictated by the specific task at hand, whether it requires classification, regression, or generative design.

Algorithms such as random forests, support vector machines, or various deep neural networks (including graph convolutional networks) are then trained on the prepared and labeled datasets. A crucial aspect of this stage is the optimization of hyperparameters, which are external configuration variables for the model. This optimization is typically performed using distinct training and validation sets, often employing techniques like cross-validation, to ensure the model achieves optimal performance and generalizability.

12. Prediction and application

Once a machine learning model has been rigorously trained and validated, it is deployed to predict properties or outcomes for unseen molecules or reactions. The applications are diverse and far-reaching, encompassing tasks such as predicting reaction products, elucidating retrosynthetic pathways, and forecasting physicochemical properties.

13. Analysis and interpretation

The final stage involves the analysis and interpretation of the model's prediction performance to extract process meaningful chemical information. For instance, models can provide insights into how specific functional groups influence reactivity or how molecular structure correlates with properties like free energy. The performance and reliability of the model are then evaluated.

13.1. Additional frequently used terms

Transfer Learning: Transfer learning involves using a network that has been trained on a related task and then tweaking its parameters to adapt to a new task, often with less data.[71,72]

Multitask learning: This involves training a model on multiple prediction tasks simultaneously to decrease the likelihood of overfitting.[72,73]

One-Shot Learning: A technique used to overcome applications with extremely limited data that uses networks to compress inputs into a continuous latent space and then compares the representation in this space to a larger, trained latent space.[72,74]

Autoencoders: A network architecture used to compress an input into a real-valued vector, commonly referred to as the latent space. A decoder network then takes this vector as its input and tries to reproduce the original input data.[72,75]

Data Augmentation: This involves expanding a data set by creating new training examples through reasonable manipulations of the data. One of the simplest demonstrations of this is rotating images in a data set while maintaining the same label[72,76] (Fig. 1).

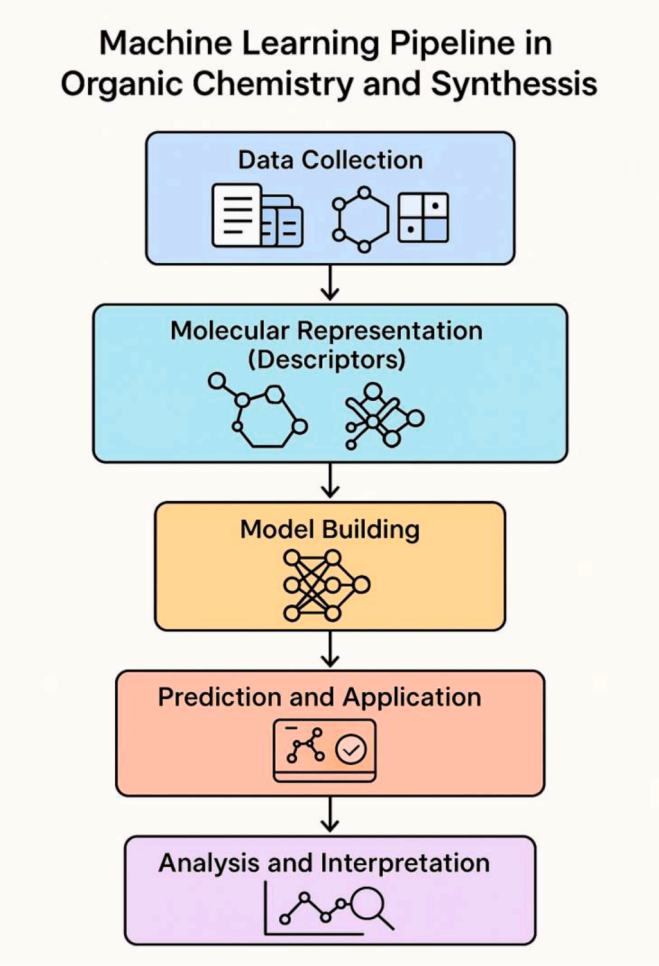


Fig. 1. ML model Pipeline.

14. Application of machine learning for free energy and kinetics estimation

14.1. Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy based on experimental data (Chung et al. 2022)

14.1.1. Background

Accurate prediction of solvation free energy (G_{solv}) and solvation enthalpy (H_{solv}) is critical for solvent selection and molecular design in diverse chemical applications and organic synthesis.[77] Traditional methods for predicting solvation properties, such as molecular dynamics simulations and quantum chemistry calculations, while theoretically rigorous, are often computationally expensive, particularly for large molecules or complex systems, and data-driven methods have emerged as promising alternatives. [77] Chung et al. (2022) introduced and evaluated machine learning models (Solute-ML, Direct-ML), and compared them to the standard Solute-GC computational model (a functional group contribution method) — for the prediction of Abraham solute parameters and crucial thermodynamic properties: solvation free energy (G_{solv}) and solvation enthalpy (H_{solv}).

Abraham's Linear Solvation Energy Relationships (LSERs), developed by Professor Michael H. Abraham and co-workers, provide a quantitative framework for understanding and predicting solute–solvent interactions by expressing a given property (like solvation free energy or enthalpy, or partition coefficients) as a linear combination of terms that quantify specific types of intermolecular forces.[78,79] The general form of an Abraham LSER equation relates a solute's property in a given solvent system to a set of solute-specific parameters and solvent-specific coefficients.

$$\log K(298K) = c + eE + sS + aA + bB + lL$$

$$\Delta G_{solv}(T) = -RT\ln K$$

$$\Delta H_{solv} = c' + eE' + sS' + aA' + bB' + lL'$$

Abraham solute (E, S, A, B, L) and solvent (c, *e*, s, a, *b*, l), and new solvent parameters (c′, e′, s′, a′, b′, l′) are used for ΔH_{solv}

Each solute parameter represents a fundamental physicochemical property of the solute, and the linear relationship offers a computationally efficient way to estimate solvation properties once the parameters are known. However, these models are limited by (1) the availability and quality of Abraham solute parameter datasets and (2) the inherent oversimplification.

14.1.2. Methods and data

Three distinct models were trained and tested by Chung et al. (2022)

to predict solvation properties, each underpinned by uniquely compiled and processed experimental datasets. Solute-GC (based on the group contribution method) leverages atom-centered functional groups with sophisticated corrections for ring and polycyclic strain. Solute-ML and Direct-ML are machine learning models built upon a directed message passing neural network architecture.

14.1.3. Data compilation

Abraham solute parameter data and thermodynamic data were gathered from different databases, literature sources, and experimental records.

Solute-DB: This comprehensive database comprises 8366 unique Abraham solute parameters (E, S, A, B, L) gathered from various literature sources.

dG$_{solv}$ -DB: Consisting of large collections of experimental solvation free energies (ΔG_{solv}), this dataset integrates data from well-established repositories like the Minnesota solvation database, CompSol, FreeSolv, and other meticulously curated published works. To evaluate data bias and model robustness, dG$_{solv}$DB was further subdivided into dG$_{solv}$DB1, dG$_{solv}$DB2, which incorporates in-house gas-water $logk_w$ partition coefficient data), and dGsolvDB3 (which includes in-house octanol−water $logp_{ow}$ and solvent−water $logp$ partition coefficient data).

dH$_{solv}$ -DB: This database contains 6322 experimental solvation enthalpies (ΔH_{solv}), primarily sourced from the Acree Enthalpy of Solvation dataset and the CompSol database

14.1.4. Solute-GC (group contribution method)

The Solute-GC method represents a refined group contribution approach for predicting Abraham solute parameters. The fundamental principle of group contribution methods is that the properties of a molecule can be estimated by summing the contributions of its constituent functional groups. The Solute-GC model calculates the solute parameters (ESABL) as shown in the equation below; specific halogen correction terms are added if a halogen atom occurs.

$$\mathrm{E,S,A,B,L} = \mathrm{SUM\ AC} + \mathrm{RSC} + \mathrm{LDI} + (\mathrm{Halogen})$$

Atom-centered (AC) functional groups: Sumed the contribution from all groups.

Ring Strain Correction (RSC): Applied for each ring cluster found in a molecule.

Long Distance Interaction (LDI) groups: Applied to each long–-distance interaction group found in a molecule.

14.1.5. Machine learning models (Solute-ML and Direct-ML)

Both Solute-ML and Direct-ML models employ the Chemprop software package, which is built upon a Directed Message Passing Neural Network (D-MPNN) architecture. D-MPNNs are particularly well-suited for molecular property prediction because they can effectively learn complex, non-linear relationships by transforming atom and bond features within a molecule into rich, learned molecular representations.

Solute-ML: This model adopts a two-step, indirect prediction strategy. First, an ensemble of D-MPNNs is trained on the Abraham solute parameters database Solute-DB to directly predict the Abraham solute parameters (E, S, A, B, L) for a given molecule. The training set uses SMILES strings as input, with the best performance achieved by concatenating 2D-RDKit molecular descriptors. Once these parameters are predicted, they are subsequently used as inputs into the established LSER equations (along with known solvent parameters) to calculate ΔG_{solv} and ΔH_{solv} .

Direct-ML: In contrast, the Direct-ML model employs a single-step, direct prediction approach. An ensemble of D-MPNNs is trained directly on the free energy dG$_{solv}$ -DB and dH$_{solv}$ -DB datasets. The SMILES strings for both the solvent and the solute are each converted into a separate latent representation by a dedicated D-MPNN. These embeddings are then concatenated. Molecular feature descriptors RDKit vectors for both the solute and solvent are concatenated with the latent representation after the D-MPNN and before the second neural network for property prediction. The concatenated embedding is subsequently fed into a neural network for the direct prediction of the solvation property (free energy or enthalpy), bypassing the intermediate prediction of Abraham parameters and the LSER equations.

14.1.6. Dataset splits and model evaluation

Approximately 90% of the data was allocated for training and validation, and the remaining 10% was reserved as an unseen test set. Furthermore, both random and substructure-based splits were employed to assess the model's interpolative capability and extrapolative power.

14.1.7. Results and discussion

Evaluation Indices: Mean Absolute Error (MAE), Median Absolute Error (MedAE), Root Mean Squared Error (RMSE) (Table 2).

Direct-ML consistently outperformed the other models across all test splits, matching the performance of physics-based models like COSMO-RS.

Solute-ML showed moderate accuracy, providing interpretable intermediate parameters.

Solute-GC was the least accurate, but easily interpretable with fragment-based information.

14.1.8. Generalizability to unseen chemical space

Direct-ML's ability to handle unseen solvents was also tested by excluding some data during training and then evaluating on those “out-of-sample” pairs. Despite novel solvent contexts, the error rose only slightly (around 0.1–0.2 kcal/mol) for random splits and stayed on par—or even improved—in some detailed substructure-based splits (Table 3).

14.1.9. Future directions and benefits of model combination

The machine learning based model consistently demonstrated superior predictive accuracy, often achieving performance comparable to advanced quantum chemistry methods, particularly on random test splits, and demonstrating the potential of a data-driven approach. Averaging the predictions from Solute-ML and Direct-ML, or even all three models, led to slight but notable improvements in overall accuracy across most test sets. Discrepancies can serve as indicators of potentially uncertain or outlier values, particularly in mitigating significant errors associated with outliers. Model combination also shows potential for the future of predictive chemistry with integrated, multimodal strategies.

Table 2
Performance of the three distinct models.

Model	ΔG_{solv} MAE (Random Split)	ΔH_{solv} MAE (Random Split)	ΔG_{solv} MAE (Scaffold Split)
Direct-ML	0.40 kcal/mol	0.50 kcal/mol	0.91 kcal/mol
Solute-ML	0.52 kcal/mol	0.64 kcal/mol	1.14 kcal/mol
Solute-GC	0.64 kcal/mol	0.84 kcal/mol	1.30 kcal/mol

Table 3
Strengths and limitations.

Model	Strengths	Limitations
Solute-GC	Interpretable, efficient, easily extensible.	Poor performance for long-range or nonlocal interactions.
Solute-ML	Better accuracy with retention of interpretable parameters.	Still limited by LSER linearity, the model depends on the quality of the parameter dataset.
Direct-ML	Best overall accuracy, generalizes to new solvents.	Black-box nature, sensitive to training data bias.

14.2. Machine learning to predict experimental solvent effects on reaction rates based on quantum chemistry calculations

14.2.1. Background

Yunsie Chung and William H. Green trained a machine learning model to predict the solvation free energy and solvation enthalpy of activation ($\Delta\Delta G_{solv}$, $\Delta\Delta H_{solv}$) for a solution phase reaction, using datasets based on the COSMO-RS calculation method comprising over 28,000 neutral reactions and 295 solvents. Data-driven approaches that combine QM calculations with machine learning models are more computationally efficient compared to traditional QM methods.[80]

14.2.2. Background on prediction targets

The primary targets for prediction in this ML model are the solvation free energy and solvation enthalpy of activation differences ($\Delta\Delta G_{solv}$ and $\Delta\Delta H_{solv}$) for a reaction–solvent pair, which can then be used to estimate relative rate constants between different solvents. The solvation free energy (ΔG_{solv}) and solvation enthalpy ((ΔH_{solv}) are the changes in Gibbs free energy and enthalpy changes when a molecule is transferred from an ideal gas to a solvent at fixed condition. The $\Delta\Delta G_{solv}$ and $\Delta\Delta H_{solv}$ of a reaction–solvent pairs are the solvation free energy and solvation enthalpy differences between a transition state (TS) and reactant(s) accordingly.

$$\Delta\Delta G_{solv} = \Delta G(TS)_{solv} - \Delta G(R)_{solv}$$

$$\Delta\Delta H_{solv} = \Delta H(TS)_{solv} - \Delta H(R)_{solv}$$

The relative reaction rate was then estimated with the following Arrhenius equation relationship.

$$K_{liq/gas} = e^{\left(\frac{\Delta\Delta Gsol}{RT}\right)}$$

$$K_{relative} = e^{\left(\frac{\Delta\Delta Gsol1}{RT} - \frac{\Delta\Delta Gsol2}{RT}\right)}$$

Fig. 2 Linear approximation method was used to calculate the $\Delta\Delta G_{solv}$ and $\Delta\Delta H_{solv}$ under different temperatures with the following equation.

$$\Delta\Delta G_{solv}(T) = \Delta\Delta H_{solv}(298\ K) - T\left(\Delta\Delta H_{solv}(298\ K) - \Delta\Delta G_{solv}(298\ K)/_{298\ K}\right)$$

14.2.3. Data generation and computational method

The data were primarily generated through *ab initio* COSMO-RS calculations. The pre-training and fine-tuning datasets were generated using COSMO-RS calculations at the BP86/def2-TZVPD level of theory with a fine grid cavity (BP-TZVPD-FINE level). The computational workflow involved:

1. Geometry Optimization: Gas-phase geometries for reactants, products, and transition states were obtained from existing datasets (Grambow et al. and Harms et al.).
2. Single-Point Energy Calculations: Performed in COSMO and gas phases using TURBOMOLE 7.5 to generate screening charge densities and energies.
3. Solvation Property Calculation: ΔG_{solv} and ΔH_{solv} of reactants, products, and TSs were computed in 295 common solvents at 298 K using COSMOtherm (release 2021) with BP_TZVPD_FINE_21 parametrization. $\Delta\Delta G_{solv}$ and $\Delta\Delta H_{solv}$ were then calculated. $\Delta\Delta H_{solv}$ was derived from the temperature gradient of $\Delta\Delta G_{solv}$ around 298 K.
4. Data Cleaning: Computed data were filtered to remove physically unreasonable values (e.g., $|\Delta\Delta G_{solv}|$ >40 kcal/mol for pre-training, $|\Delta\Delta G_{solv}|$ >10 kcal/mol for fine-tuning), which were suspected to be due to COSMO-RS calculation errors for specific reactions/geometries.

