Review

Journal of the Brazilian Chemical Society

https://dx.doi.org/10.21577/0103-5053.20250082

J. Braz. Chem. Soc. **2025**, *36*, 8, e-20250082, 1-20

A Survey of Basic Concepts and Applications of Machine Learning to Chemistry

Julio Cesar Duarte, [a,b] ***Antonio G. S. de Oliveira-Filho,*** [c] ***Matheus Máximo-Canadas,*** [d] ***Rubens C. Souza*** [b] ***and Itamar Borges Jr.*** [*,c,d]

[a]*Departamento de Engenharia da Computação, Instituto Militar de Engenharia (IME), Praça General Tibúrcio, 80, 22290-270 Rio de Janeiro-RJ, Brazil*

[b]*Departamento de Engenharia de Defesa, Instituto Militar de Engenharia (IME), Praça Gen. Tibúrcio 80, 22290-270 Rio de Janeiro-RJ, Brazil*

[c]*Instituto de Química de São Carlos, Universidade de São Paulo, Av. Trabalhador São-Carlense, 400, 13566-090 São Carlos-SP, Brazil*

[d]*Departamento de Química, Instituto Militar de Engenharia (IME), Praça General Tibúrcio, 80, 22290-270 Rio de Janeiro-RJ, Brazil*

Theoretical and computational chemistry (TCC) is a set of theories and models that, over the years, were refined to the point that it is possible to determine measurable quantities with precision, predict experimental results, and provide fundamental insights into chemical phenomena and mechanisms that may be difficult or impossible to observe experimentally. Machine Learning (ML), on the other hand, is a subfield of Artificial Intelligence (AI) that applies different types of statistical methods to a large volume of data or a smaller volume of precise data, combined with high computational power, enabling the discovery of complex patterns and production of explanations inaccessible through human deductive reasoning and intuition alone or using traditional scientific methods. Recently, ML, combined or not with TCC methods, has emerged as a transformative force, bringing significant advances in chemistry and materials science. This review surveys basic ML concepts and their applications in chemistry, focusing on supervised and unsupervised learning approaches, data preprocessing, and model development workflows, exploring the most relevant ML algorithms selected for their specific usefulness in chemical applications. Integrating ML with traditional computational chemistry methods, such as density functional theory, is highlighted as a powerful synergy for accelerating materials discovery and design. Key areas of impact discussed include High-Throughput Virtual Screening (HTVS) of molecules and materials, spectroscopy (including UV-Vis and fluorescence), organic electronics (such as solar cells and organic light-emitting diodes), potential energy surfaces, and molecular dynamics. It also addresses critical aspects of ML in chemistry, including data representation, model interpretability through explainability techniques, and the emerging role of large language models (LLM). Tips for acquiring knowledge of ML for practical applications in chemistry are given.

Keywords: machine learning in chemistry, artificial intelligence, theoretical and computational chemistry, data interpretability/explainability, molecules and materials discovery and design

1. Introduction

Machine Learning (ML) algorithms applied to various chemical and materials science problems sparked a scientific and technological revolution, which allows addressing fundamental questions and developing new applications as never before.[1-18] ML, a subfield of Artificial Intelligence (AI), belongs to the so-called fourth paradigm of science - data-driven discovery and scientific research, which emerged after the first three paradigms: experimental, theoretical (physical laws and theories such as quantum mechanics and thermodynamics), and computational, which include density functional theory (DFT) and molecular dynamics (MD).[19,20] These three paradigms are also called the three pillars of science.[21] The interplay between them is the modern foundation of the scientific method and has dramatically accelerated scientific and technological progress. In theoretical chemistry and materials science, with the advent of electronic

*e-mail: itamar@ime.eb.br
Editor handled this article: Maurício Coutinho-Neto (Guest)

digital computers a few decades ago, the computational implementation (third) of theoretical models (second) has enabled the simulation and rationalization of complex real-world phenomena, with DFT and MD being good examples. The fourth paradigm has also contributed to advancing the first three paradigms, recognized as the experimental, theoretical, and computational branches in nearly all scientific domains. The latest paradigm includes ML, statistical learning, data mining, pattern, and anomaly detection, among other approaches.[9,22] The amount of data produced by experiments and simulations forms the foundation of the fourth paradigm, which unifies the first three paradigms of science.[19]

For millennia, materials development relied on serendipity and the empirical correlation of processing and properties.[23] New molecules have also been created through this approach. Examples of this strategy include vulcanized rubber, Teflon, anesthesia, and penicillin. However, the rising costs associated with developing and producing various types of materials and synthesizing special molecules (e.g., energetic materials, whose synthesis can also be hazardous), combined with the massive number of chemically feasible materials and molecules, render traditional approaches inefficient. Approximately 10^4 synthetic materials are in use today, representing a minuscule fraction of the 10^{80} potential chemically meaningful combinations. These possible combinations of chemical composition, structure, and experimental conditions that result in stable forms of matter constitute the chemical compound space (CCS) or, simply, chemical space (CS).[24-26]

Nowadays, the discovery of new materials or molecules, especially more complex ones, requires a new approach. In recent decades, innovations have emerged from modern materials computational design techniques based on DFT and earlier theories, like statistical thermodynamics. More recently, modern data science and ML tools developed for increasingly complex problems have become an attractive alternative for rationalizing and developing new materials and molecules.[6,7,9,22,27-31] It is a fact that we are just at the beginning of research on new materials and molecules in all their complexity.[32] Consequently, ML techniques, combined with the greater efficiency of computational power at decreasing costs, suggest that this approach has significant potential to produce scientific and applied advancements at lower costs and in reduced time. Good textbooks on ML applications to chemistry already exist.[33,34] New scientific journals in the field (*Digital Discovery*, published by the RSC, and *Artificial Intelligence - Chemistry*, published by Elsevier, are two very recent examples), whose editors-in-chief are renowned researchers in this area, were just released. Virtual editions of ML applications in chemistry are increasingly frequent in scientific journals, as is editorial guidance on desired contributions.[35] Furthermore, several events, such as conferences and symposiums, even those not primarily focused on AI, now dedicate sessions and topics of interest specifically to ML applications, highlighting the growing interdisciplinary relevance of the field.

The current impact of ML in other fields of knowledge and society is well known. Technological applications include web searches, translation, natural language processing, autonomous vehicles, medical diagnostics, social media analysis, and robotics.[36] A recent example of significant importance is predicting the genetic evolution of viruses with the potential to cause the next pandemic, which could enhance vaccine resilience.[37] For similar reasons, training new generations of researchers proficient in the most modern data science and ML techniques should be included in the essential toolkit of chemists and material scientists.

The success of theoretical and computational chemistry in determining and predicting measurable quantities with great accuracy and rationalizing experimental results today is indisputable.[38-42] DFT, in particular, has played an essential role in chemistry and materials science due to its relatively high predictive power, applicability, versatility, and computational efficiency. In the area of ML and big data, using quantum chemical electronic structure models,[43] especially DFT,[44] is particularly important for generating theoretical data and designing model architectures of an ML model.[36,45] A pioneering example is the Harvard Clean Energy Project, developed to find organic photonic materials with high efficiencies based on frontier orbital energy data.[6,46] Similar projects exist in many other areas of chemistry, such as catalysis.[47,48]

All properties of organic molecules, whether physical, chemical, biological, or technological, depend on their chemical structure and vary systematically. Establishing quantitative correlations between molecular properties and chemical structure is of great social importance, as it can mitigate or resolve various environmental, medicinal, and technological issues. These correlations are expressed as quantitative structure-property relationships (QSPR).[49-51]

A major goal of QSPR studies is to determine a mathematical relationship between the property of interest and one or more descriptive parameters, known as descriptors, derived from the structure of the molecule. The basic strategy is to obtain an optimized quantitative relationship to predict the properties of compounds, including those that cannot be measured. Thus, the QSPR approach has provided insights into how molecular structure influences properties and has significantly enabled

the development of structures with desired properties in a reverse process.

QSPR descriptors can be empirical, derived from quantum chemical calculations, or classical chemical knowledge. Classical organic physical chemistry has long been concerned with correlating chemical properties and structure. Hammett and Taft[52-57] pioneered linear free energy relationships (LFERs) that produced important insights into the mechanisms of organic reactions with different substituents. Hammett's theory quantifies the electron-accepting or donating power of substituents attached to aromatic rings using sigma constants, which are dimensionless numbers - two of the present authors recently published a systematic set of different types of these constants for dozens of substituents determined using ML techniques.[57] This work was featured in *Chemical World*, an RSC publication, in July 2023.[58] Aspuru-Guzik *et al.*,[17] from the leading group in ML applied to Chemistry and Materials Science, recognize it as "able to leverage DFT and ML to successfully machine learning Hammett parameters."[17]

In short, the great relevance of Hammett's linear equation is that it allows the evaluation of the effect of electronic interactions that substituents exert on the reactive center. This effect modifies the relative energies between reactants and products, thus affecting the equilibrium position (thermodynamics). The equation also allows for an analysis of the same electronic effects in the transition state of the rate-determining step (slow step), which affects the reaction rate (kinetics).[59]

Distinct and successful applications of the original Hammett theory abound and have significantly extended the original concept. Among them, one can cite the use of nuclear magnetic resonance (NMR) signals to examine electronic effects within organic molecules[60] and CH-π interactions;[61,62] photophysical properties in solvents;[63] hydrogen bonding between solute and solvent of *para*-substituted benzoic acids;[64] 1,3-dimethyl-4,5-disubstituted imidazolylidene XNHC (X = H, Me, Cl) ligands in CdSe quantum dots;[65] photophysical properties of 2-phenylamino-1,10-phenanthrolines synthesized with different substituents;[66] characterization of differences in catalytic activity of Au nanoparticles;[67] metal-organic frameworks (MOFs)[68] and non-covalent interactions.[69] ML applications to these topics, e.g., MOFs or the similar COFs (covalent organic frameworks), are beginning to appear.[70] In our case, in 2023, we used Hammett's theory and ML-derived constants to study optoelectronic properties and effects of intramolecular charge transfer of substituted diketopyrrolopyrrole (DPP)[71] and *p*-nitroaniline derivatives[72] with potential applications in organic electronic devices. Previously, without using Hammett's theory, our group studied various families of nitroaromatic molecules, which are components of explosives, to investigate their sensitivity to impact, which we reviewed in a recent book chapter.[73] In 2023, our group employed ML models to investigate the sensitivity of nitroaromatic explosives to understand the molecular origin of this property,[74] and our group have been working on computing the properties of fluids in the liquid, gaseous, and supercritical phases using ML.[75,76]

In the case of the development and rationalization of molecules and materials, it is possible to extend the QSPR concept to include the steps of processing and performance evaluation (Figure 1).[20,23,77] In the diagram in Figure 1, deductive scientific cause-and-effect relationships (or correlations) flow from left to right (forward model). Inductive engineering relationships flow in the opposite direction (inverse model). The aforementioned ML work on nitroaromatic explosives is an example of the direct approach.[74]

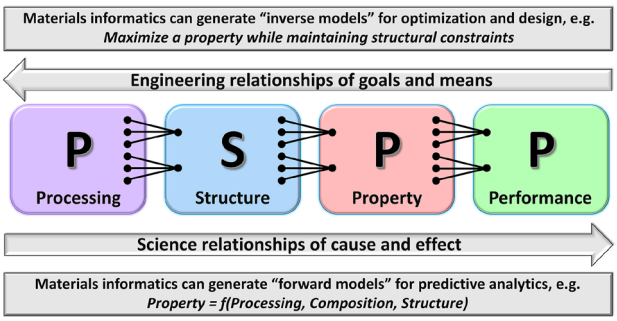

Figure 1. The relationships of processing-structure-property-performance in molecular and materials sciences and engineering, and how theoretical-computational approaches can help decipher such relationships through direct and inverse models.

