

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

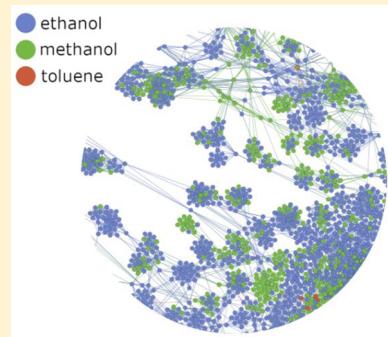
³ Eric Walker,[†] Joshua Kammeraad,^{†,10} Jonathan Goetz,[‡] Michael T. Robo,[†] Ambuj Tewari,[‡] and Paul M. Zimmerman*,^{†,10}

⁵ [†]Department of Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, Michigan 48109, United States

⁶ [‡]Department of Statistics, University of Michigan, 1085 South University Avenue, Ann Arbor, Michigan 48109, United States

7 Supporting Information

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. An example of network analysis shows where data points for a Claisen condensation reaction 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.



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 sets^{1–3} as well as supervised machine learning tools^{4–10} 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.⁶¹

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

Received: April 16, 2019

Published: August 5, 2019

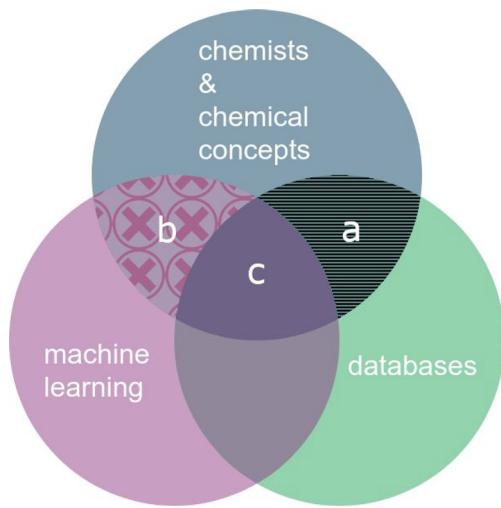


Figure 1. A conceptualization of machine learning in chemical applications. (a) Databases do not inherently form chemical concepts, but chemists can provide interpretations of their results and/or be informed by the database entries. (b) Machine learning in chemistry can make predictions without transparent 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.

72 reactions outside the scope of any database, expert system, or
 73 machine learning method. In other words, chemists are highly
 74 effective at understanding the details of chemical reactions
 75 using a set of broad physical principles, which are applicable to
 76 reactions that have never before been performed. With this
 77 consideration in mind, it becomes clear that machine learning
 78 reaction prediction computations follow an entirely different
 79 track and specialize in reaction types that most resemble their
 80 training data. By Zipf's law,²¹ databases are primarily populated
 81 by the most frequently used reactions, with a power law
 82 decrease in number of data points with rank.^{22,23} Therefore,
 83 data-driven algorithms—to no surprise—work best for the
 84 most popular reactions, and it is nontrivial to generalize these
 85 to untested, emerging, or even relatively low-population
 86 reaction classes.

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

101 In addition to these challenges, computational researchers
 102 face another key difficulty, due to the quality of available
 103 reactivity data. The largest data sets inevitably contain
 104 information from a wide variety of sources, and unlabeled or
 105 mislabeled data (i.e., "label noise") are commonplace.²⁴
 106 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

167 correctly. The full definition of this procedure is available in
168 the Supporting Information.

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

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

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

181 **2.2. Prediction Algorithms.** To provide solvent classi-
182 fication, a handful of machine learning algorithms was
183 employed. The first two, support vector machines^{36,37}
184 (SVM) and deep neural networks (NN),³⁸ are nonlinear
185 classification techniques that can predict on any number of
186 solvent identities represented by the training data set. SVM
187 classification requires a kernel function to measure similarity
188 between data points, and choices for the kernel are discussed
189 below. NNs do not require this measure, as the number of
190 input nodes can be scaled to the number of input variables.

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

199 **2.3. Similarity Measures.** The network match and SVM
200 algorithms are sensitive to the choice of similarity measure.⁴⁰
201 In typical machine learning analysis, one takes fixed-length
202 feature vectors and subjects them to standard kernels, for
203 example

$$204 K = \exp(-\gamma \|x - x'\|^2) \quad (2)$$

205 where γ is a hyperparameter that is chosen during cross-
206 validation. When a “good” kernel is chosen, the data points
207 become implicitly organized by this measure, and predictions
208 can be highly accurate. Typical kernels used in machine
209 learning, however, are not necessarily useful measures for
210 chemical structures. For example, the kernels are not size
211 consistent, so large and small molecules will receive widely
212 differing similarity scores. This leads to inconsistency in
213 making predictions for large vs small molecules, and the
214 problem only becomes worse for larger and larger molecules.⁴¹
215 Alternatively, specialized kernels might provide higher
216 accuracy by more closely representing the underlying structure
217 of the data. One such choice is the Tanimoto measure (eq
218 1),³⁴ which is particularly well-suited for use in chemical
219 problems.³⁵ While Tanimoto is frequently applied to organic
220 molecules, it can also be applied to catalyst structures. We
221 denote the product Tanimoto by T_p and the catalyst Tanimoto
222 by T_C .

