Published in final edited form as:

J Chem Inf Model. 2019 September 23; 59(9): 3645–3654. doi:10.1021/acs.jcim.9b00313.

Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst

Eric Walker[†], Joshua Kammeraad[†], Jonathan Goetz[‡], Michael T. Robo[†], Ambuj Tewari[‡], Paul M. Zimmerman^{*,†}

[†]Department of Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, Michigan 48109, United States

[‡]Department of Statistics, University of Michigan, 1085 South University Avenue, Ann Arbor, Michigan 48109, United States

Abstract

Reaction databases provide a great deal of useful information to assist planning of experiments but do not provide any interpretation or chemical concepts to accompany this information. In this work, reactions are labeled with experimental conditions, and network analysis shows that consistencies within clusters of data points can be leveraged to organize this information. In particular, this analysis shows how particular experimental conditions (specifically solvent) are effective in enabling specific organic reactions (Friedel–Crafts, Aldol addition, Claisen condensation, Diels–Alder, and Wittig), including variations within each reaction class. Network analysis shows data points for reactions tend to break into clusters that depend on the catalyst and chemical structure. This type of clustering, which mimics how a chemist reasons, is derived directly from the network. Therefore, the findings of this work could augment synthesis planning by providing predictions in a fashion that mimics human chemists. To numerically evaluate solvent prediction ability, three methods are compared: network analysis (through the *k*-nearest neighbor algorithm), a support vector machine, and a deep neural network. The most accurate method in 4 of the 5 test cases is the network analysis, with deep neural networks also showing good prediction scores. The network analysis tool was evaluated by an expert panel of chemists, who generally agreed that the algorithm produced accurate solvent choices while simultaneously being transparent in the underlying reasons for its predictions.

Graphical Abstract

*Corresponding Author: Phone: (734) 615-0191. paulzim@umich.edu.

Author Contributions

E.W. performed the calculations, created the graphs, and wrote the paper. P.Z. supervised the work and wrote the paper, and J.K. assisted with coding the algorithms. A.T. and J.G. provided guidance on designing networks, the deep neural net, and the support vector machine. M.R. organized the expert chemist panel.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00313.

Tanimoto similarities statistics, preprocessing flowchart, exception handling, catalyst dictionary, artificial neural network and *k*-nearest neighbor, accuracies statistics, Claisen condensation and Diels–Alder, human chemist focus group trials, and cluster example with chemical structure labels (PDF)

The authors declare no competing financial interest.

1. INTRODUCTION

Reaction data sets contain a wealth of information that can be used to make informed decisions in the laboratory and in preparing for large-scale production. This data shows millions of individual syntheses that transform available substrates into interesting products, all resulting from sustained efforts of the chemical community over decades. Chemists often search these data sets for insight and examples when performing novel reactions, which of course are never present in the data set. This process therefore relies on chemical know-how and inference to make a reaction plan that is relevant to the current synthesis target. In our age of modern computation with its incredible advances in data science, it is natural to ask whether these inferences and plans might be greatly improved compared to current man-machine interchanges (Figure 1).

Recent progress in computational techniques to translate reaction data sets into predictive models has generated considerable enthusiasm for computer-aided synthesis planning. In the last ten years, notable studies on reaction prediction and synthesis planning algorithms have evolved, including expert systems developed from curated data sets¹⁻³ as well as supervised machine learning tools⁴⁻¹⁰ or graph-based tools¹¹ applied to commercial reaction databases¹² or patent-harvested reactivity data.¹³ The intent to automate chemical decisions has been pursued long before contemporary interest in machine learning and artificial intelligence.¹⁴⁻¹⁸

Recently, the area has received great interest, including the modern expert system called Chematica, which created synthesis plans for 8 biologically active molecules that were successfully demonstrated in the laboratory.⁷ Chematica achieved this by manual encoding tens of thousands of reaction rules, representing many years of input from expert chemists.

On the machine learning front, reports by Segler and Waller^{19,20} have shown that a graph-driven neural network strategy can provide (without the extended human effort) synthesis plans that are equivalent in quality to literature reports, as judged by graduate-level organic chemists.

Even with these successes, serious limitations to computer-aided synthesis remain. To chemists, reactions are primarily known by their overall classification, not just specific instances of $A + B \rightarrow C$. When predicting the outcome or conditions for a reaction, chemists make decisions using generalized knowledge, called chemical intuition. Intuition is the skill gained from instruction and experience in using chemical principles—grounded in physical properties—to navigate experimental design and analysis of laboratory outcomes. The intuition of expert chemists is a powerful science that is fully applicable to reactions outside the scope of any database, expert system, or machine learning method. In other words, chemists are highly effective at understanding the details of chemical reactions using a set of broad physical principles, which are applicable to reactions that have never before been performed. With this consideration in mind, it becomes clear that machine learning reaction prediction computations follow an entirely different track and specialize in reaction types that most resemble their training data. By Zipf's law,²¹ databases are primarily populated by the most frequently used reactions, with a power law decrease in number of data points with rank.^{22,23} Therefore, data-driven algorithms—to no surprise—work best for the most popular reactions, and it is nontrivial to generalize these to untested, emerging, or even relatively low-population reaction classes.

