CHEMISTRY

Applying statistical modeling strategies to sparse datasets in synthetic chemistry

Brittany C. Haas[1], Dipannita Kalyani[2], Matthew S. Sigman[1]*

The application of statistical modeling in organic chemistry is emerging as a standard practice for probing structure-activity relationships and as a predictive tool for many optimization objectives. This review is aimed as a tutorial for those entering the area of statistical modeling in chemistry. We provide case studies to highlight the considerations and approaches that can be used to successfully analyze datasets in low data regimes, a common situation encountered given the experimental demands of organic chemistry. Statistical modeling hinges on the data (what is being modeled), descriptors (how data are represented), and algorithms (how data are modeled). Herein, we focus on how various reaction outputs (e.g., yield, rate, selectivity, solubility, stability, and turnover number) and data structures (e.g., binned, heavily skewed, and distributed) influence the choice of algorithm used for constructing predictive and chemically insightful statistical models.

INTRODUCTION

The application of data science and statistical modeling in organic chemistry has emerged as a modern approach to reaction optimization and probing structure-activity relationships. This has encouraged a continually evolving landscape of strategies and questions regarding when to deploy a certain algorithm and the applicability of various types of molecular features. These two challenges are explicitly dependent on the experimental data available, including the number of experiments, the data distribution, and the identity of the output(s) measured. There have been many reviews on machine learning (ML) in synthetic chemistry (*1*, *2*), and we have recently presented perspectives on the history of these topics (*3*) as well as protocols for designing datasets amenable to ML (*4*, *5*). Herein, we describe the tactics by which our group and others construct and interpret statistical models for chemical systems with an aim to describe the details to those that are just entering this exciting field.

In our experience, for most statistical modeling campaigns, an experimentalist has already collected much of the data intended for several downstream applications, including mechanistic interrogation, improving reaction performance, and understanding the scope and limitations of a reaction. These datasets are usually difficult to expand considerably for various practical reasons (e.g., cost, resources, and experimental burden such as measuring rates). Ideally, intentional dataset design (*3*) is implemented regardless of the amount of data anticipated to be collected, but this is not always feasible. In other words, most data collected in both academia and industry are generally sparse, whether intentionally designed or not. We describe several practical considerations in handling and analyzing real-world chemistry challenges associated with sparse experimental data using examples mainly from our group. Specifically, we detail the strategies of statistical modeling efforts to glean valuable insights from low data regimes.

Statistical modeling of chemical reactivity, most often defined by selectivity, yield, and/or rates, is dependent on three pillars (Fig. 1A): data (what is being modeled), representation (how the chemical structures involved are described), and algorithm (how the data are processed as a function of the representation). All three pillars are interdependent and must be considered together to develop the best approach for a specific objective. Furthermore, if the goal is interpretability (i.e., mechanism or hypothesis generation), then ML in chemistry will also require grounding in physical measures (e.g., quantum mechanical calculations to understand electron distributions and/or potential energy surfaces that directly influence the reaction output). Below, we describe considerations for each pillar.

Data

Many statistical modeling approaches are relevant regardless of the dataset size under study. However, the scope of this review is confined to modest to medium-sized datasets, which often limit employable methods and present unique statistical challenges (*2*, *6*). Here, we loosely define, from a data chemist's perspective, small to be fewer than 50 experimental data points, medium to be up to 1000 data points, and large to be >1000 data points. Although these ranges are not definitive, they reflect the amount of data achieved from common experimental campaigns: Small datasets typically result from substrate/catalyst scope exploration, while medium datasets usually use high-throughput experimentation (HTE), and large datasets may also use HTE or can be mined from the literature.

Furthermore, the composition of the dataset is perhaps the most important consideration and will be discussed further in the next section. However, it is worth noting that we have largely found developing statistical models using data collected under a single set of reaction conditions (e.g., solvent and temperature) to be effective for elucidating substrate and/or catalyst reactivity trends, a common precept in developing linear free energy relationships. If the reaction conditions change, then statistical models can often still be constructed if the underlying mechanism is conserved. However, when reaction conditions lead to different interactions that affect reactivity, it becomes necessary to parameterize the reaction conditions and use nonlinear fitting techniques (*7*, *8*).

Representation

The molecular representation used to construct models is a key consideration. Molecules can be represented by descriptors that range from simplistic and computationally inexpensive to specifically designed and computationally expensive. No matter the level of

[1]Department of Chemistry, University of Utah, Salt Lake City, UT 84112, USA. [2]Discovery Chemistry, Merck & Co. Inc., Rahway, NJ 07065, USA.
*Corresponding author. Email: matt.sigman@utah.edu

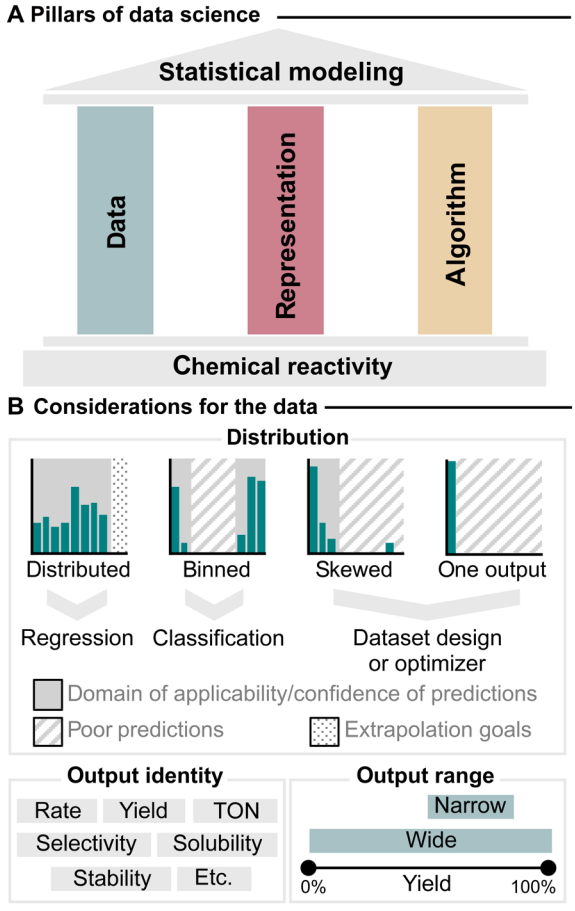

Fig. 1. Statistical modeling in chemistry. (**A**) Pillars of data science. (**B**) Considerations for the data necessary to take into account when choosing a modeling algorithm.

complexity, molecular descriptors serve to quantify molecular features using mathematical relationships. The most common include quantitative structure activity relationships (QSAR) (*9*), fingerprint (*10*–*12*), graphs (*13*, *14*), semiempirical (*15*), density functional theory (*16*–*18*), and designer descriptors (*19*), which have all been successfully used for statistical modeling. Specifically, we often use modern computationally derived molecular descriptors for modeling efforts (*17*) and emphasize that the collection of descriptors specific to the reactive moiety lends itself to more mechanistically grounded models. Collecting descriptors at an appropriate level of theory can be computationally intensive and challenging, but the community has been reporting automated workflows (*20*) as well as descriptor libraries for common substrates (*21*) and ligands (*22*, *23*). The evolution of these descriptors is tied to effectively building predictive and interpretable statistical models (*4*, *6*, *17*, *24*).

Algorithms

Last, largely dependent on its inputs (data and representation), the choice of algorithm is often nuanced and a result of the best performance for a particular objective. Herein, we address various algorithms commonly used by our group and others for analyzing sparse datasets. Often, we find that the most rudimentary statistical modeling approaches provide sufficient chemical interpretability.