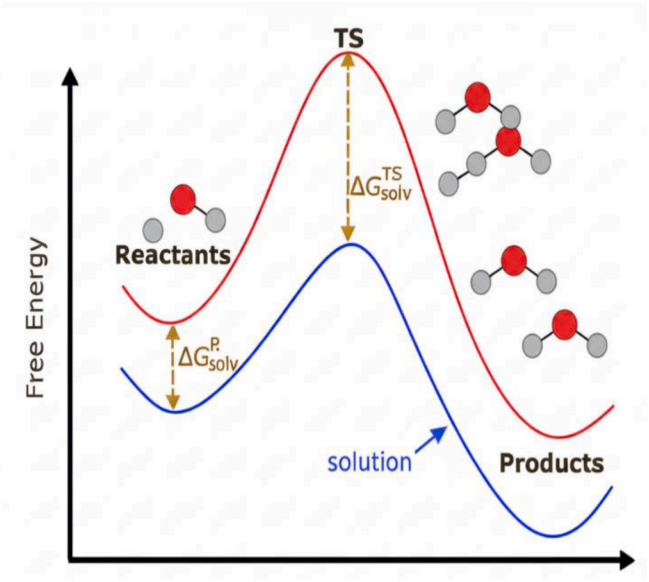


Fig. 2. solvation free energy between the transition state (TS) and reactant(s) accordingly.

14.2.4. Data sets and splitting method

Three main data sets were prepared:

- Pre-training Set: 7,796,583 reaction–solvent pairs (26,448 unique reactions, 295 solvents). Diverse neutral closed-shell and free radical reactions involving H, C, N, O atoms, with at most seven heavy atoms. Includes both forward and reverse reactions. Geometries from Grambow et al. A subset of 500,000 data points was identified as optimal for model training, ensuring all reactions and solvents appeared at least once.
- Fine-tuning Set: 542,833 reaction–solvent pairs (1,870 unique reactions, 295 solvents). A subset of 46,122 data points was sampled, with a bias towards polar solvents. Common neutral, free radical reactions, including bimolecular hydrogen abstraction, unimolecular hydrogen migration, and radical addition to multiple bonds. Involves H, C, and O atoms, with a maximum of 10 heavy atoms. Geometries from Harms et al.
- Experimental Test Set: 165 relative rate constants for 15 neutral reactions and 49 solvents (273 K–392 K). Content: None of these reactions appear in the pre-training or fine-tuning sets. Reactions are larger (up to 22 heavy atoms) and include B-scission, H-abstraction, Diels–Alder, and other types.

The pre-training set was split into 90 percent training/validation and 10 percent test sets using reaction and solvent splits. The 90% portion was further divided 80/20 for training and validation, and five folds were generated. For test sets, 5% of reactions and 5% of solvents were randomly selected to evaluate performance on unseen reactions and solvents. The fine-tuning set used an 80/20 training/validation split with one-fold, and the experimental dataset served as the final unseen test set for the fine-tuned model.

14.2.5. Machine learning model architecture

Based on atom-mapped reaction SMILES and solvent SMILES strings, the model is mainly implemented by using Chemprop with a directed message passing neural network D-MPNN (a type of GCNN that transforms atom and bond features into a latent molecular representation through bond-level message passing for predicting chemical and

reaction properties). The model uses the conformational graph representation (CGR), which superimposes the reactant and product graphs to mimic the 2D structure of the transition state for reaction encoding. The CGR is then passed into a D-MPNN to generate a reaction embedding. Separately, another D-MPNN converts the solvent graph into a solvent molecular embedding. These representations are concatenated and fed into a feedforward neural network (FNN) to predict the regression targets. Additionally, RDKit60 within Chemprop with features for atom type, bond type, formal charge, chirality, etc, was used to optimize the performance, and the model's hyperparameters were optimized through 47 iterations of Bayesian optimization using the Hyperopt package.

14.2.6. Training Protocol and ensemble approach

The transfer learning approach is used with pre-training sets and subsequently fine-tuning sets. Within each of the five folds of the pre-training set, an ensemble of 5 different models is generated by using different random initializations of model parameters. The optimized parameters from the 25 pretrained models are then used to initialize the 25 fine-tuned models that are trained on the fine-tuning set. Average predictions from an ensemble of all 25 fine-tuned models are used for the final error evaluation with the experimental set, and models are trained or validated on the tested experimental reactions.

14.2.7. Result and testing

Overall, the model achieves RT ($\ln K_{relative}$) MAE/RMSE of 0.68/0.88 kcal/mol on the experimental set and achieves the MAE/RMSE of 0.71/1.16 and 1.03/1.63 kcal/mol on unseen reactions for the prediction of solvation free energy and solvation enthalpy of activation ($\Delta\Delta G_{solv}$, $\Delta\Delta H_{solv}$), respectively, relative to the COSMO-RS calculations. RP-solv feature (solvation energies and enthalpies of reactants and products) improved model performance during pre-training. However, the fine-tuned model, notably without the RP-solv feature (possibly due to observed overfitting during fine-tuning), achieved the best overall performance. The performance gain is observed only with the small size of the fine-tuning set. Thus, transfer learning is particularly beneficial since it avoids mixing the two datasets that differ in the level of theory used for geometry optimizations and also avoids weighting the dataset in a way that favors the set considered more relevant to real liquid-phase systems. Moreover, the model can have drastically different outcomes depending on the number of fine-tuning epochs used. A portion of the prediction error stems from the inherent error of the COSMO-RS method used to generate the training data. Analysis of outliers primarily consisted of unusual reactions unlikely to occur in real systems. However, the model is shown to provide reliable predictions of relative rate constants when tested on the experimental set that contains numerous neutral reactions. Moreover, this method is constructed based on the assumption that the solvation effect does not change the 3D geometries of the reactant and TS, and the majority of the reactants are rigid, with no or fewer than four rotatable bonds.

14.3. Hybrid free energy perturbation (FEP)/machine-learning methodology for the computation of hydration free energies

14.3.1. Background

Force field parametrization is a complex process required for precise simulations. This challenge remains due to the limitations of fixed-charge force fields, which cannot fully capture molecular interactions, particularly polarization effects, and Scheen et al. believe that accuracy cannot be fully achieved due to an incomplete description of the physics of the process. This realization has led to the development of correction terms for hydration and binding free energies calculated based on classical force fields. Scheen et al. (2020) integrate traditional 'physics-driven' free energy calculations based on FEP methods with 'data-driven' machine learning methods by introducing a correction term to construct a hybrid model to predict hydration free energies of small molecules, requiring substantially smaller training sets to achieve reasonable accuracy. This hybrid approach addresses some of the inherent limitations in conventional FEP methods, such as computational cost, inefficiencies, and internal inaccuracies due to force field parametrizations.[81]

14.3.2. Hybrid FEP/ML methodology

The core idea is to train ML models to correct the systematic errors in FEP calculations, referred to as the ΔG offset.

$$\Delta G_{offset}(A) = \Delta G_{exp}(A) - \Delta G_{FEP}(A)$$

Thus, define a corrected hydration free energy as:

$$\Delta G_{FEP/ML}(A) = \Delta G_{FEP}(A) + < \Delta G_{OFFSET}(A) >_{N_{POP}}$$

14.3.3. Data set and feature preparation

A subset of 595 molecules from the FreeSolv database (aside from experimentally determined values, the database contains absolute free energies of hydration computed from alchemical simulations using GROMACS. Note: FreeSolv SAMPL4 data set was excluded) was used for training. Features were generated from all compounds present in FreeSolv using RDKit, including atom-pair fingerprints (APFP), extended-connectivity fingerprints (ECFP), topological fingerprints (TOPOL), molecular properties (MolProps), and generated noise fingerprints (X-NOISE). All fingerprint features were appended individually to MolProp, normalized, and then reduced in dimensionality using Principal Component Analysis (PCA) up to 95% of the variance.

14.3.4. Machine learning models

FEP/ML Training Set: This set used the ΔG_{offset} (the difference between the free energy prediction and the experimental value) as the label for each data point.

ML Training Set: This set used the experimentally determined hydration free energy (ΔG_{exp}) directly as the label for each data point.

14.3.5. Training Protocols

Models were developed using Scikit-Learn, including Support Vector Machines (SVM), Random Forests (RF), Deep Neural Networks (DNN), and Multiple Linear Regression (MLR) methods. For a given training set, machine-learning models were used to fit the training domain using 5-fold cross-validation over 10 replicates, resulting in a total of 50 trained models. The offset estimator was then defined as the arithmetic mean of these offset values, and the standard deviation of the mean was used as a measure of the precision of the calculated offset. Bayesian hyperparameter optimization was used to fine-tune the models, and optimal configurations were determined after approximately 50 optimization steps.

14.3.6. Result and performance

Among these models, SVM and DNN models showed better performance on the training data, converging to a mean unsigned error (MUE) of approximately 0.55 kcal/mol, while MLR and RF converged around 0.75 kcal/mol. The top-performing FEP/ML models were primarily based on the MolProps feature set for SVM, RF, and MLR, but not for DNN, which showed larger uncertainties. Specifically, evaluated on the FreeSolv-SAMPL4 test set, the top-performing FEP/ML model (SVM; MolPropsAPFP) outperformed standalone FEP in Pearson r, MUE, and RMSE statistics (0.96, 0.76 kcal/mol, and 1.28 kcal/mol, respectively for FEP/ML; 0.92, 1.16 kcal/mol, and 2.49 kcal/mol, respectively, for standalone FEP) and had higher ranking statistics (Spearman ρ and Kendall τ) than standalone FEP (0.96 and 0.86 for FEP/ML; 0.87 and 0.72 for standalone FEP). DNN model achieved similar accuracy but introduces significant uncertainties compared to FEP.

When retrospectively introduced into the SAMPL4 challenge, the FEP/ML SVM model (MolPropsAPFP) significantly improved the performance in predicting hydration free energies compared to the standalone FEP method in nearly all statistical metrics. In contrast, pure ML

models performed broadly similarly to FEP but with remarkably large uncertainties, indicating significant variability in the predicted free energies of hydration of the same compound by the ensemble of ML models. By contrast, the FEP/ML predictions are of similar precision to the FEP predictions, as the uncertainties in the offset terms are comparable to or smaller than the uncertainties in the alchemical estimates.

14.3.7. Training set size

FEP/ML models began outperforming standalone FEP when trained on only about 20 compounds and converged to their peak performance after being trained on approximately 400 compounds. Standalone ML models, in contrast, required a much larger training set of around 450 compounds to surpass standalone FEP and did not appear to have fully converged, even with the largest training set.

14.3.8. Transferability of offsets to related SAMPL4 submissions

The transferability of ML-derived offsets to other FEP-based SAMPL4 tests was assessed. The offsets generally improved or maintained performance for most assessments that featured an FEP methodology. Submissions using similar force fields (GAFF and AM1-BCC charges, GROMACS engine) showed improvements. However, transferability was influenced by the charge model, with RESP-based methods showing worsened accuracy, while VCHARGE results slightly improved. OPLS force field-based assessments were also shown to be beneficial, but GROMOS or AMOEBA submissions did not perform as well, likely due to lower correlation with AM1-BCC/GAFF hydration free energies.

14.3.9. Key features and future direction

The work by Scheen et al. successfully demonstrates that a hybrid FEP/ML methodology combines the "physics-driven" rigor of FEP with the "data-driven" efficiency of machine learning. By learning to correct the systematic errors of FEP calculations, the FEP/ML approach offers a robust, efficient, and accurate tool with short training time and low cost. However, the model is limited in its transferability due to force field and charge model differences.

14.4. Holistic prediction of the pK_a in diverse solvents based on a machine learning approach

14.4.1. Background

The acid dissociation constant, expressed as pK_a, is a fundamental parameter in physical chemistry, and pK_a is critical for predicting solubility, membrane permeability, molecular reactivity, and binding interactions, particularly in drug development. However, data on non-aqueous pK_a are relatively scarce. Yang et al. tackle the challenge of predicting pK_a values in diverse solvents using a machine learning approach, with a focus on non-aqueous environments where traditional computational chemistry methods falter.[82]

14.4.2. Methodology

To develop a robust and generalizable model for predicting pK_a values across diverse solvents, a wide range of machine learning algorithms were implemented, including Tree Regression, Random Forest, Gaussian Process, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, Bagging Tree, Extra Tree, as well as Neural Network (NN) and XGBoost algorithms using the entire curated iBonD dataset. Neural Network and XGBoost model, the latter being a gradient boosting algorithm, achieve rapid and most accurate pK_a predictions.