Each relationship from left to right is many-to-one; thus, those from the right are one-to-many. Therefore, many processing routes can produce the same material or molecular structure, and multiple structures can reach the same property. Each result from a computational simulation or an experimental observation can be considered a data point or example for a direct model. A set of data points (which forms a database) can be used in an ML approach to develop direct models much more rapidly than the time required for an experiment or simulation. This acceleration of direct models can not only guide future simulations or experiments but also allow for the implementation of inverse models, which are much more challenging and critical for discovering and designing molecules and materials. In contrast, since inverse relationships are one-to-many, a good inverse model should be able to identify multiple optimal solutions, if they exist, to have the flexibility to select the easiest and cheapest obtainable

molecular or material structure.[20] In short, inverse designs (or approaches) start from desirable properties and end in a subset of the chemical space, unlike the direct approach that explores the chemical space to yield properties.[77]

The systematic way of selecting promising molecules or materials using ML techniques is known as High-Throughput Virtual Screening (HTVS). This approach originated in the pharmaceutical industry for drug discovery.[6,77,78] Chemicals or materials are subject to simulations, whether generated on the fly or obtained from databases, to estimate specific properties for a given application. Candidates failing computational tests are rejected, assuming that the predicted performance corresponds to the experimental performance with high probability. Therefore, the HTVS approach reduces the vast candidate space to a manageable set of promising molecules or materials. It is important to highlight that HTVS in large chemical spaces is an inverse molecular design approach. Instead of directly designing structures, the designed computational tests evaluate the candidate space, leading to final hits based on the predicted properties.[2,7,77] Consequently, the current HTVS approach is a powerful accelerator, as computational simulations are significantly cheaper than experiments. Combined with this approach, DFT's increasing accuracy and efficiency have transformed materials science and chemical sciences. Aside from the computational cost, the main appeal of HTVS is the ability to "customize" functional parameters to reproduce experiments, which significantly increases DFT's predictive power.[2,77]

The area of ML applications to chemical compounds and materials is, in a certain sense, an offshoot of or related to cheminformatics. This field centers on QSAR relationships to predict the properties of compounds, especially for developing new drugs. Thus, cheminformatics collects, stores, analyzes, and manipulates chemical data.[79,80]

Therefore, modern ML applications in chemistry are the latest stage of the quest to systematically characterize, identify, rationalize, and produce molecular compounds, materials, or fluid mixtures with optimal properties for potential applications.

2. Basic Concepts of Machine Learning

ML provides a fundamentally new way to explore the chemical compound space, uncovering correlations and cause-effect relationships that are otherwise inaccessible. This approach can be defined as using data-driven, automatically improvable computational algorithms without explicitly programming for specific tasks; ML is directly related to statistical learning.[29]

In the field of ML, there are two major classes of problems: supervised and unsupervised learning.[7,9,22,36] In unsupervised learning, the goal is to identify underlying structures to achieve a higher level of understanding. In other words, it seeks latent structures that help represent the known dataset $\boldsymbol{X}$, commonly referred to as attributes or features (input feature space). In this case, raw data lack labels, meaning that no known output data exists. When the dataset $\boldsymbol{X}$ is finite, the learning process is known as clustering, in which the algorithm groups the data into a certain number of clusters based on the similarity of their properties. Unsupervised problems are not as rigorously defined as supervised problems because they may have multiple correct answers, depending on the model used.[36] For instance, our group employed the unsupervised ML algorithm MeanShift[81] to determine optimal substituents from the relationship between our Hammett constants and the exciton binding energy of substituted DPPs.[71]

Supervised, or predictive, learning, in turn, aims to learn the function $f(\boldsymbol{X})$ that maps the inputs $\boldsymbol{X}$ to the target attributes using a labeled dataset $(x_i, y_i) \in (\boldsymbol{X}, f(\boldsymbol{X}))$, known as the training set. If the output y_i produced by the algorithm is a finite categorical set (e.g., whether a molecule is fluorescent or not), the problem is known as classification. Otherwise, if the outputs are continuous scalar values ($y_i \in R$), the problem is called regression, and the algorithm will predict output values for unknown instances as we have done.[57,74] Another recent example from our group has been the prediction of the most intense electronic absorption transition for organic molecules and different solvents.[82]

A typical regression analysis reconstructs the function that passes through a set of known points with minimal error. Regression analysis in ML, however, aims to identify functions that predict interpolations between data points, thus minimizing prediction error for new data points.[36] In an ML regression approach, calculations (e.g., DFT) or experiments are conducted on a representative dataset to train the ML model to predict the remaining unseen compounds.[36] Many ML models are expected to interpolate among unseen compounds better than extrapolate, but a measure of chemical similarity or dissimilarity should be available for evaluation.

A hybrid approach that combines supervised and unsupervised classes is semi-supervised learning, which employs labeled and unlabeled data to enhance model performance. This approach is especially convenient when labeled data are scarce or costly, yet a substantial volume of unlabeled data is accessible. The labeled data support learning specific patterns, while the unlabeled data discover latent structures within the feature space, promoting

improved generalization and scalability. A recent study[83] has demonstrated the utility of semi-supervised learning in developing soft sensors for (bio-)chemical process control, improving predictions with limited-quality data. Additionally, in the field of multivariate calibration in spectroscopy, this approach reduces prediction bias when there is a large volume of unlabeled data and limited labeled data, as evidenced by simulations using near-infrared reflectance spectra.[84]

Among other ML methods relevant to chemistry and materials, reinforcement learning addresses problems by combining supervised and unsupervised learning. Reinforcement learning generally involves defining an agent within an environment that learns by receiving feedback in the form of penalties and rewards[36] without providing any input/output pairs.[9,22] For example, an unfolded protein (state) undergoes geometric modifications (actions) to approach its folded structure (future reward).[85] This approach can be used for molecular design without requiring a representative set of reference structures beforehand.[7] In chemistry, reinforcement learning is increasingly used to find molecules with desired properties in large chemical spaces.[27] A recent application of this technique was employed to optimize transition states and minimum energy pathways.[86] Another study demonstrates the use of reinforced learning for optimizing molecular geometries, reaching significant reductions in the optimization steps.[87] Two recent reviews provide an in-depth look at the growing role of reinforced learning in chemistry: one explores applications in molecule generation, geometry optimization, and retrosynthetic pathway search,[88] while the other offers an accessible summary of reinforcement learning theory and its applications in sequential decision-making tasks in chemistry.[89]

An ML problem or project, broadly speaking, is divided into four stages, also known in the area as a workflow: (*i*) problem definition, (*ii*) data collection, (*iii*) representation modeling, and (*iv*) algorithm selection and training.[6,9,90,91] Step (*i*) begins with feature selection, identifying properties that should reflect the key physicochemical processes of the phenomenon. For instance, in the case of organic photovoltaics, the fundamental photophysical processes involve the absorption of electromagnetic radiation by the donor molecule (D) in the ground state (S_0), followed by the formation of a singlet exciton (S_1), its transfer to the donor-acceptor (D/A) interface, and generation of a charge-separated state (CS); often, the exciton may fail to produce the CS state due to radiative or non-radiative recombination to the ground state.[92] In step (*i*) of the workflow, the type of ML approach (supervised, unsupervised, reinforcement learning, among others) is chosen, thus formalizing the problem within an ML framework. The remaining steps (*ii*)-(*iv*) are discussed below.

The quality and scale of the data play an essential role in developing high-performance ML models.[16,90,93] For instance, when we employ small datasets produced by computational chemistry,[94-96] their high quality will be determined by the accuracy of the theoretical and computational techniques, as we have done.[57,74] In contrast to pre-existing databases (static datasets), an iterative process for ML model development and database generation (dynamic datasets) can be employed to identify missing data and refine the model.[91] Even with a small amount of data points, careful curation is essential as this is a prerequisite for obtaining accurate results, a principle already well recognized by the Quantitative Structure-Activity Relationship (QSAR) and cheminformatics community.[97] We can also use widely available databases designed for applications in chemistry and materials. These may be based on quantum chemistry (e.g., DFT and Time Dependent Density Functional Theory (TDDFT) calculations) or experimental results, and their number and quality continue to grow,[14,90] as do screening processes for them.[98] While it is not possible to cite all of them, it can be mentioned some relevant databases based on quantum chemical data, such as the TDDFT/CC2 vertical spectra of small organic molecules,[99] Harvard's DFT and TDDFT dataset for organic photovoltaics,[100] the experimental NIST database for thermophysical properties[101] and the Cambridge Structural Database with crystal structures of small organic and organometallic molecules.[102] Prominent examples of solid-state property datasets include the Materials Project,[103] AFLOW,[104] the Materials Cloud,[105] the Open Quantum Materials Database,[106] and NOMAD.[107] Similarly, a series of computational molecular databases have been developed, focusing on small organic molecules within the GDB-17 universe, initially created for drug discovery, which contains 166.4 billion molecules.[108] From GDB-17, several QM (quantum mechanical) databases based on quantum chemical calculations have been developed. The QM9 database, with ground-state properties of over 100,000 molecules, was a pioneering achievement in this field.[109,110] The QM7-X includes 42 physicochemical properties for about 4.2 million equilibrium and non-equilibrium structures of small organic molecules with up to 7 atoms (C, N, O, S, Cl), excluding H.[111] Other molecular databases have been expanded to include conformers and non-equilibrium configurations,[111,112] static dipole polarizability,[113] ionized states,[114] and radicals,[115] to name a few. New databases of interest to the chemical community keep coming out,[116,117] as well as methods of dealing with them.[98,110] In 2025, the QCML dataset containing quantum

chemistry reference data from 33.5M DFT and 14.7B semi-empirical calculations appeared.[118] However, the indispensable necessity of ML modeling for extensive and accurate datasets has some important issues, especially for Large Language Models (LLMs) like GPT, which is behind ChatGPT.[119]

Data for an ML model can come from heterogeneous sources, be of different types, contain unknown dependencies and internal inconsistencies, have missing or unreliable parts, and may pose privacy issues, among others.[20] Therefore, careful data preprocessing is essential before executing ML algorithms, which include advanced data visualization,[120] discretization, sampling, normalization, type conversion, and feature selection techniques.[121,122]

Using the usual 3D representation of a molecule, e.g., Cartesian coordinates of all atoms, step (*iii*) of the ML workflow, is only sometimes the most efficient representation for an ML model. A more direct and appropriate representation can save computational resources in pattern learning and yield more accurate ML model performance.[77,123] An ideal molecular representation should be unique, invariant under symmetries (permutational, rotational, reflective, and translational), efficient to obtain, and capable of capturing the associated physics. The molecular representation problem remains an open research challenge: there are many, and none has proven universally effective for all properties.[77,124] Typical examples are the simplified molecular-input entry system (SMILES) strings, a 1D text format with its grammar syntax[125] that can be easily converted into 2D graphs for graph-based investigations, molecular fingerprints,[126] and Coulomb matrix representation based on the electrostatic forces between the charges of each atom.[127] SMILES strings can be used to determine fingerprints or calculate molecular properties (descriptors for an ML model) with open platforms such as RDKit[128] or ChemAxon.[129] Figure 2 illustrates some of the possible molecular representations.

To illustrate the typical workflow of an ML project, especially step (*iv*), let us consider a supervised learning problem (the most common) for predicting molecular properties.[6] The dataset contains molecules in a given representation (features) and their corresponding properties (labels). First, the dataset is split into three parts: training, validation, and testing. The model (algorithm) is trained (optimization of the coefficients or weights) on the training dataset. When only training and testing sets are used, k-fold cross-validation techniques are commonly applied, where the dataset is randomly divided into k parts, with k – 1 parts used to train the algorithm and the remaining part for testing, and this process is repeated k times with different

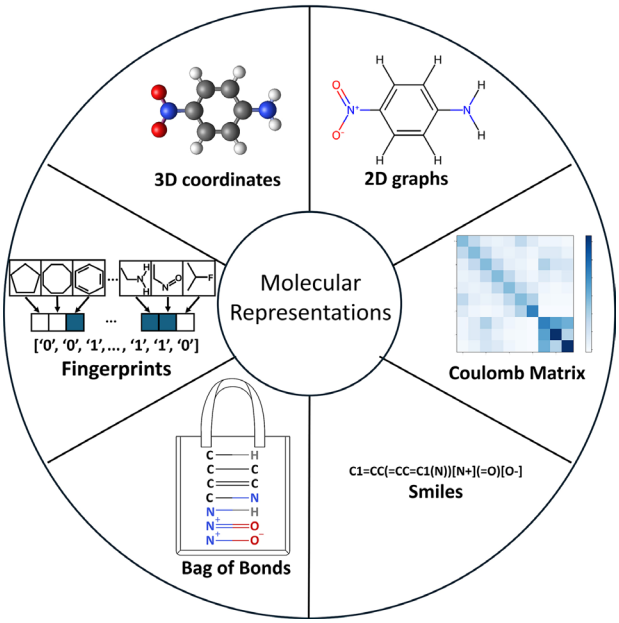

Figure 2. Possible molecular representations for ML models.[123]

test set partitions. Cross-validation is a standard evaluation method to eliminate any chance of model overfitting, which occurs when an algorithm accurately predicts the training data but fails to generalize to unseen test data.