DATA FACTORS TO CONSIDER WHEN CHOOSING A MODELING ALGORITHM

Before discussing the details of various commonly used algorithms, it is worth noting data factors that can provide important insight into a choice of algorithm for sparse datasets. The distribution of the data is arguably the most important determinant of how one deploys ML (Fig. 1B). Additionally, the reaction output itself (e.g., yield versus selectivity) and the range of the reaction output used, which are often interconnected, provide insights into the type of algorithm that best suits the objective. These factors will be recalled in discussions of algorithm selection.

Distribution of the reaction output

A histogram of the measured reaction output versus the number of examples should be used to elucidate the distribution of the dataset to be (i) reasonably distributed, (ii) binned (e.g., high versus low), (iii) skewed, or (iv) comprised of essentially one output value. These data structures are likely tied to the diversity of the inputs evaluated [e.g., substrate(s), catalyst, and conditions]. The ideal dataset for statistical modeling is well distributed and is often well suited for regression tasks. This allows the resulting model to have a wider domain of applicability, in which predictions can be made with greater confidence, and extrapolation becomes justifiable. Binned data, commonly bimodal but could have any number of data groupings, lend themselves to classification algorithms. Last, datasets that are heavily skewed or exhibit only a singular output may require the acquisition of a better-distributed dataset before modeling. Data acquisition campaigns strictly for reaction optimization purposes are outside the scope of this review, but Bayesian optimization (*25*, *26*), other active learning techniques (*27*, *28*), or dataset design principles (*3*, *29*, *30*) should be used to obtain a more distributed data structure. Particularly in cases of only poor performance data (e.g., <10% yield or poor selectivity), a search algorithm, such as a Bayesian optimizer, could improve the diversity of reaction outputs.

Identity of the measured reaction output

Many reaction measurements have been successfully modeled in recent years and include yield, rate, selectivity, solubility, stability, and turnover number (*4*, *5*, *17*, *24*). Several reaction outputs, including $\Delta\Delta G^{\ddagger}$ as a measure of selectivity and rate, are akin to linear free energy relationships and can be modeled linearly (*31*). Yield, however, is a variable that is confounded by many aspects including reactivity, purification, and product stability (*3*). Additionally, the assay by which a reaction output is measured can affect modeling efforts. In particular, the time point at which the reaction is assayed (*32*) and whether the crude reaction or the isolated product is assayed can affect the data used to train a model. These factors will be highlighted when relevant within the case studies presented below.

Range of measured reaction output

Furthermore, the range of the reaction output can affect the effectiveness of model performance especially when considering extrapolation (Fig. 1B). It is critical to have examples of both "good" and "bad" results; all accurately collected data should be used in model training. Historically, results that have been construed as negative are underreported but are essential to modeling efforts, aiming to understand the full range of reactivity/selectivity (*33*). Some reaction outputs are bounded (e.g., yield is 0 to 100%), while others are unbounded (e.g., rate and $\Delta\Delta G^{\ddagger}$). The necessary range for modeling

varies by the identity of the reaction output; in any case, even if the range is sufficient, having only a single example at extrema can substantially bias the resultant statistical model and convolute the statistical measures.

Data quality

The quality of data can be affected by assay scale (e.g., HTE, small-scale laboratory experiments, and mole-scale experiments), measurement precision (e.g., detection limit and number of significant digits), and the number of replicates performed. Additional accuracy in the data can serve to better differentiate the data and is particularly advantageous for regression modeling tasks. For example, using a rounded enantiomeric ratio (er) versus an er with an extra significant digit (if the assay is accurate) in the case of 99:1 versus 98.5:1.5 represents ~0.25 kcal/mol difference (at 25°C). Together, prospectively establishing the goals for statistical modeling can serve to guide the data collection and modeling efforts efficiently, as ideal outcomes may look different when striving specifically for mechanistic insight versus reaction optimization or predictive capacity.

COMMON ALGORITHMS

Choosing a suitable algorithm for a given dataset is critical. This review will define and provide examples of several commonly used algorithms, although not an exhaustive list, it will include those that facilitate interpretability while also highlighting the additional considerations discussed above. Given the focus on sparse datasets, the algorithms discussed here are curated to the ones that are less susceptible to overfitting: when a model is too complex and begins to fit the inherent noise in the data, severely limiting the generality of the model (*34*–*36*). There is likely more than one algorithm that would be applicable for most datasets; the study objective, statistical validation of the model, and trial and error all contribute to the model ultimately deemed "best."

STATISTICAL MODEL EVALUATION

To conduct statistical analysis of models, regardless of the algorithm used to generate it, the dataset is first divided into training, test, and, if possible, external validation sets. Often with small datasets, we use only training and test sets, and, taking into account the dataset size and diversity, it is split in any proportion (e.g., 50:50, 80:20, and even no split for very small datasets that rely on leave-one-out cross validations) using one of several common algorithms as a prevention measure for introducing bias: random, based on the distribution of data points (*y*-equidistant), or based on descriptor variance [Kennard-Stone; (*37*)]. The training set is the data used to build the model and serves to determine the dependent molecular descriptors. The test set is the data used to assess the predictive ability of the trained model, as it has not been seen in the model construction process. The most rigorous way to prevent data leakage and protect against overfitting is to select a model based on the training and test set statistics and then conduct external validation, but this can be difficult in small data regimes. Thus, for sparse datasets, we commonly rely on the training/test set split to assess the model's robustness/generalizability and predictive power. Various statistics (e.g., R^2, Q^2, and test R^2) useful for evaluating models trained on small and large datasets alike have been defined in past reviews (*6*, *17*).

CLASSIFICATION

Classification algorithms are the most rudimentary modeling approaches that can often elucidate key physical organic insights. They are particularly well suited for binned data structures and sparse datasets. Generally, reaction outputs such as yield and conversion lend themselves well to classification tasks, in addition to reaction outputs that are inherently binary (e.g., on/off and site selectivity), as demonstrated by the examples presented in Figs. 2 and 3.

Decision trees

Decision trees are nonlinear algorithms that function by classifying each data point based on a user-defined cutoff. The relationship between the reaction output being modeled and the descriptor(s) used does not need to assume a linear relationship. The data are partitioned recursively into nodes (Fig. 2B) based on values of the descriptors provided to the algorithm, with the aim of increasing the purity of each node. Decision trees can be used to accommodate more than two classes; however, binary classification will be highlighted here, as often a chemist's goal is to classify reactions as active/inactive (or high/low selectivity). The result of this analysis is represented using a confusion matrix that dissects the data into four quadrants: namely, true positives (prediction for the desired activity/selectivity matches the ground truth), true negatives (prediction for the undesired activity/selectivity matches the ground truth), false positives (prediction of the desired activity/selectivity does not match the ground truth), and false negatives (prediction of the undesired activity/selectivity does not match the ground truth). Figure 2A depicts a confusion matrix specific to single-node decision trees (vide infra) for bimodal classification. Accuracy is used to quantify the number of predictions made by the classification algorithm that match the ground-truth value. F1 score integrates both precision (accuracy of positive predictions) and recall (sensitivity and ability to find all true positives within a dataset) into a single metric that can be used to better understand model performance. Achieving an accuracy of 1.0 is ideal, but, if a data point is misclassified, then we prefer false positives, to ensure that a potential hit (e.g., reaction corresponding to high reactivity or selectivity) is not missed when the model is used for virtual screening. Although decision trees are not poised for extrapolative predictions (i.e., predictions outside the domain of the training data), they can be a helpful tool when the training data permit a sufficiently wide domain of applicability.

