Title

The Evolution of Data-Driven Modeling in Organic Chemistry

Journal

ACS Central Science, 7(10)

ISSN

2374-7943

Authors

Williams, Wendy L
Zeng, Lingyu
Gensch, Tobias
et al.

Publication Date

2021-10-27

DOI

10.1021/acscentsci.1c00535

ACS central science

http://pubs.acs.org/journal/acscii

Outlook

The Evolution of Data-Driven Modeling in Organic Chemistry

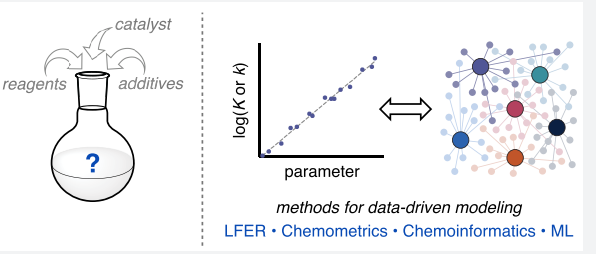
Wendy L. Williams,[‡] Lingyu Zeng,[‡] Tobias Gensch,* Matthew S. Sigman,* Abigail G. Doyle,* and Eric V. Anslyn*

Cite This: *ACS Cent. Sci.* 2021, 7, 1622−1637

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Organic chemistry is replete with complex relationships: for example, how a reactant's structure relates to the resulting product formed; how reaction conditions relate to yield; how a catalyst's structure relates to enantioselectivity. Questions like these are at the foundation of understanding reactivity and developing novel and improved reactions. An approach to probing these questions that is both longstanding and contemporary is data-driven modeling. Here, we provide a synopsis of the history of data-driven modeling in organic chemistry and the terms used to describe these endeavors. We include a timeline of the steps that led to its current state. The case studies included highlight how, as a community, we have advanced physical organic chemistry tools with the aid of computers and data to augment the intuition of expert chemists and to facilitate the prediction of structure−activity and structure−property relationships.



INTRODUCTION

In recent years, machine learning and artificial intelligence have emerged as powerful tools in organic chemistry.[1−8] As a consequence, we thought it prudent to provide the community with a timeline of events that have both inspired and contributed to the clear uptick in the applications of various data-driven strategies to the chemical sciences. These strategies are rooted in linear free energy relationships (LFERs), of which the Hammett relationship is the paradigmatic example.[9] Classically, these analyses related a single parameter, a mathematical way to describe a subunit or the entirety of a molecule, to chemical reactivity.[10] Although LFERs initially were used to gain mechanistic insight, if a model captures underlying chemical reactivity, it can in principle predict the reactivity of unknown reactions. Nevertheless, the simplicity of LFERs can limit their predictive ability, particularly in complex chemical systems.

Thus, as time evolved, multiparameter approaches to correlate chemical reactivity to structure were introduced. In addition, the advent of computers facilitated the use of increasingly larger data sets and more advanced algorithms to describe and predict the reactivity of more complex systems.[11,12] In parallel with technological advances, many new terms, such as chemometrics and chemoinformatics, were introduced to describe these endeavors. Further, the realization that structure−activity modeling is not restricted to classic substituent effects, nor to seeking linear relationships of data to experimental observables, led to another terminology evolution, e.g., that of machine learning.

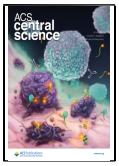
Here, we provide a synopsis of the history of data-driven approaches in organic chemistry alongside a tour through the evolution of terminology used to describe these endeavors (Figure 1). We include key historical steps that led us to the current state of machine learning in chemical synthesis (Figure 2). This Outlook does not aim to be a comprehensive review of modern work but rather will highlight advances in the field with select case studies. For a more comprehensive review of modern approaches, we refer readers to other recent reviews of machine learning in organic chemistry.[2−8]

> While machine learning and artificial intelligence have emerged as active areas in organic chemistry, these fields stem from a rich history of data-driven approaches.

LINEAR FREE ENERGY RELATIONSHIPS

Linear free energy relationships (LFERs) represent a well-established and powerful method to relate reactivity with chemical structure, historically represented by quantitative experimental parameters or descriptors (Figure 3A). Parameters describe the influence of a subunit (substituent) of a molecule,

Received: May 3, 2021
Published: October 19, 2021



https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

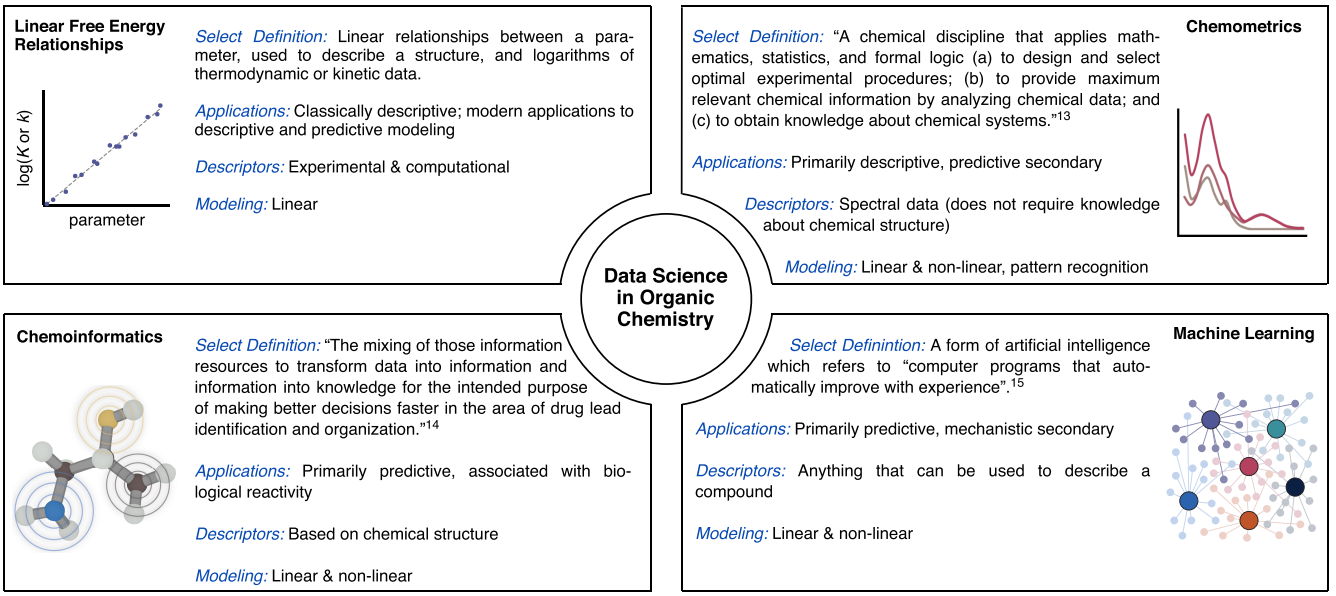


Figure 1. Fields that have contributed to the development of data science in organic chemistry.[13−15]

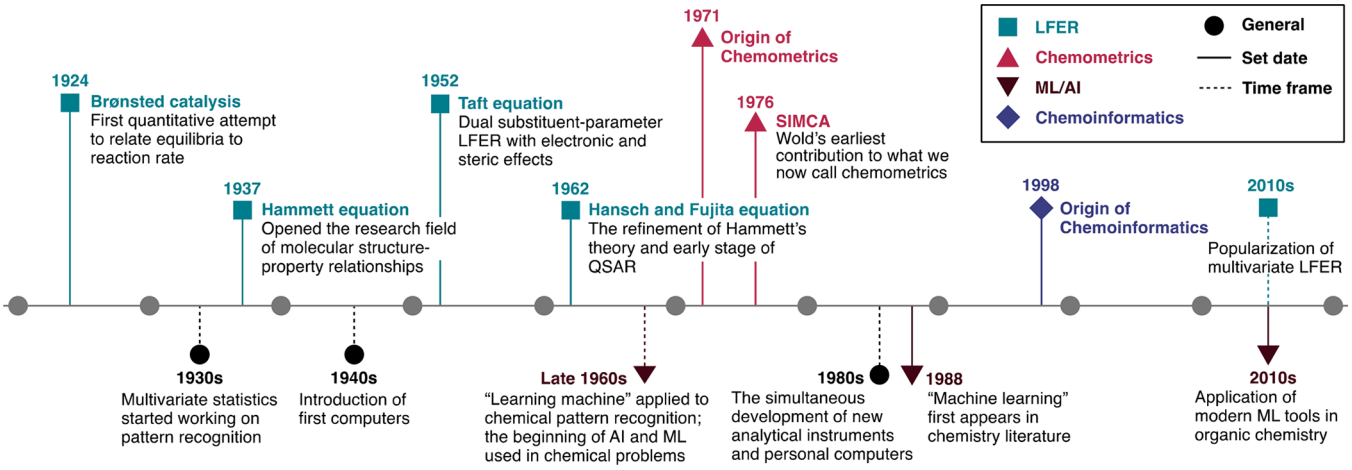


Figure 2. Timeline of major developments of data-driven modeling in organic chemistry.

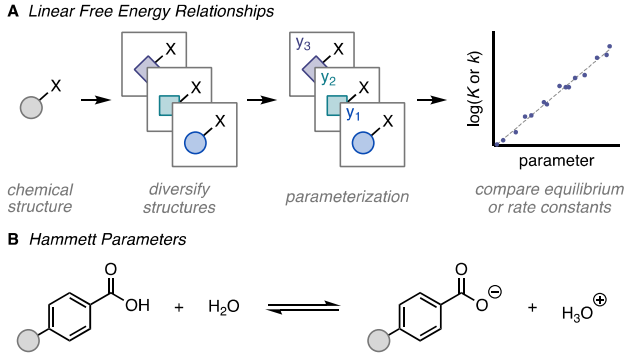


Figure 3. (A) Schematic workflow for linear free energy relationships. (B) Ionization constant of benzoic acid used to derive Hammett parameters.

an entire chemical structure, or even the solvent. The relationships are linear in energy, meaning that the changes in Gibbs free energy resulting from structural modifications are additive. Thus, these relationships involve logarithms of thermodynamic and kinetic data (e.g., equilibrium and rate constants, respectively).[10,16] This is readily understood by recalling that $\Delta G^\circ = -RT(\ln K_{eq})$.