Based on these considerations, the basic issue with computer generalization of reaction concepts might be traced to the (lack of) interpretability of the underlying algorithms. Here, we focus on machine learning techniques, where generalization is attempted by machine, rather than by experts. Machine learning techniques are notoriously “black box” in character and provide no direct relationships between the predictions that are made and the underlying reasons for the predictions. The most advanced machine learning strategies, for example the hugely popular area of “deep learning” through neural networks, fall into this category. Without interpretability, any machine learning exercise will face severe difficulty of justifying its value to chemistry, as the generalizability of the model will be suspect.

In addition to these challenges, computational researchers face another key difficulty, due to the quality of available reactivity data. The largest data sets inevitably contain information from a wide variety of sources, and unlabeled or mislabeled data (i.e., “label noise”) are commonplace.²⁴ Doubts regarding the reliability of the training data compound with the lack of interpretability of black box machine learning tools result in inevitable mistrust by seasoned experts. Even for otherwise accurate entries in a data set, deciphering the difference between “reagent” or “catalyst” can lead to confusion for data-driven learning. Especially in the area of chemical synthesis, where reactants, reagents, catalysts, solvents, and other reaction conditions must be specified precisely for the reaction to work, consistent information regarding these factors is paramount. Along these lines, neural network models for predicting suitable reaction conditions have been reported by Gao and co-workers.²⁵ In their study, a large data set of millions of reactions was used to train the models, with accuracies of about 70% in the top-ten computational predictions. As described herein, we

take a different approach where reaction data is partitioned into named organic reactions, allowing focus into predictability of specific types of reactions. This strategy is closer in spirit to what is practiced by laboratory chemists and importantly will allow improved verification of the machine predictions by making transparent models for the reaction conditions.

In this article, we explore whether machine learning techniques can be made interpretable while maintaining high accuracy in predicting a key reaction condition. The reaction condition we focus upon is the solvent, which is a deceptively simple condition because one solvent choice may allow a planned reaction to succeed, but another solvent choice may lead to no reaction at all. Indeed, reports have shown that choice of solvent can change reaction rates by orders of magnitude.²⁶ Solvent compatibility therefore represents a key question that is not only challenging but important for progressing through synthetic space in the laboratory. Ultimately, this article will show that solvents can be selected at high accuracy with a fully interpretable, statistically sound machine learning method across a testbed of five named organic reactions, totaling over 50,000 specific examples.

2. METHODOLOGY

2.1. Data Source.

All data used in this study was obtained from the Reaxys database,¹² which contains approximately 45 million reactions. To focus our study, five named organic reactions were chosen: Diels–Alder, Friedel–Crafts, Wittig, Aldol addition, and Claisen condensation. These represent a diversity of reaction conditions, including catalyst and solvent choices. The data set was limited to single-step reactions, reactants that are commercially available, interpretable solvent designation, and contained a subset of the 78 most common solvents. Data points with missing catalyst entries were treated as uncatalyzed and remain in the data sets. After collecting this data, the Diels–Alder, Friedel–Crafts, Wittig, Aldol addition, and Claisen condensation reaction data sets contained 18,394, 29,021, 9,685, 6,603, and 12,151 total useable data points, respectively.

The raw data from Reaxys required moderate amounts of preprocessing to be useful. For example, many catalysts had “aluminum”, “aluminium”, or “Al” specified: these are all equivalent. Naming was standardized using the chemical identifier resolver from the National Institutes of Health,²⁷ which transforms each name into the IUPAC convention. To capture the remaining ambiguities, catalyst names were made lowercase, metal names were replaced by their atomic symbol, and a hand-crafted dictionary was created to eliminate additional specific cases that were not otherwise handled correctly. The full definition of this procedure is available in the Supporting Information.

Molecular structures were stored as SMILES^{28,29} strings, which were provided by Reaxys. Molecular fingerprints were generated through Open Babel³⁰ using the MACCS keys,^{31–33} which contain 166 functional groups. Rather than use the reactant structures, these fingerprints were generated for the products to simplify the data structure from two reactants to just one product. Tanimoto measures^{34,35}

$$T_{A,B} = \frac{\left[\sum_{j=1}^n x_{jA} x_{jB} \right]}{\left[\sum_{j=1}^n (x_{jA})^2 + \sum_{j=1}^n (x_{jB})^2 - \sum_{j=1}^n x_{jA} x_{jB} \right]} \quad (1)$$

between all product pairs within each reaction data set provide a measure of chemical similarity between data points. Catalyst similarity can also be measured by Tanimoto or by one-hot encoding of the catalyst identity.

2.2. Prediction Algorithms.

To provide solvent classification, a handful of machine learning algorithms were employed. The first two, support vector machines^{36,37} (SVM) and deep neural networks (NN),³⁸ are nonlinear classification techniques that can predict on any number of solvent identities represented by the training data set. SVM classification requires a kernel function to measure similarity between data points, and choices for the kernel are discussed below. NNs do not require this measure, as the number of input nodes can be scaled to the number of input variables.

The *k*-nearest neighbor algorithm (kNN) is the third solvent classification technique.³⁹ The network match technique requires a similarity metric (a kernel), and makes predictions by finding *k* points in the training set that are most similar to the test point. The most frequent solvent in the *k* neighbors in the similarity network is the top solvent prediction. The Supporting Information shows a small, labeled network to demonstrate how kNN clusters similar molecules in practice.