The simplest form of a decision tree algorithm involves only a single node, which we refer to as a threshold or reactivity cliff. It has been leveraged effectively to identify and understand reactivity and selectivity cliffs for several reaction types (*38*–*40*). In an example, a reactivity cliff was identified for four distinct Ni-catalyzed Suzuki-Miyaura cross-coupling reactions, one of which is shown in Fig. 2B (*41*). These reactions were evaluated against 90 phosphine ligands that represented the chemical diversity of the recently reported monodentate phosphine descriptor library (Kraken) (*22*), and the reaction was defined by the user to be active at 10% yield (horizontal boundary or *y*-cut). The *y*-cut value can be defined on the basis of a natural split in the data, a mechanistic hypothesis (e.g., one catalytic turnover), and/or to facilitate the utility of the model (e.g., the reaction can be further optimized). In this example, the most robust (exhibiting nearly no false negatives) threshold (vertical dashed line) was determined using the percent buried volume of the ligand conformer with the lowest buried volume [$\%V_{bur}$(min)] from Kraken. At the algorithm-determined threshold value, ligands are classified

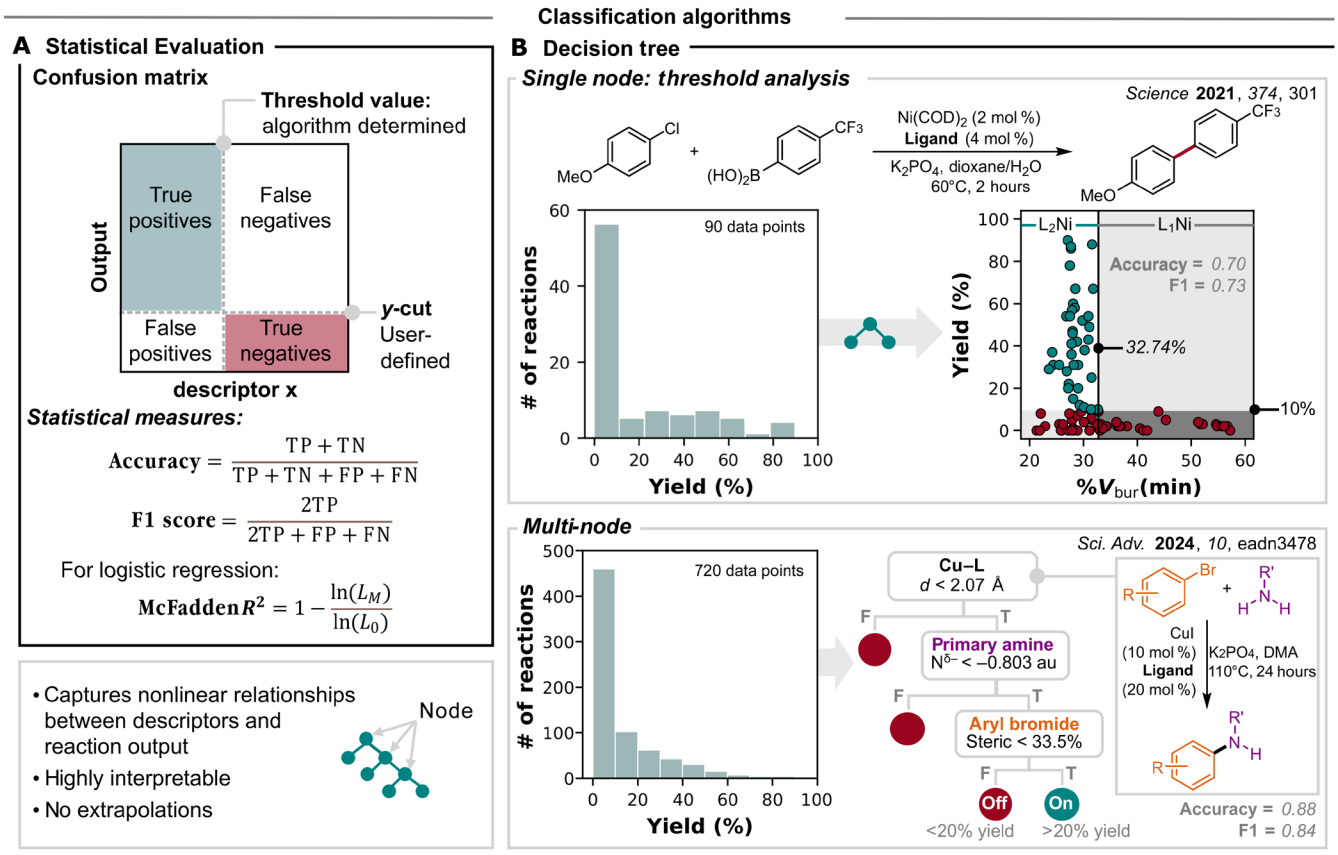

Fig. 2. Classification algorithms. (**A**) Statistical evaluation metrics and (**B**) decision trees [single node: adapted with permission from (*41*). Copyright 2021 The American Association for the Advancement of Science; multi-node: adapted from (*42*). Copyright 2024 The American Association for the Advancement of Science]. TP, true positive; TN, true negative; FP, false positive; FN, false negative.

as active or inactive. The identified descriptor led to mechanistic insights related to the Ni ligation state. Specifically, ligands with a $\%V_{bur}$(min) value of >32% exhibited monoligated (L_1Ni) complexes, while phosphines with values of <32% were bisligated (L_2Ni) complexes, resulting in high-yielding reactions. Notably, other similar descriptors, such as the Boltzmann averaged $\%V_{bur}$, did not achieve the same statistical robustness, reinforcing the importance of suitable molecular descriptors as inputs for statistical modeling.

For larger datasets, more nodes in the decision tree can be warranted (*42*), but small datasets are prone to overfitting and bias toward overrepresented reaction outcomes when using more extensive decision trees (*43*). In an effort to confront the intrinsic unpredictability of Ullman C—N coupling reactions, HTE was used to evaluate 720 reactions, resulting in a 3.5:1 ratio between off and on reactivity as defined by a 20% yield cutoff (Fig. 2B). A decision tree (nodes = 3) was used to predict the effect of substrate (i.e., amine and aryl bromide) and ligand structural features on reaction yield with 88% accuracy. This decision tree allows reactions to be classified while simultaneously accounting for three factors (i.e., ligand, amine, and aryl bromide) that do not have a direct linear relationship to the measured yield, similar to how multivariate linear regression (MLR) accounts for multiple factors when modeling a continuous variable (vide infra).

Logistic regression

In contrast to decision trees, which predict discrete outcomes based on splits in the feature space, logistic regression predicts the probability of a sample being of a certain designation (e.g., active or inactive) using a logistic (i.e., sigmoid) function based on molecular descriptor(s). The decision boundary must be linear in feature space [i.e., between the reaction output and descriptor(s) that defines the decision boundary]. The accuracy of these models is evaluated similar to decision trees by the number of samples assigned to each category in the confusion matrix. We have only recently been using logistic regression (*44*), but it is gaining traction in the field (*45*). Logistic regression was recently used to assess the regioselectivity of a Rh-catalyzed coupling of *N*-alkyl nicotinate salts with aryl boronic esters (Fig. 3A) (*46*). Twenty bisphosphine ligands were selected, evaluated, and determined to lead to either C2- or C6-arylated products, which were then required to be converted to a binary variable (1 or 0, respectively). The sigmoidal function (plotted with a black line) that resulted from this analysis revealed a correlation between the bisphosphine bite angle and site selectivity, elucidating the mechanistic underpinning that small bite angle phosphines favored C2 arylation. It is admittedly difficult to determine whether a simple decision tree or logistic regression should be used for a given dataset. However, logistic regression provides a probability of