Many of the first parameters in LFERs were derived from reaction equilibria. In 1924, Brønsted and co-workers derived the first quantitative relationship between equilibria and reaction rate.[17] The linear relationship, known now as the Brønsted catalysis law, relates the ionization constant of acids (K_a's) to the rate of general-acid catalyzed reactions via a sensitivity factor α (Figure 4, eq 1). Thus, acid/base dissociation could be a reference process that is related to the outcomes of entirely different reactions. One may consider this as the first correlation that allows for the prediction of reaction behavior based on quantitative parameters (K_a and α). The Brønsted catalysis law marked the beginning of a revolution in physical organic chemistry. Through the 1930s, many papers noted quantitative relationships between reference reactions and entirely new processes. Specifically, benzoic acid acidity was found to correlate with the rate for various reactions involving substrates, reagents, or catalysts bearing substituted aromatics as well as other fragments.[18−20]

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

eq 1	Brønsted Catalysis Law	$\log(k_{cat}) = \alpha \log(K_a) + C$	α: sensitivity factor K_a: ionization constant of acids
eq 2	Hammett Relationship	$\log(K_x / K_H) = \rho\sigma_x$ or $\log(k_x / k_H) = \rho\sigma_x$	K_x: equilibrium constant k_x: rate constant ρ: reaction constant σ_x: substituent parameter
eq 3	Taft Equation	$\log(k_s / k_{CH3}) = \rho^*\sigma^* + \delta E_S$	k_x: rate constant ρ^*: sensitivity factor to polar effects σ^*: polar substituent constant δ: sensitivity factor to inductive effects E_s: steric substituent constant
eq 4	Taft-Topsom Equation	$\log(k_x / k_H) = \rho_F\sigma_F + \rho_x\sigma_x + \rho_\alpha\sigma_\alpha + \rho_R\sigma_R$	k_x: rate constant ρ_x: sensitivity factor σ_x: substituent constant F: field x: induction α: polarizability R: resonance
eq 5	Hansch-Fujita Equation	$\log(1 / C) = k\pi + k'\pi^2 + \rho\sigma + k''$	C: IC_{50} k: constant π: difference in water partition coefficients[a] ρ: sensitivity factor σ: Hammett constant, refers to ortho position of aromatic ring
eq 6	Equation for Free-Wilson Approach	$BA = \Sigma a_i x_i + \mu$	BA: biological activity a_i: contribution of a structural feature x_i: presense[b] or absense of a fragmanet μ: average contribution of the parent molecule

Figure 4. Equations for evolution of free energy relationships. [a]Difference between the analogue as compared to the unsubstituted compound. [b]$x_i = 1$ if the fragment is present; $x_i = 0$ if it is absent.

Thus, in 1937, Hammett introduced an equation that provided a quantitative description of these relationships.[9] The Hammett parameter, σ_x, for a substituent x is derived from the ionization constant of the corresponding substituted benzoic acid (Figure 3B). The relative stability of an x-substituted benzoate ion is influenced by the electronics of the substituent; e.g., an electron-donating substituent destabilizes the benzoate ion whereas an electron-withdrawing substituent stabilizes the benzoate ion. The Hammett relationship (Figure 4, eq 2) correlates induction and resonance contributions from substituents (the σ_x-values) to the reactivity of a wide range of organic structures. The ρ-value reveals the sensitivity of a reaction to the induction and resonance changes imparted by the x-substituents relative to x = H. Nearly every class of organic reaction has been analyzed using the Hammett equation or its extended forms (Figure 4).

An illustrative example in the area of asymmetric catalysis comes from the Jacobsen group in their development of a Mn^{III} salen-catalyzed enantioselective epoxidation of alkenes (Figure 5A). Jacobsen and co-workers used an LFER to understand the impact of changing the salen ligand substituent on the enantioselectivity (related to $\Delta\Delta G^{\ddagger}$) and the mechanism of the asymmetric epoxidation reaction.[21] A Hammett plot demonstrated a linear correlation of the donating ability of the substituent, as measured by σ_p (subscript p here refers to a substituent in the *para* position), to the enantioselectivity. On the basis of this observation and other experimental evidence, the researchers concluded that the variation in enantioselectivity resulted from changes in the position of the epoxidation

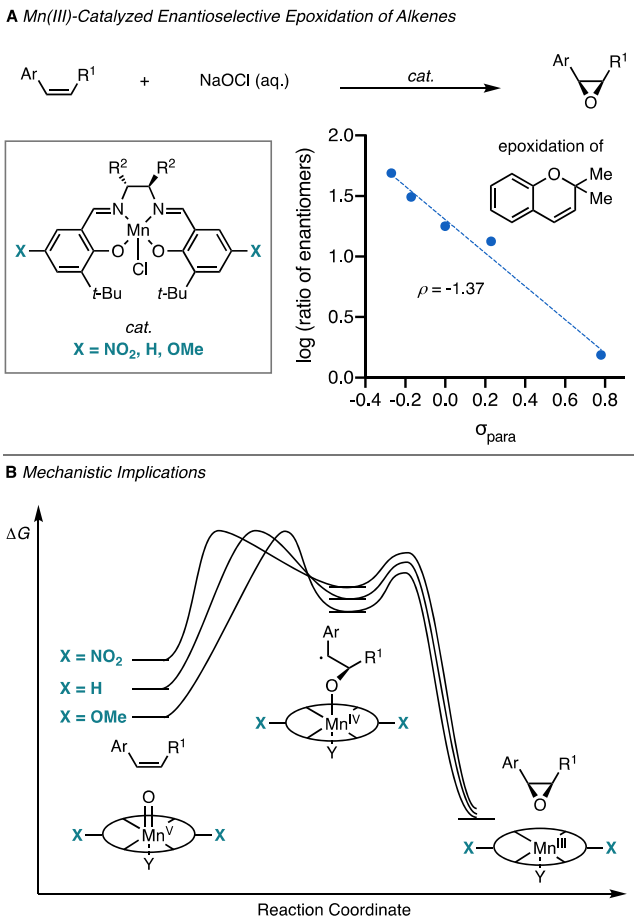


Figure 5. (A) Hammett plot for the Mn^{III} salen-catalyzed enantioselective epoxidation of alkenes. (B) Mechanistic implications for the position of the transition state along the reaction coordinate. Adapted from ref 22. Copyright 1998 American Chemical Society.

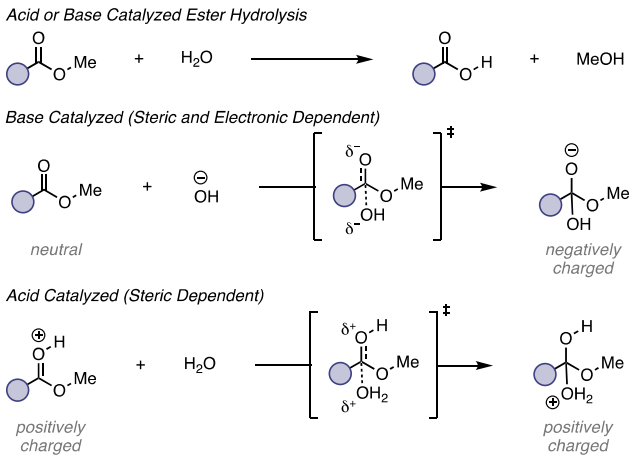


Figure 6. Mechanisms of ester hydrolysis under acid or base catalysis.

transition state relative to the reaction coordinate (Figure 5B).[22] An increase in electron density of the ligand resulted in a milder oxidant that would proceed via a more product-like transition state with greater nonbonding interactions between the catalyst and substrate due to proximity, thus increasing enantioselectivity.

Mechanistic applications of univariate LFERs, such as the Jacobsen example, have been successful in cases where

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

substrates and catalysts can be systematically modified to isolate the impact of a single molecular property and mitigate the effects of other parameters.[23] Further, breaks in linearity, or outliers of the model, often provide additional insights into the mechanism.[24] However, the simplicity of the traditional descriptors used in an LFER can limit the obtainable insight in more complex scenarios. A univariate LFER assumes a linear relationship between a single parameter and reactivity or selectivity; however, chemical reactions are generally more complex as reaction outcomes are dependent on numerous factors and often in a nonlinear manner.

The challenge of modeling reactions influenced by multiple parameters has been investigated throughout the history of LFERs. In an overly simplistic sense, all chemical reactivity can be divided into at least steric and electronic effects (of course, solvent effects are critical, but our focus is on individual structural changes to reactants). In this vein, in 1952, Taft reported a two variable approach that derived electronic and steric parameters from the rates of acid- and base-catalyzed esterification/hydrolysis (Figure 6).[25] Taft assumed that base-catalyzed hydrolysis would be influenced by both steric and electronic effects, whereas acid-catalyzed hydrolysis would only be influenced by steric effects. This assumption is founded upon the nature of the rate-determining step: formation of the tetrahedral-carbon intermediate. In the case of base catalysis, this step involves a change in charge: a neutral substrate is converted to a negatively charged intermediate implicating a dominating role for electronic effects. In contrast, under acidic conditions, a positively charged substrate is converted into a positively charged intermediate. Thus, for acid catalysis, there is no change in formal charge; therefore, electronic effects are mitigated and steric effects dominate. According to these assumptions, Taft derived electronic and steric parameters from the rates of ester hydrolysis under basic and acidic conditions, respectively. On the basis of these assumptions, he derived a dual substituent LFER that separated electronic (σ^*) and steric (E_s) effects (Figure 4, eq 3).[25−27]

This multiparameter approach inspired others to explore increasingly more parameters. One particularly well-known approach, the Taft-Topsom equation, separated several substituent effects (e.g., Figure 4, eq 4). Here, parameters for field, induction, polarizability, and resonance and their contributions to the observed reactivity (the associated σ_x-values) were defined.[28] Building on this, in 1962, Hansch and Fujita and co-workers moved the field of LFERs toward phenomena more relevant to the pharmaceutical sciences and biochemistry (Figure 4, eq 5), introducing correlations with partition coefficients (such as log P and π).[29] This advance is recognized as the origin of quantitative structure−activity relationships (QSAR) as well as the foundation for the field of chemoinformatics, a term introduced several decades later (vide infra).

COMPUTERS

The earliest examples of LFERs relied on the use of experimentally derived parameters. As quantum chemical methods have improved, the use of computationally derived parameters emerged as an alternative to experimental parameters to describe molecules.[30,31] In an early example of the application of computationally derived parameters to LFERs, Eyring and co-workers demonstrated that the computed energy of the highest occupied molecular orbital (HOMO) of phenols (Figure 7A) linearly correlated with their oxidation potential (Figure 7B).[32] On the basis of this relationship, the oxidation potentials for 180 additional phenols, some of which had never been synthesized, were predicted (Figure 7C). Although the paper does not provide validation for these values, it serves as an example of the power of computational parameters to predict the reactivity of molecules before their synthesis.

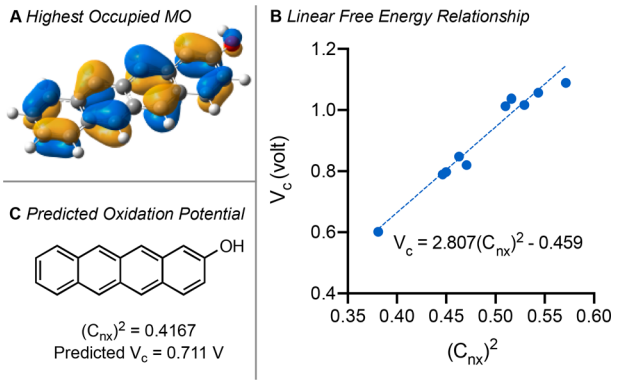


Figure 7. (A) Computed highest occupied molecular orbital. (B) LFER relating $(C_{nx})^2$, a measure of the energy of the HOMO, to oxidation potential (V_c). (C) Predicted oxidation potential based on computed $(C_{nx})^2$.[32]

Computational parameters offer many benefits over experimental ones, such as the ability to parametrize a structure preceding synthesis or access to parameters with no observable experimental equivalent. A level of automation can also be introduced with the derivation of computational parameters.[30] The relatively good accuracy at low computational cost of density functional theory (DFT)[33] has facilitated the use of computational parameters in linear free energy relationships.[31] However, when computing several structures, especially in cases where multiple conformers need to be surveyed, computational cost can become a challenge. In addition, computational parameters are sensitive to the model system used, such as functional/basis set, solvent model, or parametrization of a single conformer or several conformers.