14.4.3. Data source and dataset characteristics

The iBonD database, which contains over 30,000 experimental pK_a values for around 20,000 compounds across 46 solvents, was used as the primary training data set. pK_a data outside each solvent's pK_a window were removed, and for molecules with multiple entries, averaged values were used. One hundred errors were identified as outliers through iterative model training and were also removed. The final dataset consisted of 15,338 compounds with 19,397 pK_a values across 39 solvents. The most frequently applied solvents with more than 400 pK_a data include Water (61.7%), DMSO (9.3%), EtOH/H_2O (8.2%), MeOH (3.6%), CH_3CN (5.2%), DMF (2.2%). Structurally, the curated iBonD dataset consists mainly of N-H (47%), O-H (44%), C-H (7%), and others (2%). Each experimental pK_a was assigned to the primary responsible ionization site. The dataset is composed of over 50% neutral molecules (ca. 11,105 pK_a values), over one-third protonated molecules (ca. 7,104 pK_{aH} values), and a small fraction of negatively charged molecules (1,224 pK_{aN} values). For multi-ionizable charged molecules, the structural designation was restricted to the dominant microstate, with microscopic pK_a serving as a good approximation of the macroscopic constant.

14.4.4. Model architecture and training strategy

A holistic pK_a prediction model was constructed based on curated experimental data collected within the iBonD database and correlated with Structure Physicochemical (also denote Physical Organic Chemistry) properties combined descriptors (SPOC) based on molecules. Before model training, each molecule was annotated with its ionic form—neutral (pK_a), protonated (pK_{aH}), or deprotonated (pK_{aN})—via Ionic Status Labeling (ISL), and its molecular characteristics were encoded into SPOC descriptors, which integrate structural fingerprints including MACCS keys (a 166-bit binary vector indicating the presence or absence of predefined substructures), EState indices (which reflect the electronic and topological environment of atoms within the molecule), and Morgan fingerprints (circular, hashed representations of atom-centered environments capturing local chemical topology), along with RDKit-derived physicochemical properties (such as molecular weight, logP, and hydrogen bonding capacity), allowing fast and informative descriptor generation to support accurate machine learning-based pK_a prediction across diverse solvents.

14.4.5. Data splitting and model training

Model evaluation was conducted using 5-fold cross-validation and an 80:20 train/test split. For comparison, single solvent models (SSMs) were also built for the six most data-rich solvents (H_2O, DMSO, EtOH/H_2O (1:1), acetonitrile, MeOH, and DMF).

14.4.6. Performance analysis

Descriptor selection was key to model performance. The best results were achieved using the SPOC descriptor set, which combined MACCS fingerprints with RDKit-derived physicochemical features. Adding Ionic Status Labeling (ISL) further improved accuracy. Feature importance analysis showed physicochemical descriptors contributed 55%, structural descriptors 40%, and ionic/solvent labels 5%.

14.4.7. Model performance (note: unit in kcal/mol)

Among these models, NN (with three hidden layers and early stopping) and XGBoost emerged as the top performers due to their superior predictive accuracy and scalability. NN: MAE = 0.87, RMSE = 1.41, r^2 = 0.95; XGBoost: MAE = 0.88, RMSE = 1.50, r^2 = 0.94. The holistic model consistently outperformed the SSMs (especially in solvents like DMSO and acetonitrile). pK_a values for the top six solvents were predicted and analyzed for the intrinsic free–energy relationship between any pair, finding that "like-solvents," such as protic–protic or aprotic–aprotic pairs, show better correlations than "unlike-solvents," like protic–aprotic pairs. This outcome demonstrates that the model successfully learned the inherent relationships and distinct solvation behaviors between the solvents.

14.4.8. External validation

The holistic model (HM), particularly the HM-XGBoost variant, was demonstrated through four external validation studies. First, in the SAMPL6 blind prediction challenge for macro- and micro-pK_a values of

24 drug-like molecules in water, HM-XGBoost achieved MAE = 0.80, outperforming previous machine learning entries. A second test on 17 pharmaceutical compounds with multiple protonation sites yielded micro-pK_a predictions with MAE = 0.44, successfully identifying site-specific protonation states. In a third application, the model predicted (pK_{aS} values of 84 organocatalysts in DMSO, achieving remarkable accuracy for thioureas (MAE = 0.16) and prolinamides (MAE = 0.23), with an overall MAE of 1.01 across five organocatalyst classes. Notably, predictions outperformed DFT in some instances, though internal hydrogen bonding and zwitterionic forms introduced errors for BINOL-type and 6′-HBCA catalysts. Lastly, for aminocatalyst (pK_{aH} values in MeCN, the model achieved an MAE of 1.57, a reasonable result given the limited MeCN training data, with underestimation trends attributed to substituent effects not well represented in the dataset.

14.4.9. Key features

Yang et al. have developed a holistic pK_a prediction model (HM) based on the iBonD experimental pK_a dataset. The ML models show effective capability for predicting pK_a values across diverse solvents, remarkable accuracy comparable with DFT methods in some instances, and considerable transferability on unseen tests. Moreover, the model successfully identified "like–solvents correlation" purely based on a data-driven approach, showing the potential of chemical awareness and interpretability of the ML model developed.

14.5. Combine machine learning and thermodynamics for accurate and consistent pK_a prediction

14.5.1. Background

Prediction of pK_a remains a complex challenge, often due to the intricate and coupled chemical equilibria among the various protonated forms of a molecule. Existing computational methods usually oversimplify this complexity. Template-based methods rely on empirical fragment values derived from experimental data. Traditional Machine Learning and Deep Learning capture the interaction of molecules globally, but often lack thermodynamic consistency, treating the ionization site independently. Luo et al. introduce Uni− pK_a, a novel framework that predicts pK_a by explicitly integrating thermodynamic principles into a machine learning model. Shifting the traditional method of directly predicting pK_a values to the modeling of free energies from all possible protonation microstates, ensures that any derived pK_a values, whether macro or micro, are inherently thermodynamically consistent with the entire system.[83]

14.5.2. Methodology

The system includes a systematic microstate enumerator and a neural network, trained under a multi-fidelity strategy, based on the following theoretical foundation. The microstate enumerator proposes the protonation ensemble of the target molecules, and the machine learning part serves as the core algorithm for the microstate free energy prediction.

For a micro pK_a equilibrium:

$$K^{k,i}_{a,m} := \frac{[K - H_mA^{m+}]\,[H^+]}{[i - H_{m+1}A^{(m+1)+}]} = \exp\{ - \beta[\Delta_f G^-_m(K - H_mA^{m+}) - \Delta_f G^-_m(i - H_{m+1}A^{(m+1)+})]\}$$

For a macro pK_a equilibrium:

$$K_{a,m} := \frac{[H^+]\sum_i[i - H_mA^{m+}]}{\sum_i[i - H_{m+1}A^{(m+1)+}]} = \frac{\sum_i \exp\{ - \beta[\Delta_f G^-_m(i - H_mA^{m+})]\}}{\sum_i \exp\{ - \beta[\Delta_f G^-_m(i - H_{m+1}A^{(m+1)+})]\}}$$

$\beta = (RT)^{-1}$, $\Delta_f G^-_m$ is the standard molar Gibbs free energy change of formation.

pH dependent free energy: $\Delta_f G_m$ (H_mA^{m+}; pH) = $\Delta_f G^-_m$ (H_mA^{m+} (aq)) + $\frac{mln10}{\beta}$ pH

pH-dependent free energies give the fraction of each microstate across the ensemble under particular pH conditions:

$$w(k - H_mA^{m+}; \text{pH}) := \frac{[K - H_mA^{m+}]}{\sum_n \sum_i [i - H_nA^{n+}]} = \frac{\exp\{ - \beta[\Delta_f G^-_m(k - H_mA^{m+})]\}}{\sum_n \sum_i \exp\{ - \beta[\Delta_f G^-_m(i - H_nA^{n+})]\}}$$

The above mathematical formalism demonstrates that micro pK_a is simply a special case of macro pK_a and that all pK_a related properties, including the pH-dependent distribution of protonation states, can be derived from microstate free energies.

14.5.3. Dataset preparation: the microstate enumerator

Uni− pK_a microstate enumerator, equipped with a systematic algorithm, reconstructs the complete protonation ensemble for each molecule. The enumerator works by using a template of SMARTS patterns for common ionizable sites, which is based on an augmented version of a template from a previous work, MolG pK_a. The algorithm iteratively expands the protonation ensemble by alternating between deprotonation (A2B) and protonation (B2A) rounds, systematically generating all possible microstates. The pretraining phase leverages about 1 million molecules with empirical pK_a values in the ChEMBL database, resulting in more than 3 million protonation states after microstate enumeration.

14.5.4. The machine learning model architecture

Based on the Uni-Mol model, a transformer-based molecular encoder designed for learning 3D molecular representations, which effectively capture long-range interactions within a molecule, was modified for predicting pK_a values. Uni-Mol backbone includes invariant spatial positional encoding, a dual-level representation, atomic charge information, and a Free Energy to pK_a (FE2pKa) module, which acts as a crucial bridge between the neural network's raw output and the thermodynamic equations to the final pK_a prediction. This module ensures that the model's predictions are physically consistent and interpretable within the context of the protonation ensemble theory.

14.5.5. Model training method and strategy

The pretraining phase uses a large-scale, low-fidelity dataset from the ChEMBL database, which contains millions of computationally predicted pK_a values.

Supervised pK_a Prediction: The model is trained to predict pK_a values using the ChEMBL data. Crucially, it was trained to predict the microstate free energies, which are then used to calculate an ensemble-aware loss-based function on the macro pK_a data and micro free energy prediction as shown in the below formula, allowing the model to learn the underlying free energy landscape.

$$L_{pK_a}(g^A, g^B;\ pK_a) = \frac{1}{2}\left[pK_a + \frac{\sum_i e^{-g^B_i}}{\sum_i e^{-g^A_i}}\right]$$

g^A_i, g^B_i are microstates free energy output in A and B macro states.

Self-Supervised Masked Atom Prediction: The model learns to predict the identity and fundamental properties of masked atoms by utilizing the cross-entropy loss function.

$$L_{atom} = -\sum_{i \in M} \log p(y_i|x_i)$$

M denotes the set of masked atoms, y_i is the true type of atom, and x_i is the input containing the token.

Self-Supervised Masked Charge Prediction: The model predicts the original formal charges of masked atoms by utilizing the cross-entropy loss function.

$$L_{charge} = -\sum_{i \in M} \log p(C_i|x_i)$$

M denotes the set of masked atoms, C_i is the true charge, and x_i is the input containing the token.

Self-Supervised 3D Position Recovery: The model learns to recover the true 3D coordinates of masked atoms from corrupted inputs with two additional heads using smooth loss.

$$L_{\text{coord}} = -\sum_{i\in M} Smooth_l(\widehat{r_i} - r_i)$$

$$L_{\text{dist}} = -\sum_{i\in M}\sum_j Smooth_l\left(\widehat{d_{i,j}} - d_{i,j}\right)$$

M denotes the set of masked atoms, $\widehat{r_i}$ and r_i are the predicted and true coordinates of the atom, respectively. $\widehat{d_{i,j}}$ and $d_{i,j}$ are the predicted and true pairwise distances between masked atoms, respectively.

14.5.6. Final composition of the loss function

$$L_{\text{all}} = L_{\text{pK}_a} + 2L_{\text{charge}} + L_{\text{atom}} + 2L_{\text{coord}} + L_{\text{dist}}$$

After pretraining, the model is fine-tuned on a smaller, high-fidelity experimental dataset from the DataWarrior and iBonD.

14.5.7. Performance

Uni− pK_a has been rigorously evaluated on a variety of external datasets and consistently achieves the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) on key benchmarks like the Novartis, SAMPL6, and SAMPL7 datasets, outperforming other cheminformatics models (Table 4).

14.5.8. Thermodynamic consistency: glycine case tests

The ionization of glycine involves a thermodynamic cycle where the molecule can be deprotonated through two different paths. A fundamental thermodynamic relation dictates that the sum of pK_a values along one path must equal the sum along the other. For example, the ratio of the zwitterion form (2−HA) to the neutral form (1−HA) can be derived in two ways:

From the acidic ionization of the dicationic form, $1-\ H_2A^+$, giving $\frac{c(2-HA)}{c(1-HA)} = \frac{K_3}{K_1}$

From the basic ionization of the anionic form, $1-\ A^-$, giving $\frac{c(2-HA)}{c(1-HA)} = \frac{K_2}{K_4}$

The cyclic relation $pK_1 + pK_2 = pK_3 + pK_4$ must hold true. When tested with a graph-neural-network-based model, pK_a solver, which predicts each micro pK_a value independently, shows that pK_a solver's predicted values for glycine violate this fundamental principle, with its predicted $pK_1 + pK_2$ (8.00) not equaling $pK_3 + pK_4$ (13.64). This inconsistency leads to a more than five orders of magnitude difference when calculating the same quantity, $\frac{c(2-HA)}{c(1-HA)}$, from the two different pathways. In contrast, Uni− pK_a, by modeling the underlying free energies from which all pK_a values are derived, automatically satisfies the cyclic relation, reporting $pK_1 + pK_2 = pK_3 + pK_4$ (11.99). This allows consistency regardless of the pathway.

14.5.9. Comparison to quantum chemistry methods

Schrödinger's Jaguar, DFT-based software, often achieved the highest accuracy in pK_a prediction. A comparative case study on a family of tertiary amines demonstrates that Uni− pK_a can achieve an MAE of 0.52, outperforming Jaguar's accuracy of 0.69. However, accuracy challenges remain for certain complex systems, where Jaguar's accuracy improves significantly with exhaustive conformational modeling (from MAE = 1.07 down to 0.20), while Uni-pK_a cannot match (MAE = 0.70).

14.5.10. Limitations and future directions

By establishing free energy as a central pivot, this innovative framework is a shift from a direct correlational approach to a mechanistic, physics-informed model. However, the current framework may struggle with complex systems that exhibit subtle stereoelectronic effects, and the explicit handling of tautomerism is not yet fully integrated into the model, due to the scarcity of experimental data. Future research might benefit from combining these strengths by using first-principles training data from QC calculations to address the limitations of the ML model, particularly for challenging cases like stereoelectronic effects and tautomerism (Table 5).