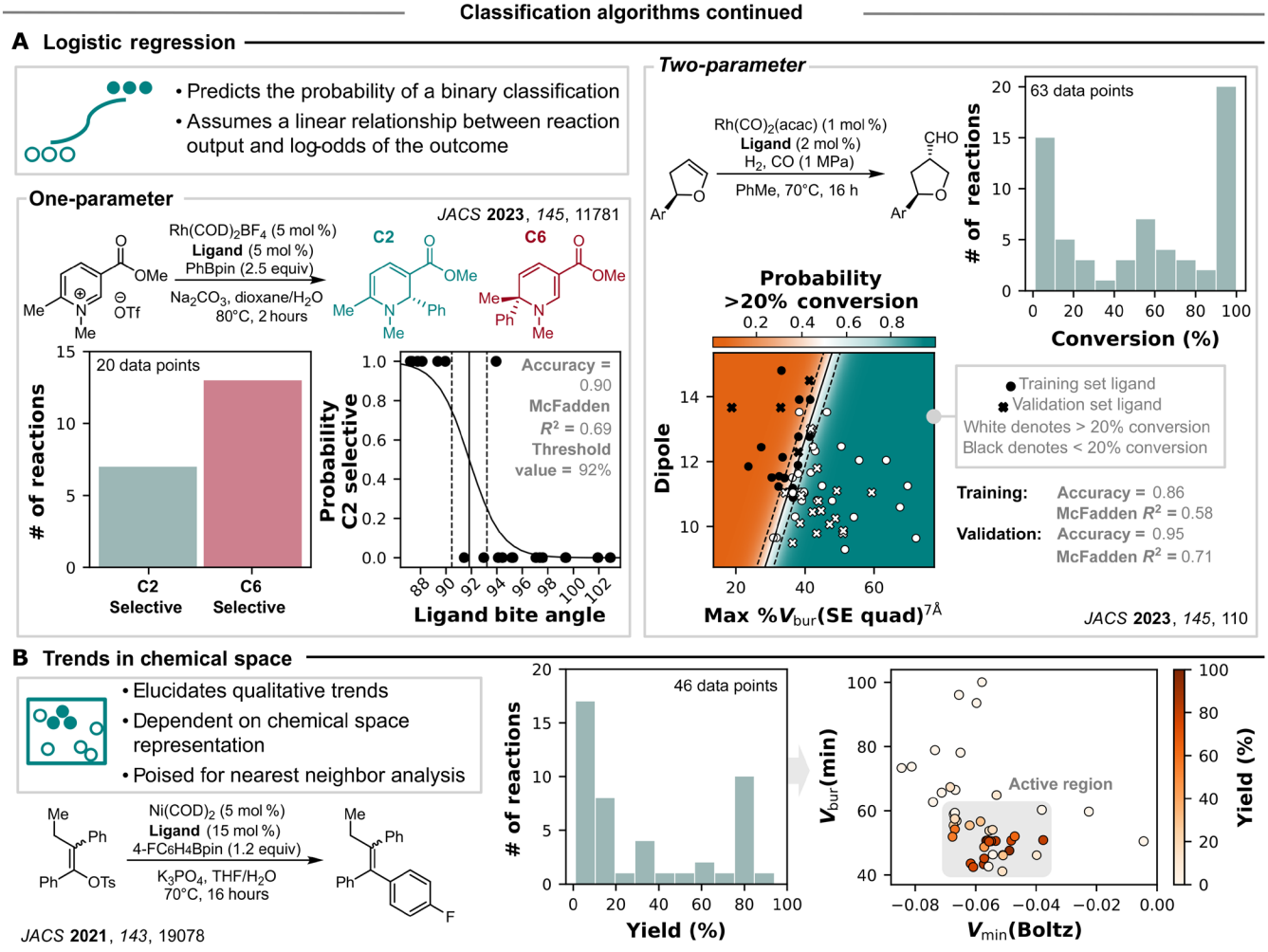

Fig. 3. Classification algorithms continued. (**A**) Logistic regression [one-parameter: adapted with permission from (*46*). Copyright 2023 American Chemical Society; two-parameter: adapted with permission from (*23*). Copyright 2023 American Chemical Society] and (**B**) trends in chemical space [adapted with permission from (*52*). Copyright 2021 American Chemical Society]. h, hours.

success associated with its predictions, which can be very useful to convey a degree of confidence for the success/failure of a suggested reaction. Thus, the evaluation of both model architectures, together with the goals for the project, can facilitate the model selection.

Logistic regression is also not restricted to one parameter, such as the example illustrated in Fig. 3A. In an example, a two-parameter logistic regression was used to identify the bisphosphine ligands that induced high conversion for hydroformylation of a dihydrofuran (*23*). A two-parameter logistic regression model involving ligand dipole and a quadrant buried volume [%V_{bur}(SE)] elucidated the most probabilistically optimal ligand features to catalyze the reaction in high conversion. Ligands falling in the teal region have a higher probability of being active, while those falling in the orange region are estimated to be inactive.

Trends in chemical space

A visual classification of reaction outcomes can be achieved by overlaying the reaction outcome on a representation of chemical space. Qualitative trends can be elucidated using a two-descriptor plot [informed by a decision tree (*47*) or other algorithms] or a map of many descriptors that have been visualized using a dimensionality reduction technique [e.g., principal components analysis (*48*) and uniform manifold approximation and projection (*49*)]. This often lends itself to nearest neighbor analysis to glean further insights or make predictions (*50*, *51*). An example of a small-dataset–guided reaction optimization campaign was reported in the context of a Ni-catalyzed diastereoconvergent cross-coupling of enol tosylates with pinacol boronates (Fig. 3B) (*52*). HTE was used to evaluate 47 representative monophosphine ligands selected from the Kraken library (*22*). Overlaying the observed yield on a ligand steric [V_{bur}(min)] versus an electronic [V_{min}(Boltz)] descriptor plot revealed a region of the chemical space that was active (i.e., >10% yield). Subsequent analysis of the active region allowed for visualization of the (*E*)- and (*Z*)-selective regions. Effective data visualization (*43*) of the reaction output on a chemical space representation can provide valuable mechanistic insight and facilitate rapid reaction optimization.

REGRESSION

To model continuous reaction outputs that are well distributed, it is better to define a regression task to make continuous predictions.

Regression is commonly used to model unbounded reaction outputs like selectivity and reaction rate, as demonstrated by the examples presented in Fig. 4. Linear regression is more sensitive to outliers, especially in small/sparse datasets, further emphasizing the importance of well-designed datasets and quality experimental data. We will also briefly mention nonlinear regression methods and their challenges when considering sparse datasets.