The introduction of computers also made an epochal shift in physical organic chemistry by facilitating the use of large data sets with more extensive computational approaches.[11] Along with the Free-Wilson approach developed in 1964 (Figure 4, eq 6), the early QSAR models used multivariate regression to relate biological activity to the presence or absence of certain substructures in a molecule.[34,35] Also, in this same decade, pattern recognition approaches born from the field of applied mathematics from the 1930s entered the chemistry literature (Figure 2), giving rise to the origin of chemometrics.[36] We note that even LFERs are a form of pattern recognition when σ, σ^+, and σ^- Hammett values are compared to find the best linear fit, which thereby imparts insight into charge and resonance effects. Importantly, as delineated here, we see a gradual progression of single variable linear free energy relationships to multivariate algorithms, leading ultimately to large-data chemometric methods.

DATA SETS

Another consideration in the progression of data science in organic chemistry is the availability of large, high-quality experimental data sets.[37,38] These data sets have been compiled from either high-throughput experimentation (HTE) or the combination of disparate data sets from the literature. Although

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

HTE has a rich history in the field of biology, its adoption into the field of chemistry only occurred recently and has mainly been adopted in industrial settings.[39] Alternatively, the organic literature contains a large volume of data, but it is often stored in different unstructured formats. Pioneering work from Lowe introduced an open-access database that extracted data from USPTO.[40] Other efforts, such as the Open Reaction Database, are seeking to expand access to experimental data through open access schema and a centralized repository.[a] While compiling data from the literature has enabled the use of larger data sets, a challenge with this approach is the bias of literature reactions toward positive results, such that only reactions with high yields or selectivity are reported. However, negative results provide important insight into a chemical system and are necessary to build predictive models.

The ability to access structured data sets that represent the chemical space of a reaction and provide more information on reaction progress and outcomes beyond yield and selectivity will continue to aid in the development of data science in organic chemistry.

CHEMOMETRICS

Chemometrics emerged as a discipline partially due to the ability to use computers in chemistry.[41−43] In the 1960s, many

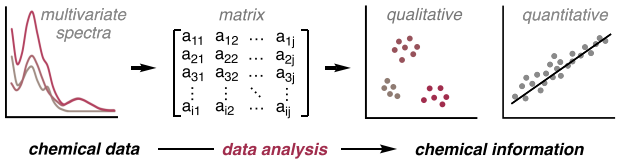


Figure 8. General workflow for chemometrics.

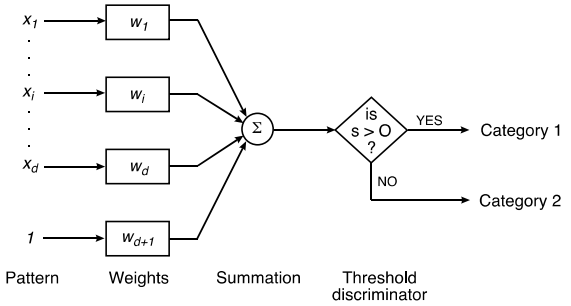


Figure 9. Schematic representation of the binary pattern classifier, the result of which is multicategory pattern classification by least-squares. Reproduced from ref 48. Copyright 1969 American Chemical Society.

branches of chemistry were generating large data sets from spectroscopy, chromatography, kinetics, and other experimental methods. However, no statistical methods at the time were available to cope with data sets containing many variables (often several hundred). With timing that coincides with the advent of

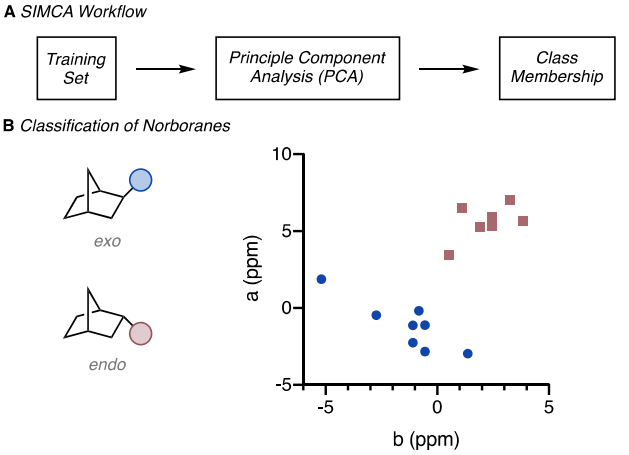


Figure 10. (A) Workflow for Soft Independent Modeling of Class Analogy (SIMCA). (B) SIMCA used for classifying exo and endo norboranes.[60]

chemometrics, pattern recognition techniques or classification methods, now referred to as machine learning, were introduced to the chemical sciences.[44,45] In fact, chemical pattern recognition is now regarded, in part, as an origin of chemometrics, and the two terms are often synonymous.

In 1971 (Figure 2), Kowalski and Wold coined the word "chemometrics" (Figure 8) and shortly after founded the International Chemometrics Society in 1974. Their definition of chemometrics is very broad: "*the application of mathematical and statistical tools to chemistry*".[41] While other definitions of chemometrics have been published,[46,47] one that we believe is particularly contemporary comes from Massart et al.[13] in 1998: "*A chemical discipline that applies mathematics, statistics and formal logic (a) to design and select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems.*" These definitions are so broad in order to encompass machine learning or artificial intelligence in any chemical endeavor, including synthetic methodology development.

From 1969 onward (Figure 2), a series of papers[48−51] applied a "computerized learning machine" (a historical term for machine learning methods) to chemical problems. For example, Jurs, Kowalski, and Isenhour applied a learning machine to the interpretation of low-resolution mass spectra of organic compounds, initiating an area of research that would later culminate in the fully fledged chemical data analysis software ARTHUR.[41] They used a single threshold logic unit (TLU) for binary classification (Figure 9). The TLU is an early model of an artificial neuron that returns +1 or −1 based on the sign of the sum of all vector elements after a linear transformation of the input. The model is trained by a gradient descent method now known as "delta rule" from iterative observations of individual training set members.[52] The iterative training method means that this model truly "learns from experience", a notion commonly associated with machine learning.

In this study, each compound was represented by the scaled intensity at integer m/z ratios in the fixed range of 12−132 u. The interpretation task was then transformed into a series of 26 binary classifications to determine the number of carbon, hydrogen, oxygen, and nitrogen atoms in a molecule. This method was further refined to recognize substructures or

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

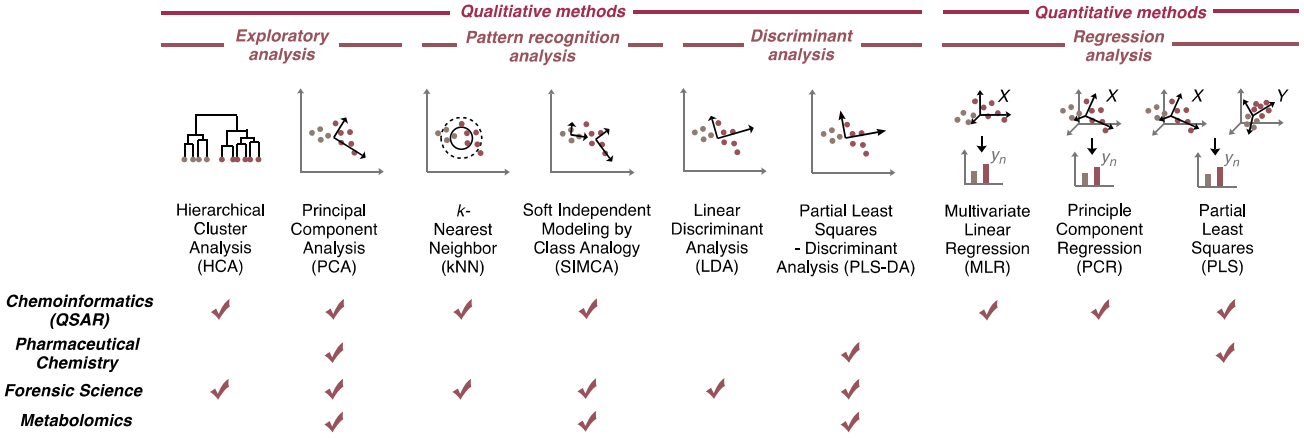


Figure 11. Representative data analysis methods used in chemometrics and their contributions to other disciplines. Exploratory analysis summarizes the main characteristics in multidimensional data. For examples, HCA clusters data by distance; PCA projects data to the first few principal components to seek the largest variance. Pattern recognition analysis and discriminant analysis can both classify data into different groups. Pattern recognition analysis creates general patterns and classifies new objects into groups. For examples, kNN uses a plurality vote of its *k*-nearest neighbors (e.g., inside solid/dashed line circle); SIMCA calculates the residual distance from the disjoint PCA models for each group. Discriminant analysis requires the label of independent variables (X) for classification. LDA projects X data to seek the greatest separation between the different groups; PLS-DA is a PLS variant based on categorical dependent variables (Y). As quantitative methods, regression analysis builds the model to give continuous prediction. MLR regresses Y on the X directly. PCR regresses Y on a subset of the principle components of X. PLS projects both X and Y to a new space, in which X explains the maximum variance in Y. Black and gray axes describe original and new data spaces, respectively. y_n stands for the n^{th} dependent variable.

combine information from mass and infrared spectra in the interpretation tasks.[53,54] Aspects of the modeling procedure, such as training set design and feature selection, remain relevant to data driven predictive analysis today (see below).[49]

Another example of a learning algorithm that was used for pattern recognition is Cora ("cortex"), which consists of feature selection and a voting scheme.[55] Ioffe and co-workers utilized this tool for modeling catalytic reactivity in two case studies:[56] (1) the activity of oxides as heterogeneous catalysts for CO oxidation, in which the components were represented by several physicochemical properties, and (2) the use of V_2O_5 as a catalyst for oxidation of various hydrocarbons, in which the starting materials and products were described by quantum chemical calculations.

Shortly later, Kowalski and Bender applied unsupervised learning to visualize high-dimensional chemical feature spaces in two-dimensional plots using linear projections as well as nonlinear manifold learning techniques.[57,58] They first demonstrated the utility of such two-dimensional representations for subsequent clustering using a divisive hierarchical clustering method and classification by the *k*-nearest neighbor method. As an example, they demonstrated the clustering and classification of the acid/base character of element oxides on the basis of six physicochemical properties of the elements themselves, thus predicting the reactivity with the parameters indicative of chemical structure. This is a clear direct tie to the use of multiparameter LFERs for predictive purposes.