2.3. Similarity Measures.

The network match and SVM algorithms are sensitive to the choice of similarity measure.⁴⁰ In typical machine learning analysis, one takes fixed-length feature vectors and subjects them to standard kernels, for example

$$K = \exp(-\gamma \|x - x'\|^2) \quad (2)$$

where γ is a hyperparameter that is chosen during cross-validation. When a “good” kernel is chosen, the data points become implicitly organized by this measure, and predictions can be highly accurate. Typical kernels used in machine learning, however, are not necessarily useful measures for chemical structures. For example, the kernels are not size consistent, so large and small molecules will receive widely differing similarity scores. This leads to inconsistency in making predictions for large vs small molecules, and the problem only becomes worse for larger and larger molecules.⁴¹

Alternatively, specialized kernels might provide higher accuracy by more closely representing the underlying structure of the data. One such choice is the Tanimoto measure (eq 1),³⁴ which is particularly well-suited for use in chemical problems.³⁵ While Tanimoto is frequently applied to organic molecules, it can also be applied to catalyst structures. We denote the product Tanimoto by T_p and the catalyst Tanimoto by T_c .

Two similarity measures will be examined in this work

$$K^{(1)} = T_p \delta_C \quad (3)$$

$$K^{(2)} = T_p T_C \quad (4)$$

where δ_C gives 1.0 for a catalyst match, where the two catalysts are the same between the two reaction data points and 0.0 otherwise.

2.4. Cross-Validation and Computational Details.

Within each reaction class, training and predictions were performed using 5-fold cross-validation, and accuracy results are reported only for data points outside of the training set. This procedure was repeated 10 times, shuffling the data randomly with each training-test cycle. The mean, maximum, and minimum errors can be found in the Supporting Information, and the mean accuracies are reported in the main text.

The NN training was performed by the MLPClassifier algorithm⁴² with a ReLU activation function in the scikit-learn package.⁴³ The NN is 4 layers deep, with each of the two inner layers having the same dimension as the input vector. The output layer has the dimension equal to the number of solvents in the neural network training data. For example, the training data for Diels–Alder contains 15 solvents. Therefore, for each of those solvents a predicted weight is assigned, and this weight is normalized across the output layer. The predicted solvent is the one with the top weight, and the second most likely solvent has the second highest weight, and so on. The NN is trained and cross-validated using the product features and the catalyst fingerprints or with the product features and the catalyst identity. Catalyst identities are encoded to number categories based on the catalyst name, and therefore each type represents one input feature to the neural network.

Support vector classification (SVC) was also performed using scikit-learn. SVC was fed a concatenated vector of a fingerprint of the catalyst and a fingerprint of the products. SVC was tested for two types of kernels, the default radial-basis function, for which the formula is displayed in eq 2, and a linear (or dot-product) kernel. Recalling that a MACCS fingerprint contains exclusively 1's and 0's, the linear kernel counts the total number of functional groups common between two fingerprints. Although Tanimoto kernels may be applied to SVCs,^{37,40} for the data sets in this work Tanimoto kernels did not increase solvent prediction accuracy, and therefore the results from the Tanimoto kernel are provided in the Supporting Information. The Supporting Information also contains a description of error handling when generating the fingerprints required for these kernels.

The k -NN algorithm for solvent prediction and network visualization was created in Python by our group. The two similarity measures (eqs 3 and 4) were tested, but measure 3 is used exclusively for the graphs shown in this work (with $k = 10$). Tests involving the similarity measure of eq 4 are shown in the Supporting Information. The number of neighbors was tested to understand their effect on kNN performance, where the k parameter is the only tunable parameter for this method. The solvents predicted for each reaction are the most

frequent or popular solvents among the neighbors, and the second most frequent solvent was the second solvent prediction, etc. To break ties in frequency of solvents, the similarity measure itself was used for sorting. Visualization of the reaction networks was performed using the Force Atlas 2 algorithm⁴⁴ as implemented in the Gephi software package.⁴⁵ The Python codes are freely available at the repository located at https://bitbucket.org/ericawalk/solvent_selection/.

3. RESULTS AND DISCUSSION

3.1. Statistical Results.

The five data sets were subjected to the above-described learning techniques for solvent prediction. Multiple feature sets and similarity measures were considered, and representative choices are shown in Table 1. In particular, 6 unique algorithms are presented: kNN with two different similarity measures, neural networks using two types of raw features, and SVC using radial basis functions or a Tanimoto similarity measure. All of these methods employ the full fingerprint feature set for the reactant molecules and either one-hot encoding of the catalyst identity or catalyst fingerprints. At a basic level, this means that the kNN and support vector classification use similarity measures as their basic variable for making models, whereas the neural network processes the full set of raw features to generalize the factors responsible for solvent choice. As explained in the computational details, all results come from cross-validation, with errors reported only for points that were not used in training.