Linear regression

The most basic form of linear regression is directly correlating molecular descriptors to the experimental reaction output, resulting

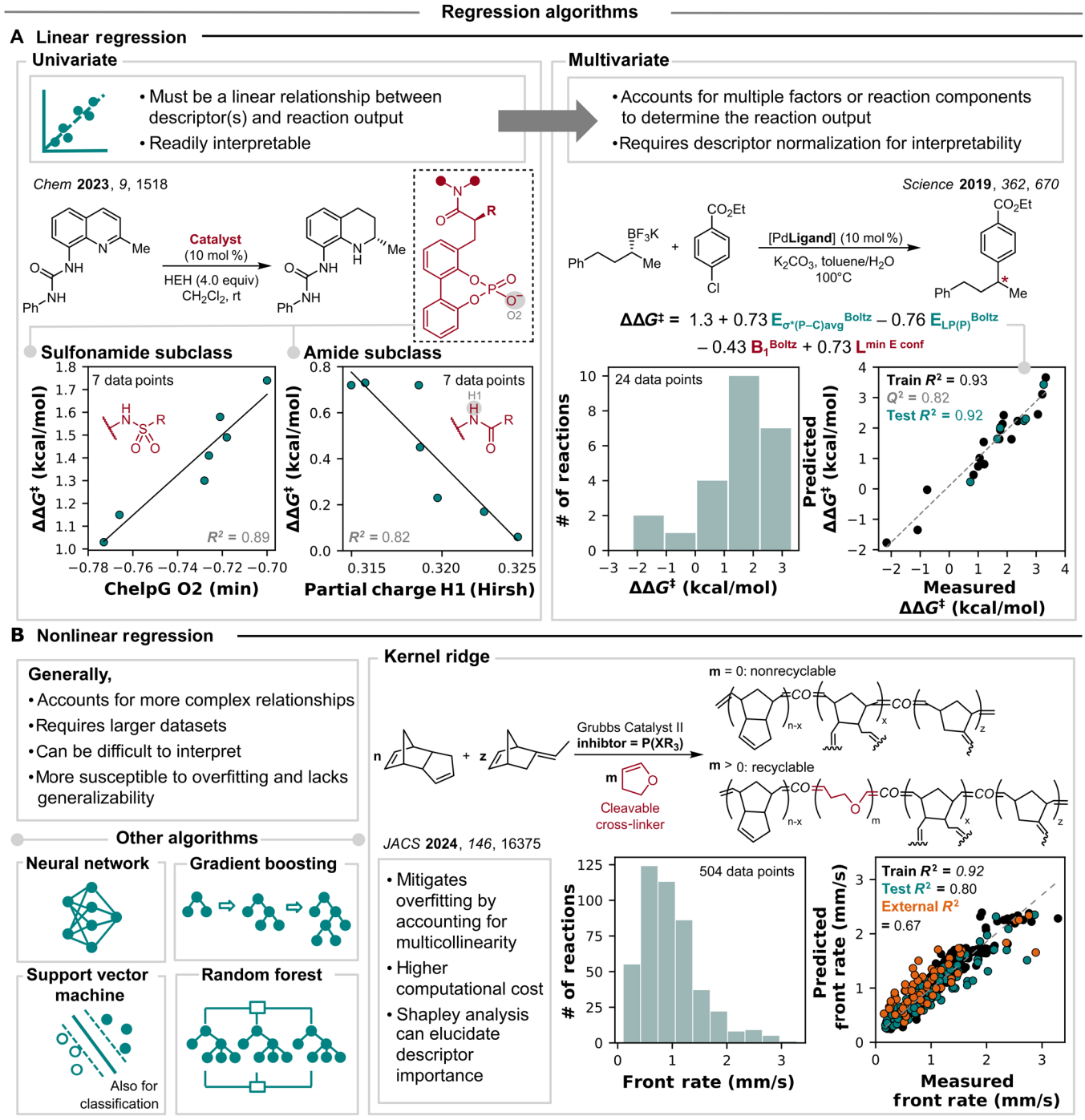

Fig. 4. Regression algorithms. (**A**) Linear regression, both univariate [adapted with permission from (*39*). Copyright 2023 Elsevier Inc.] and multivariate [adapted from (*56*). Copyright 2018 The American Association for the Advancement of Science] and (**B**) nonlinear regression approaches highlighting kernel ridge [adapted with permission from (*64*). Copyright 2024 American Chemical Society]. rt, room temperature.

in a univariate correlation. The most notable and historical univariate correlation in organic chemistry is Hammett analysis (*31*). Similar to Hammett values (*22*), publicly available tabular data like descriptor libraries [e.g., Kraken (*22*), carboxylic acids (*53*), alkyl amines (*53*), and aryl bromides (*21*)] allow for facile univariate analysis without the need for computational resources or model training. Univariate trends are particularly helpful if the dataset can be fragmented into groups that systematically modulate only one reaction component (e.g., substrate or catalyst) (*5*). This allows for interpretation of the impact that a single molecular descriptor has on the reaction output, isolated from the remainder of the reaction variables. For example, in analyzing adaptable chiral phosphoric acid catalysts for the enantioselective transfer hydrogenation of 8-aminoquinoline, unique univariate selectivity correlations were found for the sulfonamide and amide catalyst subclasses (Fig. 4A) (*39*). These distinct correlations for each catalyst subclass suggested the differences in the hydrogen bonding networks of the active catalyst conformer that are ultimately responsible for selectivity.

Multivariate linear regression

When the reaction output being modeled is too complex to be described by only one descriptor, a situation often encountered in the modern chemistry enterprise, MLR can be used. Even in cases with one independent variable (e.g., substrate or catalyst), MLR may be necessary to capture several steric and/or electronic factors at play that cannot be adequately conveyed by one descriptor (*54*). More often, MLR is used when several reaction components [e.g., substrate(s) and catalyst] simultaneously affect the reaction output, requiring multiple molecular descriptors to sufficiently correlate the observation. MLR has been used to successfully build statistical models for many reaction outputs, including reaction rate (*55*), enantioselectivity (*56*, *57*), regioselectivity (*58*–*60*), and solubility (*61*).

When constructing MLR models with a forward stepwise algorithm, molecular descriptors are added sequentially to generate bi-, tri-, and multivariate correlations. Models that minimize the sum of squared errors are kept at each step (*17*). Alternatively, a backward stepwise algorithm ensures that all possible models are built, but it is much more computationally demanding. Models for which the training Q^2 or test R^2 is substantially diminished compared to the training R^2 indicate that the model is not generalizable, will lack predictive power, and is not statistically valid. We caution these are indications of overfitting (*34*–*36*) and note that spurious correlations can result from including an excessive number of descriptors for the algorithm to choose from (*62*). Models can also be evaluated for interpretability by examining the molecular descriptors included in the model equation, which are connected to the underlying phenomena that lead to effective reactions. Given descriptors are first normalized, the relative signs and magnitudes of the descriptor coefficients in the model equation can also provide insight.

In Fig. 4A, MLR was used to correlate the ligand-dependent selectivity of an enantiodivergent Pd-catalyzed Suzuki cross-coupling using an enantioenriched alkylboron nucleophile and aryl electrophile (*56*). The four monophosphine ligand parameters in this model were found to be mechanistically insightful, as the electronic descriptors (teal) discriminate between stereoretentive and stereoinvertive transmetalation pathways, while the steric descriptors (red) indicate the extent of competitive β-hydride elimination (red).

Nonlinear regression

As the size of a dataset increases, nonlinear modeling becomes a viable and sometimes necessary approach to capture all factors contributing to the observed reaction output. Most are familiar with fitting the rate of enzymatic reactions with the Michaelis-Menten model, which is an example of nonlinear fitting of data. Applications of nonlinear regression methods to sparse chemistry datasets are less common in reported studies.

Kernel ridge regression

Kernel ridge regression (*63*) is suited to handle nonlinear relationships between reaction output and molecular descriptors and is uniquely poised to prevent overfitting by accounting for multicollinearity in the loss function (*63*). As a recent example, the rate of frontal ring-opening metathesis polymerization using the Grubbs generation II catalyst and various phosphine inhibitors was successfully modeled using nonlinear kernel ridge regression (Fig. 4B) (*64*). Notably, this is a larger dataset (504 data points), but the diminished performance of the external validation set (orange) indicates the model is on the cusp of overfitting. However, the small number of descriptors (*4*) used in the model to predict front rate in tandem with Shapley analysis (*65*) allowed for chemical interpretation of the model.