In 1976, Wold[59] reported the method of Soft Independent Modeling of Class Analogy (SIMCA). SIMCA separates data into classifications by first performing a principal component analysis (PCA) on a data set to determine key features and then separates the data into classes on the basis of these features (Figure 10A). SIMCA is considered to be the origin of modern chemometrics as opposed to simple curve fitting, such as used in LFERs.[36,59,60] As the first example, Wold and co-workers performed SIMCA analysis of ^{13}C NMR data of norbornanes. The data were analyzed to determine if the structure of a norbornane is exo or endo and whether there existed consistent patterns for each type of molecule (Figure 10B).[60] In fact, most of the early advances involving SIMCA were for classification.[36] As Figure 11 displays, a number of chemometrics approaches have been developed for a variety of disciplines:[36] chemoinformatics,[61] metabolomics,[62−64] medicinal and pharmaceutical chemistry,[65,66] forensic science,[47] and food science.[67,68] In Figure 11, we depict a breakdown of the common chemometrics methods and their most common graphical results as well as their general utility. The application of Bayesian statistics has also been explored.[69]

CHEMOINFORMATICS

With the increasing reliance on informatics in many scientific fields, in silico chemistry has significantly expanded the areas of

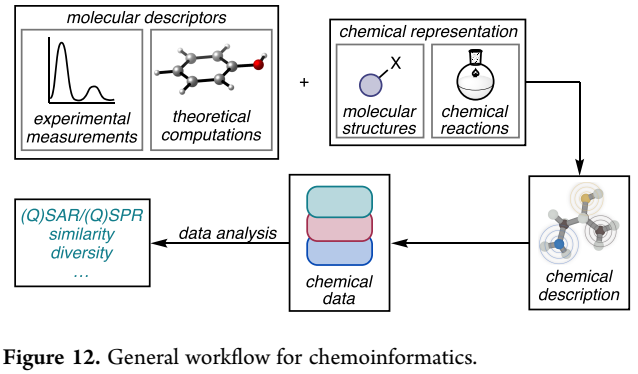


Figure 12. General workflow for chemoinformatics.

possible chemical investigations, i.e., chemoinformatics (Figure 12).[70] Even though the term "chemoinformatics" took shape in late 1990s, the field originated from several beginnings.[71] Brown first defined the term in 1998: "*Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and*

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

organization".[14] However, modern definitions no longer imply that chemoinformatics is necessarily only linked to drug discovery.[70,72−74] For example, Gasteiger and Engel generally defined this discipline as "*the application of [the] informatics method to solve chemical problems*".[75] Chemoinformatics can be described as a theoretical chemistry discipline complementary to quantum chemistry and force-field molecular modeling,[1,76] which focuses on describing molecular structure in a favorable format (for example, as matrices) for use in statistical modeling. Irrespective of this broader definition, chemoinformatics is primarily associated with QSAR or quantitative structure−property relationships (QSPRs) focused upon drug-lead identification.[76]

Early QSAR models, such as Hansch and Free-Wilson analysis, were generally based on multivariate regression with limited features.[11,34] Although groundbreaking, these approaches were only valid for closely related compounds and when linear modeling was applicable. Modern QSAR has increased the use of global models, which are trained on a broad range of compounds, even those lacking structural similarity. Also, the application of sophisticated computational algorithms, embodied in machine learning techniques (discussed below), makes chemoinformatics capable of handling large-scale data sets.[76,77] Chemoinformatics covers a broad range of scientific strategies from chemical data collection and analysis to the exploration of structure−activity relationships and prediction of in vivo compound activities.[78]

A general chemoinformatics model often has a "two-part process" to convert molecules to features and then to properties: (1) encode a compound as feature vectors; (2) map the feature vectors to the property of interest by applying chemoinformatics methods (Figure 12).[76] Compared with other branches of computational chemistry, chemoinformatics involves data processing that cannot be done without in silico mathematics and depends on large data sets that cannot be compressed to standard mathematical models.[70] As with chemometrics, chemoinformatics depends upon mathematical, statistical, and machine learning methods to translate chemical data into chemical information with the assistance of a computer. The two fields have borrowed heavily from each other and use many of the same methods.[79] The difference is that chemometrics uses multivariate data from instruments (e.g., spectral data), which often requires no information about chemical structure, while chemoinformatics concentrates on generating data from the description of the chemical structure. Although these two disciplines have a different focus on solving problems in chemistry, some literature[70,79] regards chemometrics as part of chemoinformatics.

Chemoinformatics can be considered as a very specific application of machine learning with an emphasis on modeling structure−property relationships for molecules. Similar to chemometrics, knowledge external to chemistry (e.g., graph theory for developing chemical descriptors) can be integrated into the workflow before machine learning methods are applied.[1]

■ ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial intelligence is a general term for the study and construction of "intelligent agents": devices or programs capable of cognitive functions such as learning, problem solving, and decision making upon perceiving stimuli.[80] This encompasses, but is not limited to, the field of machine learning, which refers to programs that improve with experience at performing a task.[15]

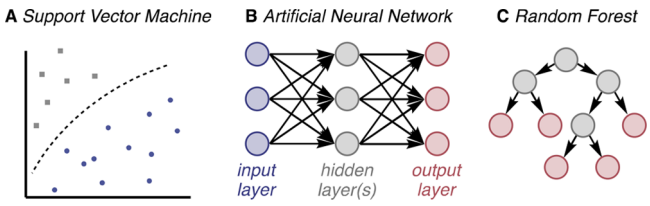


Figure 13. Machine learning methods. (A) SVM models represent data as points in space and cluster the data by hyperplanes. (B) ANN models contain a series of artificial neurons that receive, process, and output data. (C) RF models analyze data by a series of decision trees.

With these notions in mind, the application of artificial intelligence and machine learning in a general sense to chemical problems began in the late 1960s. Several seminal papers were published in 1969, including two highly influential projects based on heuristics, and thus belong to artificial intelligence in a broader sense: Dendral and Logic and Heuristics Applied to Synthetic Analysis (LHASA).

Referred to as the first expert system, the Dendral project led by Feigenbaum, Buchanan, Lederberg, and Djerassi made extensive use of heuristics with the aim of scientific hypothesis generation.[81,82] Its utility for chemical questions was first demonstrated by the enumeration of isomers of organic molecules given a molecular formula[81] as well as the interpretation of mass spectral data of ketones.[82] Corey and Wipke's LHASA[83] was the first implementation of the formalized rules of retrosynthesis that Corey had published two years prior.[84] This marked the beginning of the ongoing and active development of computer-assisted synthesis planning software.[85] Other groups pursued this goal early on,[86] including Dugundji and Ugi's use of a matrix representation of molecules[87] or Gelernter et al. applying another heuristics-based approach.[88]

Around 1988, the term "machine learning" (ML) started appearing in the titles of chemistry literature (Figure 2).[89−92] The introduction of machine learning techniques in the early 1990s marked a pivotal point in the evolution of chemical analysis methodology.[93] This led to a blur between what is considered chemometrics or machine learning, but we believe a subtle distinction has evolved: a reliance on linear relationships is now more associated with chemometrics, whereas nonlinear relationships and large data sets are more commonly considered ML.[70] In actuality, there is no sharp distinction between the statistical methods of chemometrics and machine learning. In both cases, computers are used to generate models, which have the capacity to cope with advanced model selection algorithms that are increasingly more sophisticated as the machine learning/chemometric community improves their approaches.

While chemometricians will claim support vector machines (SVMs), artificial neural networks (ANNs), and forest methods (such as random forest, RF) for their field, organic chemists generally consider these "advanced chemometrics" methods as machine learning (Figure 13). There is literature[94−97] that compares the results from traditional chemometric methods to what is now commonly termed as machine learning methods (e.g., SVM, ANN, RF). However, we believe these methods ("traditional" or "advanced") should not be simply compared by their performance. The performance of the model relies on whether the algorithm is suitable for the data, which means that methods should be selected according to the properties of the data and the hypothesis to be analyzed. For example, as Brereton

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

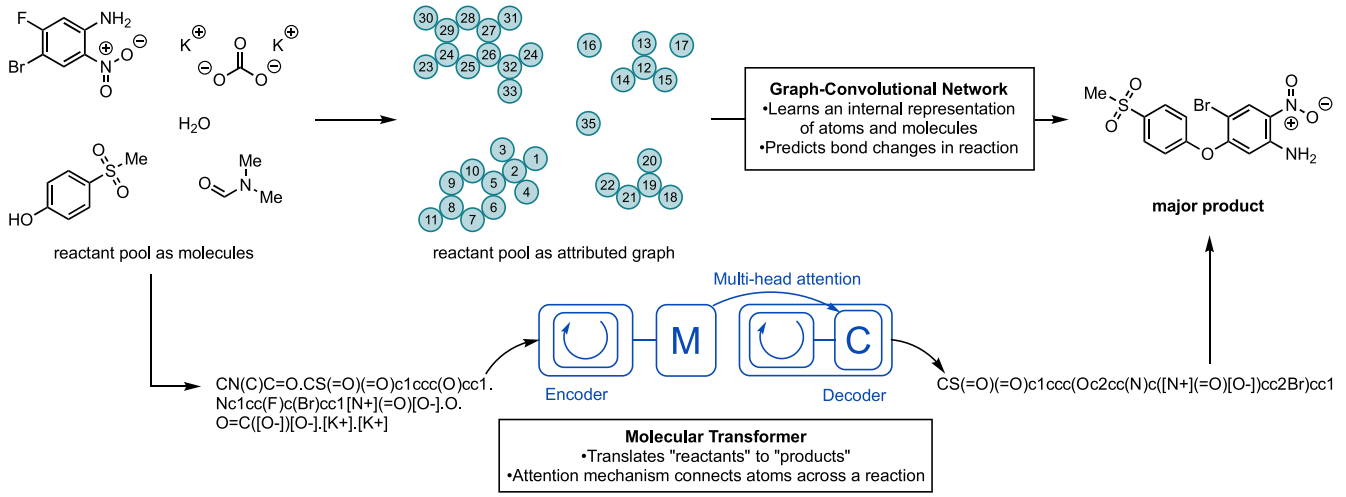


Figure 14. Graph-convolutional network (GCN) and a molecular transformer as applied to modeling molecules and predicting reaction outcomes. Adapted from ref 112 with permission from the Royal Society of Chemistry, Copyright 2019. Adapted from ref 114. Copyright 2019 American Chemical Society.

and Lloyd articulated in their review, most applications of SVM are on data sets with small numbers of variables in analytical chemistry.[98] However, there is no inherent reason they cannot be extended to highly multivariable data sets. Both chemometrics and machine learning evolved from the fields of pattern recognition and computational learning theory by applying statistical methods to improve model performance.[99] For relatively small or sparse data sets, simple machine learning algorithms (e.g., multiple linear regression, linear discriminant analysis (LDA), PCA, and PLS) may work well. With larger amounts of data and with higher complexity, especially in high-throughput screening (HTS) assays, the sought-after predictions often benefit from more sophisticated algorithms (e.g., SVM, RF, ANN, etc.).[3]

Similarly, the early applications of SVMs have rapidly developed since the late 1990s in several research areas, including bioinformatics and biometrics. SVMs map data as points in higher-dimensional spaces that can be used for classification by identifying hyperplanes that separate clusters in the data. In the early 2000s, SVMs were introduced to chemistry for QSAR and protein structure studies.[100−103] Also, SVM can be applied to both classification and predictive problems.[98] ANNs are highly complex methods that pass information through interconnected layers of mathematical transformations, thereby generating internal representations of the original data.[104] In chemistry, the interest in neural-network computing has grown rapidly since 1986,[107] and different aspects of ANN methods have been investigated in QSAR studies since the 1990s.[108] More recently, the complex cognitive capacities of some ANN architectures have enabled applications beyond the prediction of numerical targets, as described below. RFs are an ensemble method to improve prediction accuracy based on a majority-voting scheme, which is an extension of decision tree algorithms.[105,106]

The use of chemometrics and ML in chemical applications corresponds mainly to either descriptive or predictive settings, respectively. In descriptive modeling, the focus is to find a quantitative relationship between probability distributions that can then be interpreted and applied to make predictions. Conversely, in a predictive setting, we find that machine learning is the more commonly used terminology. For example, being able to make predictions, such as the outcome of a reaction, is the primary purpose of these mathematical applications, and the interpretation is perhaps secondary. Impressive applications have become possible using machine learning techniques, some of which autonomously generate scientific hypotheses (see below). Not only can this streamline the chemical discovery process, but also it can lead to experiments and discoveries that might not have been considered on the basis of human intuition or reasoning alone. This has been illustrated for several tasks relevant to organic chemistry including molecular design, synthesis planning, and reaction optimization and discovery. Thus, we now briefly highlight some representative examples, but a full discussion of the more recent achievements is out of the scope of this Outlook.[109]

REACTION OPTIMIZATION AND CATALYST DESIGN

Machine learning has been applied to several aspects of reaction optimization, including: (a) the qualitative prediction of what reaction occurs between a set of starting materials and reagents, (b) the quantitative prediction of reaction outcomes given examples of a known reaction with variations of reaction conditions, reagents, or catalysts, and (c) autonomous reaction exploration, which requires one to select reaction conditions to try in each successive test iteration.