Among the evaluated models, kNN stands out as the best at predicting the experimental solvent, with success rates of 69 to 80% using the one-hot encoding of the catalyst identity. The same model using catalyst fingerprints drops in accuracy to 43 to 66% across the 5 reaction classes. The decrease in solvent selection accuracy might be attributed to the treatment of the catalyst by the MACCS keys. One aspect is the noise handling, where the kNN fingerprint vector doubles in length, in a sense diluting the quality of the reactant features toward less impactful catalyst features. An increase in catalyst features is not especially helpful for kNN because catalysts typically include metals, which are not particularly well represented in the MACCS keys, where MACCS keys are most useful for main-group compounds. Neural network models also perform reasonably well, reaching accuracies between 48 and 76% using the catalyst identity features. The neural network improves substantially when using the catalyst fingerprints, with improvements over kNN for the Diels–Alder and Wittig reactions by a small amount (<1%), while performing less well on the Friedel–Crafts, Aldol addition, and Claisen condensation reaction classes. The NN, relative to kNN, is not as influenced by the increased size of the catalyst fingerprint as kNN and intrinsically avoids irrelevant features in the fingerprint. Support vector methods generally underperformed in comparison with the kNN and neural network models.

Since more than one solvent may be equally applicable to a given reaction, but databases report only the single solvent used for a particular experiment, testing the computational models for “top N” performance is a natural procedure. Table 2 shows the accuracies of the kNN and neural network models for top-3 prediction accuracy, i.e., whether the experimental solvent is in the top 3 predicted by the model. The best choices for the features

from Table 1 are used for each algorithm, respectively. Accuracies improve across the board, as expected, with the kNN slightly outperforming the neural network. For 3 of the 5 reaction classes, the two models give accuracies within 1% of each other, and the kNN shows clearly a better performance for the Friedel–Crafts data set (which is the most challenging of the 5 as reflected by the worst-case scenario accuracy in Table 2). Overall, the prediction accuracies of 91 to 98% of the kNN tool show it to be highly capable of selecting good solvents.

These results of Table 2 show that the kNN and neural network models significantly outperform the baseline scenario, where the most prevalent solvents in the data sets are chosen. When simply taking the statistically most common solvents as the top 3 predictions, accuracies of 62 to 95% can be achieved. Interestingly, the models are unable to significantly outperform the worst case prediction for the Wittig reaction. On the other hand, the Claisen condensation worst case accuracy is 95%, but the 2 computational models give 98 to 99% accuracy, showing a significant improvement. Most dramatically, the Friedel–Crafts reaction worst case prediction gives 62% accuracy on the top 3, making it a more challenging case than the other 4 reactions. The kNN technique shines brightest in this reaction class, with 93% accuracy, which is markedly better than the neural network at 89%.

3.2. Visualizing the Reaction Landscape.

The high performance of kNN suggests that the underlying similarity measure provides a strong means for organization of the data sets. The specific similarity measure given in eq 3 uses a combination of catalyst identity and reactant Tanimoto, so the overall closeness between two data points depends only on these two factors. To better understand the relationships that lead to successful predictions, two-dimensional graphs of the data for the Friedel–Crafts reaction are shown in Figures 2 and 3. Analogous plots are shown in the SI for the Claisen condensation and Diels–Alder data sets.

Figure 2 shows that a small fraction of the catalysts appears frequently within the data, while many others appear sparsely. For a small cluster consisting of a single catalyst, often only one solvent is used for all of the reactions (note that the clustering is performed *without* knowledge of solvent). These data points are easy to predict to high accuracy, with the solvent label matching one-to-one with the catalyst species. For the larger clusters, the same catalyst might appear with several different solvents, as is obvious in Figure 3 (e.g., the Al(III) region of (a), with a colorful mixture of solvents shown in (b)). For the kNN algorithm, these clusters are harder to accurately predict compared to the more uniform, isolated smaller clusters. At the same time, these regions of the graph do contain subclusters that are consistent, so the ordering is not random. The overlaps between these subclusters lead to unavoidable errors, as classification algorithms cannot easily distinguish overlapping data points. In these cases, solvent predictions through the top-3 classification (Table 2) will gain considerably in accuracy compared to top-1 classification (Table 1), as multiple solvents may perform just as well for similar substrate/catalyst combinations. In total, however, Figures 2 and 3 show that there is a considerable degree of order in clustering, with many small clusters having highly consistent solvent designations.

Having the reactions ordered by solvent suggests an interesting possibility: does the reactant/catalyst ordering imply order in the solvent *properties*. Figure 4 suggests that this is indeed the case, where the Friedel–Crafts reaction network is labeled with solvent descriptors. The three solvent properties—dielectric constant, boiling point, and protic/aprotic—form clusters of consistency, showing that the solvents are chemically similar to one another. Importantly, this ordering in solvent properties is found in the product/catalyst network and is not an ordering predetermined by the solvent properties. Instead, the product/catalyst network implies necessary traits of solvents, which then in turn are neatly represented in the graphs.

In addition to the graphical analysis, Figure 5 shows that Zipf's law²¹ approximates the distribution of catalysts in all five reaction data sets. The fact that Zipf's law generally applies to the catalyst distribution of all 5 data sets suggests a significant trend: there are a large number of catalyst identities which appear only once, forming a non-negligible slice of the data where predictions cannot be easily made. In the interpretable kNN solvent prediction algorithm, single-catalyst reactions have no neighbors. Not only does this greatly limit the ability of kNN to make predictions, it implies that there is not enough data to training machine learning algorithms in general for these important “outlier” cases. For the 5 reaction classes, 0.76% to 7.30% of the data points are single-catalyst.