14.6. AI-based prediction of organic reaction products using graph-convolutional neural networks

14.6.1. Background and motivation

While there is a rich history of using machine learning for structure−activity models in chemistry, it is only now that it is being successfully applied more broadly to organic synthesis and synthesis design. Traditional retrosynthetic software often relies on manually curated reaction templates that are brittle and limited in scope. A major obstacle has been the lack of reliable forward reaction prediction models, especially in the absence of contextual data (e.g., temperature, solvent), and publicly available reaction databases are biased towards high-yielding reactions, making it difficult to train models to classify unproductive reactions.[84] Coley et al. (2017) introduced a machine learning (ML) approach that augments template-based reaction enumeration with a neural network trained to rank the plausibility of candidate products for predicting organic reaction outcomes.

14.6.2. Methodological innovations

Coley et al. introduced a two-stage architecture for forward reaction prediction. The model first enumerates all chemically plausible product candidates via automatically extracted forward synthesis templates, and then ranks them using a neural network classifier that identifies the reaction core associated with the most likely reaction modes, and ranks them by the possibility of the major product.

14.6.3. Forward enumeration

Data was augmented by supplementing reaction databases with

Table 4
Performance of Uni− pK_a and selected models on various prediction tasks.

Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
	Novartis						SAMPL6		SAMPL7		SAMPL8	
	acid		base		total							
Schrödinger Epik Classical	0.990	1.531	0.876	1.175	0.830	1.160	0.784	0.962	1.121	1.648		
ChemAxon Marvin	0.808	1.144	0.835	1.145	0.860	1.170	1.007	1.248	0.559	0.708	1.300	1.511
ACD/Labs									0.550	0.783		
SPOC + XGBoost									0.767	1.011	1.476	1.622
SPOC + NN									0.832	1.141	0.932	1.156
MolG pK_a	0.849	1.287	0.789	1.064	0.870	1.270	0.522	0.773	0.797	0.980	0.835	1.150
Graph pK_a									0.594	0.726	0.758	0.934
pK_a solver							0.710	0.930				
MF−SuP− pK_a	0.850	1.090	0.610	0.790	0.710	0.920	0.687	0.751	0.656	0.816		
Schrödinger Epik v7										0.920		
Uni-pK_a	0.810	1.061	0.493	0.653	0.620	0.840	0.554	0.716	0.570	0.735	0.631	0.878

Table 5
Comparative summary of ai/ml models for free energy and kinetics prediction.

Model/Reference	Primary Methodology	Target Property	Key Performance
Chung et al. (2022)	Hybrid GC/ML (D-MPNN)	Abraham Solute Parameters, ΔG_{solv} ΔH_{solv}	Direct-ML ΔG_{solv} (random solute split): MAE 0.40 kcal/mol; ΔH_{solv} (random solute split): MAE 0.50 kcal/mol
Chung & Green (2024)	Hybrid QM/ML (COSMO-RS + D-MPNN)	$\Delta\Delta G_{solv}$ $\Delta\Delta H_{solv}$ Reaction Rate: $K_{relative}$	MAE/RMSE of 0.71/1.16 and 1.03/1.63 kcal/mol for the prediction of ($\Delta\Delta G_{solv}$, $\Delta\Delta H_{solv}$)
Scheen et al. (2020)	Hybrid FEP/ML (FEP + SVM/DNN)	Hydration Free Energies	MUE: as low as 0.55 kcal/mol; Outperformed standalone FEP
Yang et al. (2020)	Neural Networks (NN) and XGBoost models (SPOC descriptors, ISL)	pK_a diverse solvents	MAEs: 0.87 (NN), 0.88 (XGBoost); r^2: 0.95 (NN), 0.94 (XGBoost); External validations: SAMPL6 MAE = 0.80, pharmaceutical micro-pKa MAE = 0.44, Organocatalysts MAE = 1.01 Note: Unit in kcal/mol
Luo et al. (2024)	Machine Learning pK_a model based on thermodynamic principles	Micro and Macro pK_a prediction integrated on thermodynamic principles	MAE 0.62, RMSE 0.84 on Novartis tests, MAE around 0.59, RMSE 0.77 on SAMPL tests

chemically plausible negative reaction examples. This framework avoids relying on reaction yield data and leverages negative information by considering non-selected candidates as implicit negatives. From 1.1 million atom-mapped reactions in the USPTO database, 140,284 unique forward reaction templates were extracted, encoding core structural changes via reaction SMARTS strings. To balance coverage and computational efficiency, only the top 1,689 templates (those with ≥ 50 precedents) were retained. These were applied to all molecules in the reaction (reactants, reagents, solvents, and catalysts) to form a unified reactant pool. The templates generated an average of around 350 candidate products per reaction.

14.6.4. Candidate ranking

Three models were built and used to evaluate the scores of each candidate product, estimating its likelihood of being the major product.

(1) Baseline Model

A control model evaluates candidates based solely on product structure. Each candidate is represented by a radius-2 Morgan fingerprint (1024 bits). A shallow neural network (one hidden layer, tanh activation) maps this to a scalar score. This model disregards any information about reactants or transformation mechanisms.

(2) Edit-Based Model

The primary model captures mechanistic information through an edit-based representation of the reaction core. Each reaction is decomposed into four edit types: atom gain/loss of hydrogen, and bond formation/cleavage. Each edit is encoded with physicochemical features of the involved atoms and bonds (e.g., atomic number, hybridization, partial charge, TPSA), yielding vectors in $\mathbb{R}^{32}$ (for atoms) or $\mathbb{R}^{68}$ (for bonds). Separate subnetworks process each edit type via three dense layers, and the resulting latent vectors are summed and passed through a final network to output a scalar reactivity score. Scores are normalized across candidates using a softmax layer to approximate product distributions.

(3) Hybrid Model

This model combines the edit-based and baseline models by summing their scalar scores before softmax normalization. This fusion improves predictive accuracy by integrating both structural and mechanistic information.

14.6.5. Training

With 1,689 reaction templates, negative reaction data were used to augment the dataset, yielding a total of 5,335,669 reaction–product pairs for training and evaluation. The models were trained with 5-fold cross-validation using an optimizer with a 70/10/20 train/validation/test split in each fold.

14.6.6. Results

- The hybrid model achieved the best performance, assigning the major product to rank 1 in 71.8% of cases, rank ≤ 3 in 86.7% of cases, and rank ≤ 5 in 90.8% of cases under 5-fold cross-validation. The edit-based model significantly outperformed the baseline model (68.5% accuracy vs. 33.3%).
- Model confidence, evaluated via the categorical cross-entropy loss ($-\log p_{true}$), further distinguished performance: the edit-based model achieved a loss of 1.34, versus 3.28 for the baseline. The hybrid model yielded the lowest loss at 1.21, indicating not only greater accuracy but also greater certainty in predictions.
- Input feature analysis revealed that the model relies on a combination of functional (e.g., TPSA, Estate index, Gasteiger partial charge) and structural (e.g., atomic number, number of neighbors, aromaticity) descriptors. No single atom-level descriptor dominated prediction; instead, the model drew from both structural and functional descriptors (Table 6).

14.6.7. Qualitative predictions

In reactions involving typical functional group interconversions (e.g., chlorinations, amide formations, Grignard additions), the hybrid model accurately identified the true product with high confidence (often $> 90\%$) without contextual information (e.g., catalyst or solvent). Misranked predictions were usually associated with missing or misrepresented contextual features—such as steric hindrance, reagent-specific reactivity, or directing group effects—which are not fully captured by the edit-based featurization. For example, the model failed to correctly predict reactions where selectivity was governed by subtle steric shielding or reagent identity.

14.6.8. Limitations and future directions

The model uses overgeneralized reaction templates, with the top 1,689 templates covering a majority of reactions; novel transformations absent from the template library remain out of reach. The lack of

Table 6
Three models' performance on prediction summary.

Model	Top-1 Accuracy	Top-3	Top-5	Cross-Entropy Loss
Baseline	33.3%	48.2%	55.8%	3.28
Edit-Based	68.5%	84.8%	89.4%	1.34
Hybrid	71.8%	86.7%	90.8%	1.21

temperature, solvent, and time data restricts applicability to context–sensitive reactions. Molecule representation through strings allows mechanistic insight, but it may neglect information about non-core atoms, and subtle effects (e.g., steric, electronic) are only partially captured by atomic features. However, importantly, the integration of reaction-specific features (edits) with product-based features (fingerprints) yielded the best overall performance, emphasizing the complementary value of both approaches.

14.7. Prediction organic reaction outcome using graph-convolutional neural networks

14.7.1. Background

Coley et al. further present a supervised learning framework that leverages graph-convolutional neural networks (GCNNs) for the prediction of chemical reactivity and the major products of organic reactions based solely on the graph representations of reactants, reagents, and solvents, addressing the issue of minor structural differences at the reaction center being ignored in the traditional approach.[85] By adopting a two-stage architecture mirroring expert chemist reasoning, the model achieves remarkable accuracy in reaction product prediction.

14.7.2. Model architecture and representation

Coley et al. embed human-inspired reasoning directly into a graph-convolutional neural network (GCNN) architecture for reaction product prediction, facilitated by a global attention mechanism that reveals chemically consistent insights into reactivity. Broadly speaking, the model operates on atom-mapped reaction inputs and formulates reaction outcome prediction as a two-step process: (1) perceiving likely bond changes (potential reactive sites) and (2) ranking candidate products—mirroring the two-step workflow of expert chemists.

Perception of Reactive Sites: The first stage involves identifying potential sites within the reactant molecules that are most likely to change connectivity, achieved using a Weisfeiler-Lehman Network (WLN, a type of GCNN). The WLN takes the reactant molecules, represented as a single attributed graph (nodes as atoms, edges as bonds), as input. Atoms are featured by properties such as atomic number, formal charge, and aromaticity, while bonds are featured by bond order and ring status. The WLN iteratively updates atom-level representations by incorporating information from adjacent atoms and employs a global attention mechanism to account for long-range effects, including those from reagents. This network predicts likelihood scores for bond order changes.

Evaluation and Ranking of Candidate Outcomes: Focused enumeration of possible product candidates is performed, constrained by chemical valence rules. The second stage utilizes a Weisfeiler-Lehman Difference Network (WLDN) to refine the preliminary rankings of these candidate products. The WLDN embeds both reactants and candidate outcomes as attributed graphs. A numerical reaction representation is calculated based on differences in atom representations between the candidate and reactant atoms. This representation is then processed through a neural network, and its score is combined with the preliminary scores from the WLN to yield a final ranking. The model is trained to assign the highest score to the true (recorded) product.

Data: Trained on 480 k USPTO reactions (410 k train/30 k validation/40 k test).

14.7.3. Results and performance

The model's performance was rigorously evaluated on a test set of approximately 40,000 reactions from the USPTO database.

Predictive Accuracy: The model achieved a top-1 accuracy of 85.6%, a top-2 accuracy of 90.5%, and a top-3 accuracy of 92.8%. For rare reaction prediction performance (those with fewer than five precedents), the model achieved over 60% top-1 accuracy and showed a 5–6% improvement over prior models, such as template-matching methods and sequence-to-sequence translation networks (Table 7).

Table 7
GCNN model prediction performance compared with selected models.

Method	Size	Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]
WLN/WLDN	3.2 M	79.6	–	87.7	89.2
Sequence-to-sequence	30 M	80.3	84.7	86.2	87.5
This work	2.6 M	85.6	90.5	92.8	93.4

Computational Efficiency: Requiring approximately 100 ms per example on a single consumer GPU.

Human Benchmarking: In a study involving graduate students, postdocs, and faculty, the model's performance matched or exceeded human expert chemists in predicting the correct major products across a set of 80 reactions.

14.7.4. Interpretability and chemical insight

A key strength of this model is its interpretability: its global attention mechanism generates atom-level attention maps that reveal which atoms—through both their local context and long-range interactions—drive the predicted reactivity at a given site. In addition, the model predicts a probability distribution over multiple product species, allowing for the prediction of side products or impurities. These insights, while not providing full mechanistic rationales, reflect a qualitative chemical understanding purely from the data-driven approach.

14.8. Neural-symbolic machine learning for retrosynthesis and reaction prediction

14.8.1. Background

Segler and Waller (2017) introduced a framework that integrates hand-coded or algorithmically extracted reaction rules with deep neural networks to perform reaction prediction and retrosynthesis. This pioneering work introduced a neural-symbolic framework that addresses critical shortcomings of traditional rule-based systems, namely their inability to account for molecular context and to prioritize applicable transformation rules effectively.[86] Training on millions of reactions from the Reaxys database, their model learns to rank and select rules based on the full molecular environment, thereby resolving reactivity conflicts. The approach demonstrated top-10 accuracies of 98% (forward prediction) and 95% (retrosynthesis) using 8720 automatically extracted rules—far surpassing classical expert systems and logistic regression models.

14.8.2. Neural-symbolic integration methodology

Symbolic component: Hand-coded or algorithmically extracted reaction rules.

Neural component: Deep learning models prioritize rules by learning context-dependent applicability from reaction data.

14.8.3. Data curation and splitting

Reaxys database (1771–2015), filtered to single–product reactions with ≤ 3 reactants. Final sets contain 3 million reactions for hand-coded rules, and 4.9 million reactions for extracted rules (Table 8).

Table 8
Reaction rule generation methods.

Rule Type	Generation Method	Size	Filtering Criteria
Hand-coded	Expert-curated	103	N/A
Algorithmic	Shell-based scheme (reaction center + neighbors)	8,720	Frequency ≥50
High-frequency	Shell-based scheme (reaction center + neighbors)	137	Frequency ≥5,000

14.8.4. Molecular representation

Descriptor: Counted ECFP4 fingerprints (CDK 1.5.13) and summed fingerprint vectors for multi-reactant systems for order-invariant encoding.

14.8.5. Neural networks framework

Architectures: Fully Connected (FC512 ELU) NN and Highway Networks for multiclass classification.