The hyperparameters (preset parameters of the algorithm, as opposed to those learned during training) represent the choice of features, training set, and model architecture (algorithm) and impact the performance of the model. These hyperparameters are optimized by maximizing the predictive accuracy of the model for the validation set, and the final model (trained algorithm) can be used to predict the properties of molecules not included in the training set ("unlabeled molecules"). The workflow of a complete ML project is illustrated in Figure 3, and the resulting knowledge can be represented as QSPR relationships, which may be invertible, facilitating the discovery and design of molecules and materials.

Once the data have been preprocessed, they are ready for ML modeling. At this stage, careful attention must be given to the separation and validation of the different subsets of the data. Suppose the target attribute is numerical (e.g., a transition energy). In that case, regression algorithms can be used for predictive modeling. In contrast, classification techniques are suitable for categorical data (e.g., whether a compound is metallic or not).[20] For a detailed tutorial on using neural networks for regressions, see the literature.[130]

Sometimes, training just one ML algorithm may not be enough to get superior results, as these algorithms can have inherent limitations depending on the characteristics and complexity of the data. While individual algorithms

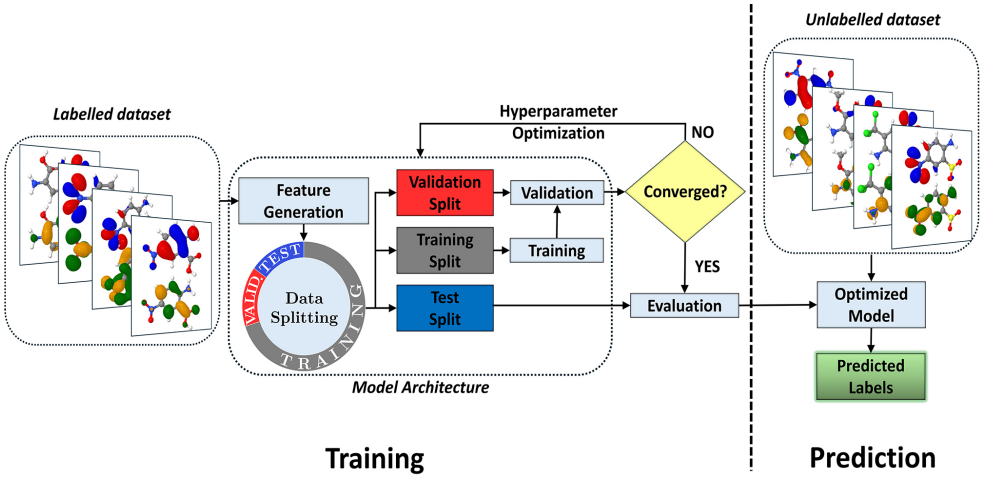

Figure 3. Workflow for supervised learning of molecular properties. A known (labeled) dataset is used to optimize the algorithm and subsequently used to estimate the molecular properties of an unknown (unlabeled) dataset. The Test Split, also known as the "Holdout Split", refers to the portion of data in the test set.[6]

may perform well in certain aspects, they often struggle with particular patterns or noise within data. Even when algorithms exhibit similar overall performance, they might specialize in different regions of the datasets, meaning that they might be complementary if combined.

Ensemble, or Committee learning, combines multiple models to make more accurate and robust predictions than a single model.[121] These models can be either combined with entirely different algorithms (heterogeneous approach) or from variations of the same algorithm applied in different ways (homogeneous). One of the simplest yet effective ensemble methods is using a voting strategy, where different classifiers are trained independently, and their predictions are combined through a majority or weighted vote. This method helps mitigate the bias of any single model, ensuring that the final prediction is less likely to be influenced by the weaknesses of individual classifiers.

In addition to simple voting mechanisms, more sophisticated ensemble techniques exist that focus on improving model performance iteratively. For instance, one popular approach is boosting, where the same algorithm is trained multiple times, each time focusing on the areas where the previous model performed poorly, such as instances with higher training errors.[131] This iterative process allows the model to progressively learn from its mistakes and improve its performance, particularly on the more challenging parts of the dataset. Similarly, bagging, or bootstrap aggregation methods like Random Forests train multiple instances of the same algorithm on different subsets of the dataset and then aggregate their predictions, which reduces variance and helps prevent overfitting. To sum up, while bagging techniques combine multiple models trained on different subsets of data, boosting trains the model sequentially concentrating on the errors made by the previous model.

By combining the strengths of different models or different iterations of the same model, ensemble learning techniques significantly enhance the final trained performance of the model.

Table 1 illustrates some popular ML algorithms in chemistry and materials science.

In the Supplementary Information section, it is provided a more detailed description of the most important ML algorithms for applications in chemistry and materials science. Error metrics for evaluating the performance of the algorithms and explainability in each case are also presented and discussed.

An area of artificial intelligence that has recently begun to show relevant applications in chemistry and materials science is that of LLMs.[119,151-161] Despite this promising start, the potential and scope of results from LLMs do not yet compare to those obtained with other ML methods.[162] Prompt engineering, as old as chatbots, is a key discipline for effectively using tools like ChatGPT. As AI becomes increasingly pervasive, it will emerge as a unique and indispensable skill across numerous professions.[163] In the case of this helpful tool for several different tasks, in particular, proprietary aspects of its engineering pose challenges for scientific research (e.g., what datasets were used for its training?).

The explainability or interpretability of results obtained with ML algorithms, known as Explainable Artificial Intelligence (XAI),[164-167] is fundamental for the utility of this approach in chemistry. The concept of XAI is that ML models are not limited to numerical predictions or classifications but also employ techniques to identify patterns, correlations, and cause-effect relationships in the results. Various possible approaches exist to identify the correlations in the numerical predictions and classifications.[168,169]

Table 1. Some popular machine learning algorithms,[121,122] especially in chemistry and materials science

Modeling technique	Capability	Short description
I Bayes[132,133]	classification	probabilistic classifier based on Bayes' theorem
Logistic regression[134]	regression	fits the data to a logistic curve with sigmoidal S shape
Linear regression[135]	regression	least-squares fit of the data with respect to the input features
MeanShift[81]	classification	locates arbitrarily shaped regions with uniform density in data sampling
Artificial Neural Networks[136]	both	uses hidden layer(s) of neurons to connect inputs and outputs
Support Vector Machines[137,138]	both	builds a feature space as multidimensional hyperplanes based on the minimization of structural risk
Decision table[139]	both	builds rules involving different combinations of attributes
Alternating decision tree[140]	classification	a tree with alternating prediction and decision knots in which an instance goes through every applicable path
Logistic model tree[141]	classification	a classification tree with logistic regression functions on the leaves
Random Tree	both	uses a subset of randomly chosen attributes
Reduced error pruning tree[142]	both	builds a tree by information gain/variance and prunes it through error reduction to avoid overfitting
Decision tree regressor[121,143]	both	a decision tree that on each knot divides the data into classes based on its features to find non-linear relationships
Gradient Boosting[144]	committee	a boosting[a] algorithm in which each stage of the model uses gradient descent optimization to reduce the residual errors from the previous stage
AdaBoost[145]	committee	in contrast with gradient boosting, each iteration optimizes the weights associated with each instance[b]
Bagging[146]	committee	builds many models from resampled subsets of the training to improve the stability of the model through variance reduction
Random subspace[147]	committee	systematically builds multiple trees by the pseudo-random selection of features subsets
Random Forest[148]	committee	a set of multiple random trees
LASSO[149]	regression	selects and regularizes variables to increase prediction accuracy and interpretability
LASSO LARS IC (LLIC)[150]	regression	combines the LASSO method with the LARS algorithm that finds the aspect of the input features that provides the best match with the target feature

[a]Boosting sequentially trains multiple weak learners (base learners) to produce a strong learning algorithm. [b]An instance in ML is an example within the training set, described by a certain number of attributes (features). [c]Techniques to calibrate ML models aim to minimize a predefined loss function to prevent overfitting and underfitting in predictions.[33]

One of them is SHAP (Shapley Additive exPlanations) plots, derived from game theory,[168,169] which have been used to provide explanations in several molecular prediction models.[170-174] Our group used SHAP plots to demonstrate how specific molecular properties contributed to increasing or decreasing its sensitivity depending on the impact sensitivity of a certain nitroaromatic explosive.[74]

Many of the tools exemplified in Table 1 have their way of evaluating the inputs used during training by providing tools to evaluate the weights of each algorithm input or the degree of importance for the generated ML model.

3. Representative Literature on Machine Learning

ML techniques are particularly well-suited for accurately recognizing and quantifying non-linear relationships, especially in inferring decision rules or categorizing complex patterns based on data. These tasks are highly challenging, even for researchers with great scientific intuition.[36] However, it is essential to emphasize that a critical requirement in ML projects is combining expertise and (physical-) chemical intuition since an ML model will hardly automate this process.[2]

Disruptive advances in materials and molecules can emerge from unexpected regions of the chemical space and outliers present in the datasets used in ML models.[77,175] Moreover, it is challenging to train ML algorithms to study materials and molecules, as the data is generally sparse and noisy, which is problematic in this research area.[32,75,176] One possibility to address this problem is to develop and curate databases of physicochemical properties determined by

quantum chemical methods and molecular dynamics (MD), combined with experimental data.

The theoretical study of excited states of molecules is essential for rationalizing and complementing experiments and elucidating many fundamental processes in nature, such as photosynthesis and human vision.[177] Experimental spectroscopic techniques (e.g., UV-Vis spectroscopy) do not directly describe the exact mechanisms of photoinduced reactions, making theoretical studies especially valuable. However, calculating excited states is highly complex and costly, often requiring specialized knowledge.[178] Our group recently showed how ML can address the problem of computing excited states in absorption[82] and fluorescence spectra,[179] and discuss these examples in detail in section 4.

Functional materials are capable of performing multiple functions due to their specific properties. They can exist naturally or be artificial. In the case of photoconductive conjugated organic polymers, there are several important applications in electronic devices and sensor technology. Optoelectronics applications of particular interest are organic solar cells (OSCs), sensors, and organic light-emitting diodes (OLEDs).[180,181] The most active areas in organic photovoltaics (OPVs) research are designing more efficient materials for the active layer[182] and investigating the underlying mechanisms.[183] Our group have made some contributions to the study of the electronic properties of these materials through quantum chemical calculations, with particular emphasis on charge transfer processes and states, an important property in such devices,[71,184-187] as well as a phenomenon of intrinsic interest in different areas of chemistry, physics, medicine, and engineering.

There is a wide range of ML applications to photoconductive conjugated polymers for organic electronics, particularly in predicting the power conversion efficiencies (PCEs) of organic solar cells (OSCs).[188-191] Greenstein and Hutchison[192] developed models that combine semiempirical methods, ML, a dataset of 84 donor-acceptor pairs, and 47 property descriptors to determine PCEs, short-circuit current (J_{sc}), and open-circuit voltage (V_{OC}) for non-fullerene acceptors starting from molecules with PCEs > 9%. The result was a linear equation involving some of these descriptors that makes good PCE predictions, with the largest contribution coming from the transition energy with the highest oscillator strength of the first donor transitions. Sun *et al.*[193] prepared a dataset with 1719 donor materials from the literature to build QSPR models for OSCs. Using ML, specifically the Random Forest technique, they obtained accurate predictions of PCEs measured in OSCs. Sahu *et al.*[194] used different ML algorithms on a dataset of 280 OPV molecules, quantum chemical calculations, and 13 descriptors to predict PCEs of new D (donor)/A (acceptor) molecules for OSCs. Munshi *et al.*[195] used a deep neural network (DNN) based transfer learning model to generate SMILES fingerprints of candidate polymers for OPV devices, validated with calculations of other properties such as PCE. This approach allowed in a second stage to work with a relatively small dataset (ca. 1400 conjugated polymers) to generate new polymer repeat units. Other authors sought to optimize OSC PCEs by focusing on certain classes of polymers.[196] Malhotra *et al.*[197] used graph neural networks (GNN) starting from 1318 D/A pairs selected by Miyake and Saeki[188] to predict OSC PCEs successfully. The discussed examples are just a tiny sample of the potential of ML, combined with or without quantum chemistry, to advance the field of OSCs.