3.3. Human Chemist Focus Group Trials.

To provide feedback on the algorithm, a small group of chemists was assembled for evaluation and trials. The purpose of gathering these chemists was 3-fold: 1. to provide comparisons between computer and expert solvent predictions, 2. to evaluate whether computer solvent predictions were within reason on unlabeled data points, and 3. to give a general discussion of strengths and weaknesses of the algorithm. See the Supporting Information for a complete description of the focus group procedure (Section VIII).

The focus group evaluated a set of Friedel–Crafts and Claisen condensation reactions and was asked to predict the correct solvent from a list of 78 possibilities. The chemists were given the reaction in ChemDraw format along with the specific catalyst used in the Reaxys data entry. Of the 26 reactions in the evaluation set, 17 were labeled reactions (i.e., solvent listed in the Reaxys data entry), and 9 of the set of reactions were unlabeled. The chemists were asked to select solvents over a time frame of 30 min, equating to a little over 1 min per reaction on average.

The 17 labeled reactions in the test set were Friedel–Crafts reactions. Since the labels are available, accuracy of the computer and chemist can be evaluated on this subset, while for the 9 unlabeled points, “accuracy” is much more qualitative and will be discussed in the subsequent paragraph. In the labeled subset the computer's first solvent choice matched the label in 9 occurrences (via the kNN algorithm), giving correct predictions of dichloromethane 3 times,^{46–48} water 3 times,^{49–51} dichloroethane 1 time,⁵² and chloroform 2 times.^{53,54} The second choice of the computer matched the Reaxys entry twice for dichloroethane and once for water. The chemists, however, performed at a lower success rate than the computer: an average of 2 matches per chemist was found with the 17 solvent labels. As discussed above, the Friedel–Crafts reaction is a particularly difficult one to make solvent choices, leading to apparent disagreement between chemists and available solvent

labels. This disagreement was explicitly discussed after the expert testing to give additional insight.

After the human chemists completed their solvent selections, the computer solvent selections were revealed for labeled as well as unlabeled data points (26 in total). The human chemists were then asked to rate their level of agreement with the computer's first solvent choice. The rating scale of 1 (no disagreement) to 5 (full disagreement) allowed the chemists to give subjective feedback about the performance of the computer. As shown in Figure 6c, the chemists more often sided with the algorithm than against, with a mean agreement level of 2.3 on the 1 to 5 scale. The human chemists therefore regarded the algorithm as generally accurate, although one exception is shown below with a consistent, strong disagreement.

Figure 6c reveals a trend of the chemists disagreeing more regarding the unlabeled data points rather than the labeled. This may be explained by the unlabeled data point predictions from the computer being less informed than for labeled data points. This is evidenced by the computer not selecting second nor third solvents for 7 of the 9 unlabeled data points, due to the sparsity of neighbor connections from which to select solvents. Among the labeled data points, a lower fraction of 4 missing second solvents and 14 missing third solvents out of 17 allowed a better populated region of data and more informed solvent choices.

One example showed high levels of disagreement between the focus group and the computer (all chemists rated this point with a 5, full disagreement). In this case, the computer selected water for a reaction in which one of the reactants was acetic anhydride.⁵⁵ The chemists noted that water reacts with acetic anhydride, causing an undesired side reaction (resulting in a lower yield than an unreactive solvent). This side reaction could not be identified with the solvent prediction algorithms, leading to a knowledge gap that was swiftly noticed by the panel. In this particular case one would hope that a machine learning algorithm could be trained to learn the incompatibility with water as a solvent. As shown in Table S15 in the Supporting Information, however, using water as a solvent alongside reactant anhydride is based upon actual data in the literature. While the expert panel would object to this combination, the data-driven algorithm makes predictions based on the data it has available and is not able to learn from the expert objections.

A number of observations were made by the chemists regarding the solvent selection process, some of which are noted in Scheme 1. In addition to these points, the kNN algorithm was felt to be more transparent than other popular machine learning algorithms such as an artificial neural network. In other words, the concept of similarity measured by the kNN algorithm was acceptable to this group of experts. The algorithm "thought chemically" up to a point because the functional groups of the fingerprints were driving the similarity measures, and therefore the neighbor choices and ultimately the solvent predictions were seen as made through acceptable chemical reasoning.

4. CONCLUSIONS

Networks of named organic reactions show clusters of consistency that are predictive of experimental conditions. Techniques that exploit this consistency were shown to accurately

select solvents for these reactions, an essential component of experimental conditions. The kNN metric, a Tanimoto similarity with a catalyst label, was shown to be not only particularly effective in solvent classification but also rich in relevant conceptual chemical information. The raw molecular structure information was insufficient to accurately predict solvent, but the inclusion of a catalyst label provided the necessary chemical information to fill in this gap. Additionally, the solvent choices through this metric mimicked those of a human chemist and were visually interpretable, allowing an expert panel of chemists to view the algorithm favorably in critical testing.

While the kNN method was tested on five common named reactions, it may be expanded to less-well-known classes of reactions based on the results of the present work. For example, a more general prediction algorithm could classify reactions by SMIRKS,⁵⁶ the reaction analogy to SMILES. SMIRKS would allow systematic classification of reactions by their mechanisms and the chemical substructure involved in those mechanisms, allowing treatment of essentially any class of reaction where the reactant/product pairs are known.