There are many other nonlinear algorithms for both classification and regression including support vector machines, *k*-nearest neighbors, random forest (ensemble of decision trees), gradient boosting (builds an ensemble of trees sequentially), and neural networks (*66*). Most nonlinear algorithms tend to be more data hungry, difficult to glean mechanistic insight from, and highly susceptible to overfitting and often exhibit poor generalizability. When applied to sparse datasets, although nonlinear algorithms may account for more complex relationships, these pitfalls are amplified. Examples of nonlinear algorithms can be found in the literature, including Doyle and coworkers' employment of a random forest model for the prediction of C—N cross-coupling reaction yields (*67*). However, this model was trained on 4608 data points, a much larger dataset than defined in the scope of this Review. Alternatively, neural networks were used by Newhouse and coworkers, who reported a feed-forward neural network trained on 17 computationally expensive transition states for the prediction of enantioselectivity of a Negishi cross-coupling reaction with *P*-chiral–hindered phosphines (*68*). We emphasize that a great deal of caution needs to be exercised when using these more data-hungry nonlinear algorithms. This means paying particular attention to the evaluation metrics that probe the extent of the model's generalizability (e.g., Q^2 and test R^2).

SEQUENTIAL APPROACHES

Sequential deployment of classification and regression algorithms has been successfully leveraged for a range of downstream tasks. Classification is often used first to curate the dataset, making it more amenable to regression tasks. Classification can serve to intentionally reduce the number of zeros in a dataset attributed to poor solubility, excessively slow rate, incompatible functional groups, etc. Alternatively, it can provide a mechanistically grounded curation of the data, when a reactivity cliff is identified to group the data into reactions that are hypothesized to follow the same reaction mechanism. For example, if a reaction is only viable with monoligated complexes, then a ligand steric descriptor may remove the ligands

that are more likely to form bisligated complexes, which enforces a different mode of reactivity. Provided that the remaining data are still distributed well across the range of the measured reaction outcome, then regression algorithms can be used to glean more nuanced mechanistic insight into the reaction.

Work by Dotson *et al.* (*23*) demonstrated two examples of applying this strategy with the aim of optimizing two steps of a process chemistry route to an active pharmaceutical ingredient for the treatment of asthma. A reactivity cliff for the catalyst was first identified for a Hayashi-Heck reaction, and, then, MLR was used to build a predictive model for selectivity. As a second example, logistic regression was used to classify ligands that achieved >20% conversion in a site selective hydroformylation reaction before building a MLR model for selectivity. In cases like this, preliminary classification modeling blurs the line between dataset design and statistical modeling, emphasizing the malleability of a general data-driven workflow.

UNMODELABLE DATA

When many algorithms have been deployed without success, a dataset can be deemed unmodelable with the available tools. It can often be attributed to poor dataset distribution, meaning the dataset was undistributed (i.e., all one reaction output value) or the dataset spanned a very narrow range of the reaction output. For example, a data science workflow was recently used to optimize a chiral bisphosphine ligand and determine "diverse" substrates for scope evaluation in a Pd-catalyzed aryl-carbonylation of sulfonimidamides. Statistical modeling efforts were performed to elucidate the structure-activity relationship as a function of the aryl iodide substrate. This was not successful likely due to data skewed toward generally high performing examples, even with proper substrate selection using the principles of dataset design (*69*). We often also consider another pillar of data science when we reach an impasse in modeling that insufficient molecular representation or molecular features are being deployed, which can inspire next-generation descriptor development (*4*). Furthermore, the other reaction variables (i.e., solvent, temperature, and additives), which are present in many datasets collected without the intent of modeling, can affect model performance.

At the outset of projects, initial data collection campaigns, especially when not intentionally designed, often exhibit attributes of unmodelable datasets. In this case, additional data can be collected to create a more well-distributed dataset that makes statistical modeling achievable. For the Ullmann coupling reaction presented above, a form of active learning was used to build a more comprehensive dataset that was better suited for predictive purposes (*42*). The data from the original HTE screen were highly skewed toward inactive ligands. Therefore, classification models were used to determine an important parameter for reaction success. The model was subsequently applied for the prediction of other active ligands to supplement further data collection, improving the overall distribution of data.

Although more data are a common suggestion to improve models, using more informed molecular descriptors can also serve to reduce the number of model parameters necessary and increase the model interpretability. In more extreme cases, advancements in descriptors, modeling approaches, and/or collection of more data may warrant re-evaluation of statistical models years later. For example, the enantioselectivity of an oxidative amination of tetrahydroisoquinolines under chiral-anion phase-transfer catalysis originally modeled in 2015 (*70*) was remodeled in 2017 using transition state surrogates (*71*) that better quantified important noncovalent interactions, resulting in a more simplistic and readily interpretable model.

CONSIDERATIONS FOR INTERPRETABILITY

If model interpretability is an objective, then there are additional considerations besides statistics that are necessary for model selection. Statistics of models trained on small datasets, although important to determine generalizability, may be substituted for greater chemical interpretability. In other words, a model with modestly reduced statistical performance may provide the researcher opportunities to better formulate a mechanistic hypothesis. In this case, the molecular descriptors from which the model is built must be chemically interpretable themselves. As a result, when constructing a model, we tend to favor simplicity of the parameters used in the model (e.g., exclude cross terms). An interesting and related aspect is how to navigate changes in mechanism in statistical modeling. This has been previously discussed in another review (*4*). However, the project objective may assist in determining when to model data together to develop a general model versus separating a dataset to build mechanism-specific models.

Another important factor to consider when modeling sparse datasets, no matter the algorithm used, is how many model parameters are warranted to avoid overfitting (*34*, *36*). This is subjectively determined by the number of training data points. Based on our experience, ~8 to 10 data points per parameter in a model (e.g., MLR and decision tree) is reasonable in the sparse data regime. However, this can be increased or decreased depending on the complexity of the reaction under study (e.g., a second order reaction should need at least two parameters to describe, one for each reaction component), the interpretability of the model, and the stage of the project (e.g., we are much more willing to accept overfit preliminary models and use the descriptor in active learning pursuits). We have found that, for decision trees, it is more important to exercise prudence in cases where only a small percentage of the data points belong to one class, regardless of the size of the dataset. Additionally, pruning can be used for tree-based methods to lessen the number of nodes and improve model generalizability. Some algorithms (e.g., random forest) use more descriptors, but the number of descriptors should not exceed the number of data points, to avoid overfitting. To achieve the most robust statistical model with the fewest number of descriptors, we reemphasize the importance of quality, interpretable molecular descriptors (vide supra).

Last, valuable mechanistic insights can be gleaned by connecting statistical models to transition state structures (*4*). The computational cost of computing transition states makes statistical models especially helpful for translating interactions of one representative transition state to that of different substrate-catalyst combinations, for which it may be prohibitive to compute transition states (*57*, *72*). Recently, we reported a mechanistic picture for an enantioselective cinchona alkaloid catalyzed sulfonimidamide acylation by deconstructing a MLR model in conjunction with transition state structures (*57*). By plotting the continuous descriptors (i.e., non-classifiers) in the MLR model, the catalyst-substrate steric matching necessary for enantioinduction was easily deciphered and readily mapped onto the transition state structures.

CONCLUSION

In summary, as detailed herein, many considerations are necessary to model sparse datasets, and neither the process of statistical modeling

nor the selection of the best model is trivial. Often, our decisions are determined with the practicing chemist "in the loop" by applying their chemical intuition; depending on the project objective, we may opt to sacrifice statistical performance or predictive capacity of a model to ease interpretability. Although statistical modeling for chemical reactivity has largely been used by academic labs, it can be advantageous to industrial chemists, as demonstrated by many successful industrial-academic collaborations (*23*, *55*, *57*, *67*, *69*). Anecdotally, we have observed that the generality of classification lends itself to discovery chemistry, whereas regression can be more precise and used by process chemists. Ultimately, achieving the best possible statistical model for a sparse dataset requires a balance of pushing the bounds of traditional statistical modeling (i.e., ML) approaches while staying grounded in physical organic chemistry.