Activation: Exponential Linear Unit (ELU).

Output Layer: Softmax for rule probability distribution.

Framework: Keras (Theano/TensorFlow backend).

14.8.6. Splitting and Training Protocol

Random 7:1:2 split (training/development/validation). Models were trained on the training set and tested on the development set, with hyperparameters optimized. The two best-performing models were then tested on the validation set.

14.8.7. Key results and performance

Best trained neural-symbolic models, a neural network with one hidden layer (FC512 ELU), a logistic regression model, and a deep highway network, were compared to a purely rule-based expert system operating with the same rule set. Their performance was evaluated based on the rules used.

- **Top-n accuracy**: Correct rule ranked among the top n predictions
- **Mean Reciprocal Rank (MRR)**: Measures ranking quality (1.0 = perfect)
- **Weighted Precision (W. Prec.)**: Minimizes false positives (Table 9 and 10)

Moreover, the authors find that a model with 137 algorithmically extracted rules performed surprisingly better than the system with 103 hand-coded rules.

14.8.8. Advantages and limitations

The neural-symbolic model exhibited a remarkable speed advantage, operating approximately 150 times faster than the rule-based expert system, and achieved an accuracy of over 95% in predicting reaction outcome. Despite high performance, limitations include the reliance on predefined reaction rules (limiting extrapolation to novel chemistries) and a lack of stereochemical awareness.

14.9. Planning chemical syntheses with deep neural networks and symbolic AI

14.9.1. Background

Inspired by their previous work, Segler and colleagues introduced a more advanced system, 3 N-MCTS, by combining Monte Carlo tree search (MCTS) with symbolic artificial intelligence, enhanced by deep neural networks.[86] The methodology employs two key neural networks:

Table 9
Performance on 103 hand-coded rules.

Task/Model	Acc	Top 3-Acc	MRR	W. Prec.
Reaction Prediction				
Random	0.03	0.12	0.04	0.03
Expert System	0.07	0.33	0.12	0.46
Logistic Regression	0.86	0.97	0.91	0.86
Highway Network	0.92	0.99	0.96	0.92
FC512 ELU	0.92	0.99	0.96	0.92
Retrosynthesis				
Random	0.03	0.12	0.04	0.03
Expert System	0.05	0.30	0.06	0.11
Logistic Regression	0.64	0.95	0.77	0.62
Highway Network	0.77	0.98	0.86	0.77
FC512 ELU	0.78	0.98	0.87	0.78

Table 10
Performance on 8,720 extracted rules.

Task/Model	Acc	Top 10-Acc	MRR	W. Prec.
Reaction Prediction				
Random	0.00	0.00	0.00	0.00
Expert System	0.02	0.18	0.02	0.06
Logistic Regression	0.41	0.65	0.49	0.31
Highway Network	0.78	0.98	0.86	0.77
FC512 ELU	0.77	0.97	0.85	0.76
Retrosynthesis				
Random	0.00	0.00	0.00	0.00
Expert System	0.02	0.19	0.01	0.06
Logistic Regression	0.31	0.59	0.41	0.23
Highway Network	0.64	0.95	0.75	0.63
FC512 ELU	0.62	0.94	0.74	0.62

an expansion policy network to guide the search and a filter network to pre-select the most promising retrosynthetic steps. These networks were trained on a comprehensive dataset encompassing nearly all published organic chemistry reactions, enabling the system to learn complex patterns and prioritize effective synthetic routes. The integration of MCTS allows for efficient exploration of the reaction space.

14.9.2. Methodology

The 3 N-MCTS system integrates MCTS, a search technique well-suited for sequential decision problems in which random search steps are performed until a solution or a maximum depth is reached, with three specialized deep neural networks for retrosynthesis analysis.

14.9.3. Data set and rule extraction

The network guides the MCTS search, having been trained on 12.4 million single-step reactions from the Reaxys database (pre-2015 reactions for training, and post-2015 reactions for testing). It learns to prioritize transformations, capture functional group tolerance, and apply extracted transformation rules.

14.9.4. Policy networks (3 neural networks)

Expansion Rules: Only reaction center atoms/bonds, with ≥3 occurrences of reactions. A total of 301,671 rules (79% coverage of post-2015 reactions) were included. Employed a deep highway network with exponential linear unit nonlinearities.

Rollout Rules: Include atoms/bonds in the reaction center and first-degree neighbors, with≥ 50 occurrences of reactions. A total of 17,134 rules (52% coverage of post-2015 reactions) were included. A neural network with one hidden layer is trained in the same way as the expansion policy (Table 11).

In-Scopes Filter Network: A deep neural network trained as a binary classifier, designed to predict the feasibility of proposed transformations. Reaction databases lack "failed reaction" data, and this network was trained on a massive dataset of generated "negative" examples using an implicit data augmentation method.

14.9.5. MCTS integration pipeline

Workflow for Node Expansion.

1. **Fingerprint Encoding:** Convert molecule to ECFP4 Fingerprint Encoding.
2. **Selection**: Sequentially selects the most promising position within the tree until a leaf node is reached. The algorithm balances the selection of high-value positions and unexplored positions. If a leaf

Table 11
Policy network summary.

Policy Type	Architecture	Rule Set Size	Inference Time
Expansion	5-layer Highway Network (ELU)	301,671	90 ms
Rollout	1 Hidden Layer (ELU)	17,134	10 ms

node is visited for the first time, it proceeds to rollout. If it is visited for the second time, it is expanded by the expansion policy.
3. **Expansion:** Generates precursor options to identify the top 50 transformations, filtering these precursors through the in-scope network to retain only feasible reactions, and add valid candidates as child nodes.
4. **Rollout:** Deploys the fast rollout policy for random decomposition of unsolved nodes until reaching building blocks or the 25-step depth limit, assigning rewards (+1 for complete solve, partial for incomplete, −1 for failure).
5. **Update:** Backpropagates results by rollout rewards, terminates upon reaching time/iteration limits, traces the highest-value path from root to terminal nodes, and returns either the solved route or marks failure if maximum depth is exceeded.

The MCTS algorithm iteratively refines the search tree through four distinct phases: selection, expansion (generating new nodes via the expansion policy and in-scope filter), rollout, and update, forming the 3 N-MCTS network for retrosynthesis prediction.

14.9.6. Performance

Quantitative Evaluation: The 3 N-MCTS system was evaluated against baseline methods using 497 molecules first reported from the post-2015 Reaxys reactions test set. Compared to heuristic BFS (using hand-coded SMILES cost functions) and neural BFS (policy-network cost), 3 N-MCTS demonstrated superior efficiency and coverage, solving 80% of targets within 5 s and 92% within 60 s—outperforming neural BFS (40% and 71%) and heuristic BFS (0% and 4%) decisively. Further studies revealed that all MCTS components contributed critically to its success: Monte Carlo search (without tree building) solved only 89.54%, while UCT—an MCTS variant ignoring policy probabilities—achieved 87.12%, and neural BFS reached just 84.24%, confirming the necessity of tree search, policy-guided selection, and rollouts. The complete 3 N-MCTS system achieved peak performance (95.24% of targets solved, 13.0 s per molecule), demonstrating a speedup and nearly doubled coverage versus traditional methods (Table 12).

Qualitative Evaluation (Double-Blind AB Tests): In a rigorous double-blind study involving 45 organic chemists, routes generated by 3 N-MCTS were considered equivalent to reported literature routes, with no significant preference for the latter (57.0% preference for MCTS routes vs. 43.0% for literature routes). Furthermore, chemists significantly preferred 3 N-MCTS-generated routes (68.2%) over those from a baseline heuristic BFS system (31.8%).

Potential Emergent Chemical Intelligence: The in-scope filter network rediscovers fundamental electronic properties based solely on the training data set.

- LUMO correlations in Diels–Alder reactions ($r^2 = 0.74$)
- Hammond parameters in electrophilic substitutions ($r^2 = 0.78$)

Limitations: Reliance on a training dataset of reactions with multiple occurrences limits the system's ability to handle novel or underrepresented reactions, such as natural product synthesis, due to data sparsity. Notably, this system, in its current form, does not perform reaction condition prediction, focusing solely on the retrosynthetic pathway itself. It ignores reaction equilibria, tautomers and 3D conformations, since it lacks enantioselectivity data for prediction.

Table 12
Comparative Performance of 3 N-MCTS Against a few Baselines models.

Method	Policy	Solved (%)	Time/Molecule (s)
3 N-MCTS	Neural	95.24	13.0
Monte Carlo	Neural	89.54	275.7
UCT	Neural	87.12	30.4
Neural BFS	SMILE3/23/2	84.24	39.1
Heuristic BFS	SMILE5/25/2	55.53	422.1

14.9.7. Comparative analysis

The 2017 paper established the foundational concept of neural networks learning chemical context and prioritizing outcomes, while the 2018 paper scaled this up to a 3 N-MCTS system, further enhancing its performance. These two works by Segler, Preuss, and Waller collectively underscore the transformative potential of integrating deep learning and AI methods with advanced search algorithms in synthesis prediction. However, challenges remain, particularly in handling truly novel reactions, stereochemistry, and integrating a complete mechanistic understanding.

14.10. Predict organic chemical reactions by machine learning based molecular orbital reaction model

14.10.1. Background

Determining the major products of chemical reactions by existing computational approaches falls into three main categories: quantum mechanical (QM) simulations, rule-based expert systems, and machine learning (ML) methods. QM simulations are accurate, but impractical for large-scale prediction. Rule-based systems suffer from limited scalability and generalizability. ML approaches have historically been hindered by poor data quality and a lack of mechanistic detail. Baldi et al. (2011) introduced a novel hybrid framework that integrates concepts from quantum mechanics, rule-based systems, and machine learning. Representing molecules using graph-based idealized molecular orbitals (MOs) and modeling reactions as interactions between electron donors and acceptors enables the enumeration and ranking of elementary steps.[87] This two-stage ML pipeline—filtering unreactive sites and ranking productive mechanistic reactions—demonstrated unprecedented performance and generalizability, representing a powerful balance between physical realism and computational efficiency.

14.10.2. Methodology

The system's foundation is an MO-based reaction unit model that enumerates primary idealized MO interactions. A molecule m is a labeled molecular graph G_m (A_m, B_m). Each atom is augmented with labels for approximate electron-filled and electron-unfilled MOs. An electron-filled MO (donor)/electron-unfilled MO (acceptor) is represented as: $F/U = (\mathrm{a}, t_f/\, t_u, n_f/\, n_u, c_f/\, c_u)$. Note: a: central atom, $t \in \{\mathrm{n}, \pi, \sigma\}$: orbital type (lone pairs, double/triple bonds, single bonds). n: neighbor atom (possibly null) for bond orbital. c: recursively defined adjacent chained filled orbital (possibly null). A chemical reaction is the interaction between a single filled MO and a single unfilled MO, where product generation (altering charges/bonds) uses a recursive algorithm.

14.10.3. Two-stage machine learning approach

Chemical reaction prediction is framed as identifying productive reactions for a query $(r|c)$, where r refers to the reactants and c refers to the condition.

14.10.3.1. Stage 1: reactive site filtering. $I_r = (F_{m1} \cup F_{m2}) \times (U_{m1} \cup U_{m2})$ defined as set of interactions, where F_m is the set of all electron-filled Mos, U_m is the set of all electron-unfilled Mos. Productive reactions are denoted as $I^{+}_{(r|c)}$, and unproductive reactions as $I^{-}_{(r|c)}$. Manage combinatorial complexity by filtering unreactive filled and unfilled MOs at the atom level. An atom in molecule m under conditions c is labeled "filled reactive" if it's the central atom of any MO in a productive filled orbital interaction; otherwise, it's "filled unreactive." Similar labels apply to unfilled MOs. Separate classifiers are trained for these labels to filter out most unreactive atoms while preventing productive reactions from being discarded. This ensures $I^{+}_{(r|c)}$ remains intact while drastically reducing $I^{-}_{(r|c)}$.

14.10.3.2. Stage 2: orbital interaction ranking. After filtering, identifying productive reactions from the reduced set is formulated as a statistical

ranking problem. A pairwise formulation is used, achieved by training artificial neural networks on ordered pairs of productive and unproductive orbital interactions, learning a relative productivity function $(i^+|i^-) \in I^+_{(r|c)} \times I^-_{(r|c)}$.

14.10.4. Data set construction and representation

Reaction Explorer was used as a dataset that covers fundamental organic chemistry with over 1500 elementary rules in 80 reagent models. Each rule describes a single mechanistic step via SMIRKS patterns and electron flow, and also includes over 4000 multistep "test case" syntheses. Data derivation involved inputting 1630 multistep test cases, focusing on general ionic reactions with C, H, N, O, Li, and halides from 31 Reaction Explorer reagent models. Processing these cases yielded mechanistic steps, each considered a productive elementary reaction. Reaction conditions (temperature, cation/anion solvation potential, 0–1 scale) were assigned to reactions from each reagent model, quantitatively mapping qualitative solvent ideas (e.g., polar protic, nonpolar). Grouping favorable orbital interactions over the same reactants and conditions formed the set of productive interactions $I^+_{(r|c)}$. In addition, with labeled MO interactions $(r|c)$ queries, each possible atom and conditions $(a|c)$ are provided.

14.10.5. Model architecture

Two separate classifiers are trained to predict whether an atom is reactive (filled) or non-reactive (unfilled), using the same features and ML methods but different labels. Each data point is represented as $\langle a|c|l\rangle$ (a, c, l): atom a, conditions c, and labeled l (1 if reactive, 0 if not).

14.10.6. Reactive site ML classifier features representation

Reaction Conditions: Temperature, cation/anion solvation potential (real-valued). Atom and Neighborhood Features (14 real-valued): Formal/partial charges (atom, 1–2 bond neighborhoods), exponentially decaying sum of atomic radii, molecular weight, and smallest ring size. Neighborhood Fingerprints (Count Features): Topological depiction around the atom. Standard Molecular Graph (743 features): Counts of paths (depth 3) and subtrees (depth 2, out-degree ≤ 2) rooted at the atom. Pharmacophore Point Graph (759 features): Molecular graph mapped to an isomorphic graph with restricted labels grouping chemical motifs by reactivity. Paths/subtrees computed to depth 6. Dimensionality Management: Features with a nonzero in < 25 data points disregarded. Total Features: 1516.