In addition to having strong interpretive value, the kNN algorithm was the most accurate technique on 4 of the 5 reaction data sets, with accuracies of 91.0 to 98.0% on top-3 predictions. The highly popular technique, a deep neural network, was found to be numerically useful as well and slightly outperformed kNN on 1 of the 5 data sets (98.7% vs 98.0% accuracy), performed similarly on 3 of the 5, and performed somewhat worse on the most challenging data set, the Friedel–Crafts reaction (89.0% vs 92.8%), cf. Table 2. In total, kNN was found to be a strong technique, having interpretive value as well as statistical accuracy, making it outcompete neural networks from our point of view. Continued studies using kNN are likely to provide a route for machine-expert interfaces, where feedback from the machine learning technique allows the expert chemist to learn deep insights from large data sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work is supported by the NSF (CHE1551994) and the NIH (R35GM128830). Matthew Hannigan, Kirk Shimkin, and Grayson Ritch are acknowledged as the members of the expert chemist panel.

REFERENCES

- (1). Fooshee D; Mood A; Gutman E; Tavakoli M; Urban G; Liu F; Huynh N; Van Vranken D; Baldi P Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng* 2018, 3, 442–452.
- (2). Kayala MA; Azencott CA; Chen JH; Baldi P Learning to Predict Chemical Reactions. *J. Chem. Inf. Model* 2011, 51, 2209–2222. [PubMed: 21819139]
- (3). Chen JH; Baldi P No Electron Left Behind: A Rule-Based Expert System to Predict Chemical Reactions and Reaction Mechanisms. *J. Chem. Inf. Model* 2009, 49, 2034–2043. [PubMed: 19719121]
- (4). Segler MHS; Preuss M; Waller MP Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* 2018, 555, 604–610. [PubMed: 29595767]

- (5). Fitzpatrick DE; Battilocchio C; Ley SV Enabling Technologies for the Future of Chemical Synthesis. *ACS Cent. Sci* 2016, 2, 131–138. [PubMed: 27163040]
- (6). Kowalik M; Gothard CM; Drews AM; Gothard NA; Weckiewicz A; Fuller PE; Grzybowski BA; Bishop KJM Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angew. Chem., Int. Ed* 2012, 51, 7928–7932.
- (7). Segler MHS; Kogej T; Tyrchan C; Waller MP Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci* 2018, 4, 120–131. [PubMed: 29392184]
- (8). Raccuglia P; Elbert KC; Adler PDF; Falk C; Wenny MB; Mollo A; Zeller M; Friedler SA; Schrier J; Norquist AJ Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* 2016, 533, 73–76. [PubMed: 27147027]
- (9). Gómez-Bombarelli R; Duvenaud D; Hernández-Lobato JM; Aguilera-Iparraguirre J; Hirzel TD; Adams RP; Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci* 2018, 4, 268–276. [PubMed: 29532027]
- (10). Ma J; Sheridan RP; Liaw A; Dahl GE; Svetnik V Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model* 2015, 55, 263–274. [PubMed: 25635324]
- (11). Lin AI; Madzhidov TI; Klimchuk O; Nugmanov RI; Antipin IS; Varnek A Automated Assessment of Protective Group Reactivity: A Step Toward Big Reaction Data Analysis. *J. Chem. Inf. Model* 2016, 56, 2140–2148. [PubMed: 27783508]
- (12). Reaxys, Elsevier.: <https://www.reaxys.com> (accessed January 2018).
- (13). Coley CW; Barzilay R; Jaakkola TS; Green WH; Jensen KF Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci* 2017, 3, 434–443. [PubMed: 28573205]
- (14). Corey EJ; Wipke WT Computer-Assisted Design of Complex Organic Syntheses. *Science* 1969, 166, 178–192. [PubMed: 17731475]
- (15). Corey EJ General Methods for the Construction of Complex Molecules. *Pure Appl. Chem* 1967, 14, 19–38.
- (16). Pensak DA; Corey EJ Computer-Assisted Organic Synthesis. *ACS Symp. Ser* 1977, 61, 1–32.
- (17). Sello G Reaction Prediction-the Suggestions of the Beppe Program. *J. Chem. Inf. Model* 1992, 32, 713–717.
- (18). Jorgensen WL; Laird ER; Gushurt AJ; Fleischer JM; Gothe SA; Helson HE; Paderes GD; Sinclair S CAMEO – A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem* 1990, 62, 1921–1932.
- (19). Segler MHS; Waller MP Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J* 2017, 23, 5966–5971. [PubMed: 28134452]
- (20). Segler MHS; Waller MP Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J* 2017, 23, 6118–6128. [PubMed: 27862477]
- (21). Zipf GK Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology; Addison-Wesley Press: Cambridge, MA, 1949.
- (22). Grzybowski BA; Szymkuc S; Gazewska EP; Molga K; Dittwald P; Wolos A; Klucznik T Chematica: A Story of Computer Code That Started to Think like a Chemist. *Chem.* 2018, 4, 390–398.
- (23). Szymkuc S; Gajewska EP; Klucznik T; Molga K; Dittwald P; Startek M; Bajczyk M; Grzybowski BA Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed* 2016, 55, 5904–5937.
- (24). Fefilatyev S; Shreve M; Kramer K; Hall L; Goldgof D; Kasturi R; Daly K; Remsen A; Bunke H Label-Noise Reduction with Support Vector Machines. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*.
- (25). Gao H; Struble TJ; Coley CW; Wang Y; Green WH; Jensen KF Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci* 2018, 4, 1465–1476. [PubMed: 30555898]
- (26). Struebing H; Ganase Z; Karamertzanis PG; Sioumkrou E; Haycock P; Piccione PM Computer-Aided Molecular Design of Solvents for Accelerated Reaction Kinetics. *Nat. Chem* 2013, 5, 952–957. [PubMed: 24153374]