Organic light-emitting diodes (OLEDs), especially those based on materials subject to thermally activated delayed fluorescence (TADF), have high performance and thus have attracted much attention.[198,199] These highly efficient, inexpensive, and flexible materials are used in displays and lighting. In TADF materials, an inversion of the ordering of triplet (T_1) and singlet (S_1) states and a small energy gap ΔE_{ST} (< 2.0 eV) are determinants of their efficiency.[200] The classic work on TADF OLEDs combining quantum chemistry and ML, in addition to experimental validation, is that of Gómez-Bombarelli *et al.*,[201] who explored a chemical space of 1.6 million molecules, screening 400k molecules with TDDFT to arrive at a few thousand with potential for efficient OLEDs. Furukori *et al.*[202] performed a virtual screening with neural networks to rapidly collect photoluminescence decay profiles of TADF materials. Bu and Peng[203] combined TDDFT calculations and neural networks to identify 384 promising TADF molecules from 44,470, resulting in molecules with efficiencies superior to existing ones. Kim *et al.*[204] proposed a quantitative score (TADF-likeness) to evaluate the TADF potential of molecules using a data-driven chemical similarity concept with known TADF molecules using neural networks and the k-means clustering method to divide the datasets.

Fluorescent probes, or chemical-fluorescent sensors, are organic molecules emitting characteristic fluorescence in the UV-Vis region. Since the fluorescent wavelength and emission intensity are significantly affected by molecular interactions, fluorescent probes can detect molecules in the environment and biomolecules with high sensitivity.[205] In the field of luminescent sensors, including fluorescent ones, ML, as in organic photovoltaics (OPVs), has a significant impact - see the references in Mousavizadegan *et al.*[206]

In the area of electronic structure, ML approaches can be used to develop composite methods, similar to the G_n

family,[207] as done by Cameron *et al*.[208] and Holm *et al*.[209] with neural networks. Using the Δ-learning approach, ML techniques can be employed to fit the difference between a lower-level calculation and a more expensive high-level calculation. An interesting Δ-learning approach started from DFT energies to obtain CCSD(T) gold-standard energies.[210,211] Another application of ML approaches is the construction of potential energy surfaces for dynamical studies. In this case, neural networks,[212] kernel ridge,[213] or Gaussian process regressions[214] typically replace the classical least-squares fits with physically motivated analytical functions with significant data efficiency, accuracy gains, and greater applicability.[215] Another advantage of ML methods over classical potential energy surface fitting is that the ML methods can simultaneously fit the surface and enforce the symmetries of the system, and perform transformations such as diabatization conveniently and automatically, allowing faster dynamical calculations.[216-218] Advances in the field of photochemistry/excited states of molecular systems already include the existence of datasets calculated with multireference methods, with most applications based on neural networks, particularly the prediction and analysis of spectra, the determination of excited state properties, assistance in choosing active spaces for multireference methods, and Highest Occupied Molecular Orbital (HOMO)-Lowest Unoccupied Molecular Orbital (LUMO) gap predictions, as discussed in recent reviews.[177,219]

Thermophysical properties (thermodynamic and transport properties) of simple fluids and mixtures near the critical point and in the supercritical region are fundamental problems that are still not fully understood and have critical applications in different industries. For instance, the conditions of the Brazilian pre-salt petroleum are supercritical, which poses different challenges to extraction and separation. Pre-salt petroleum is naturally multiphasic and composed of mixtures of different fluids under supercritical thermodynamic conditions at high pressures and temperatures, making simulation or even semiempirical equations of state a great challenge. Obtaining and rationalizing the Pressure-Volume-Temperature (PVT) diagram of single- and multicomponent systems is thus essential. CO_2 (which is very present in pre-salt wells) and natural gas composed of methane (CH_4) with contents above 70% are followed by significantly smaller proportions of ethane (C_2H_6) and propane (C_3H_8) - the proportion of the mixture depends on the geological region. Therefore, studying simple fluids and their mixtures is of special scientific and practical relevance, as experimental and theoretical data on the thermophysical properties of binary and ternary mixtures are very scarce.[220] In the area of supercritical fluids, ML applications are beginning to emerge.[221,222] Representative works include that of Zhu and Müller,[223] who used multilayer neural networks trained with pseudo-data generated by equations of state to determine fluid properties such as critical temperatures and pressures, subcritical vapor-liquid equilibrium, and supercritical density. Zhao, Kuo, and Jin[221] used a deep neural network (DNN) to predict the diffusion coefficients of supercritical water mixtures from molecular dynamics data. They studied binary and ternary mixtures of H_2, CH_4, CO, O_2, and CO_2 under supercritical conditions using transfer learning: the DNN model was trained on a larger dataset (binary mixtures) and then applied to a smaller dataset (ternary mixtures). In a similar but earlier approach, Liu *et al*.[224] used the Support Vector Regression (SVR) ML model with input data for H_2O, CO_2, and H_2 generated by molecular dynamics and collected from NIST to accurately determine PVT diagrams of these fluids and their mixtures in the supercritical region and surroundings.

Another area where ML techniques can have an impact is in heterogeneous catalysis research. They can help correlate catalyst performance with its physicochemical properties by using high-throughput methods that either rely on experimental data alone or combine it with quantum chemistry results. A recent review of machine learning for the experimental and computational development of heterogeneous catalysts has just been published as a preprint from the lead and pioneer group in applying ML to materials science.[225]

4. Machine Learning Applications in Chemistry

In this section, it is illustrated three different ML applications developed according to the four typical steps in the workflow presented above. In the first, the purpose was to identify the most intense absorption peak of organic molecules.[82] In the second, a general approach was developed for dealing with the thermophysical properties of substances in different regions of the phase diagram.[75] In the third, it is explored neural networks for potential energy surface fitting.

4.1. Machine learning prediction of the most intense peak of the absorption spectra of organic molecules

This work aims to predict important properties of the most intense absorption peak of organic molecules and rationalize them. These properties include the oscillator strengths, excitation energies, and the transition orbitals involved in the most intense singlet valence transition (step *i*).

Having defined the target properties that the developed ML models can predict, the next step, (*ii*), is to identify a

suitable database that meets these requirements or create a new one from scratch. In this case, the QM-symex database[226] provided the most suitable set of target properties (oscillator strengths, excitation energies, transition orbitals, and others). The next step, (*iii*), is to choose a representation of the modeling. In this case, it was used the SMILES representation to identify the molecular descriptors (molecular properties) to be used as input features for the model. The RDKit tool is used to generate the molecular descriptors;[127] it uses a SMILES representation of a geometry of the molecule to determine a wide range of molecular electronic and structural descriptors. Since QM-symex stores molecular geometries in Cartesian coordinates, each molecule must be converted into its SMILES representation. After completing this conversion and selecting the target properties, a dataset for each target property is prepared to train the ML models.

In step (*iv*), the algorithms are selected, and their training is performed. Given the vast array of available ML models, the LazyPredict tool is employed using the prepared datasets in step (*iii*). This tool systematically evaluates dozens of models simultaneously to identify the most suitable one for a given dataset. In this case, the input features consisted of various molecular RDkit properties, including molecular weights, the number of aromatic rings, the number of heavy atoms, and molecular fingerprints, among several others. The target, or output, properties are oscillator strengths, excitation energies, and the transition orbitals involved in the most intense singlet valence transition.

The three top-performing models selected by LazyPredict are evaluated using cross-validation techniques along with systematic training, testing, and validation set partitioning. The two data sets are also used in an artificial neural network (ANN) model to compare the performances. The evaluation process of the four ML models utilizes multiple metrics, including the mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2), discussed in the Supplementary Information section.

It was found that the predictive performance of the ML models is comparable to computational quantum chemistry techniques. To rationalize the results, or in other words, to bring interpretability to the ML models, it was employed SHAP plots[168] since they provide valuable insights into the contribution of each molecular descriptor to the target properties. This analysis establishes meaningful relationships between input features and the investigated excited-state target properties. A similar approach was also used for investigating fluorescence peak wavelengths and quantum yields.[179]

4.2. Using raw thermophysical data to develop consistent data sets

This application aims to develop ML models to produce, from raw experimental thermophysical data, a consistent dataset by focusing on the thermal conductivity of methane (step *i*). This approach addresses challenges arising from variability and noise inherent in experimental measurements carried out with different apparatuses, under different conditions, and subject to various environmental factors.

The second step is collecting raw thermal conductivity data for methane from multiple experimental sources in several papers, covering the liquid, vapor, and supercritical phases. The choice of the input features for the ML algorithms in step (*iii*) includes temperature and pressure, while the target (output) property (variable) is the thermal conductivity value. Given the different physical laws governing each phase, data must be separated into three sets, one for each phase (liquid, vapor, and supercritical). This separation is done since the ML models struggled to predict properties accurately when trained on datasets encompassing multiple phases simultaneously.

Once the dataset is structured, different ML models are trained to predict thermal conductivities for each phase (step *iv*). The ML models include AdaBoost, Extra Trees, Gradient Boosting, Random Forest, Extreme Gradient Boosting, Support Vector Regression, and an ANN. Hyperparameter optimization is performed using the GridSearchCV technique with k-fold cross-validation to enhance model performance. This process involves constructing a grid of possible hyperparameter combinations and evaluating each configuration based on the negative mean squared error.

Cross-validation techniques and k-fold validation ensure the reliability and generalizability of the models. Once trained, the ML models are assessed by comparing their predictions with experimental data and a reference dataset from NIST, which presents statistically treated data. The evaluation process employed the same error metrics of the previous example.

Results showed that ML models can effectively handle noisy and heterogeneous experimental data, producing predictions consistent and comparable to the NIST-refined values without requiring extensive preprocessing, even without removing outliers before running the models. Therefore, this ML-based approach offers a faster and more cost-effective alternative to traditional data processing methods, providing a robust solution for systematically integrating diverse experimental datasets.

4.3. Potential energy surface fitting using neural networks

Potential energy surfaces are the effective potential for nuclear motion, as defined by the Born-Oppenheimer approximation that separates the electronic and nuclear degrees of freedom in the molecular Hamiltonian. The availability of accurate and computationally efficient potential energy surfaces is the bottleneck for the investigation of chemical dynamics, especially when using quantum dynamics[227,228] or *quasi*-classical trajectories.[229,230] The usual procedure to generate potential energy surfaces involves fitting the parameters of an adequate function to accurate electronic structure calculations using a least-squares method,[231] which involves many different geometries. The dimensionality of such surfaces is given by $3N - 6$, in which N is the number of atoms in the molecule. For the simplest case of a diatomic molecule, one has a potential energy curve, i.e., a potential energy as a function of the internuclear separation. For these curves, many physically motivated analytical functional forms can be used;[232] however, as the system size increases and the dimensionality of the potential energy surface increases accordingly, more flexible non-linear regression methods, such as neural networks, are more advantageous than the traditional fixed-functional form fitting. The following description is not specific to any single use of neural networks in potential energy surface fitting, but is representative of many applications found in the literature.[233-237] For a full tutorial with a sample code for potential energy surface fitting, see the literature.[129]

After defining the problem (step *i*), a data set is created that covers the relevant configuration space (molecular geometries) of the system (step *ii*). Typically, this involves configurations around significant minima (intermediates), asymptotic regions (reactants and products), transition states, and reaction pathways. A high-level electronic structure calculation is conducted for each molecular geometry, producing an electronic energy, which represents the value of the potential energy surface for that configuration. In this example, the target properties are these energies, and the relevant descriptors are the internal coordinates that define the configuration or functions derived from such coordinates, such as the distance between atoms, r_{ij}, Morse coordinates $e^{-r_{ij}}$, or permutationally symmetrized linear combinations of coordinates (step *iii*). Grid-based sampling, random displacements, direct dynamics at basic levels of theory, or a combination of various methods can be used to explore the configuration space.

The next step is selecting and training the model, which is essentially the determination of the weights and biases that minimize the deviation between the predicted energies and the energies in the training set (step *iv*). Typical applications have datasets on the order of 10k points and use two hidden layers with 20 to 70 neurons on each layer, with the hyperbolic tangent as the activation function.

The root-mean-square error of the predictions can be used to evaluate the quality of the model. Typical values are in the order of 10 meV (0.2 kcal mol^{-1}), which indicates that the fitting does not introduce significant errors to the underlying electronic structure method used for the dataset. Besides the small fitting error, the requirements for an accurate potential model include smoothness and the absence of non-physical behavior, such as unphysical holes or barriers. Furthermore, the ML model for a potential energy surface is rarely the final goal of a given investigation: the model is used in dynamical and/or spectroscopy calculations whose results can be validated against available experimental data. If a given model is unsatisfactory, it can be improved by including additional data in training or by modifying the hyperparameters of the neural network.