Qualitative Prediction. Deciding if a reaction occurs between certain starting materials in the presence of certain reagents and predicting the product are key intuitive skills that chemists learn during their training. This skill also serves as the basis for suggesting novel reactions. Work toward computer models with such capabilities has been carried out all throughout the history of AI applications in chemistry.[87,110,111] Most early approaches utilized expert-coded reaction templates to map reactions to starting materials, a daunting task given the sheer quantity of possible reactions.

An alternative to expert-coded reaction templates is to learn the chemical reactivity from a large reaction database with appropriate ML models. One approach to this is the use of graph-convolutional networks (GCN). In a recent example, Coley et al. achieved this by representing molecules as annotated graphs and using a GCN to learn an internal representation of the atoms and molecules and, finally, predicting the bond

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

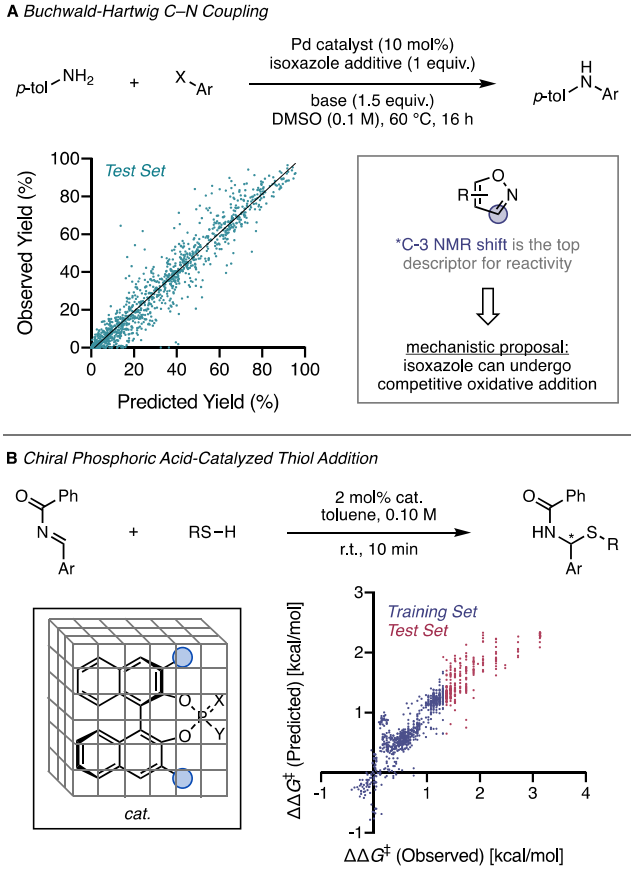


Figure 15. Machine learning for the quantitative prediction of reaction outcomes. (A) Prediction of the reaction yield for Buchwald–Hartwig C–N couplings.[117] (B) Prediction of the enantioselectivity of chiral phosphoric acid-catalyzed thiol additions to *N*-acylimines.[118]

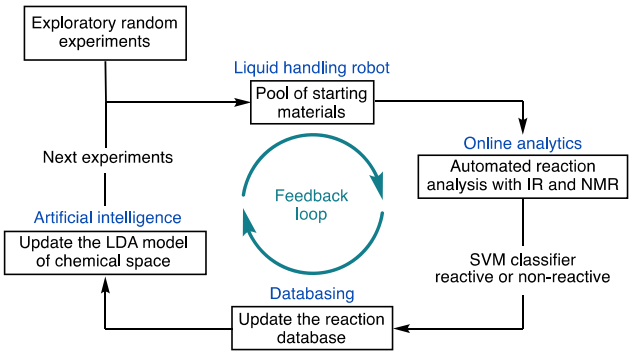


Figure 16. Automated synthesis platform. Figure adapted with permission from ref 123. Copyright 2018 Springer Nature.

changes happening in a reaction (Figure 14).[112] Another recent approach to template-free reaction prediction utilizes techniques originally used in natural language processing. Formally, the prediction task is treated as a "translation" of the language of reactants/reagents to the language of the products. In practice, reaction information is already stored in text form, most commonly the SMILES (Simplified Molecular-Input In-Line System) strings.[113] Schwaller et al. showed that a transformer model with a multihead attention mechanism, termed the molecular transformer, was able to perform this translation task (Figure 14).[114] Both groups employed data sets from the USPTO patent database[40] to train and test their models. In a common subset consisting of ca. 400k and 40k reactions for training and testing, respectively, the GCN model predicted the highest ranked product correctly in ca. 86% of the test cases and the molecular transformer correctly in ca. 90%. Both model outputs ranked lists of possible products, which can further be used to predict possible side products in a reaction. In the molecular transformer model, a visualization of the attention mechanism revealed a fascinating finding. The model learned to perform atom mapping by correctly connecting the atoms in the products to the corresponding atoms in the reagents without having been trained on mechanistic information.[115] In fact, many other models for reactivity prediction require the atom mapping as input information along with the reagents and products of a reaction, which is a major drawback because atom mapping is a very tedious and error-prone procedure for large reaction data sets. In some reactions, even the determination of the correct atom mapping can involve difficult mechanistic studies.

Quantitative Prediction. The quantitative prediction of reaction outcomes when changing individual reactants or reaction conditions is a well-established use of statistics in chemistry with the previously discussed linear free energy relationships and the widely used engineering tool of Design of Experiments[116] as prominent examples. These methods work well when continuous reaction parameters such as temperature or concentration are changed or when a single molecular component, such as the catalyst structure, is varied. The change of several discrete parameters throughout a reaction optimization can necessitate high-dimensional molecular representations and complex machine learning models to predict the reaction outcomes and potentially discover better-performing catalysts. For example, Dreher and the Doyle group investigated the impact of various reaction conditions on the yield of a Buchwald–Hartwig C–N coupling reaction (Figure 15A).[117] They considered changes to the catalyst, base, and one of the substrates as well as the effect of a potentially reactive additive to mimic functional group tolerance. Each molecule was represented by quantum-chemically obtained properties such as vibrational frequencies or atomic partial charges. A random forest model was trained to predict the yields of 4608 reactions in total. The effects of the training set size and composition were investigated and the interpretation of important features in the final model led to a mechanistic hypothesis concerning the challenging electrophilic side reactivity of the functional additive.

In another example, the Denmark group investigated the enantioselectivity of chiral phosphoric acid-catalyzed thiol additions to *N*-acylimines as a function of the catalyst and both the substrates (Figure 15B).[118] The catalysts were represented by their average steric occupancy on a three-dimensional grid in order to reflect conformational flexibility in the molecular representation. Using a total of 2150 reactions, they found that support vector regression and deep feed-forward neural networks were best suited to predict the enantioselectivity of each reaction. Although the training set for the deep feed-forward neural network only comprised ligands that gave less than 80% ee, the model was able to predict the enantioselectivity of ligands that gave higher than 80% ee.

Many other approaches have been taken to represent molecules to predictive algorithms without the need for DFT computed descriptors.[119] This includes representations rooted in chemoinformatics such as molecular fingerprints[120] or text-

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622–1637

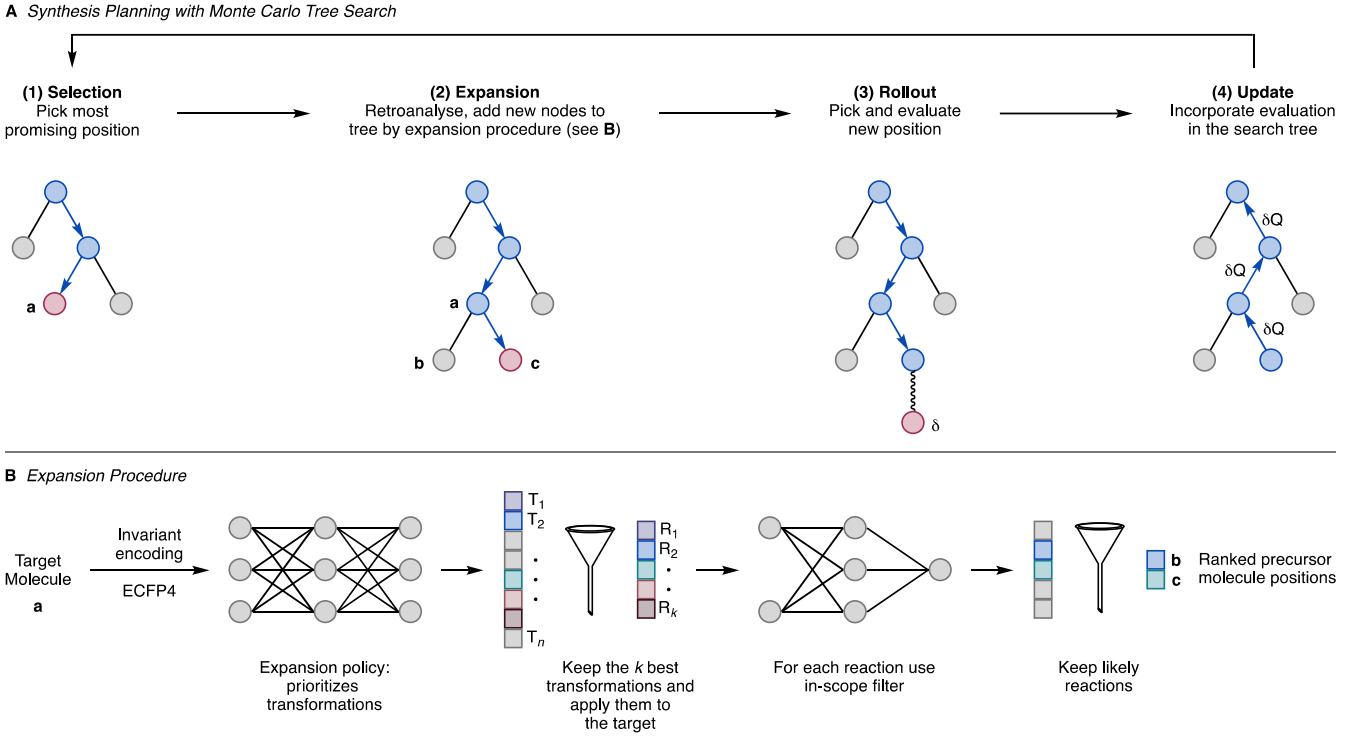


Figure 17. Retrosynthesis algorithm. (A) Schematic overview of the Monte Carlo tree search. (B) Schematic overview of the expansion procedure. Figure adapted with permission from ref 125. Copyright 2018 Springer Nature.