- (27). Chemical Identifier Resolver. National Cancer Institute, National Institutes of Health. <https://cactus.nci.nih.gov/chemical/structure> (accessed January 2018).
- (28). Weininger D SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Model 1988, 28, 31.
- (29). Weininger D SMILES. 2. Algorithm for Generation of Unique SMILES Notation. J. Chem. Inf. Model 1989, 29, 97.
- (30). O'Boyle NM; Banck M; James CA; Morley C; Vandermeersch T; Hutchinson GR Open Babel: An Open Chemical Toolbox. J. Cheminf 2011, 3, 33.
- (31). Leach AR; Gillet VJ An Introduction to Chemoinformatics; Springer: Dordrecht, The Netherlands, 2007; DOI: 10.1007/978-1-4020-6291-9.
- (32). Fernández-de Gortari E; García-Jacas CR; Martínez-Mayorga K; Medina-Fraco JL Database Fingerprint (DFP): An Approach to Represent Molecular Databases. J. Cheminf 2017, 9, 9.
- (33). Durant JL; Leland BA; Henry DR; Nourse JG Reoptimization of MDL Keys for Use in Drug Discovery. J. Chem. Inf. Comput. Sci 2002, 42, 1273–1280. [PubMed: 12444722]
- (34). Rogers DJ; Tanimoto TT A Computer Program for Classifying Plants. Science 1960, 132, 1115–1118. [PubMed: 17790723]
- (35). Bajusz D; Racz A; Heberger K Why is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? J. Cheminf 2015, 7, 20.
- (36). Suykens JAK; Vandewalle J Least Squares Support Vector Machine Classifiers. Neural Proc. Lett 1999, 9, 293–300.
- (37). Balfer J; Bajorath J Visualization and Interpretation of Support Vector Machine Activity Predictions. J. Chem. Inf. Model 2015, 55 (6), 1136–1147. [PubMed: 25988274]
- (38). Goodfellow I; Bengio Y; Courville A Deep Learning; MIT Press: 2016 <http://www.deeplearningbook.org> (accessed Aug 9, 2019).
- (39). Cover T; Hart P Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory 1967, 13, 21–27.
- (40). Ralaivola L; Swamidass SJ; Saigo H; Baldi P Graph Kernels for Chemical Informatics. Neural Netw. 2005, 18, 1093–1110. [PubMed: 16157471]
- (41). Collins CR; Gordon GJ; von Lilienfeld OA; Yaron DJ Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. J. Chem. Phys 2018, 148, 241718. [PubMed: 29960361]
- (42). Hinton GE Connectionist Learning Procedures. Artif. Intell 1989, 40, 185–234.
- (43). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V; Vanderplas J; Passos A; Cournapeau D; Brucher M; Perrot M; Duchesnay E Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res 2011, 12, 2825–2830.
- (44). Jacomy M; Venturini T; Heymann S; Bastian M ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS One 2014, 9, No. e98679. [PubMed: 24914678]
- (45). Bastian M; Heymann S; Jacomy M Gephi: An Open Source Software for Exploring and Manipulating Networks. International AAAI Conference on Weblogs and Social Media; 2009.
- (46). Hutait S; Biswas S; Batra S Efficient Synthesis of Maxonine Analogues from N-Substituted Benzyl-1-Formyl-9H-Carbolines. Eur. J. Org. Chem 2012, 2012, 2453–2462.
- (47). Yeh M-CP; Lin M-N; Chou Y-S; Lin T-C; Tseng L-Y Synthesis of the Phenanthrene and Cyclohepta[a]naphthalene Skeletons via Gold(I)-Catalyzed Intramolecular Cyclization of Unactivated Cyclic 5-(2-Arylethyl)-1,3-Dienes. J. Org. Chem 2011, 76, 4027–4033. [PubMed: 21500865]
- (48). Xu X-L; Li Z Catalytic Electrophilic Alkylation of p-Quinones through a Redox Chain Reaction. Angew. Chem 2017, 129, 8308–8312.
- (49). Meshram HM; Kumar DA; Goud PR; Reddy BC Triton B-Assisted, Efficient, and Convenient Synthesis of 3-Indolyl-3-Hydroxy Oxindoles in Aqueous Medium. Synth. Commun 2009, 40, 39–45.