Overall, neural networks can effectively generate high-fidelity, low-cost (compared to the electronic structure calculations performed to create the dataset) regression models for multidimensional potential energy surfaces that can accurately compute many properties, such as scattering cross-sections, rate constants, branching ratios, and spectra.

5. How to Start Using Machine Learning from Scratch

One essential first recommendation for researchers interested in incorporating ML techniques into their work is to learn the Python programming language. While other languages such as R, Julia, Java, and C++ are also widely used in ML applications, Python is prominent as the most suitable choice due to its simple syntax and extensive ecosystem of ML and related libraries or tools, including NumPy,[238] Pandas,[239] Scikit-Learn,[240] TensorFlow,[241] Keras,[242] and PyTorch.[243] Additionally, Python is extensively adopted within the scientific and technological communities, ensuring continuous support through online forums and access to up-to-date resources. There is also a wide availability of sources on the internet for learning Python, and the official Python documentation site[244] remains one of the most reliable sources, offering current information and practical examples.

Furthermore, online courses provided by paid platforms or free resources are practical for a dynamic and structured learning experience. The questions and answers

communities in the Stack Exchange Network[245] also provide important information and learning opportunities. For people entering the ML field from physics or chemistry, Data Science,[246] Cross Validated,[247] and Matter Modeling[248] Stack Exchange sites are handy. For those who prefer textbooks, two particularly significant works are the books by Matthes[249] and Deitel.[250]

While theoretical understanding is important, we strongly recommend learning Python through hands-on practice. Upon mastering programming, the next step involves delving deeper into the concepts of ML. Classical texts such as Russel and Norvig's book[251] and Morenney[252] provide a solid theoretical foundation combined with practical guides. For specific applications in the field of chemistry, the books by Dral[34] and by Janet and Kulik[33] are highly recommended. Moreover, comprehensive reviews like the present text are very convenient. These articles provide an up-to-date overview of the topic, synthesizing the most recent findings. They particularly benefit those seeking rapid and relevant information without exploring numerous individual studies.

The field of ML is vast and fast-growing. Therefore, any attempt to deliver a fully comprehensive and up-to-date report, even restricting solely to applications in chemistry, would be impossible and overwhelming to newcomers. With that in mind, we presented the interested researcher with a coherent path through the vast ML landscape to allow them to appreciate some applications and hopefully integrate this knowledge into their toolbox.

6. Concluding Remarks

The rapid progress in ML is remarkable, and its impact extends significantly into chemistry and materials science. This review highlights transformative potential of ML in traditional theoretical and computational approaches. ML offers powerful tools to address challenges associated with high-dimensional chemical spaces, the computational demands of quantum mechanical methods, and the intricate interplay of molecular properties.

The integration of ML into chemistry is more than an improvement in computational tools: it is a paradigm shift. ML provides novel methods for exploring the chemical space, uncovering patterns and relationships that traditional approaches might overlook. Combining ML with quantum chemistry calculations, particularly DFT, has proven especially powerful for generating extensive databases, accelerating calculations and workflows, improving accuracy and predictive capabilities, and innovating model architectures.

Due to its interdisciplinary nature, successful ML applications in chemical physics and physical chemistry increasingly require collaboration between chemists, physicists, and computer scientists. This interdisciplinarity also offers multiple entry points into the field, accommodating researchers with varying primary backgrounds. Beyond the specifics of ML algorithms and their details, data quality is essential to avoid the well-known concept of "garbage in, garbage out" from computer science. This field continues to benefit from growing databases of computational and experimental results, many of which are published as open access, making them increasingly accessible to researchers at all levels.

The convergence of ML and chemistry unlocks remarkable opportunities to address demanding scientific and societal challenges. From advancing green chemistry initiatives to optimizing energy storage systems and accelerating pharmaceutical development, ML is reshaping entire sectors of chemical research, offering novel solutions to complex old problems.

As computational infrastructure and algorithmic sophistication continue to evolve, ML-driven approaches will allow the exploration of previously uncharted regions of chemical space. Combining traditional chemical expertise with modern data science techniques will be essential for navigating and contributing to new chemical discoveries and innovations while mastering these skills will be invaluable for empowering the next generation of researchers.

Supplementary Information

Supplementary information (a more detailed description of the most important ML algorithms for applications in chemistry and materials science, error metrics for evaluating the performance of the algorithms and explainability in each case are also presented and discussed) is available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgments

I. B. thanks the Brazilian agencies CNPq (grant numbers 304148/2018-0 and 409447/2018-8) and FAPERJ (grant number E-26/204.294/2024) for funding this research. Support for this research also came from the National Institute of Science and Technology on Molecular Sciences (INCT-CiMol) grant CNPq 406804/2022-2, and Nano Network grant FAPERJ E-26/200.008/2020. A.G.S.d.O.-F. acknowledges CNPq for grant 309572/2021-5 and FAPESP for grant 2021/00675-4. M.M.-C. thanks Capes and R.C.S. for a PhD scholarship. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Author Contributions

Julio C. Duarte, Antonio G. S. de Oliveira-Filho, Matheus Máximo-Canadas, and Rubens C. Souza were responsible for writing, review, and editing; Itamar Borges, Jr. for the conceptualization, funding acquisition, project administration, writing original draft, writing, review, and editing.

Julio Cesar Duarte *is a computer engineer who graduated from the Military Engineering Institute (IME) in 1998. He earned an MSc in Informatics from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2003 and a PhD from PUC-Rio in 2009. In 2021, he had a postdoctoral fellowship at PUC-Rio. Currently, he serves as a professor in both the Graduate Program in Systems and Computing and the Defense Engineering Graduate Program while holding the position of Pro-Rector of Teaching and Research at IME. His work spans artificial intelligence, machine learning, and natural language processing, with recent research focusing on multimodal media processing, large language models, and malware analysis.*

Antonio G. S. de Oliveira-Filho *holds a Bachelor's (2009) and PhD (2013) in Chemistry from the University of São Paulo (USP). He completed postdoctoral research at the same institution and Emory University (2013-14). Since 2015, he is a professor at USP, currently at the São Carlos Institute of Chemistry (IQSC-USP). His research focuses on molecular spectroscopy, chemical kinetics, and dynamics, employing advanced computational chemistry methods. He has experience in electronic-structure approaches, transition metal complexes, catalytic systems, and neural networks for potential energy surface fitting and diabatization.*

Matheus Máximo-Canadas *is a PhD candidate in Chemistry at IME, supervised by professors Itamar, Julio Cesar, and Jakler Nichele (IME). His PhD research focuses on applying machine-learning techniques to fluids in different thermodynamic phases. He holds an MSc from IME (awarded Chemistry Outstanding Student, 2023) in ab initio and DFT calculations of substituent and solvent effects on excited states. He also holds a BSc in Chemistry from the Instituto Federal do Rio de Janeiro (IFRJ, 2021), where he studied experimentally electron impact ionization and fragmentation of astrophysical molecules using TOF-MS. Additionally, he conducted undergraduate research at Embrapa (2018), focusing on GC-MS analysis of tilapia off-flavors.*

Rubens Caio de Souza *holds a degree in mathematics from Fundação Educacional Unificada Campograndense (2016) and a MSc in Mathematics from Universidade Federal do Estado do Rio de Janeiro (2021). He is currently a PhD student in the Defense Engineering Program, supervised by Professors Itamar Borges Jr. and Julio Cesar Duarte at IME. He has experience using artificial intelligence for data analysis in chemistry and materials science, focusing on the molecular properties of photon absorption and emission, as well as developing software and educational materials for teaching and scientific communication.*

Itamar Borges Jr. *is a full professor of Physical Chemistry at IME and, since 2003, has been a research fellow of the Brazilian Council for Scientific Research (CNPq). He was chair of the Brazilian Symposium of Theoretical Chemistry (SBQT) in 2013. Since 2020, he has been a permanent associate editor of the Journal of Molecular Modeling. At IME, he has taught and supervised undergraduate and graduate students on various topics in Physical Chemistry, Theoretical Chemistry, and Machine Learning, including organic photovoltaics, materials science, spectroscopy, supercritical fluids, chemical defense, and energetic materials. He has published over 100 peer-reviewed research papers in these fields.*

References

1. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A.; *Nature* **2018**, *559*, 547. [Crossref]
2. Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P.; *Adv. Sci.* **2019**, *6*, 1900808. [Crossref]
3. Dral, P. O.; *J. Phys. Chem. Lett.* **2020**, *11*, 2336. [Crossref]
4. Dral, P. O. In *Advances in Quantum Chemistry*, vol. 81; Ruud, K.; Brändas, E. J., eds.; Academic Press, 2020, p. 291. [Crossref]
5. Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q.; *J. Chem. Inf. Model.* **2021**, *61*, 3197. [Crossref]

6. Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A.; *Acc. Chem. Res.* **2021**, *54*, 849. [Crossref]
7. Westermayr, J.; Gastegger, M.; Schütt, K. T.; Maurer, R. J.; *J. Chem. Phys.* **2021**, *154*, 230903. [Crossref]
8. Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A.; *Nat. Mater.* **2021**, *20*, 750. [Crossref]
9. Schleder, G. R.; Fazzio, A.; *Rev. Bras. Ensino Fis.* **2021**, *43*, e20200407. [Crossref]
10. Zhang, J.; Chen, D.; Xia, Y.; Huang, Y.-P.; Lin, X.; Han, X.; Ni, N.; Wang, Z.; Yu, F.; Yang, L.; Yang, Y. I.; *J. Chem. Theory Comput.* **2023**, *19*, 4338. [Crossref]
11. Bhat, V.; Callaway, C. P.; Risko, C.; *Chem. Rev.* **2023**, *123*, 7498. [Crossref]
12. Hu, Y.; Buehler, M. J.; *APL Mach. Learn.* **2023**, *1*, 010901. [Crossref]
13. Xu, Y.; Ge, J.; Ju, C.-W.; *Energy Adv.* **2023**, *2*, 896. [Crossref]
14. Margraf, J. T.; *Angew. Chem., Int. Ed.* **2023**, *62*, e202219170. [Crossref]
15. Bustillo, L.; Laino, T.; Rodrigues, T.; *Chem. Sci.* **2023**, *14*, 10378. [Crossref]
16. Back, S.; Aspuru-Guzik, A.; Ceriotti, M.; Gryn'ova, G.; Grzybowski, B.; Gu, G. H.; Hein, J.; Hippalgaonkar, K.; Hormázabal, R.; Jung, Y.; Kim, S.; Kim, W. Y.; Moosavi, S. M.; Noh, J.; Park, C.; Schrier, J.; Schwaller, P.; Koji Tsuda, K.; Vegge, T.; Anatole von Lilienfeld, O.; Walsh, A.; *Digital Discovery* **2024**, *3*, 23. [Crossref]
17. Cheng, A. H.; Ser, C. T.; Skreta, M.; Guzmán-Cordero, A.; Thiede, L.; Burger, A.; Aldossary, A.; Leong, S. X.; Pablo-García, S.; Strieth-Kalthoff, F.; Aspuru-Guzik, A.; *Faraday Discuss.* **2024**, *256*, 10. [Crossref]
18. Aspuru Guzik, A.; Bechtel, T.; Bernales, V.; Biggin, P. C.; Bigi, F.; Borges Jr., I.; Briling, K. R.; Cheung, J.; Collins, C. M.; Darmawan, K. K.; David, N.; Day, G. M.; Deringer, V. L.; Draxl, C.; Dyer, M.; Annabel Eardley-Brunt, Evans, R.; Fairlamb, I.; Franklin, B. A.; George, J.; Goulding, M.; Grundy, J.; Hafizi, R.; Hakkennes, M.; Hickey, N.; James, G.; Veronika Juraskova, V.; Kalikadien, A. V.; Kapil, V.; Heather J. Kulik, Kumar, V.; Kuttner, C.; Lederbauer, M.; Lou, Y.; Mante, E.; Marsh, L.; Martin, J.; Middleton, C.; Nematiaram, T.; Pare, C. W. P.; Pasca, B.; Pickard, C. J.; Ruscic, B.; Ryder, M. R.; Savoie, B. M.; Sun, W.; Szczypiński, F. T.; Taniguchi, T.; Torrisi, S.; Vishnoi, S.; Walsh, A.; Shirui Wang, S.; *Faraday Discuss.* **2024**, *256*, 177. [Crossref]
19. Hey, T.; Tansley, S.; Tolle, K.; Gray, J.; *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft Research: 2009. [Crossref]
20. Agrawal, A.; Choudhary, A.; *APL Mater.* **2016**, *4*, 053208. [Crossref]
21. Weinzierl, T.; In *Principles of Parallel Scientific Computing: A First Guide to Numerical Concepts and Programming Methods*; Weinzierl, T., ed.; Springer International Publishing, 2021, p. 3. [Crossref]
22. Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A.; *J. Phys: Mater.* **2019**, *2*, 032001. [Crossref]
23. Olson, G. B.; *Science* **1997**, *277*, 1237. [Crossref]
24. Kirkpatrick, P.; Ellis, C.; *Nature* **2004**, *432*, 823. [Crossref]
25. Mullard, A.; *Nature* **2017**, *549*, 445. [Crossref]
26. von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A.; *Nat. Rev. Chem.* **2020**, *4*, 347. [Crossref]
27. Tkatchenko, A.; *Nat. Comm.* **2020**, *11*, 4125. [Crossref]
28. Yao, Z.; Lum, Y.; Johnston, A.; Mejia-Mendoza, L. M.; Zhou, X.; Wen, Y.; Aspuru-Guzik, A.; Sargent, E. H.; Seh, Z. W.; *Nat. Rev. Mater.* **2023**, *8*, 202. [Crossref]
29. Duan, C.; Nandy, A.; Kulik, H. J.; *Annu. Rev. Chem. Biomol. Eng.* **2022**, *13*, 405. [Crossref]
30. Boonpalit, K.; Kinchagawat, J.; Namuangruk, S.; *ChemElectroChem* **2024**, *11*, e202300681. [Crossref]
31. Jung, S. G.; Jung, G.; Cole, J. M.; *J. Chem. Inf. Model.* **2024**, *64*, 1187. [Crossref]
32. Raabe, D.; Mianroodi, J. R.; Neugebauer, J.; *Nat. Comput. Sci.* **2023**, *3*, 198. [Crossref]
33. Janet, J. P.; Kulik, H. J.; *Machine Learning in Chemistry*; American Chemical Society, 2020. [Crossref]
34. Dral, P. O.; *Quantum Chemistry in the Age of Machine Learning*; Elsevier, 2023. [Crossref]
35. Ferguson, A. L.; Pfaendtner, J.; *J. Phys. Chem. C* **2023**, *127*, 5197. [Crossref]
36. Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A.; *Chem. Rev.* **2021**, *121*, 9816. [Crossref]
37. Thadani, N. N.; Gurev, S.; Notin, P.; Youssef, N.; Rollins, N. J.; Ritter, D.; Sander, C.; Gal, Y.; Marks, D. S.; *Nature* **2023**, *622*, 818. [Crossref]
38. Carter, E. A.; *Science* **2008**, *321*, 800. [Crossref]
39. Martínez, T. J.; *Acc. Chem. Res.* **2017**, *50*, 652. [Crossref]
40. Houk, K. N.; Liu, F.; *Acc. Chem. Res.* **2017**, *50*, 539. [Crossref]
41. Grimme, S.; Schreiner, P. R.; *Angew. Chem., Int. Ed.* **2018**, *57*, 4170. [Crossref]
42. Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S.; *Angew. Chem., Int. Ed.* **2022**, *61*, e202205735. [Crossref]
43. Marzari, N.; Ferretti, A.; Wolverton, C.; *Nat. Mater.* **2021**, *20*, 736. [Crossref]
44. Teale, A. M.; Helgaker, T.; Savin, A.; Adamo, C.; Aradi, B.; Arbuznikov, A. V.; Ayers, P. W.; Baerends, E. J.; Barone, V.; Calaminici, P.; Cancès, E.; Carter, E. A.; Chattaraj, P. K.; Chermette, H.; Ciofini, I.; Crawford, T. D.; De Proft, F.; Dobson, J. F.; Draxl, C.; Frauenheim, T.; Fromager, E.; Fuentealba, P.; Gagliardi, L.; Galli, G.; Gao, J.; Geerlings, P.; Gidopoulos, N.; Gill, P. M. W.; Gori-Giorgi, P.; Görling, A.; Gould, T.; Grimme,

S.; Gritsenko, O.; Jensen, H. J. A.; Johnson, E. R.; Jones, R. O.; Kaupp, M.; Köster, A. M.; Kronik, L.; Krylov, A. I.; Kvaal, S.; Laestadius, A.; Lewin, M.; Liu, S.; Loos, P.-F.; Maitra, N. T.; Neese, F.; Perdew, J. P.; Pernal, K.; Pernot, P.; Piecuch, P.; Rebolini, E.; L. Reining, L.; Romaniello, P.; Ruzsinszky, A.; Salahub, D. R.; Scheffler, M.; Schwerdtfeger, P.; Staroverov, V. N. Sun, J.; Tellgren, E.; Tozer, D. J.; Trickey, S. B.; Ullrich, C. A.; Vela, A.; Vignale, G.; Wesolowski, T. A.; Xu, X.; Yang, W.; *Phys. Chem. Chem. Phys.* **2022**, *24*, 28700. [Crossref]

45. Huang, B.; von Rudorff, G. F.; von Lilienfeld, O. A.; *Science* **2023**, *381*, 170. [Crossref]
46. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A.; *J. Phys. Chem. Lett.* **2011**, *2*, 2241. [Crossref]
47. Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; C. Lawrence Zitnick, C.; Ulissi, Z.; *ACS Catal.* **2021**, *11*, 6059. [Crossref]
48. Tran, R.; Lan, J.; Shuaibi, M.; Wood, B. M.; Goyal, S.; Das, A.; Heras-Domingo, J.; Kolluru, A.; Rizvi, A.; Shoghi, N.; Sriram, A.; Therrien, F.; Abed, J.; Voznyy, O.; Sargent, E. H.; Ulissi, Z.; Lawrence Zitnick, C.; *ACS Catal.* **2023**, *13*, 3066. [Crossref]
49. Kiralj, R.; Ferreira, M. M. C.; *J. Braz. Chem. Soc.* **2009**, *20*, 770. [Crossref]
50. Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A.; *Chem. Rev.* **2010**, *110*, 5714. [Crossref]
51. Polishchuk, P.; *J. Chem. Inf. Model.* **2017**, *57*, 2618. [Crossref]
52. Hammett, L. P.; *Chem. Rev.* **1935**, *17*, 125. [Crossref]
53. Johnson, C. D.; *The Hammett Equation*; Cambrige University Press: Cambridge, UK, 1973.
54. Anslyn, E. V.; Dougherty, D. A.; *Modern Physical Organic Chemistry*; University Science Books: Herndon, 2006.
55. Jezuita, A.; Ejsmont, K.; Szatylowicz, H.; *Struct. Chem.* **2021**, *32*, 179. [Crossref]
56. Hansch, C.; Leo, A.; Taft, R. W.; *Chem. Rev.* **1991**, *91*, 165. [Crossref]
57. Monteiro-de-Castro, G.; Duarte, J. C.; Borges, I.; *J. Org. Chem.* **2023**, *88*, 9791. [Crossref]
58. Barsted, G.; *86-Year Old Hammett Equation Gets a Machine Learning Update*, https://www.chemistryworld.com/news/86-year-old-hammett-equation-gets-a-machine-learning-update/4017798.article, accessed in March 2025.
59. Costa, P.; Ferreira, V.; Esteves, P. M.; Vasconcellos, M.; *Ácidos e Bases em Química Orgânica*; Bookman Companhia Editora: Porto Alegre, 2005.
60. Iguchi, D.; Ravelli, D.; Erra-Balsells, R.; Bonesi, S. M.; *Molecules* **2020**, *25*, 2061. [Crossref]
61. Emenike, B. U.; Spinelle, R. A.; Rosario, A.; Shinn, D. W.; Yoo, B.; *J. Phys. Chem. A* **2018**, *122*, 909. [Crossref]
62. Tian, Y.; Wang, L.; Fu, G.; Zhang, C.; Lu, R.; Dong, X.; *Theor. Chem. Acc.* **2020**, *139*, 78. [Crossref]
63. Porobić, S. J.; Božić, B. Đ.; Dramićanin, M. D.; Vitnik, V.; Vitnik, Ž.; Marinović-Cincović, M.; Mijin, D. Ž.; *Dyes Pigm.* **2020**, *175*, 108139. [Crossref]
64. Rachuru, S.; Skelton, A. A.; Vandanapu, J.; *Comput. Theor. Chem.* **2020**, *1190*, 113024. [Crossref]
65. Aruda, K. O.; Amin, V. A.; Thompson, C. M.; Lau, B.; Nepomnyashchii, A. B.; Weiss, E. A.; *Langmuir* **2016**, *32*, 3354. [Crossref]
66. Teixeira, R. I.; da Silva, R. B.; Gaspar, C. S.; de Lucas, N. C.; Garden, S. J.; *Photochem. Photobiol.* **2021**, *97*, 47. [Crossref]
67. Kumar, G.; Tibbitts, L.; Newell, J.; Panthi, B.; Mukhopadhyay, A.; Rioux, R. M.; Pursell, C. J.; Janik, M.; Chandler, B. D.; *Nat. Chem.* **2018**, *10*, 268. [Crossref]
68. Zalomaeva, O. V.; Evtushok, V. Y.; Ivanchikova, I. D.; Glazneva, T. S.; Chesalov, Y. A.; Larionov, K. P.; Skobelev, I. Y.; Kholdeeva, O. A.; *Inorg. Chem.* **2020**, *59*, 10634. [Crossref]
69. Mati, I. K.; Cockroft, S. L.; *Chem. Soc. Rev.* **2010**, *39*, 4195. [Crossref]
70. Oliveira-Filho, F. L.; Esteves, P. M.; *J. Chem. Inf. Model.* **2024**, *64*, 3278. [Crossref]
71. Monteiro-de-Castro, G.; Borges Jr., I.; *J. Comput. Chem.* **2023**, *44*, 2256. [Crossref]
72. Máximo-Canadas, M.; Borges, I.; *J. Mol. Model.* **2024**, *30*, 120. [Crossref]
73. Borges Jr., I.; Oliveira, R. S. S.; Oliveira, M. A. S. In *Theoretical and Computational Chemistry*, vol. 22; Mathieu, D., ed.; Elsevier, 2022, ch. 4, p. 81. [Crossref]
74. Duarte, J. C.; Rocha, R. D.; Borges, I.; *Phys. Chem. Chem. Phys.* **2023**, *25*, 6877. [Crossref]
75. Maximo-Canadas, M.; Duarte, J. C.; Nichele, J.; Pereira, K.; Ramos, R.; Borges, I.; *ACS Engineering Au* **2025** *in press*. [Crossref]
76. Maximo-Canadas, M.; Duarte, J. C.; Nichele, J.; Alves, L. S.; Vieira Pereira, L. O.; Franco, L. G.; Borges Jr., I.; *ChemRxiv* **2024**. [Crossref]
77. Sanchez-Lengeling, B.; Aspuru-Guzik, A.; *Science* **2018**, *361*, 360. [Crossref]
78. Melville, L. J.; Burke, K. E.; Hirst, D. J.; *Comb. Chem. High Throughput Screening* **2009**, *12*, 332. [Crossref]
79. Wishart, D. S.; *Curr. Protoc. Bioinf.* **2007**, *18*, 14.1.1. [Crossref]
80. Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A.; *Nat. Rev. Drug Discovery* **2024**, *23*, 141. [Crossref]
81. Comaniciu, D.; Meer, P.; *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *24*, 603. [Crossref]
82. Souza, R. C.; Duarte, J. C.; Goldschmidt, R.; Borges Jr., I.; *J. Braz. Chem. Soc.* **2025**, *36*, e-20250037. [Crossref]