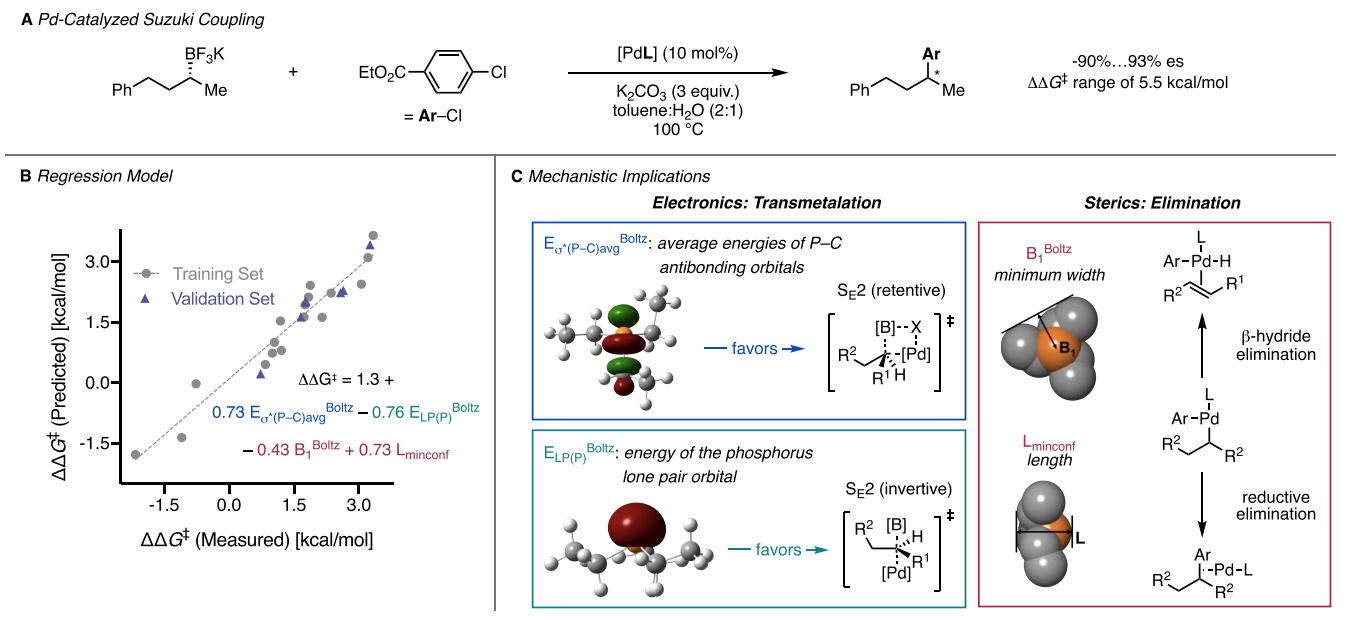


Figure 18. (A) Pd-catalyzed Suzuki reaction. (B) Multivariate regression model for the full data set. (C) Contributing parameters and mechanistic implications. Adapted from ref 134 with permission from AAAS, Copyright 2018.

based fingerprints[121] as well as representations that are intended to provide a physical description of the molecules such as the coulomb matrix and its evolution SLATM,[122] but a full discussion of this field is beyond the scope of this Outlook.

Autonomous Reaction Exploration. Reaction automation has also been used to search for entirely new reactions. Cronin's group used an automated synthesis platform to carry out experiments in a limited chemical space defined by a certain number of molecules as potential starting materials (Figure 16).[123] A support vector machine classifier was trained to detect if a reaction has occurred in each experiment, using the result to populate a reaction database. Using this database, the chemical space was modeled by linear discriminant analysis (LDA) in order to suggest successive experiments with a higher probability of a reactive combination of starting materials. Using this workflow, four new reactions were identified and reproduced in separate batch experiments.

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

An aspect that sets the field of machine learning apart from the pure statistical modeling discussed in other parts of this Outlook is the ability to generate specific scientific hypotheses, enable autonomous reaction performance, or enact reaction/synthesis planning.

As discussed at the beginning of this section, computer-assisted synthesis planning was one of the earliest applications of artificial intelligence to chemical problems. However, a complete retrosynthesis is complicated by the number of synthetic steps that need to be considered, as at each intermediate, multiple alternative disconnections might be considered before arriving at the simple starting materials.[124] This necessitates a ranking of individual branches before fully traversing the complete search tree. In many cases, CASP (Computer Assisted Synthesis Planning) algorithms rely on heuristics to tackle this. The Waller group has combined three ANNs with a Monte Carlo tree search to obtain a retrosynthesis algorithm that does not rely on hand-coded reaction rules, utilizing 12 million reactions from the Reaxys database for training (Figure 17A).[125] The individual neural networks are used to suggest precursors at each step, check if that step is in-scope of a known transformation, and rank that step within the search tree (Figure 17B). Much progress has also been achieved with CASP models based on heuristics by other groups that has led to several commercially available products,[126] including Chematica that was developed by the Grzybowski group using carefully expert-coded reaction templates.[127]

A common question about machine learning models concerns their potential for creativity. In domains outside of chemistry, models have been developed that are capable of generating images, text, or music by sampling from a learned latent space of their domain of applicability.[128,129] The utility of such generative models has also been explored in the context of chemical discovery, most commonly in the context of medicinal chemistry where molecules with specific physiological and physicochemical properties need to be designed.[130,131] One approach is to represent molecules as text that encodes the full structure, for example, by SMILES strings[113] and adapt text-generating models to generate SMILES strings corresponding to new molecules, for example, using techniques developed in the context of natural language processing.[132] This can be used to generate potential drug candidates by applying desired properties such as biological activity or solubility as constraints on the generated SMILES strings[133] and, historically, is called chemoinformatics (see above).

This can lead to the discovery of experiments, catalysts, or reactions that might not have been considered on the basis of human intuition alone. In most cases, such hypotheses do not consist of a single suggested molecule or reaction but rather a large number of suggestions. Using such tools, the task of a chemist can shift from generating hypotheses to ranking and choosing them, thus possibly selecting from a wider pool of ideas.

MODERN EXAMPLES OF LFERs

As presented in this Outlook, ways to describe and predict chemical reactivity have greatly expanded in the field of chemistry. Since the initial introduction of linear free energy relationships, models capable of describing much more complex problems have emerged. However, these new methods do not take away from the power of the descriptive and predictive ability of classic models like linear regression. In a recent report from the Biscoe and Sigman groups, multivariate LFERs were highlighted as a way to analyze a reaction mechanism and predict reactivity (Figure 18).[134] The authors found that the enantiospecificity (*es*) of a Pd-catalyzed alkyl-Suzuki reaction could be described by the computed orbital energy of the phosphorus ligand lone pair ($E_{LP(P)}$) and the computed energy of the P−C σ^* orbitals ($E_{\sigma^*(P-C)}$), a measure of the π-backbonding ability of the ligand. This correlation suggests that the stereoinvertive transmetalation proceeds through a coordinatively unsaturated intermediate that would be stabilized by a strong sigma donation from the ligand. In contrast, the stereoretentive transmetalation is stabilized by π-backbonding, indicating that this transformation involves the precoordination of a donor on the substrate, likely OH^-. The addition of two steric parameters, the Sterimol parameters B_1^{Boltz} and the length of the ligand (L), to account for competitive β-hydride elimination further improved the fit of the model. The model, which was based on a series of ligands in a training set, gave an excellent fit (R^2 = 0.94). It could also predict the *es* for a validation set of ligands not included in the original model ($R^2_{(EV)}$ = 0.87).

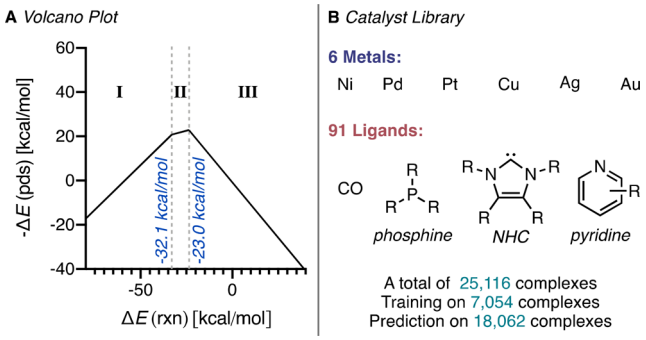


Figure 19. (A) Volcano plot showing ΔE(pds) vs ΔE(rxn). ΔE(pds) is the energy difference of the potential-determining step; region I is reductive elimination, region II is transmetalation, region III is oxidative addition. ΔE(rxn) is the energy difference of oxidative addition. (B) Catalyst library. Adapted from ref 137. Copyright 2021 American Chemical Society.

Developments in machine learning methods have also aided in the implementation of classic models. For example, volcano plots have been used in heterogeneous and homogeneous catalysis to estimate catalyst performance on the basis of Sabatier's principle, which states that an active catalyst should bind substrate neither too tightly nor too loosely (the plateau of the volcano plot).[135,136] In the context of transition metal-catalyzed cross-coupling, the Corminboeuf group demonstrated that a descriptor value such as the relative energies (ΔE) of oxidative addition can determine if a catalyst falls into this active range (Figure 19A).[137] While this descriptor value can be computed by DFT, the computational cost to do so for thousands of catalysts is intractable. Instead, machine learning can be utilized to estimate ΔE values of oxidative addition, thus

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

circumventing costly DFT computations.[138] The catalyst library (Figure 19B) studied in this work consisted of combinations of 91 ligands (CO, phosphines, N-heterocyclic carbenes, and pyridines) and 6 transition metals (Ni, Pd, Pt, Cu, Ag, and Au) for a total of 25 116 possible species for each intermediate. A kernel ridge regression (KRR) model, trained on 7054 complexes, predicted the ΔE values of 18 062 additional complexes. Using a preconstructed volcano plot, 557 complexes were identified to have ΔE descriptors that would be within the active window for catalysis. This work highlights the ability of machine learning to readily screen thousands of possible catalyst/ligand combinations without the need for costly DFT computations.

CONCLUSION

It is clear from the timeline embodied in Figure 2 that the use of data-driven modeling in chemistry has had a long and rich history. Over a period of nearly 100 years, chemists have created a multitude of approaches for examining experimental data to make mechanistic conclusions and predictions of reactivity. The earliest correlations evaluated differences in free energies and took the form of linear univariate (or sometimes multivariate) relationships involving parameters, mainly derived from experimental measurements arising from substituent effects that are dictated by systematic changes in chemical structures (e.g., pK_a, σ, E, etc.). These linear free energy relationships correlated substituent effects (induction, resonance, sterics, etc.) to reactivity and are primarily used to explore reaction mechanisms, but, as we have noted, it is important to realize that the correlations can also be predictive. If parameters for new chemical structures are known, the linear relationships will reveal where the new structures will fall in a spectrum of reactivities. By the early 21st century, this predictive power compelled chemists to become increasingly more sophisticated in the kinds of correlations used, resulting in the use of nonlinear kernel functions, multilayered neural nets, or random forest trees (Figure 13) as well as other mathematical and statistical approaches commonly referred to as "machine learning". In these studies, the parameters have become far broader and often include spectral or computational data, while still retaining elements of electronic and steric substituent effects.