- (50). Sobral AJFN; Rebanda NGCL; Da Silva M; Lampreia SH; Ramos S; Matos B; Paixao JA; Rocha Gonsalves Antonio M. D'A. One-Step Synthesis of Dipyrromethanes in Water. *Tetrahedron Lett.* 2003, 44, 3971–3973.
- (51). Jiang H; Zhang J; Xie J; Liu P; Xue M Water-Soluble (salicyladimine)₂Cu Complex as an Efficient and Renewable Catalyst for Michael Addition of Indoles to Nitroolefins in Water. *Synth. Commun* 2017, 47, 211–216.
- (52). Chatterjee PN; Roy S Allylic Activation Across an Ir-Sn Heterobimetallic Catalyst: Nucleophilic Substitution and Disproportionation of Allylic Alcohol. *Tetrahedron* 2012, 68, 3776–3785.
- (53). Chen X; Jiang H; Hou B; Gong W; Liu Y; Cui Y Boosting Chemical Stability, Catalytic Activity, and Enantioselectivity of Metal-Organic Frameworks for Batch and Flow Reactions. *J. Am. Chem. Soc* 2017, 139, 13476–13482. [PubMed: 28870069]
- (54). Xu B; Guo Z-L; Jin W-Y; Wang Z-P; Peng Y-G; Guo Q-X Multistep One-Pot Synthesis of Enantioenriched Polysubstituted Cyclopenta[b] Indoles. *Angew. Chem., Int. Ed* 2012, 51, 1059–1062.
- (55). Morkved EH; Andreassen T; Froehlich R; Mo F; Gonzalez SV Thiophen-2-Yl and Bithienyl Substituted Pyrazine-2,3-Dicarbonitriles as Precursors for Tetrasubstituted Zinc Azaphthalocyanines. *Polyhedron* 2013, 54, 201–210.
- (56). SMIRKS – A Reaction Transformation Language. Daylight Chemical Information Systems, Inc. <https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html> (accessed January 2018).

Figure 1.

A conceptualization of machine learning in chemical applications. (a) Databases do not inherently form chemical concepts, but chemists can provide interpretations of their results and/or be informed by the database entries. (b) Machine learning in chemistry can make predictions without transparent reasoning and does not typically inform expert chemists about new chemical concepts. (c) The future of the field of chemistry research will seamlessly integrate machine learning to the current foundation of chemical concepts with established databases. Chemists will regularly utilize interpretable and predictive machine learning tools.

Figure 2. Friedel–Crafts reaction network. The network is chromatically labeled by a frequent catalyst with the scheme provided in the legend of Figure 3(a).

Figure 3.

Two network cutouts chromatically labeled by property. One cutout is red, and the other cutout is black: (a) frequent catalysts chromatic label, (b) frequent solvents label, and (c) solvent prediction accuracy label.

Figure 4.

Frequent solvents cutout of the Friedel–Crafts network. The network is able to cluster reactions by solvent without any prior knowledge of solvent. Beyond solvent identity, solvent properties reveal clusters of consistency at least as large as solvent identity: (a) solvents, (b) solvent prediction accuracy, (c) dielectric constant, (d) boiling point, and (e) protic or aprotic.

Figure 5.

Zipf's law describes catalyst distributions. For all 5 reaction data sets, catalysts appear with a frequency described by a power law known as Zipf's law.

Figure 6.

Human chemist focus group trials of computer solvent selection. (a) The computer ranks its top 3 solvents, and human chemists select solvents from a reaction test set. (b) The number of matches between the computer and the human chemists is counted for solvent-labeled reactions. (c) The spectrum of agreement to disagreement of the human chemists to the computer is totaled.

Focus Group Observations

- Chemists agreed that the computer algorithm was providing reasonable results. The chemists did not detect biases in solvent choice from the computer.
- The computer had access to a large data set of reactions which gave it an advantage over a human, who could look up reactions in the same database given sufficient time and effort.
- The computer made solvent predictions based upon the most popular solvents among its neighbors, but chemists did not believe it was thinking critically about whether the solvent was good or not.
- The nearest neighbor algorithm provides a list of data entries used in a particular solvent prediction, and this data points to original articles that allow for the prediction to be checked and interpreted.

Scheme 1. Focus Group Observations

Table 1.
Accuracy of 6 Solvent Prediction Methods Across Five Reaction Data Sets (Top 1)^a

reaction	kNN with catalyst labels	kNN using catalyst fingerprints	NN with catalyst labels	NN using catalyst fingerprints	SVC with radial basis function	SVC with custom kernel
Friedel–Crafts	79.0	43.4	56.8	70.1	45.9	54.9
Aldol addition	78.0	47.4	47.8	67.0	58.8	66.9
Claisen condensation	80.1	66.0	76.1	78.2	66.2	66.2
Diels–Alder	79.9	58.7	68.5	80.5	59.9	66.8
Wittig	68.8	45.1	59.6	69.4	49.6	58.4

^aValues are in %.

Table 2.Comparison of Prediction Accuracies if Any of the Top 3 Predicted Solvents Match the Database Entry^a

reaction	kNN top 3 allowed	NN top 3 allowed	worst case scenario
Friedel–Crafts	92.8	89.0	62.3
Aldol addition	94.5	92.1	71.5
Claisen condensation	98.0	98.7	94.5
Diels–Alder	93.9	93.5	80.5
Wittig	91.3	91.3	90.8

^aThe worst case scenario is always predicting the 3 most common solvents from the training data. Values are in %.