14.10.7. Reactive site filtering machine learning algorithms and architectures

ANNs were trained for filled and unfilled reactivity labels.

Algorithm and Architecture: Artificial Neural Networks with a single hidden layer with sigmoid activation, a single output node, and 10 hidden nodes.

Training Protocol: Features were normalized to [0,1], and gradients were computed via standard back-propagation with L2-regularized cross-entropy error. The final weights were optimized by stochastic gradient descent with per-weight adaptive learning rates, and optimization stopped after 100 epochs.

Decision Threshold Fitting: Model was fitted using internal cross-validation to achieve a False Negative Rate (FNR) of 0, which prioritizes sensitivity.

Validation: 10 repeats of 10-fold cross-validation over distinct $(m|c)$ tuples.

The result is a neural network model which, given an input feature vector, outputs a probability of the $(a|c)$ tuple being labeled reactive.

14.10.8. Orbital interaction ranking ML classifier features representation

Reaction Conditions: Temperature, cation/anion solvation potential (real-valued). Molecular Difference Features (124 features): Reactant features subtracted from product features (counts of bonds, charges, rings, aromaticity, rotatable bonds), indicating "created" or "destroyed" features. Forward Filled and Unfilled MO Features (251 features): Orbital type, chained orbital type, involved atoms, formal/partial charges, exponentially decaying sum of atomic radii around orbital. Inverse Filled and Unfilled MO Features (199 features): Same orbital features for inverse MOs (products yielding reactants) to account for reversibility. Reactive Site Features (Subset, 1103 features): Nonredundant real-valued features and pharmacophore point graph features (path depth 3, tree depth 2, π depth 4). Total Features: 1677.

14.10.9. Orbital interaction ranking machine learning algorithms and architectures

A pairwise approach using shared-weight ANNs was employed.

Algorithm and Architecture: Each network had a single sigmoidal hidden layer and a linear output node. Outputs were tied to a sigmoidal output layer with fixed weights $(+1, -1)$ to ensure consistent scoring, with 20 hidden nodes.

Training Protocol: Features normalized to [0,1]. Five subsplits of each training set were created (all productive interactions, random partition of unproductive). Final test predictions averaged from five models. Training used simultaneous forward propagation, L2 regularized cross-entropy error, and weight gradients via back-propagation (left then right), with the same updates applied to both. Optimization stopped after 25 epochs.

Validation: 10 repeats of 10-fold cross-validation over $(r|c)$ tuples.

14.10.10. Performance and results

The performance of the model was evaluated using two primary metrics for ranking quality: Normalized Discounted Cumulative Gain (NDCG) and percent within n.

- NDCG Perfect Ranking: 89.5% of the queries, the top-ranked reaction was a correct (productive) one.
- Percent Within n: The system recovered all productive reactions with at most four incorrect reactions ranked higher (percent within 4) in 99.86% of cases.

14.10.11. Generalization

The system demonstrates an ability to make reasonable predictions for reactions it has not been explicitly trained on and shows successful multistep reaction prediction. Demonstrated by a successful three-step Claisen condensation prediction, which is a test case the model had not seen during training. When tested on the formation of seven and eight membered ring reactions for which the original rule-based system had no rules, the model was able to generalize and predict plausible mechanisms.

14.10.12. Flexibility for alternative routes

Providing a ranked list of outcomes allows the identification of other plausible, though less favorable, reaction pathways. For example, in several ring–forming reactions, the model correctly identified the most productive reaction and also correctly ranked a reasonable alternative mechanism as the second-best option.

14.10.13. Errors

Most errors are close and chemically reasonable—often plausible mechanisms that differ only because the ground truth excludes intermediates or reversible steps. Some misranking, such as preferring a five-member ring over a six-member ring product, reflects partial learning of carbocation stability but not subsequent rearrangements. Besides, reaction conditions were found to have little impact on model performance.

14.10.14. Advantage and future direction

The novel approach, which combines a physically inspired molecular orbital (MO) model with a two-stage machine learning ranking system,

is a successful concept. It moves away from the limitations of traditional rule-based expert systems, which require extensive human effort to create and maintain. The system achieved strong performance (89.05% perfect ranking, 99.86% accuracy for top 4), and strong generalizability when predicting multistep syntheses. The paper lays the groundwork for future enhancements, including extending the model to handle radicals, pericyclic reactions, and stereochemistry.

14.11. Machine learning for organic reaction condition prediction and optimization: a comprehensive approach

14.11.1. Background

Prior computational efforts were fragmented, addressing only isolated elements (e.g., solvents) or specific reaction classes, and lacked large-scale quantitative evaluation. Gao et al. (2018) addressed these limitations by creating a neural network model for predicting complete reaction conditions for significant and arbitrary organic transformations, specifically a hierarchical neural network—including catalyst(s), solvent(s), and reagent(s)—along with the optimal temperature for a given organic reaction, trained on 10 million reactions from the Reaxys database.[88]

14.11.2. Methodology and model design

The model utilizes a sequential multi-output architecture, enabling information from earlier predictions to inform subsequent ones. This design captures the inherent interdependence among reaction conditions. A feedforward neural network is employed to predict up to five chemical components—catalyst, two solvents, and two reagents—alongside the reaction temperature. Each chemical component is predicted as a classification task over a predefined set of species, while the temperature is handled as a regression problem. However, pressure and concentration were ignored due to data limitations.

14.11.3. Data curation

Source: 53 M reactions from Reaxys. Retained only single-step, single–product reactions (excluded multistep/multiproduct entries). Removed reactions with unparsable structures (via RDKit), missing conditions, or temperatures outside the range of −100 °C to 500 °C. Applied a minimum frequency threshold (more than 100 occurrences) to chemical species, significantly reducing the number of distinct classes for solvents (from 11,246 to 232), reagents (from 151,214 to 2,247), and catalysts (from 10,323 to 803). Transition-metal-containing reagents were reclassified as catalysts.

Final Dataset: 11.4 M high–quality reactions.

14.11.4. Molecular representation

Reactants/Products: Encoded as Morgan fingerprints (radius 2; 16,384 bits), capturing structural features.

Chemical Context: Catalysts/solvents/reagents represented as one-hot vectors (fixed-class classification). A NULL class corresponding to a missing element is not recorded.

Reaction Fingerprint: Difference between product and reactant fingerprints, highlighting transformed substructures.

14.11.5. Neural network model architecture

Step 1: Dense Fingerprint Representation (Dense FP): The reaction fingerprint and product fingerprint, each comprising 16,384 bits, are concatenated to form a 32,768-bit input vector. This combined vector is then fed into the initial layers of the network. These inputs pass through two fully connected (FC) layers, both employing ReLU activation functions and each consisting of 1000 neurons. The output of these layers constitutes a dense, learned representation of the fingerprints, referred to as Dense FP.

Step 2: Catalyst Prediction: The Dense FP is subsequently passed through two additional FC layers. The first uses a ReLU activation (with 300 neurons), and the second employs a Softmax activation function. This final layer has 803 output neurons, corresponding to the number of distinct catalyst classes. This step predicts the most suitable catalyst, or its absence, for the given reaction.

Step 3: First Solvent Prediction: The one-hot vector representing the predicted catalyst from Step 2 is concatenated with the Dense FP. This combined vector then serves as the input to two more FC layers (ReLU, 300 neurons; Softmax, 232 neurons for the first solvent classes). This sequential integration is crucial, allowing the model's catalyst choice to inform and influence the subsequent solvent prediction directly.

Step 4: Subsequent Chemical Context Predictions (Second Solvent, First Reagent, Second Reagent): Step 3 is iteratively repeated for the prediction of the second solvent, the first reagent, and the second reagent. For each subsequent prediction, the one-hot vector of the most recently predicted chemical species (e.g., the first solvent for the second solvent prediction, or the second solvent for the first reagent prediction) is concatenated with the Dense FP and then passed through dedicated FC layers with Softmax outputs. The number of output classes varies for each of these predictions (second solvent: 228; first reagent: 2240; second reagent: 1979), reflecting the specific chemical species present in each category within the filtered dataset.

Step 5: Temperature Prediction: In the final step, all the one-hot vectors corresponding to the predicted chemical context elements (catalyst, first solvent, second solvent, first reagent, and second reagent) are concatenated together with the initial Dense FP. This comprehensive vector is then fed into two FC layers (ReLU, 300 neurons; Linear, 1 neuron) to predict the continuous temperature value for the reaction.

14.11.6. Training Protocol

Split: 80/10/10 (train/validation/test).

Optimization: Training stopped after five epochs of no validation loss improvement.

14.11.7. Predictive performance

The model was tested for performance on a test set of approximately 1 million reactions. Accuracy was assessed by determining if the proper conditions are within the top-N predictions and for temperature prediction within a specific range for the quantitative test. Qualitative insights were also provided for the model's predictive capabilities.

14.11.8. Accuracy metrics for chemical context prediction (catalyst, solvents, reagents)

Close match predictions were defined based on chemical similarity: for solvents, this involved the Euclidean distance between their Abraham parameters (e.g. methanol–ethanol pair for similarity); for catalysts and reagents, close matches were considered if they shared the same metal atoms (for organometallic compounds) or possessed identical feature-based Morgan fingerprints (based on RDKit's feature definitions, such as halogen or hydrogen bond donor).

Individual Matches: The model exhibited high accuracy for predicting individual components. For catalysts (C), a top-10 close match accuracy of 96.4% was achieved. For the first solvent (s1), it was 85.4%, and for the second solvent (s2), 91.9%. Similarly, for the first reagent ($r1$), the accuracy was 84.9%, and for the second reagent ($r2$), it was 92.1%.

Combined Chemical Context Matches: When predicting the combination of catalyst, first solvent, and first reagent (c, s1, $r1$), the model achieved a top-10 close match accuracy of 69.6%. For the full condition recommendation, encompassing catalyst, two solvents, and two reagents (c, s1, s2, $r1$, $r2$), the top-10 close match accuracy was 60.3% (Table 13).

Table 13
Outcome Prediction result summary of the comprehensive model.

Prediction Task	Top-3 Exact Matches (%)	Top-10 Exact Matches (%)	Top-3 Close Matches (%)	Top-10 Close Matches (%)
Catalyst (C)	93.6	94.9	94.9	96.4
Solvent 1	75.8	83.0	78.2	85.4
Solvent 2	90.1	91.7	90.2	91.9
Reagent 1	73.2	83.1	74.8	84.9
Reagent 2	89.3	91.8	89.3	92.1
C, s1, *r*1	57.3	66.0	60.4	69.6
C, s1, s2, *r*1, *r*2	50.1	57.3	53.2	60.3

14.11.9. Performance evaluation for temperature prediction

Prediction Accuracy: The top-one predicted temperature fell within a ±10 °C range of the recorded temperature in 36.7% of all test cases. This accuracy increased to 57.7% for predictions within a ±20 °C range. Notably, temperature prediction improved when the predicted chemical context precisely matched the recorded chemical context. In such cases, the accuracy within ±10 °C rose to 42.6%, and within ±20 °C, it reached 65.9%. The mean absolute error (MAE) for temperature prediction across all reactions in the dataset was 25.5 °C. When the correct chemical context was identified, the MAE for temperature prediction improved to 19.4 °C (Table 14).

14.11.10. Comparative analysis with baseline methods and demonstrated improvements

A baseline model that selected the top-10 combinations based solely on the frequencies of individual catalysts, solvents, and reagents performed significantly worse. For instance, the top-3 accuracy for predicting the correct combination of chemical context in the trained model was 50.1%, whereas the frequency-based baseline model achieved a mere 4.7%. A baseline model that predicted the most frequent temperature (20 °C) for every reaction resulted in a mean absolute error of 35.3 °C, which is larger than the 25.5 °C MAE achieved by the trained model (Table 15).

Efficiency: The model completes predictions in less than 100 ms per reaction, making it over 10,000 times faster than nearest-neighbor methods.

14.11.11. Qualitative insights and Chemical Embeddings

The authors selected representative reactions from common organic reaction types, including hydrolysis, esterification, alkylation, epoxidation, Wittig, reduction, oxidation, deprotection, Suzuki–Miyaura coupling, Grubbs metathesis, and Buchwald-Hartwig amination, among others. For instance, it successfully recommends functionally equivalent alternatives (e.g., piperidine in place of morpholine) and differentiates between acidic vs. basic deprotection strategies. However, poorly predicted cases exist. The proposed contexts are not close to the published procedures and might only be “chemically reasonable” for many cases.

14.11.12. Implicit learning-solvent and reagent embeddings

Moreover, the model implicitly learns solvent and reagent functional similarity without explicit chemical property inputs. Using t-SNE projections, solvents with similar physicochemical roles (e.g., polar aprotic, nonpolar, protic) cluster naturally. Similarly, reagent embeddings cluster by functional class, with incompatible reagents (e.g., reductants, oxidants, acids, and bases) distinctly segregated, indicating the model's capacity to learn solvent and reagent functionality.

Table 14
Temperature Prediction result summary of the comprehensive model.

Temperature Prediction Metric	Performance (All Reactions)	Performance (Correct Chemical Context)	Baseline Model Value
Top-1 within ±10∘C	36.7%	42.6%	N/A
Top-1 within ±20∘C	57.7%	65.9%	N/A
MAE	25.5 °C	19.4 °C	35.3 °C

Table 15
Comparative prediction Performance with Baseline Methods.

Baseline Comparison Metric	Trained Model Value	Baseline Model Value
Top-3 Accuracy (C, s1, *r*1)	50.1%	4.7%

Strengths, Limitations, and Outlook: The holistic model offers broad coverage, comprehensive and interdependent prediction capabilities to suggest specific chemical reaction conditions with learned functional similarity, and demonstrates exceptional speed. The model does not rely on explicit mechanistic or physicochemical features of the chemicals but instead learns functional similarity embeddings from the dataset alone, demonstrating the potential of utilizing high-dimensional data-driven representations. Besides the strengths of the model, limitations still exist: the model still relies on curated reaction data and may suffer from biases or gaps inherent in Reaxys. Predictions for rare or multistep reactions remain untouched; the model does not process pressure or concentration information and cannot handle stereochemistry details, due to the data limitation issue. In addition, the accuracy is only moderate as a comprehensive model when compared with models designed for specific tasks, and reaction mechanism details are beyond the current model's scope (Table 16).