Along this 100-year journey, new terminology was introduced into the literature to differentiate the applications and advent of mathematical techniques as well as experimental analyses and predictive approaches. To follow this evolution of terms, we return to the timeline of Figure 2 and the definitions we choose for this Outlook given in Figure 1. Chemometrics, as originally defined (see discussion of Figure 8 above), is so broad that it encompasses the use of any kind of mathematical and/or statistical approaches involving structural changes, experimental data, or computational parameters to understand and predict a chemical phenomenon. This would include computer analysis, and if being able to predict is dependent upon having learned, it would include the use of machine learning in chemistry. We have emphasized that the exact same protocols are the tools used in both chemometrics and machine learning, i.e., PCA, SVM, RF, ANN, etc. The definition of chemoinformatics is similarly broad, encompassing the use of "informatics" to solve chemical problems of any kind (see discussion of Figure 12 above), where informatics is defined as describing a molecular structure in a computer-readable format, such as a matrix of values. While chemoinformatics was, and still is, primarily associated with drug discovery, the terminology used in this field is only subtly different than that used in chemometrics and, therefore, also machine learning. This brings us to the terminology associated with machine learning, where an additional feature is explicit irrespective of the field in which it is applied, that of automatically improving with experience. If performed with a computer, this implies an artificial intelligence, where there is a cognitive function such as learning, problem solving, and decision making upon perceiving stimuli. When used in organic chemistry for reaction discovery or optimization, the application of computational and statistical methods involving computer analysis to perform these cognitive functions is a subtle, but important, difference from chemometrics and chemoinformatics. The upshot is there is a tangled web of interrelationships of terminology as the field of data-driven science in organic chemistry has evolved over the past 100 years.

With all of these traditional tools in play, there is no wonder that we have seen a significant uptick in machine learning reports in organic chemistry. This has been aligned with questions of how to more effectively use available data, especially in industrial settings, and how to design data acquisition with the intention of using machine learning techniques from the outset. This focus on the "data" aspect is aligned with all types of exciting directions to streamline the goal of the synthetic endeavor by integrating more modern and sophisticated data/computer science algorithms, such as molecular/catalyst design, complex molecule synthesis, reaction optimization, and mechanistic interrogation. Each of these areas will also be aided by updates to parallel reaction screening technologies that integrate data rich outputs (e.g., temporal and kinetic measures[139]), likely resulting in fully automated reaction discovery and optimization workflows. Finally, the entire premise of this field, providing understanding of chemical processes through quantitative featurization, is foundational in how one can imagine using data science in everyday mechanistic investigations and reaction methodology development.

Thus, with such an appreciation and celebration, we can appropriately place any future approaches and applications in catalyst design into perspective with a historical lens, as we have done herein.

> The combination of mathematics in any form with chemical parameters of any form with the use of a machine can and will continue to allow predictions of catalyst kinetics and thermodynamics, which in turn dictate reaction yield as well as enantio- and diastereoselectivity, i.e., the experimental outcomes that synthetic chemists care about the most.

AUTHOR INFORMATION

Corresponding Authors

Eric V. Anslyn − *Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, United States;*

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

orcid.org/0000-0002-5137-8797; Email: anslyn@austin.utexas.edu

Abigail G. Doyle − *Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States; Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-6641-0833; Email: agdoyle@chem.ucla.edu

Tobias Gensch − *Department of Chemistry, TU Berlin, 10623 Berlin, Germany;* orcid.org/0000-0002-1937-0285; Email: tobias.gensch@tu-berlin.de

Matthew S. Sigman − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* orcid.org/0000-0002-5746-8830; Email: matt.sigman@utah.edu

Authors

Wendy L. Williams − *Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States; Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States*

Lingyu Zeng − *Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c00535

Author Contributions

‡W.L.W. and L.Z. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Jose A. Garrido and Dr. Andrzej M. Żurański for helpful discussions and Dr. Jose A. Garrido for assistance with the graphics. A.G.D. and M.S.S. gratefully acknowledge the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607) for support. E.V.A. thanks the Welch Regents Chair (F-0046) for support. T.G. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy − EXC 2008/1 − 390540038 and by a Liebig Fellowship of the Fonds der Chemischen Industrie.

■ ADDITIONAL NOTE

[a]https://docs.open-reaction-database.org/.

■ REFERENCES

(1) Varnek, A.; Baskin, I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inf.* **2011**, *30* (1), 20−32.

(2) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38* (16), 1291−1307.

(3) Panteleev, J.; Gao, H.; Jia, L. Recent Applications of Machine Learning in Medicinal Chemistry. *Bioorg. Med. Chem. Lett.* **2018**, *28* (17), 2807−2815.

(4) Carpenter, K. A.; Cohen, D. S.; Jarrell, J. T.; Huang, X. Deep Learning and Virtual Drug Screening. *Future Med. Chem.* **2018**, *10* (21), 2557−2567.

(5) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59* (6), 2545−2559.

(6) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119* (18), 10520−10594.

(7) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589−604.

(8) Haywood, A. L.; Redshaw, J.; Gaertner, T.; Taylor, A.; Mason, A. M.; Hirst, J. D. Machine Learning in Chemistry: The Impact of Artificial Intelligence. In *Machine Learning for Chemical Synthesis*; The Royal Society of Chemistry, 2020; Chapter 7, pp 169−194; DOI: 10.1039/9781839160233-00169.

(9) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96−103.

(10) Wells, P. R. Linear Free Energy Relationships. *Chem. Rev.* **1963**, *63* (2), 171−219.

(11) Martin, Y. C. Hansch Analysis 50 Years On. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 435−442.

(12) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure−Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120* (3), 1620−1689.

(13) Massart, D. L.; Vandeginste, B. G.; Buydens, L.; Jong, S. D.; Lewi, P. J.; Smeyers-Verbeke, J.; Mann, C. K. Handbook of Chemometrics and Qualimetrics: Part A. *Appl. Spectrosc.* **1998**, *52*, 302A.

(14) Brown, F. K. Chemoinformatics: What Is It and How Does It Impact Drug Discovery. *Annu. Rep. Med. Chem.* **1998**, *33*, 375−384.

(15) Mitchell, T. *Machine Learning*; McGraw Hill, 1997.

(16) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books, 2005.

(17) Brønsted, J.; Pedersen, K. Die Katalytische Zersetzung Des Nitramids Und Ihre Physikalisch-Chemische Bedeutung. *Z. Phys. Chem.* **1924**, *108U*, 185−235.

(18) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17* (1), 125−136.

(19) Dippy, J. F. J.; Watson, H. B. 105. Relationships between Reaction Velocities and Ionisation Constants. *J. Chem. Soc.* **1936**, 436−440.

(20) Burkhardt, G. N.; Ford, W. G. K.; Singleton, E. 4. The Hydrolysis of Arylsulphuric Acids. Part I. *J. Chem. Soc.* **1936**, 17−25.

(21) Jacobsen, E. N.; Zhang, W.; Guler, M. L. Electronic Tuning of Asymmetric Catalysts. *J. Am. Chem. Soc.* **1991**, *113* (17), 6703−6704.

(22) Palucki, M.; Finney, N. S.; Pospisil, P. J.; Güler, M. L.; Ishida, T.; Jacobsen, E. N. The Mechanistic Basis for Electronic Effects on Enantioselectivity in the (Salen)Mn(III)-Catalyzed Epoxidation Reaction. *J. Am. Chem. Soc.* **1998**, *120* (5), 948−954.

(23) Jaffe, H. H. A Reëxamination of the Hammett Equation. *Chem. Rev.* **1953**, *53* (2), 191−261.

(24) Schreck, J. O. Nonlinear Hammett Relationships. *J. Chem. Educ.* **1971**, *48* (2), 103.

(25) Taft, R. W. Polar and Steric Substituent Constants for Aliphatic and O-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. *J. Am. Chem. Soc.* **1952**, *74* (12), 3120−3128.

(26) Taft, R. W. Linear Free Energy Relationships from Rates of Esterification and Hydrolysis of Aliphatic and Ortho-Substituted Benzoate Esters. *J. Am. Chem. Soc.* **1952**, *74* (11), 2729−2732.

(27) Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **1953**, *75* (18), 4538−4539.

(28) Taft, R. W.; Topsom, R. D. The Nature and Analysis of Substitutent Electronic Effects. *Prog. Phys. Org. Chem.* **2007**, *16*, 1−83.

(29) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178−180.

(30) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119* (11), 6561−6594.

(31) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398−2412.

(32) Fueno, T.; Ree, T.; Eyring, H. Quantum-Mechanical Studies on Oxidation Potentials and Antioxidizing Action of Phenolic Compounds. *J. Phys. Chem.* **1959**, *63* (11), 1940−1948.