REFERENCES AND NOTES

1. F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, F. Glorius, Machine learning the ropes: Principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **49**, 6154–6168 (2020).
2. B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang, G.-W. Wei, Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **123**, 8736–8780 (2023).
3. P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman, C. W. Coley, Dataset design for building models of chemical reactivity. *ACS Cent. Sci.* **9**, 2196–2204 (2023).
4. J. M. Crawford, C. Kingston, F. D. Toste, M. S. Sigman, Data science meets physical organic chemistry. *Acc. Chem. Res.* **54**, 3136–3148 (2021).
5. W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, The evolution of data-driven modeling in organic chemistry. *ACS Cent. Sci.* **7**, 1622–1637 (2021).
6. H. Shalit Peleg, A. Milo, Small data can play a big role in chemical discovery. *Angew. Chem. Int. Ed.* **62**, e202219070 (2023).
7. J. P. Reid, M. S. Sigman, Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
8. J. Werth, M. S. Sigman, Connecting and analyzing enantioselective bifunctional hydrogen bond donor catalysis using data science tools. *J. Am. Chem. Soc.* **142**, 16382–16391 (2020).
9. A. U. K. Danishuddin, Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. *Drug Discov. Today* **21**, 1291–1302 (2016).
10. H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator. *J. Cheminf.* **10**, 4 (2018).
11. RDKit: Open-Source Cheminformatics Software. www.rdkit.org/.
12. A. Capecchi, D. Probst, J.-L. Reymond, One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J. Cheminf.* **12**, 43 (2020).
13. C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
14. X. Pan, H. Wang, C. Li, J. Z. H. Zhang, C. Ji, MolGpka: A web server for small molecule pK_a prediction using a graph-convolutional neural network. *J. Chem. Inf. Model.* **61**, 3159–3165 (2021).
15. C. Bannwarth, S. Ehlert, S. Grimme, GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
16. A. V. Brethomé, S. P. Fletcher, R. S. Paton, Conformational effects on physical-organic descriptors: The case of sterimol steric parameters. *ACS Catal.* **9**, 2313–2323 (2019).
17. C. B. Santiago, J.-Y. Guo, M. S. Sigman, Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
18. K. Jorner, L. Turcani, kjelljorner/morfeus, v0.7.2., Zenodo (2022).
19. R. C. Cammarota, W. Liu, J. Bacsa, H. M. L. Davies, M. S. Sigman, Mechanistically guided workflow for relating complex reactive site topologies to catalyst performance in C–H functionalization reactions. *J. Am. Chem. Soc.* **144**, 1881–1898 (2022).
20. J. V. Alegre-Requena, S. Sowndarya S. V., R. Pérez-Soto, T. M. Alturaifi, R. S. Paton, AQME: Automated quantum mechanical environments for researchers and educators. *WIREs Comput. Mol. Sci.* **13**, e1663 (2023).
21. S. K. Kariofillis, S. Jiang, A. M. Żurański, S. S. Gandhi, J. I. Martinez Alvarado, A. G. Doyle, Using data science to guide aryl bromide substrate scope analysis in a Ni/photoredox-catalyzed cross-coupling with acetals as alcohol-derived radical sources. *J. Am. Chem. Soc.* **144**, 1045–1055 (2022).
22. T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, A. Aspuru-Guzik, A comprehensive discovery platform for organophosphorus ligands for catalysis. *J. Am. Chem. Soc.* **144**, 1205–1217 (2022).
23. J. J. Dotson, L. van Dijk, J. C. Timmerman, S. Grosslight, R. C. Walroth, F. Gosselin, K. Püntener, K. A. Mack, M. S. Sigman, Data-driven multi-objective optimization tactics for catalytic asymmetric reactions using bisphosphine ligands. *J. Am. Chem. Soc.* **145**, 110–121 (2023).
24. M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Acc. Chem. Res.* **49**, 1292–1301 (2016).
25. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
26. M. Christensen, L. P. E. Yunker, F. Adedeji, F. Häse, L. M. Roch, T. Gensch, G. dos Passos Gomes, T. Zepel, M. S. Sigman, A. Aspuru-Guzik, J. E. Hein, Data-science driven autonomous process optimization. *Commun. Chem.* **4**, 112 (2021).
27. N. S. Eyke, W. H. Green, K. F. Jensen, Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **5**, 1963–1972 (2020).
28. B. Settles, Active Learning Literature Survey, https://burrsettles.com/pub/settles.activelearning.pdf (accessed 19 August 2024).
29. E. Shim, A. Tewari, T. Cernak, P. M. Zimmerman, Machine learning strategies for reaction development: Toward the low-data limit. *J. Chem. Inf. Model.* **63**, 3659–3668 (2023).
30. R. A. Fisher, *The Design of Experiments* (Oliver & Boyd, 1935).
31. L. P. Hammett, The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **59**, 96–103 (1937).
32. A. D. Matthews, E. Peters, J. S. Debenham, Q. Gao, M. D. Nyamiaka, J. Pan, L.-K. Zhang, S. D. Dreher, S. W. Krska, M. S. Sigman, M. R. Uehling, Cu oxamate-promoted cross-coupling of α-branched amines and complex aryl halides: Investigating ligand function through data science. *ACS Catal.* **13**, 16195–16206 (2023).
33. M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P. O. Norrby, O. Wiest, Negative data in data sets for machine learning training. *J. Org. Chem.* **88**, 5239–5241 (2023).
34. J. Lever, M. Krzywinski, N. Altman, Model selection and overfitting. *Nat. Methods* **13**, 703–704 (2016).
35. C. Aliferis, G. Simon, in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, G. J. Simon, C. Aliferis, Eds. (Springer International Publishing, 2024), pp. 477–524.
36. D. M. Hawkins, The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
37. R. W. Kennard, L. A. Stone, Computer aided design of experiments. *Technometrics* **11**, 137–148 (1969).
38. K. E. Gardner, L. de Lescure, M. A. Hardy, J. Tan, M. S. Sigman, R. S. Paton, R. Sarpong, Modular synthesis of aryl amines from 3-alkynyl-2-pyrones. *Chem. Sci.* **15**, 15632–15638 (2024).
39. J. P. Liles, C. Rouget-Virbel, J. L. H. Wahlman, R. Rahimoff, J. M. Crawford, A. Medlin, V. S. O'Connor, J. Li, V. A. Roytman, F. D. Toste, Data science enables the development of a new class of chiral phosphoric acid catalysts. *Chem* **9**, 1518–1537 (2023).
40. A. R. Pancoast, S. L. McCormack, S. Galinat, R. Walser-Kuntz, B. M. Jett, M. S. Sanford, M. S. Sigman, Data science enabled discovery of a highly soluble 2,2′-bipyrimidine anolyte for application in a flow battery. *Chem. Sci.* **14**, 13734–13742 (2023).
41. S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman, A. G. Doyle, Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301–308 (2021).
42. M. H. Samha, L. J. Karas, D. B. Vogt, E. C. Odogwu, J. Elward, J. M. Crawford, J. E. Steves, M. S. Sigman, Predicting success in Cu-catalyzed C–N coupling reactions using data science. *Sci. Adv.* **10**, eadn3478 (2024).
43. D. M. Lustosa, A. Milo, Mechanistic inference from statistical models at different data-size regimes. *ACS Catal.* **12**, 7886–7906 (2022).
44. Y. T. Boni, R. C. Cammarota, K. Liao, M. S. Sigman, H. M. L. Davies, Leveraging regio- and stereoselective C(sp^3)–H functionalization of silyl ethers to train a logistic regression classification model for predicting site-selectivity bias. *J. Am. Chem. Soc.* **144**, 15549–15561 (2022).
45. C.-C. Chen, K. Mondal, P. Vervliet, A. Covaci, E. P. O'Brien, K. J. Rockne, J. L. Drummond, L. Hanley, Logistic regression analysis of LC-MS/MS data of monomers eluted from aged dental composites: A supervised machine-learning approach. *Anal. Chem.* **95**, 5205–5213 (2023).
46. K. G. Ortiz, J. J. Dotson, D. J. Robinson, M. S. Sigman, R. R. Karimov, Catalyst-controlled enantioselective and regiodivergent addition of aryl boron nucleophiles to N-alkyl nicotinate salts. *J. Am. Chem. Soc.* **145**, 11781–11788 (2023).
47. A. LeSueur, N. Tao, A. Doyle, M. Sigman, Multi-threshold analysis for chemical space mapping of Ni-catalyzed Suzuki-Miyaura couplings. *Eur. J. Org. Chem.* **27**, e202400428 (2024).