14.11.13. Advantage, drawback, and future directions

Scalability and Computational Efficiency: The ability of machine learning to process and learn from vast chemical data and operate at scale is a key differentiator from conventional computational methods. Neural-symbolic frameworks operate approximately 150 times faster than rule-based expert systems. At the same time, reaction condition prediction models complete predictions in less than 100 ms per reaction—over 10,000 times faster than traditional methods. Graph-convolutional neural networks (GCNNs) for reaction outcome prediction require only approximately 100 ms per example on consumer -grade hardware. This scalability is crucial for exploring expansive chemical spaces effectively.[89]

Predictive Accuracy: Machine learning models, particularly those leveraging deep learning architectures like Deep Neural Networks, consistently achieve superior accuracy in prediction, matching, and even surpassing traditional computational chemistry methods in some cases. However, the performance is strongly influenced by the training dataset. There are still many instances where ML models fall behind in accuracy compared to quantum chemical calculations or *ab initio* methods. For example, the Direct-ML model, as presented by Chung et al. (2022), demonstrated a Mean Absolute Error (MAE) of 0.40 kcal/mol for solvation free energy (ΔG_{solv}) in random solute splits and even matches quantum chemistry-based methods like COSMO-RS calculations. The Hybrid Free Energy Perturbation (FEP)/Machine-Learning (FEP/ML) methodology developed by Scheen et al. (2020) significantly improved upon standalone FEP calculations for hydration free energies, achieving a Mean Unsigned Error (MUE) as low as 0.55 kcal/mol, a substantial improvement over the 1.16 kcal/mol MUE observed for FEP alone. Graph-convolutional neural networks achieve top-1 accuracies of 85.6% for reaction outcome prediction. Neural-symbolic frameworks demonstrate top-10 accuracies of 98% for forward prediction and 95% for retrosynthesis. These reaction predictions match human expert levels on blind tests. Furthermore, hierarchical neural networks, as developed by Gao et al. (2018), predict comprehensive reaction conditions—including catalyst(s), solvent(s), reagent(s), and optimal temperature—interdependently achieving high “close match” accuracies (around 69.6% for the combination of catalyst, first solvent, and first reagent in the top-10 predictions). For large molecular systems or when extensive conformational sampling is required, a machine learning method addresses this bottleneck by offering a favorable trade-off between accuracy and computational cost.

Table 16
Key ai/ml advancements in organic synthesis analysis.

Model	Primary Task	Approach	Data Source	Key Performance
Coley et al. (2017)	Reaction Outcome Prediction	Neural Network (template-based)	1,689 templates (5.3 M reaction pairs)	Top-1 accuracy: 71.8%; Top-3: 86.7%; Top-5: 90.8%;
Coley et al. (2019)	Reaction Outcome Prediction	GCNNs (WLN, WLDN)	410 k USPTO reactions	Top-1 accuracy: 85.6%; Top-3: 92.8%; Top-5: 93.4%; >60% accuracy on rare reactions;
Segler & Waller (2017)	Retrosynthesis & Forward Reaction Prediction	Neural-Symbolic Framework (DNN + Rules)	3.5 M Reaxys reactions	Top-10 accuracy: 98% (forward), 95% (retrosynthesis);
Segler, Preuss, & Waller (2018)	Retrosynthesis	MCTS + Multiple DNNs (3 N-MCTS)	12.4 M Reaxys reactions	Chemists' double-blind study (57% pref. for MCTS over literature)
Baldi et al.	Reaction Outcome Prediction	MO theory + ANNs	Reaction Explorer 1500 rules	89.05% perfect ranking, 99.86% accuracy for top 4
Gao et al. (2018)	Reaction Condition Prediction	Hierarchical Neural Network	10 M Reaxys reactions	Top-10 close match accuracy: 69.6% (C, S1, *r*1); Individual: 96.4% (C), 85.4% (S1), 84.9% (*r*1); Temp within ±20 °C: 57.7%;

Generalizability: Models trained on millions of reactions from comprehensive databases like Reaxys and USPTO show robust performance on previously unseen chemical transformations. The holistic pKa prediction model achieves consistent accuracy across 39 different solvents.[82] At the same time, retrosynthetic planning systems successfully generate routes for molecules containing up to 22 heavy atoms—substantially larger than their training examples.[86] Overall, models exhibit acceptable scalability and generalizability, tested on unseen datasets. However, due to the data-driven nature of ML models, ML methods are generally incapable of inferring reactions that differ significantly from the training set.

Data-Driven Discovery and Interpretability of Chemical Insight: While some models are characterized as "black-box," others offer valuable interpretability. For example, Coley et al. (2019) demonstrated that their Graph-Convolutional Neural Network (GCNN) model, equipped with a global attention mechanism, generates atom-level attention maps. These maps reveal which specific atoms drive reactivity, reflecting a qualitative chemical understanding derived purely from a data-driven approach. The 3 N-MCTS system's in-scope filter network rediscovers LUMO correlations in Diels–Alder reactions ($r^2 = 0.74$) and Hammond parameters in electrophilic substitutions ($r^2 = 0.78$) purely from reaction data. Furthermore, Gao et al. (2018) showed that their model for reaction condition prediction implicitly learns solvent and reagent functional similarity (e.g., polar aprotic, nonpolar, protic) through dense embeddings. The move towards "chemically aware" and interpretable AI models is crucial not only for user acceptance but also for enabling AI to contribute to fundamental chemical theory and discovery, extending beyond mere correlation to reveal underlying causal relationships.[90]

14.11.14. Current drawbacks and limitations

Data Quality and Availability Constraints: A fundamental prerequisite for the successful application of AI/ML in chemistry is the availability of reliable data, which must be derived from meticulously performed experimental measurements or highly accurate theoretical calculations. The quality, quantity, and diversity of the collected data directly influence the model's performance.[91] Current models suffer from inherent biases present in reaction databases, which are skewed toward high-yielding, well-established transformation rules. This limits model performance on novel or underrepresented reactions, particularly in specialized areas like natural product synthesis, where limited data is available. The scarcity of failed-negative reaction examples hampers the development of models that can reliably predict reaction feasibility. Furthermore, inconsistencies in experimental conditions, measurement protocols, and data curation across different sources introduce systematic errors that propagate through model predictions.

Interpretability and Black Box Nature: Despite advances in attention mechanisms and interpretable architectures, many AI/ML models remain "black boxes" that provide limited chemical insight into their decision-making processes. While some models generate attention maps or feature importance rankings, most often fail to provide the mechanistic understanding that chemists require for rational synthetic design.[92,93]

Restricted Chemical Space Coverage: Indicated by the Interpretability and Black Box Nature, AI/ML models demonstrate excellent interpolative capabilities within their training domains. Still, extrapolation to truly novel chemical spaces remains challenging.[94] Models trained on specific reaction types or molecular size ranges often fail when applied to significantly larger molecules, exotic functional groups, or unprecedented reaction mechanisms. The reliance on predefined reaction templates in many systems, while providing initial high accuracy, limits their ability to predict entirely novel transformations that fall outside established chemical knowledge.[84,86]

Limited Stereochemical and Mechanistic Understanding: Stereochemistry is a fundamental aspect of organic chemistry, dictating molecular properties, biological activity, and reactivity.[95] Most AI/ML model approaches lack explicit stereochemical awareness, failing to predict enantioselectivity or diastereoselectivity in reactions where these factors are critical. Models' understanding of reaction mechanisms remains largely implicit and statistical, rather than based on fundamental chemical principles.[96] This limitation becomes particularly pronounced in reactions where steric hindrance, electronic effects, or catalyst–substrate interactions govern selectivity. Addressing this limitation requires moving beyond 2D graph representations to incorporate explicit 3D information.

Context Sensitivity and Experimental Factors: Current models often struggle with context-dependent effects that are crucial in practical organic synthesis.[97] Factors such as reaction atmosphere, conditions, addition order, and subtle impurities—all critical in experimental practice—are rarely captured in training data or model architectures. Temperature and solvent effects, while partially addressed by some models, often lack the nuanced understanding required for complex, multi-component systems. Chung and Green (2024) observed that outliers in their machine learning model for solvent effects primarily consisted of unusual reactions unlikely to occur in real systems, indicating a potential weakness in handling rare or atypical chemical scenarios. The sensitivity of AI/ML models to subtle contextual factors and their struggle with unusual outliers indicate that while they excel at general patterns, they may lack the nuanced understanding needed for highly specific or complex scenarios.

Computational Resource Requirements for Training: While inference is typically fast, the training of state-of-the-art AI/ML models requires substantial computational resources and expertise,[98] especially for the scientific discovery domain where experimental practices and technologies evolve continuously.

15. Future opportunities for AI chemistry models

15.1. Algorism and modeling design level

Integration of Comprehensive Contextual and Mechanistic Information: A major thrust for future development is to move beyond

simplified representations and incorporate the full complexity of chemical environments and molecular structures. Most current models developed do not perform complete reaction condition prediction and lack enantioselectivity data, which significantly limits models' applicability. Future work needs to integrate more comprehensive contextual data, including temperature, solvent, reaction time, explicit 3D conformations, and stereochemical information. Integrating 3D and stereochemistry will allow AI to predict subtle effects like steric hindrance or enantioselectivity, which are critical for drug design and catalyst development.

Enhanced Understanding and Interpretability: The move towards "chemically aware" and interpretable AI models with more understandability is crucial not only for user acceptance but also for enabling AI to contribute to fundamental chemical theory and discovery.

Integration of Multi-Scale Physical Understanding: Future AI/ML systems will likely achieve superior performance through the integration of multi-scale physical understanding, combining quantum mechanical insights at the electronic level with thermodynamic and kinetic understanding at the molecular level. The development of physics-informed neural networks (PINNs), which incorporate fundamental chemical principles as constraints, will enhance both accuracy and interpretability, especially in areas with low data availability, such as certain biological and engineering problems.[99]

15.2. Data collection and dataset representation level

Advanced Representation Learning and Chemical Embeddings: More sophisticated molecular representations will enable AI/ML models to capture subtle chemical features that are currently inaccessible to conventional descriptors. The integration of structural, electronic, and spectroscopic information will provide a richer chemical understanding.

Overcoming Data Limitations and Bias: Addressing the foundational issue of data quality, quantity, and inherent biases is paramount for the continued advancement and widespread applicability of AI/ML in chemistry.[91,100] Overcoming data limitations is not solely a technical challenge for machine learning engineers but also a collaborative endeavor requiring chemists to generate more diverse and high-quality data, and for the broader community to establish better data sharing practices. Future work should also focus on developing methods to mitigate biases and fill data gaps inherent in large databases, perhaps through active learning or more diverse data collection strategies. This includes effectively combining disparate data types, such as experimental results, quantum chemical calculations, and molecular dynamics simulations,[101] or employing advanced techniques like transfer learning or hierarchical modeling, to achieve both high accuracy and broad generalizability. Techniques such as few-shot learning and data augmentation can further enable models to learn from limited data, making AI applicable to areas where experimental data is inherently scarce or expensive.

Real-Time Adaptive Learning and Autonomous Experimental Integration: By integrating predictive models with robotic experimentation in closed-loop optimization systems, researchers can autonomously explore vast chemical spaces. Real-time analytical techniques further enhance this process by providing immediate feedback, enabling both experimental refinement and dynamic model updating.[102] Active learning can be used to guide experiments toward generating the most informative data points, creating a virtuous cycle between AI and experimentation, partly addressing the data sparsity issue under certain circumstances.[103] Continual learning systems will ensure that models can incorporate new experimental data while retaining prior knowledge, allowing them to evolve with the growing body of chemical information actively.

15.3. Application and strategic level

***Specialized Domain Applications and Model Combination Strategies*:** The expansion of AI/ML into specialized chemical domains will address current limitations in coverage and accuracy. Ensemble methods and multi-modal strategies have proven highly effective in enhancing the robustness and accuracy of predictions, particularly in complex chemical spaces where no single model may be universally perfect.[104] Furthermore, collaborative learning approaches such as federated learning can facilitate model development across research institutions without requiring centralized data sharing, thus preserving data privacy. Meta-learning techniques will also play a crucial role by enabling models to adapt to new chemical domains with limited training data. Together, these strategies will drive the development of intelligent, self-improving systems for chemical research.

16. Conclusion

The integration of artificial intelligence and machine learning into organic chemistry has ushered in a new era of computational capability. These data-driven approaches offer significant advantages, including enhanced accuracy and efficiency in thermodynamic property prediction, reaction outcome prediction, and retrosynthetic planning with unprecedented scalability for handling large and complex systems. Furthermore, the benefits of the hybrid approach and model combination enhance the ability of AI-driven chemical applications. The integration of supervised learning for property prediction, reinforcement learning for synthetic route optimization, unsupervised learning for chemical space exploration, and neural-symbolic approaches for mechanistic understanding creates robust hybrid systems that leverage the strengths of each methodology.

Perhaps most significantly, data-driven AI/ML systems have demonstrated the potential for generating chemical insights and discovering fundamental chemical principles from data alone. For example, LUMO correlations, Hammond parameters, and functional group compatibility and solvent–reagent relationships discovered by a neural network solely based on training data represent a profound advancement in our ability to capture and utilize chemical intuition computationally, demonstrating the potential of a data-driven AI approach in discovering novel theories in this field.

Moreover, the scalability and speed advantages of AI/ML approaches have practical implications. The ability to screen millions of potential synthetic routes in seconds, predict reaction conditions in milliseconds, and optimize multi-component systems in real-time opens unprecedented opportunities for high-throughput synthesis planning and automated chemical discovery. The convergence of these high-speed, high-accuracy capabilities points directly toward the realization of autonomous chemical research platforms, which envision a closed-loop system where AI can be implemented. The phrase "fully automated chemical discovery" is no longer a theoretical concept but an active, tangible area of development, replacing traditional, laborious, hypothesis-driven experimental cycles.