(33) Kohn, W. Nobel Lecture: Electronic Structure of Matter—Wave Functions and Density Functionals. *Rev. Mod. Phys.* **1999**, *71* (5), 1253−1266.
(34) Kubinyi, H. Free Wilson Analysis. Theory, Applications and Its Relationship to Hansch Analysis. *Quant. Struct.-Act. Relat.* **1988**, *7*, 121−133.
(35) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7* (4), 395−399.
(36) Brereton, R. G. Pattern Recognition in Chemometrics. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 90−96.
(37) Lusher, S. J.; McGuire, R.; Schaik, R. C.; van Nicholson, C. D.; Vlieg, J. de Data-Driven Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today* **2014**, *19* (7), 859−868.
(38) Jin, X.; Wah, B. W.; Cheng, X.; Wang, Y. Significance and Challenges of Big Data Research. *Big Data Res.* **2015**, *2* (2), 59−64.
(39) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8* (6), 601−607.
(40) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Thesis, University of Cambridge, 2012.
(41) Kowalski, B. R. Chemometrics: Views and Propositions. *J. Chem. Inf. Comp. Sci.* **1975**, *15* (4), 201−203.
(42) Hopke, P. K. The Evolution of Chemometrics. *Anal. Chim. Acta* **2003**, *500* (1−2), 365−377.
(43) Lavine, B. K.; Brown, S. D.; Booksh, K. S. *40 Years of Chemometrics−From Bruce Kowalski to the Future*; ACS Publications, 2015.
(44) Fu, K. *Sequential Methods in Pattern Recognition and Machine Learning*; Academic Press, 1968; Vol. *52*.
(45) Wold, S.; Sjöström, M. Chemometrics and Its Roots in Physical Organic Chemistry. *Acta Chem. Scand.* **1998**, *52*, 517−523.
(46) Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*; John Wiley & Sons, 2016.
(47) Kumar, R.; Sharma, V. Chemometrics in Forensic Science. *TrAC, Trends Anal. Chem.* **2018**, *105*, 191−201.
(48) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machines Applied to Chemical Problems. Molecular Formula Determination from Low Resolution Mass Spectrometry. *Anal. Chem.* **1969**, *41* (1), 21−27.
(49) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N. Computerized Learning Machines Applied to Chemical Problems. Convergence Rate and Predictive Ability of Adaptive Binary Pattern Classifiers. *Anal. Chem.* **1969**, *41* (6), 690−695.
(50) Kowalski, B.; Jurs, P.; Isenhour, T. L.; Reilley, C. Computerized Learning Machines Applied to Chemical Problems. Interpretation of Infrared Spectrometry Data. *Anal. Chem.* **1969**, *41* (14), 1945−1949.
(51) Kowalski, B.; Jurs, P.; Isenhour, T. L.; Reilley, C. Computerized Learning Machines Applied to Chemical Problems. Multicategory Pattern Classification by Least Squares. *Anal. Chem.* **1969**, *41* (6), 695−700.
(52) Nilsson, N. J. *Learning Machines*; McGraw-Hill Company: New York, 1965.
(53) Jurs, P. C.; Kowalski, B. R.; Isnehour, T. L.; Reilley, C. N. Investigation of Combined Patterns from Diverse Analytical Data Using Computerized Learning Machines. *Anal. Chem.* **1969**, *41* (14), 1949−1953.
(54) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N. Computerized Learning Machines Applied to Chemical Problems. Molecular Structure Parameters from Low Resolution Mass Spectrometry. *Anal. Chem.* **1970**, *42* (12), 1387−1394.
(55) Bongard, M. M. *Pattern Recognition*; Spartan Books, 1970.
(56) Ioffe, I. I.; Fedorov, V. S.; Mukhenberg, K. M.; Fuks, I. S. Forecasting Chemical Reactions by the Methods of the Statistical Recognition Theory. *Dokl. Akad. Nauk SSSR* **1969**, *189* (6), 1290−1293.
(57) Kowalski, B.; Bender, C. Pattern Recognition. Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94* (16), 5632−5639.
(58) Kowalski, B. R.; Bender, C. F. Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data. *J. Am. Chem. Soc.* **1973**, *95* (3), 686−693.
(59) Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognition* **1976**, *8* (3), 127−139.
(60) Wold, S.; Sjostrom, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. *ACS Symp. Ser.* **1977**, *52*, 243−282.
(61) Ferreira, M. M. Multivariate QSAR. *J. Braz. Chem. Soc.* **2002**, *13* (6), 1−38.
(62) Holmes, E.; Antti, H. Chemometric Contributions to the Evolution of Metabonomics: Mathematical Solutions to Characterising and Interpreting Complex Biological NMR Spectra. *Analyst* **2002**, *127* (12), 1549−1557.
(63) Trygg, J.; Holmes, E.; Lundstedt, T. Chemometrics in Metabonomics. *J. Proteome Res.* **2007**, *6* (2), 469−479.
(64) Madsen, R.; Lundstedt, T.; Trygg, J. Chemometrics in Metabolomics–a Review in Human Disease Diagnosis. *Anal. Chim. Acta* **2010**, *659* (1−2), 23−33.
(65) Biancolillo, A.; Marini, F. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. *Front. Chem.* **2018**, *6*, 576.
(66) Rajalahti, T.; Kvalheim, O. M. Multivariate Data Analysis in Pharmaceutics: A Tutorial Review. *Int. J. Pharm.* **2011**, *417* (1−2), 280−290.
(67) Marini, F. *Chemometrics in Food Chemistry*; Newnes, 2013; Vol. *28*.
(68) Munck, L.; Nørgaard, L.; Engelsen, S. B.; Bro, R.; Andersson, C. Chemometrics in Food Science—a Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance. *Chemom. Intell. Lab. Syst.* **1998**, *44* (1−2), 31−60.
(69) Chen, H.; Bakshi, B. R.; Goel, P. K. Toward Bayesian Chemometrics—A Tutorial on Some Recent Advances. *Anal. Chim. Acta* **2007**, *602* (1), 1−16.
(70) Polanski, J. Chemoinformatics: From Chemical Art to Chemistry In Silico. *Encyclopedia of Bioinformatics and Computational Biology* **2019**, *2*, 601−618.
(71) Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46* (6), 2267−2277.
(72) Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Springer Science & Business Media, 2004; Vol. *275*.
(73) Engel, T.; Gasteiger, J. *Chemoinformatics: Basic Concepts and Methods*; John Wiley & Sons, 2018.
(74) Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21* (2), 151.
(75) Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; John Wiley & Sons, 2006.
(76) Mitchell, J. B. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4* (5), 468−481.
(77) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23* (8), 1538−1546.
(78) Vogt, M.; Bajorath, J. Chemoinformatics: A View of the Field and Current Trends in Method Development. *Bioorg. Med. Chem.* **2012**, *20* (18), 5317−5323.
(79) Engel, T.; Gasteiger, J. *Applied Chemoinformatics: Achievements and Future Opportunities*; John Wiley & Sons, 2018.
(80) Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson, 2020.
(81) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. I. Number of Possible Organic Compounds. Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *J. Am. Chem. Soc.* **1969**, *91* (11), 2973−2976.
(82) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. Applications of Artificial Intelligence for Chemical Inference. II. Interpretation of Low-

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

Resolution Mass Spectra of Ketones. *J. Am. Chem. Soc.* **1969**, *91* (11), 2977−2981.
(83) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166* (3902), 178−192.
(84) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14* (1), 19−38.
(85) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23* (6), 1241−1250.
(86) Bersohn, M.; Esack, A. Computers and Organic Synthesis. *Chem. Rev.* **1976**, *76* (2), 269−282.
(87) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Computers in Chemistry* **1973**, *39* (1), 19−64.
(88) Gelernter, H.; Sridharan, N. S.; Hart, A. J.; Yen, S.-C.; Fowler, F. W.; Shue, H.-J. The Discovery of Organic Synthetic Routes by Computer. *Top. Curr. Chem.* **1973**, *41*, 113−150.
(89) Appel, R.; Hochstrasser, D.; Roch, C.; Funk, M.; Muller, A. F.; Pellegrini, C. Automatic Classification of Two-dimensional Gel Electrophoresis Pictures by Heuristic Clustering Analysis: A Step toward Machine Learning. *Electrophoresis* **1988**, *9*, 136−142.
(90) Wilcox, G.; Poliac, M.; Liebman, M. In *Prediction of 3-Dimensional Protein-Structure From Sequence Using Neural Networks (Machine Learning, Neural Networks)*; American Chemical Society: Washington, DC, 1989; Vol. *198*, p 13-COMP.
(91) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (4), 492−504.
(92) Sternberg, M. J. E.; Lewis, R. A.; King, R. D.; Muggleton, S. Modelling the Structure and Function of Enzymes by Machine Learning. *Faraday Discuss.* **1992**, *93*, 269−280.
(93) Salin, E.; Winston, P. Machine Learning and Artificial Intelligence: An Introduction. *Anal. Chem.* **1992**, *64* (1), 49A−60A.
(94) Conroy, J.; Byrne, H. J.; Ryder, A. G.; Lewis, E.; MacCraith, B. D.; Leger, M. N.; Hennessey, K.; McGlynn, E.; McLaughlin, J. A.; Madden, M. G.; O'Sullivan, G. D.; Ryder, A. G.; Walsh, J. E. Qualitative and Quantitative Analysis of Chlorinated Solvents Using Raman Spectroscopy and Machine Learning. *Proc. SPIE* **2005**, *5826*, 131−142.
(95) Amendolia, S. R.; Cossu, G.; Ganadu, M. L.; Golosio, B.; Masala, G. L.; Mura, G. M. A Comparative Study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia Screening. *Chemom. Intell. Lab. Syst.* **2003**, *69* (1−2), 13−20.
(96) Thissen, U.; Pepers, M.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Comparing Support Vector Machines to PLS for Spectral Regression Applications. *Chemom. Intell. Lab. Syst.* **2004**, *73* (2), 169−179.
(97) Czekaj, T.; Wu, W.; Walczak, B. About Kernel Latent Variable Approaches and SVM. *J. Chemom.* **2005**, *19* (5−7), 341−354.
(98) Brereton, R. G.; Lloyd, G. R. Support Vector Machines for Classification and Regression. *Analyst* **2010**, *135* (2), 230−267.
(99) (a) Kowalski, B. R.; Bender, C. F. Pattern Recognition. Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94* (16), 5632−5639. (b) Kowalski, B. R. Measurement Analysis. *Anal. Chem.* **1975**, *47* (13), 1152A−1162A. (c) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer, 2006.
(100) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2001**, *26* (1), 5−14.
(101) Czermiński, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR. *Quant. Struct.-Act. Relat.* **2001**, *20* (3), 227−240.
(102) Cai, Y.-D.; Liu, X.-J.; Xu, X.; Chou, K.-C. Prediction of Protein Structural Classes by Support Vector Machines. *Comput. Chem.* **2002**, *26* (3), 293−296.
(103) Cai, Y.-D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein. *J. Comput. Chem.* **2002**, *23* (2), 267−274.
(104) Marini, F. Artificial Neural Networks in Foodstuff Analyses: Trends and Perspectives A Review. *Anal. Chim. Acta* **2009**, *635* (2), 121−131.
(105) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.
(106) Breiman, L. Random forests. *Machine Learning* **2001**, *45* (1), 5−32.
(107) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* **1991**, *248* (1), 1−30.
(108) Yousefinejad, S.; Hemmateenejad, B. Chemometrics Tools in QSAR/QSPR Studies: A Historical Perspective. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 177−204.
(109) (a) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59* (51), 22858−22893. (b) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59* (52), 23414−23436.
(110) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163−168.
(111) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62* (10), 1921−1932.
(112) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370−377.
(113) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(114) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572−1583.
(115) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Acv.* **2021**, *7* (15), eabe4166.
(116) Weissman, S. A.; Anderson, N. G. Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications. *Org. Process Res. Dev.* **2015**, *19* (11), 1605−1633.
(117) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186−190.
(118) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), eaau5631.
(119) Pattanaik, L.; Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem.* **2020**, *6* (6), 1204−1207.
(120) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem.* **2020**, *6* (6), 1379−1390.
(121) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2* (1), 015016.
(122) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, *12* (20), 6879−6889.
(123) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature* **2018**, *559* (7714), 377−381.
(124) Ihlenfeldt, W.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1996**, *34* (23−24), 2613−2633.

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637

(125) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604−610.
(126) Wang, Z.; Zhao, W.; Hao, G.; Song, B. Mapping the Resources and Approaches Facilitating Computer-Aided Synthesis Planning. *Org. Chem. Front.* **2021**, *8*, 812.
(127) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904−5937.
(128) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661.
(129) Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *arXiv* **2020**, arXiv:2001.06937.
(130) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(131) Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11* (8), 1496−1505.
(132) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. Exploring Chemical Space Using Natural Language Processing Methodologies for Drug Discovery. *Drug Discovery Today* **2020**, *25* (4), 689−705.
(133) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60* (12), 5682−5698.
(134) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-Catalyzed C−C Bond Formation Enabled through Ligand Parameterization. *Science* **2018**, *362*, 670−674.
(135) Sabatier, P. Hydrogénations et Déshydrogénations Par Catalyse. *Ber. Dtsch. Chem. Ges.* **1911**, *44* (3), 1984−2001.
(136) Sabatier, P. *La Catalysise En Chimie Organique*; Hachette Livre-BNF; 1913.
(137) Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. The Genesis of Molecular Volcano Plots. *Acc. Chem. Res.* **2021**, *54* (5), 1107−1117.
(138) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9* (35), 7069−7077.
(139) Shi, Y.; Prieto, P. L.; Zepel, T.; Grunert, S.; Hein, J. E. Automated Experimentation Powers Data Science in Chemistry. *Acc. Chem. Res.* **2021**, *54* (3), 546−555.

https://doi.org/10.1021/acscentsci.1c00535
ACS Cent. Sci. 2021, 7, 1622−1637