48. I. T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments. *Phil. Trans. R. Soc. A* **374**, 20150202 (2016).
49. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 (2020).
50. J.-L. Reymond, The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
51. T. Gensch, S. R. Smith, T. J. Colacot, Y. N. Timsina, G. Xu, B. W. Glasspoole, M. S. Sigman, Design and application of a screening set for monophosphine ligands in cross-coupling. *ACS Catal.* **12**, 7773–7780 (2022).
52. D. Zell, C. Kingston, J. Jermaks, S. R. Smith, N. Seeger, J. Wassmer, L. E. Sirois, C. Han, H. Zhang, M. S. Sigman, F. Gosselin, Stereoconvergent and -divergent synthesis of tetrasubstituted alkenes by nickel-catalyzed cross-couplings. *J. Am. Chem. Soc.* **143**, 19078–19090 (2021).
53. B. Haas, M. Hardy, S. Sowndarya S. V., K. Adams, C. Coley, R. Paton, M. Sigman, Rapid prediction of conformationally-dependent DFT-level descriptors using graph neural networks for carboxylic acids and alkyl amines. ChemRxiv 26434 [Preprint] (2024). https://doi.org/10.26434/chemrxiv-2024-m5bpn.
54. H. Huang, H. Zong, G. Bian, L. Song, Constructing a quantitative correlation between N-substituent sizes of chiral ligands and enantioselectivities in asymmetric addition reactions of diethylzinc with benzaldehyde. *J. Org. Chem.* **77**, 10427–10434 (2012).
55. B. C. Haas, A. E. Goetz, A. Bahamonde, J. C. McWilliams, M. S. Sigman, Predicting relative efficiency of amide bond formation using multivariate linear regression. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2118451119 (2022).
56. S. Zhao, T. Gensch, B. Murray, Z. L. Niemeyer, M. S. Sigman, M. R. Biscoe, Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **362**, 670–674 (2018).
57. B. C. Haas, N.-K. Lim, J. Jermaks, E. Gaster, M. C. Guo, T. C. Malig, J. Werth, H. Zhang, F. D. Toste, F. Gosselin, S. J. Miller, M. S. Sigman, Enantioselective sulfonimidamide acylation via a cinchona alkaloid-catalyzed desymmetrization: Scope, data science, and mechanistic investigation. *J. Am. Chem. Soc.* **146**, 8536–8546 (2024).
58. J. Li, S. Grosslight, S. J. Miller, M. S. Sigman, F. D. Toste, Site-selective acylation of natural products with BINOL-derived phosphoric acids. *ACS Catal.* **9**, 9794–9799 (2019).
59. J. D. Griffin, D. B. Vogt, J. Du Bois, M. S. Sigman, Mechanistic guidance leads to enhanced site-selectivity in C–H oxidation reactions catalyzed by ruthenium bis(bipyridine) complexes. *ACS Catal.* **11**, 10479–10486 (2021).
60. E. N. Bess, R. J. DeLuca, D. J. Tindall, M. S. Oderinde, J. L. Roizen, J. Du Bois, M. S. Sigman, Analyzing site selectivity in Rh2(esp)2-catalyzed intermolecular C–H amination reactions. *J. Am. Chem. Soc.* **136**, 5783–5789 (2014).
61. J. D. Griffin, A. R. Pancoast, M. S. Sigman, Interrogation of 2,2′-bipyrimidines as low-potential two-electron electrolytes. *J. Am. Chem. Soc.* **143**, 992–1004 (2021).
62. M. Bylesjö, O. Cloarec, M. Rantalainen, in *Comprehensive Chemometrics*, S. D. Brown, R. Tauler, B. Walczak, Eds. (Elsevier, 2009), pp. 109–127.
63. M. Rupp, Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **115**, 1058–1073 (2015).
64. T. P. McFadden, R. B. Cope, R. Muhlestein, D. J. Layton, J. J. Lessard, J. S. Moore, M. S. Sigman, Using data science tools to reveal and understand subtle relationships of inhibitor structure in frontal ring-opening metathesis polymerization. *J. Am. Chem. Soc.* **146**, 16375–16380 (2024).
65. S. M. Lundberg, S.-I. Lee, in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates Inc., 2017), pp. 4765–4774.
66. I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160 (2021).
67. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
68. A. E. Cuomo, S. Ibarraran, S. Sreekumar, H. Li, J. Eun, J. P. Menzel, P. Zhang, F. Buono, J. J. Song, R. H. Crabtree, V. S. Batista, T. R. Newhouse, Feed-forward neural network for predicting enantioselectivity of the asymmetric Negishi reaction. *ACS Cent. Sci.* **9**, 1768–1774 (2023).
69. L. van Dijk, B. C. Haas, N.-K. Lim, K. Clagg, J. J. Dotson, S. M. Treacy, K. A. Piechowicz, V. A. Roytman, H. Zhang, F. D. Toste, S. J. Miller, F. Gosselin, M. S. Sigman, Data science-enabled palladium-catalyzed enantioselective aryl-carbonylation of sulfonimidamides. *J. Am. Chem. Soc.* **145**, 20959–20967 (2023).
70. A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737–743 (2015).
71. M. Orlandi, F. D. Toste, M. S. Sigman, Multidimensional correlations in asymmetric catalysis through parameterization of uncatalyzed transition states. *Angew. Chem. Int. Ed.* **56**, 14080–14084 (2017).
72. S. K. Nistanaki, C. G. Williams, B. Wigman, J. J. Wong, B. C. Haas, S. Popov, J. Werth, M. S. Sigman, K. N. Houk, H. M. Nelson, Catalytic asymmetric C–H insertion reactions of vinyl carbocations. *Science* **378**, 1085–1091 (2022).

Acknowledgments: We thank C. Coley for insightful edits. **Funding:** This work was supported by the National Science Foundation (CHE-2154502), National Science Foundation Center for Computer-Assisted Synthesis (C-CAS) (CHE-2202693), and National Institutes of Health (1R35GM136271-01) grants given to M.S.S. **Author contributions:** Conceptualization: B.C.H. and M.S.S. Visualization: B.C.H. Writing—original draft: B.C.H. and D.K. Writing—review and editing: B.C.H., D.K., and M.S.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the cited materials.

Submitted 20 September 2024
Accepted 20 November 2024
Published 1 January 2025
10.1126/sciadv.adt3013