83. Esche, E.; Talis, T.; Weigert, J.; Brand Rihm, G.; You, B.; Hoffmann, C.; Repke, J.-U.; *Chem. Eng. Sci.* **2022**, *251*, 117459. [Crossref]
84. Thomas, E. V.; *Chemom. Intell. Lab. Syst.* **2019**, *195*, 103868. [Crossref]
85. Shamsi, Z.; Cheng, K. J.; Shukla, D.; *J. Phys. Chem. B* **2018**, *122*, 8386. [Crossref]
86. Barrett, R.; Westermayr, J.; *J. Phys. Chem. Lett.* **2024**, *15*, 349. [Crossref]
87. Chang, Y.-C.; Li, Y.-P.; *J. Chem. Theory Comput.* **2023**, *19*, 8598. [Crossref]
88. Sridharan, B.; Sinha, A.; Bardhan, J.; Modee, R.; Ehara, M.; Priyakumar, U. D.; *J. Comput. Chem.* **2024**, *45*, 1886. [Crossref]
89. Gow, S.; Niranjan, M.; Kanza, S.; Frey, J. G.; *Digital Discovery* **2022**, *1*, 551. [Crossref]
90. Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D.; *Chem. Mater.* **2020**, *32*, 4954. [Crossref]
91. Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A.; *Nature Chem.* **2021**, *13*, 505. [Crossref]
92. Fu, L.; Hu, H.; Zhu, Q.; Zheng, L.; Gu, Y.; Wen, Y.; Ma, H.; Yin, H.; Ma, J.; *Nano Res.* **2023**, *16*, 3588. [Crossref]
93. Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O.; *Org. Lett.* **2023**, *25*, 2945. [Crossref]
94. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W.; *Chem. Rev.* **2023**, *123*, 8736. [Crossref]
95. Tynes, M.; Gao, W.; Burrill, D. J.; Batista, E. R.; Perez, D.; Yang, P.; Lubbers, N.; *J. Chem. Inf. Model.* **2021**, *61*, 3846. [Crossref]
96. Xu, P.; Ji, X.; Li, M.; Lu, W.; *npj Comput. Mater.* **2023**, *9*, 42. [Crossref]
97. Fourches, D.; Muratov, E.; Tropsha, A.; *J. Chem. Inf. Model.* **2010**, *50*, 1189. [Crossref]
98. Klarich, K.; Goldman, B.; Kramer, T.; Riley, P.; Walters, W. P.; *J. Chem. Inf. Model.* **2024**, *64*, 1158. [Crossref]
99. Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A.; *J. Chem. Phys.* **2015**, *143*, 084111. [Crossref]
100. Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A.; *Sci. Data* **2016**, *3*, 160086. [Crossref]
101. Linstrom, P. J.; Mallard, W. G.; *NIST Chemistry WebBook*, *NIST Standard Reference Database Number 69*; NIST: Gaithersburg, MD, 2019.
102. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C.; *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171. [Crossref]
103. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A.; *APL Mater.* **2013**, *1*, 011002. [Crossref]
104. Taylor, R. H.; Rose, F.; Toher, C.; Levy, O.; Yang, K.; Buongiorno Nardelli, M.; Curtarolo, S.; *Comput. Mater. Sci.* **2014**, *93*, 178. [Crossref]
105. Talirz, L.; Kumbhar, S.; Passaro, E.; Yakutovich, A. V.; Granata, V.; Gargiulo, F.; Borelli, M.; Uhrin, M.; Huber, S. P.; Zoupanos, S.; Adorf, C. S.; Andersen, C. W.; Schütt, O.; Pignedoli, C. A.; Passerone, D.; VandeVondele, J.; Schulthess, T. C.; Smit, B.; Pizzi, G.; Nicola Marzari, N.; *Sci. Data* **2020**, *7*, 299. [Crossref]
106. Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C.; *npj Comput. Mater.* **2015**, *1*, 15010. [Crossref]
107. Draxl, C.; Scheffler, M.; *J. Phys.: Mater.* **2019**, *2*, 036001. [Crossref]
108. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L.; *J. Chem. Inf. Model.* **2012**, *52*, 2864. [Crossref]
109. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A.; *Sci. Data* **2014**, *1*, 140022. [Crossref]
110. Valdés, J. J.; Tchagang, A. B.; *J. Comput. Chem.* **2024**, *45*, 1193. [Crossref]
111. Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr., R. A.; Tkatchenko, A.; *Sci. Data* **2021**, *8*, 43. [Crossref]
112. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E.; *J. Chem. Phys.* **2018**, *148*, 241733. [Crossref]
113. Yang, Y.; Lao, K. U.; Wilkins, D. M.; Grisafi, A.; Ceriotti, M.; DiStasio, R. A.; *Sci. Data* **2019**, *6*, 152. [Crossref]
114. Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H.; *Sci. Data* **2020**, *7*, 58. [Crossref]
115. Stocker, S.; Csányi, G.; Reuter, K.; Margraf, J. T.; *Nat. Comm.* **2020**, *11*, 5505. [Crossref]
116. Zhu, Y.; Li, M.; Xu, C.; Lan, Z.; *Sci. Data* **2024**, *11*, 948. [Crossref]
117. Rebello, N. J.; Arora, A.; Mochigase, H.; Lin, T.-S.; Shi, J.; Audus, D. J.; Muckley, E. S.; Osmani, A.; Olsen, B. D.; *J. Chem. Inf. Model.* **2024**, *64*, 6464. [Crossref]
118. Ganscha, S.; Unke, O. T.; Ahlin, D.; Maennel, H.; Kashubin, S.; Müller, K.-R.; *Sci. Data* **2025**, *12*, 406. [Crossref]
119. Listgarten, J.; *Nat. Biotechnol.* **2024**, *42*, 371. [Crossref]
120. Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; Csanyi, G.; *Acc. Chem. Res.* **2020**, *53*, 1981. [Crossref]
121. Géron, A.; *Hands-on Machine Learning with Scikit-learn, Keras and Tensor Flow: Concepts, Tools and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: Sebastopol, CA, 2017.
122. Facelli, K.; Lorena, A. C.; Gama, J.; Almeida, T. A.; Carvalho, A. C. P. L. F.; *Inteligência Artificial: uma Abordagem de Aprendizado de Máquina*, 2nd ed.; LTC: Rio de Janeiro, 2022.
123. Joshi, R. P.; Kumar, N.; *Molecules* **2021**, *26*, 6761. [Crossref]

124. von Lilienfeld, O. A.; *Int. J. Quantum Chem.* **2013**, *113*, 1676. [Crossref]
125. Weininger, D.; *J. Chem. Inf. Comput.* **1988**, *28*, 31. [Crossref]
126. Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W.; *Sci. Rep.* **2018**, *8*, 12. [Crossref]
127. Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A.; *Phys. Rev. Lett.* **2012**, *108*, 058301. [Crossref]
128. Landrum, G.; *RDKit: Open-Source Cheminformatics Software*, https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4, accessed in April 2025.
129. ChemAxon, https://chemaxon.com, accessed in April 2025.
130. Vicentini, E. D.; de Oliveira-Filho, A. G. S.; *Quim. Nova* **2021**, *44*, 229. [Crossref]
131. Freund, Y.; Schapire, R. E.; *J. Jpn. Soc. Artif. Intell.* **1999**, *14*, 771. [Crossref]
132. Liang, W.; Dai, H. In *Quantum Chemistry in the Age of Machine Learning*; Dral, P. O., ed.; Elsevier, 2023, ch. 10, p. 233. [Crossref]
133. John, G. H.; Langley, P. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada, 1995. [Crossref]
134. Hosmer, D. W.; Lemeshow, S.; *Applied Logistic Regression*; John Wiley & Sons, Inc.: New York, 2000. [Crossref]
135. Weher, E.; *Biom. J.* **1977**, *19*, 83. [Crossref]
136. Fausett, L.; *Fundamentals of Neural Networks, Architectures, Algorithms, and Applications*; Prentice-Hall: Englewood Cliffs, New Jersey, 1994.
137. Vapnik, V. N.; *The Nature of Statistical Learning*; Springer: New York, 1995.
138. Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; Vapnik, V.; *Adv. Neural Inf. Process. Syst.* **1996**, *1*, 155. [Link] accessed in May 2025
139. Kohavi, R.; *European Conference on Machine Learning*; Springer: Berlin Heidelberg, 1995, p. 174. [Crossref]
140. Freund, Y.; Mason, L. In *International Conference on Machine Learning*, Bled, 1999. [Crossref]
141. Landwehr, N.; Hall, M.; Frank, E.; *Mach. Learn.* **2005**, *59*, 161. [Crossref]
142. Witten, I. H.; Frank, E.; Hall, M. A.; *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann Publishers Inc.: Burlington, 2011.
143. Kingsford, C.; Salzberg, S. L.; *Nat. Biotechnol.* **2008**, *26*, 1011. [Crossref]
144. Friedman, J. H.; *Ann. Statist.* **2001**, *29*, 1189. [Crossref]
145. Freund, Y.; Schapire, R. E.; In *International Conference on Machine Learning*, 1996.
146. Breiman, L.; *Mach. Learn.* **1996**, *24*, 123. [Crossref]
147. Tin Kam, H.; *IEEE Trans. Pattern Analysis and Machine Intelligence* **1998**, *20*, 832. [Crossref]
148. Breiman, L.; *Mach. Learn.* **2001**, *45*, 5. [Crossref]
149. Wu, L.; Yang, Y.; Liu, H.; *Comput. Stat. Data Analysis* **2014**, *70*, 116. [Crossref]
150. Tibshirani, R.; *J. Royal Stat. Society: Series B* **1996**, *58*, 267. [Crossref]
151. Dias, A. L.; Rodrigues, T.; *Nature* **2023**, *624*, 530. [Crossref]
152. Castro Nascimento, C. M.; Pimentel, A. S.; *J. Chem. Inf. Model.* **2023**, *63*, 1649. [Crossref]
153. Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G.; *Nature* **2023**, *624*, 570. [Crossref]
154. Spirling, A.; *Nature* **2023**, *616*, 413. [Crossref]
155. White, A. D.; *Nat. Rev. Chem.* **2023**, *7*, 457. [Crossref]
156. Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; de Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B.; *Digital Discovery* **2023**, *2*, 1233. [Crossref]
157. Pan, J.; *Nat. Comput. Sci.* **2023**, *3*, 5. [Crossref]
158. Thorp, H. H.; *Science* **2023**, *379*, 313. [Crossref]
159. Sanderson, K.; *Nature* **2023**, *615*, 773. [Crossref]
160. Deb, J.; Saikia, L.; Dihingia, K. D.; Sastry, G. N.; *J. Chem. Inf. Model.* **2024**, *64*, 799. [Crossref]
161. Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B.; *Nature Machine Intelligence* **2024**, *6*, 161. [Crossref]
162. Hatakeyama-Sato, K.; Yamane, N.; Igarashi, Y.; Nabae, Y.; Hayakawa, T.; *Sci. Technol. Adv. Mater.:Methods* **2023**, *3*, 2260300. [Crossref]
163. Borges, P. V.; Ladele, A. S.; Cunha, Y. M. B. G.; Moraes, D. S.; Costa, P. B.; Santos, P. T. C.; Rocha, R.; Busson, A. J. G.; Duarte, J. C.; Colcher, S.; *Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web*; SBC: Juiz de Fora, Brazil, 2024. [Crossref]
164. Krenn, M.; Pollice, R.; Guo, S. Y.; Aldeghi, M.; Cervera-Lierta, A.; Friederich, P.; dos Passos Gomes, G.; Häse, F.; Jinich, A.; Nigam, A.; Yao, Z.; Aspuru-Guzik, A.; *Nat. Rev. Phys.* **2022**, *4*, 761. [Crossref]
165. Wellawatte, G. P.; Gandhi, H. A.; Seshadri, A.; White, A. D.; *J. Chem. Theory Comput.* **2023**, *19*, 2149. [Crossref]
166. Wu, Z.; Chen, J.; Li, Y.; Deng, Y.; Zhao, H.; Hsieh, C.-Y.; Hou, T.; *J. Chem. Inf. Model.* **2023**, *63*, 7617. [Crossref]
167. Lamens, A.; Bajorath, J.; *ChemMedChem* **2024**, *19*, e202300586. [Crossref]
168. Lundberg, S. M.; Lee, S.-I.; *31st Conference on Neural Information Processing Systems (NIPS 2017)*; Long Beach, CA, USA, 2017. [Link] accessed in May 2025