Despite these remarkable advancements, the field still faces notable challenges. Primary among these are the persistent limitations related to data quality, quantity, and inherent biases. The widespread lack of robust stereochemical prediction and the limited ability to fully incorporate 3D conformational information remain critical barriers to the comprehensive application of AI in complex synthetic contexts. Moreover, the black-box nature of many high-performing deep learning models and their current inability to provide explicit mechanistic understanding hinder trust and limit their capacity to generate novel chemical hypotheses. The reliance on predefined rules or historical data in the hybrid and model combination approach also restricts AI's ability to discover truly novel chemical transformations. At the same time, sensitivity to subtle contextual factors and outliers can compromise the robustness of predictions in real-world scenarios.

The future trajectory of AI and ML in organic chemistry should focus on generating and collecting large and high quality data, developing more chemically aware and interpretable models with deeper

mechanistic insights, the comprehensive integration of critical contextual information such as temperature, solvent, and stereochemistry, innovative strategies to overcome data sparsity and bias, and more sophisticated hybrid systems that seamlessly integrate multiple scales of chemical understanding. Ultimately, the goal is to advance AI's generalizability to truly novel and complex reactions, moving beyond interpolation to enable the autonomous discovery of unprecedented chemical transformations, thereby offering transformative solutions to critical global challenges in medicine, materials science, and energy.

Ethics compliance

This article is a review of previously published literature and does not involve new studies with human participants or animals; therefore, ethics approval, patient consent, and clinical trial registration are not applicable. All data supporting the conclusions are derived from published sources, as cited in the reference list. The authors declare no competing interests. All reproduced figures, tables, or text are properly cited and used with permission from the copyright holders where required. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

No new data were created or analyzed during this study; data sharing is not applicable.

Acknowledgement

This article is a review of previously published literature and does not involve new studies with human participants or animals; therefore, ethics approval, patient consent, and clinical trial registration are not applicable. All data supporting the conclusions are derived from published sources, as cited in the reference list. The authors declare no competing interests. All reproduced figures, tables, or text are properly cited and used with permission from the copyright holders where required. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Chung Y, Green WH. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chem Sci*. 2024;15:2871–2883.
2. Keith JA, Vassilev-Galindo V, Cheng B, et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem Rev*. 2021;121(16):9816–9872.
3. Wang S, Zhang J. Artificial intelligence is empowering chemistry research. *National Science Open*. 2024;3(2).
4. Chemical Society Reviews. Cross-disciplinary perspectives on the potential for artificial intelligence across chemistry. *Chem Soc Rev*. 2025;54. Article in press.
5. Mobley DL, Gilson MK. Predicting binding free energies: frontiers and benchmarks. *Annu Rev Biophys*. 2017;46:531–558.
6. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–1250.
7. Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. *Acc Chem Res*. 2018;51(5):1281–1289.
8. Carter EA, Rossky PJ. *Acc Chem Res*. 2006;39:71–72.
9. Lu Y, Deng G, Shuai Z. *Pure Appl Chem*. 2021;93:1425–1434.
10. Joseph M. *J Theor Comput Sci*. 2024;10:213.
11. Beveridge DL, Berne BJ. *Proc Natl Acad Sci USA*. 2005;102:6679–6680.
12. Piccini G, Lee M, Yuk SF, et al. Ab initio molecular dynamics with enhanced sampling in heterogeneous catalysis. *Catal Sci Technol*. 2022;12:12–37.
13. Wang Q, Sciabola S, Barreiro G, et al. Conformational sampling in computational studies of natural products: why is it important? *J Med Chem*. 2023;66:39315508.
14. Liu X, Bussi G, Gervasio FL. Enhanced sampling in molecular dynamics: a review of recent developments in the field of biomolecular simulations. *Comput Struct Biotechnol J*. 2023;21:2912–2925.
15. Rowland RS, Taylor R. Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *J Phys Chem*. 1996;100:7384–7391.
16. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3(3):210–229.
17. von Lilienfeld OA, Müller KR, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning. *Nat Rev Chem*. 2020;4(7):347–358.
18. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature*. 2018;559(7715):547–555.
19. Hu L, Chen J, Vaughan J, et al. Supervised machine learning techniques: an overview with applications to banking. *arXiv*. 2020:04059. abs/2008.
20. Edwards JK, Cole SR, Troester MA, Richardson DB. Supervised machine learning: a brief primer. *Am J Epidemiol*. 2013;177(9):904–912.
21. Unsupervised learning. ScienceDirect Topics.
22. Wang Z, Li M, Liu Y, Liu Z. A comprehensive survey on autoencoder-based representation learning for dimensionality reduction. *Artif Intell Rev*. 2024;57:28.
23. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *arXiv:12065538*. 2012. arXiv.
24. Kelleher, J. D. "Unsupervised machine learning techniques." KDnuggets.
25. Rami A. Deep learning techniques for fault diagnosis in interconnected systems: a comprehensive review and future directions. *Appl Sci*. 2025;15(11):6263.
26. Dimensionality reduction algorithms archives - Buff ML.
27. Zhu X. *Semi-Supervised Learning Literature Survey*. University of Wisconsin-Madison; 2006.
28. Van Engelen JE, Hoos HH. A survey on semi-supervised learning. *Mach Learn*. 2020;109:373–440.
29. Lee D-H. *Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop on Challenges in Representation Learning*. 2013.
30. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press; 2018.
31. Skrobek D, Skrobek D, Krzywanski J, et al. Artificial intelligence for energy processes and systems: applications and perspectives. *Energies*. 2023;16(8):3441.
32. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res*. 1996;4:237–285.
33. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550:354–359.
34. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
35. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
36. Deep Learning Key Terms. Explained - KDnuggets.
37. Xie Y, Wang L, Gu Y, et al. *Prediction of Rock's Brittleness and Dynamic Properties Utilizing Effective Artificial Intelligence Approaches*. 2024.
38. Łącki M. Determining the level of threat in maritime navigation based on the detection of small floating objects with deep neural networks. *Sensors*. 2024;24(23):7505.
39. Vázquez S. Ultrasound reconstruction tomography using neural networks trained with simulated data: a case of theoretical gradient damage in concrete. *Appl Sci*. 2025;15(8):4273.
40. Alla S, Bheesetty N, Komaragiri S, et al. *Integrative machine learning approaches for enhanced classification of genomic sequences: a next-generation sequencing perspective. Proceedings of the International Annual Conference of the American Society for Engineering Management*. 2024.
41. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998–6008.
42. Tay NC, Connie T, Ong TS, Teoh ABJ, Teh PS. *Abnormal behavior recognition using SRU with attention mechanism*. 2024.
43. Wu Z, Wang J, Du H, Jiang D, Kang Y, Hou T. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun*. 2023;14:2585.
44. Xiong Z, Wang D, Liu X, et al. Comparative study of descriptor-based models and graph neural networks (GCN, GAT, MPNN, and attentive FP) for molecular property prediction. *J Cheminf*. 2020;12:45.
45. Chen B, Carlsson J, Wernér E, et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking (SME). *Nat Commun*. 2023;14:2585.
46. Chemical Abstracts Service. SciFinder; American chemical society: columbus, OH. (accessed July 2025).
47. Reaxys. Elsevier. (accessed July 2025).
48. Ghiandoni G, Brown D. Predictive retrosynthesis in reaxys. *J Chem Inf Model*. 2020;60(8):3899–3909.
49. L. C., Reymond J-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J Am Chem Soc*. 2009;131(25):8732–8733.
50. GDB databases.
51. Krishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data*. 2014;1:140022.
52. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci*. 2004;44(3):1000–1005.
53. Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies. *J Comput Aided Mol Des*. 2014;28(7):711–720.
54. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model*. 2012;52(7):1757–1768.
55. St John PC, Guan Y, Kim Y, Kim S, Paton RS. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat Commun*. 2020;11:2328.
56. ChemAxon Ltd. *Marvin Suite*. ChemAxon; 2021. version 21.15.
57. Landrum G. RDKit: open-Source cheminformatics. 2021.
58. Steinbeck C, Han Y, Kuhn S, et al. The chemistry development kit (CDK): an open-source java library for Chemo- and bioinformatics. *J Chem Inf Comput Sci*. 2003;43:493–500.

59. Diller DJ, Hobbs DW. Deriving knowledge through data mining high-throughput screening data. *J Med Chem*. 2004.
60. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42(6):1273–1280.
61. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50:742–754.
62. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*. 2019;47(D1):D1102–D1109.
63. Weininger A. SMILES: the simplified molecular input line entry specification. *J Chem Educ*. 1986;63:86.
64. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31–36.
65. Heller SR, McNaught A, Pletnev I, et al. InChI, the IUPAC international chemical identifier. *J Cheminf*. 2015;7:23.
66. Daylight theory: SMARTS.
67. Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminf*. 2020;12:12.
68. Himanen L, Jäger MOJ, Morooka EV, et al. DScribe: library of descriptors for machine learning in materials science. *Comput Phys Commun*. 2020;247:106949.
69. Gilmer J, Schoenholz SS, Riley PF, et al. *Neural message passing*. 2017;70:1263–1272.
70. Jiménez J, Skalic M, Martinez-Rosell G, et al. KDEEP: protein–ligand absolute binding affinity prediction via 3D-Convolutional neural networks. *J Chem Inf Model*. 2018;58(2):287–296.
71. Pratt LY. Discriminability-based transfer between neural networks. In: *Advances in Neural Information Processing Systems 5, [NIPS Conference]*. Morgan Kaufmann Publishers, Inc.; 1993:204–211.
72. Mater A, Coote M. Deep learning in chemistry. *J Chem Inf Model*. 2020.
73. Caruana R. Multitask learning. *Mach Learn*. 1997;28:41–75.
74. Li F-F, Fergus R, Perona P. One-shot learning of object categories. *IEEE Trans Pattern Anal Mach Intell*. 2006;28:594–611.
75. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv:13126114*. 2013.
76. Wong SC, Gatt A, Stamatescu V, McDonnell MD. In: *Understanding Data Augmentation for Classification: When to Warp? 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 30 Nov.-2 Dec. 2016 2016:1–6.
77. Chung Y, Vermeire FH, Wu H, Walker PJ, Abraham MH, Green WH. Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. *J Am Chem Soc*. 2023;145(4):2382–2396.
78. Abraham MH, Ibrahim A, Zissimos AM. Determination of sets of solute descriptors from chromatographic measurements. *J Chromatogr A*. 2004;1037(14):29–47.
79. Abraham MH, Acree WE. Correlation and prediction of partition coefficients between the gas phase and water, and the solvents dodecane and undecane. *New J Chem*. 2004;28:1538–1543.
80. Chung Y, Green WH. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chem Sci*. 2024;15(9):3185–3205.
81. Scheen J, Wu W, Mey ASJS, Tosco P, Mackey M, Michel J. Hybrid alchemical free Energy/machine-learning methodology for the computation of hydration free energies. *J Chem Inf Model*. 2020;60(11):5331–5339.
82. Yang Q, Li Y, Yang J-D, et al. Holistic prediction of the pK_a in diverse solvents based on a machine learning approach. *Angew Chem Int Ed*. 2020;59(44):19282–19291.
83. Luo W, Zhou G, Zhu Z, et al. Bridging machine learning and thermodynamics for accurate pK_a prediction. *JACS Au*. 2024;4(9):3451–3465.
84. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent Sci*. 2017;3(5):434–443.
85. Coley CW, Jin W, Rogers L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem Sci*. 2019;10(2):370–377.
86. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. 2018;555(7698):604–610.
87. Kayala MA, Azencott C-A, Chen JH, Baldi P. Learning to predict chemical reactions. *J Chem Inf Model*. 2011;51(9):2209–2222.
88. Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent Sci*. 2018;4(11):1465–1476.
89. Luttens A, et al. Rapid traversal of vast chemical space using machine learning and molecular docking. *Nat Mach Intell*. 2025.
90. von Lilienfeld OA, Müller KR, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning. *Nat Rev Chem*. 2020;4:347–358.
91. Jiang J, Zhang C, Ke L, et al. A review of machine learning methods for imbalanced data challenges in chemistry. *Chem Sci*. 2025;16:7637–7658.
92. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell*. 2020;2:573–584.
93. Stärk H, Feller M, Laino T, Reiher M. Neural-network approach to reaction mechanism prediction and interpretability. *J Chem Phys*. 2020;153:244118.
94. Majewski M, Pérez A, Thölke P, et al. Machine learning coarse-grained potentials of protein thermodynamics. *Nat Commun*. 2023;14:5739.
95. Eliel EL, Wilen SH. *Stereochemistry of Organic Compounds*. New York: Wiley; 1994.
96. Jorner K, Tomberg A, Bauer C, Sköld C, Norrby P-O. Organic reactivity from mechanism to machine learning. *Nat Rev Chem*. 2021;5:240–255.
97. Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. *Acc Chem Res*. 2018;51(5):1281–1289.
98. Probst D. Growing pains: reacting to negative impacts of deep learning on machine learning for chemistry. *Chem*. 2022. Preprint.
99. Ji W, Qiu W, Shi Z, Pan S, Deng S. Stiff-PINN: physics-informed neural network for stiff chemical kinetics. *J Phys Chem A*. 2021;125(36):8098–8106.
100. Sivilotti SL, Friday DM, Jackson NE. Active learning of high-coverage sets of complementary reaction conditions. *Digital Discovery*. 2025;4:846–852.
101. Keith JA, Vassilev-Galindo V, Cheng B, et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem Rev*. 2021;121(16):9816–9872.
102. MacLeod BP, Parlane FGL, McAulay K, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv*. 2020;6(20):eaaz8867.
103. Liu Y, Yang Q, Li Y, Zhang L, Luo S. Application of machine learning in organic chemistry. *Chin J Org Chem*. 2020;40(11):3812–3827.
104. Chung Y, Hu W, Sundmacher K, et al. Ensemble and multi-modal machine learning approaches to improve property predictions in chemistry. *J Chem Inf Model*. 2022;62(12):2934–2947.