169. Štrumbelj, E.; Kononenko, I.; *Knowledge and Information Systems* **2014**, *41*, 647. [Crossref]
170. Hochuli, J.; Helbling, A.; Skaist, T.; Ragoza, M.; Koes, D. R.; *J. Mol. Graphics Modell.* **2018**, *84*, 96. [Crossref]
171. Rodríguez-Pérez, R.; Bajorath, J.; *J. Med. Chem.* **2020**, *63*, 8761. [Crossref]
172. Wojtuch, A.; Jankowski, R.; Podlewska, S.; *J. Cheminf.* **2021**, *13*, 74. [Crossref]
173. Deng, Q.; Hu, J.; Wang, L.; Liu, Y.; Guo, Y.; Xu, T.; Pu, X.; *Chemom. Intell. Lab. Syst.* **2021**, *215*, 104331. [Crossref]
174. Tian, T.; Li, S.; Fang, M.; Zhao, D.; Zeng, J.; *J. Chem. Inf. Model.* **2024**, *64*, 2236. [Crossref]
175. Albornoz, R. V.; Antypov, D.; Blanke, G.; Borges Jr., I.; Bran, A. M.; Cheung, J.; Collins, C. M.; David, N.; Day, G. M.; Deringer, V. L.; Draxl, C.; Eardley-Brunt, A.; Evans, M. L.; Fairlamb, I.; Fieseler, K.; Franklin, B. A.; George, J.; Grundy, J.; Johal, J.; Kalikadien, A. V.; Kapil, V.; Kotopanov, L.; Kumar, V.; Kuttner, C.; Lederbauer, M.; Ojeda-Porras, A. C.; Pang, J.; Parkes, M.; Miles Pemberton, M.; Ruscic, B.; Ryder, M. R.; Sakaushi, K.; Saleh, G.; Savoie, B. M.; Schwaller, P.; Skjelstad, B. B.; Sun, W.; Taniguchi, T.; Taylor, C. R.; Torrisi, S.; Vishnoi, S.; Walsh, A.; Wu, R.; *Faraday Discuss.* **2024**, *256*, 520. [Crossref]
176. Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H.; *J. Chem. Inf. Model.* **2023**, *63*, 4012. [Crossref]
177. Westermayr, J.; Marquetand, P.; *Chem. Rev.* **2021**, *121*, 9873. [Crossref]
178. Gonzalez, L.; Escudero, D.; Serrano-Andres, L.; *ChemPhysChem* **2012**, *13*, 28. [Crossref]
179. Souza, R.; Duarte, J. C.; Goldschmidt, R.; Borges, I.; *J. Chem. Inf. Model.* **2025**, *65*, 3270. [Crossref]
180. Hung, L. S.; Chen, C. H.; *Mater. Sci. Eng., R* **2002**, *39*, 143. [Crossref]
181. Mazzio, K. A.; Luscombe, C. K.; *Chem. Soc. Rev.* **2015**, *44*, 78. [Crossref]
182. Greenstein, B. L.; Hutchison, G. R.; *J. Phys. Chem. Lett.* **2022**, *13*, 4235. [Crossref]
183. Ostroverkhova, O.; *Chem. Rev.* **2016**, *116*, 13279. [Crossref]
184. Borges Jr., I.; Aquino, A. J. A.; Köhn, A.; Nieman, R.; Hase, W. L.; Chen, L. X.; Lischka, H.; *J. Am. Chem. Soc.* **2013**, *135*, 18252. [Crossref]
185. Borges Jr., I.; Uhl, E.; Modesto-Costa, L.; Aquino, A. J. A.; Lischka, H.; *J. Phys. Chem. C* **2016**, *120*, 21818. [Crossref]
186. Modesto-Costa, L.; Borges, I.; Aquino, A. J. A.; Lischka, H.; *J. Chem. Phys.* **2018**, *149*, 6. [Crossref]
187. Borges Jr., I.; Guimarães, R. M. P. O.; Monteiro-de-Castro, G.; Rosa, N. M. P.; Nieman, R.; Lischka, H.; Aquino, A. J. A.; *J. Comput. Chem.* **2023**, *44*, 2424. [Crossref]
188. Miyake, Y.; Saeki, A.; *J. Phys. Chem. Lett.* **2021**, *12*, 12391. [Crossref]
189. Mahmood, A.; Irfan, A.; Wang, J.-L.; *Chin. J. Polym. Sci.* **2022**, *40*, 870. [Crossref]
190. Zhao, Z.-W.; Geng, Y.; Troisi, A.; Ma, H.; *Adv. Intelligent Systems* **2022**, *4*, 2100261. [Crossref]
191. Westermayr, J.; Gilkes, J.; Barrett, R.; Maurer, R. J.; *Nat. Comput. Sci.* **2023**, *3*, 139. [Crossref]
192. Greenstein, B. L.; Hutchison, G. R.; *J. Phys. Chem. C* **2023**, *127*, 6179. [Crossref]
193. Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; *Sci. Adv.* **2019**, *5*, eaay4275. [Crossref]
194. Sahu, H.; Rao, W.; Troisi, A.; Ma, H.; *Adv. Ener. Mater.* **2018**, *8*, 1801032. [Crossref]
195. Munshi, J.; Chen, W.; Chien, T.; Balasubramanian, G.; *J. Chem. Inf. Model.* **2021**, *61*, 134. [Crossref]
196. Kranthiraja, K.; Saeki, A.; *ACS Appl. Mater. Interfaces* **2022**, *14*, 28936. [Crossref]
197. Malhotra, P.; Biswas, S.; Sharma, G. D.; *ACS Appl. Mater. Interfaces* **2023**, *15*, 37741. [Crossref]
198. Hong, G.; Gan, X.; Leonhardt, C.; Zhang, Z.; Seibert, J.; Busch, J. M.; Bräse, S.; *Adv. Mater.* **2021**, *33*, 2005630. [Crossref]
199. Kumar, K.; *React. Chem. Eng.* **2024**, *9*, 496. [Crossref]
200. Zou, S.-J.; Shen, Y.; Xie, F.-M.; Chen, J.-D.; Li, Y.-Q.; Tang, J.-X.; *Mater. Chem. Front.* **2020**, *4*, 788. [Crossref]
201. Gómez-Bombarelli, R.; Aguilera-lparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; *Nat. Mater.* **2016**, *15*, 1120.
202. Furukori, M.; Nagamune, Y.; Nakayama, Y.; Hosokai, T.; *J. Mater. Chem. C* **2023**, *11*, 4357. [Crossref]
203. Bu, Y.; Peng, Q.; *J. Phys. Chem. C* **2023**, *127*, 23845. [Crossref]
204. Kim, H.; Lee, K.; Kim, J. H.; Kim, W. Y.; *J. Chem. Inf. Model.* **2024**, *64*, 677. [Crossref]
205. Li, G. Y.; Han, K. L.; *Wiley Interdiscip. Rev.:Comput. Mol. Sci.* **2018**, *8*, e1351. [Crossref]
206. Mousavizadegan, M.; Firoozbakhtian, A.; Hosseini, M.; Ju, H.; *TrAC, Trends Anal. Chem.* **2023**, *167*, 117216 (and references therein). [Crossref]
207. Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A.; *J. Chem. Phys.* **1989**, *90*, 5622. [Crossref]
208. Cameron, A. R.; Proud, A. J.; Pearson, J. K.; *J. Chem. Theory Comput.* **2023**, *19*, 51. [Crossref]
209. Holm, S.; Unzueta, P. A.; Thompson, K.; Martínez, T. J.; *J. Chem. Theory Comput.* **2023**, *19*, 4474. [Crossref]
210. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A.; *J. Chem. Theory Comput.* **2015**, *11*, 2087. [Crossref]
211. Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K.; *Nat. Comm.* **2020**, *11*, 5223. [Crossref]
212. Manzhos, S.; Carrington Jr., T.; *Chem. Rev.* **2021**, *121*, 10187. [Crossref]
213. Marquetand, P. In *Comprehensive Computational Chemistry*, 1st ed.; Yáñez, M.; Boyd, R. J. eds.; Elsevier: Amsterdam, 2024, p. 413. [Crossref]

214. Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z.; *J. Phys. Chem. Lett.* **2018**, *9*, 2725. [Crossref]
215. Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M.; *J. Phys. Chem. Lett.* **2021**, *12*, 4902. [Crossref]
216. Shu, Y.; Truhlar, D. G.; *J. Chem. Theory Comput.* **2020**, *16*, 6456. [Crossref]
217. Shu, Y.; Varga, Z.; Sampaio de Oliveira-Filho, A. G.; Truhlar, D. G.; *J. Chem. Theory Comput.* **2021**, *17*, 1106. [Crossref]
218. Shu, Y.; Varga, Z.; Parameswaran, A. M.; Truhlar, D. G.; *J. Chem. Theory Comput.* **2024**, *20*, 7042. [Crossref]
219. Westermayr, J.; Dral, P. O.; Marquetand, P. In *Quantum Chemistry in the Age of Machine Learning*; Dral, P. O., ed.; Elsevier: Amsterdam, 2023, ch. 20, p. 467. [Crossref]
220. Jirasek, F.; Hasse, H.; *Annu. Rev. Chem. Biomol. Eng.* **2023**, *14*, 31. [Crossref]
221. Zhao, X.; Luo, T.; Jin, H.; *Ind. Eng. Chem. Res.* **2022**, *61*, 8542. [Crossref]
222. Roach, L.; Rignanese, G.-M.; Erriguible, A.; Aymonier, C.; *J. Supercrit. Fluids* **2023**, *202*, 106051. [Crossref]
223. Zhu, K.; Müller, E. A.; *J. Phys. Chem. B* **2020**, *124*, 8628. [Crossref]
224. Liu, Y.; Hong, W.; Cao, B.; *Energy* **2019**, *188*, 116091. [Crossref]
225. Bozal-Ginesta, C.; Pablo-García, S.; Choi, C.; Tarancón, A.; Aspuru Guzik, A.; *ChemRxiv.* **2025**. [Crossref]
226. Liang, J.; Ye, S.; Dai, T.; Zha, Z.; Gao, Y.; Zhu, X.; *Sci. Data* **2020**, *7*, 400. [Crossref]
227. de Oliveira-Filho, A. G. S.; Ornellas, F. R.; Peterson, K. A.; Mielke, S. L.; *J. Phys. Chem. A* **2013**, *117*, 12703. [Crossref]
228. Wang, Y.; Wang, D.; *J. Chem. Phys.* **2018**, *149*, 034302. [Crossref]
229. de Oliveira-Filho, A. G. S.; Ornellas, F. R.; Bowman, J. M.; *J. Phys. Chem. Lett.* **2014**, *5*, 706. [Crossref]
230. Han, S.; de Oliveira-Filho, A. G. S.; Shu, Y.; Truhlar, D. G.; Guo, H.; *ChemPhysChem* **2022**, *23*, e202200039. [Crossref]
231. Murrell, J. N.; Carter, S.; Farantos, S. C.; Huxley, P.; Varandas, A. J. C.; *Molecular Potential Energy Functions*; Wiley Interscience: Chichester, West Sussex, UK, 1984.
232. Araújo, J. P.; Ballester, M. Y.; *Int. J. Quantum Chem.* **2021**, *121*, e26808. [Crossref]
233. Jiang, B.; Li, J.; Guo, H.; *Int. Rev. Phys. Chem.* **2016**, *35*, 479. [Crossref]
234. Li, J.; Jiang, B.; Guo, H.; *J. Chem. Phys.* **2013**, *139*, 204103. [Crossref]
235. Jiang, B.; Guo, H.; *J. Chem. Phys.* **2013**, *139*, 054112. [Crossref]
236. Jiang, B.; Li, J.; Guo, H.; *J. Phys. Chem. Lett.* **2020**, *11*, 5120. [Crossref]
237. Jiang, B.; Guo, H.; *J. Chem. Phys.* **2014**, *141*, 034109. [Crossref]
238. NumPy, https://numpy.org, accessed in April 2025.
239. Pandas, https://pandas.pydata.org, accessed in April 2025.
240. Scikit-Learn, https://scikit-learn.org, accessed in April 2025.
241. TensorFlow, https://www.tensorflow.org, accessed in April 2025.
242. Keras, https://keras.io, accessed in April 2025.
243. PyTorch, https://pytorch.org, accessed in April 2025.
244. Python, https://python.org, accessed in April 2025.
245. StackExchange, https://stackexchange.com, accessed in April 2025.
246. Data Science, https://datascience.stackexchange.com, accessed in April 2025.
247. CrossValidated, https://stats.stackexchange.com, accessed in April 2025.
248. Matter Modeling, https://mattermodeling.stackexchange.com, accessed in April 2025.
249. Matthes, E.; *Python Crash Course: A Hands-On, Project-Based Introduction to Programming*, 2nd ed.; No Starch Press: San Francisco, 2019.
250. Deitel, P.; Deitel, H.; *Intro to Python for Computer Science and Data Science*; Pearson: New Jersey, 2016.
251. Russell, S. J.; Norvig, P.; *Artificial Intelligence: A Modern Approach*; Pearson: New Jersey, 2016.
252. Morenney, L.; *AI and Machine Learning for Coders*; O'Reilly Media: Sebastopol, 2020.

Submitted: December 12, 2024
Published online: May 14, 2025