223 Two similarity measures will be examined in this work

$$K^{(1)} = T_p \delta_C \quad (3) \quad 224$$

$$K^{(2)} = T_p T_C \quad (4) \quad 225$$

where δ_C gives 1.0 for a catalyst match, where the two catalysts
226 are the same between the two reaction data points and 0.0
227 otherwise.
228

2.4. Cross-Validation and Computational Details.

229 Within each reaction class, training and predictions were
230 performed using 5-fold cross-validation, and accuracy results
231 are reported only for data points outside of the training set.
232 This procedure was repeated 10 times, shuffling the data
233 randomly with each training-test cycle. The mean, maximum,
234 and minimum errors can be found in the Supporting
235 Information, and the mean accuracies are reported in the
236 main text.
237

The NN training was performed by the MLPClassifier²³⁸ algorithm⁴² with a ReLU activation function in the scikit-learn²³⁹ package.⁴³ The NN is 4 layers deep, with each of the two inner²⁴⁰ layers having the same dimension as the input vector. The²⁴¹ output layer has the dimension equal to the number of solvents²⁴² in the neural network training data. For example, the training²⁴³ data for Diels–Alder contains 15 solvents. Therefore, for each²⁴⁴ of those solvents a predicted weight is assigned, and this weight²⁴⁵ is normalized across the output layer. The predicted solvent is²⁴⁶ the one with the top weight, and the second most likely solvent²⁴⁷ has the second highest weight, and so on. The NN is trained²⁴⁸ and cross-validated using the product features and the catalyst²⁴⁹ fingerprints or with the product features and the catalyst²⁵⁰ identity. Catalyst identities are encoded to number categories²⁵¹ based on the catalyst name, and therefore each type represents²⁵² one input feature to the neural network.
253

Support vector classification (SVC) was also performed²⁵⁴ using scikit-learn. SVC was fed a concatenated vector of a²⁵⁵ fingerprint of the catalyst and a fingerprint of the products.²⁵⁶ SVC was tested for two types of kernels, the default radial-basis²⁵⁷ function, for which the formula is displayed in eq 2, and a²⁵⁸ linear (or dot-product) kernel. Recalling that a MACCS²⁵⁹ fingerprint contains exclusively 1’s and 0’s, the linear kernel²⁶⁰ counts the total number of functional groups common²⁶¹ between two fingerprints. Although Tanimoto kernels may²⁶² be applied to SVCS,^{37,40} for the data sets in this work²⁶³ Tanimoto kernels did not increase solvent prediction accuracy,²⁶⁴ and therefore the results from the Tanimoto kernel are²⁶⁵ provided in the Supporting Information. The Supporting²⁶⁶ Information also contains a description of error handling when²⁶⁷ generating the fingerprints required for these kernels.
268

The *k*-NN algorithm for solvent prediction and network²⁶⁹ visualization was created in Python by our group. The two²⁷⁰ similarity measures (eqs 3 and 4) were tested, but measure 3 is²⁷¹ used exclusively for the graphs shown in this work (with $k = 272$ 10). Tests involving the similarity measure of eq 4 are shown²⁷³ in the Supporting Information. The number of neighbors was²⁷⁴ tested to understand their effect on kNN performance, where²⁷⁵ the *k* parameter is the only tunable parameter for this method.²⁷⁶ The solvents predicted for each reaction are the most frequent²⁷⁷ or popular solvents among the neighbors, and the second most²⁷⁸ frequent solvent was the second solvent prediction, etc. To²⁷⁹ break ties in frequency of solvents, the similarity measure itself²⁸⁰ was used for sorting. Visualization of the reaction networks was²⁸¹ performed using the Force Atlas 2 algorithm⁴⁴ as implemented²⁸² in the Gephi software package.⁴⁵ The Python codes are freely²⁸³

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 %.

284 available at the repository located at [https://bitbucket.org/](https://bitbucket.org/ericawalk/solvent_selection/)
285 ericawalk/solvent_selection/.

3. RESULTS AND DISCUSSION

286 **3.1. Statistical Results.** The five data sets were subjected
287 to the above-described learning techniques for solvent
288 prediction. Multiple feature sets and similarity measures were
289 considered, and representative choices are shown in Table 1.
290 In particular, 6 unique algorithms are presented: kNN with two
291 different similarity measures, neural networks using two types
292 of raw features, and SVC using radial basis functions or a
293 Tanimoto similarity measure. All of these methods employ the
294 full fingerprint feature set for the reactant molecules and either
295 one-hot encoding of the catalyst identity or catalyst finger-
296 prints. At a basic level, this means that the kNN and support
297 vector classification use similarity measures as their basic
298 variable for making models, whereas the neural network
299 processes the full set of raw features to generalize the factors
300 responsible for solvent choice. As explained in the computa-
301 tional details, all results come from cross-validation, with errors
302 reported only for points that were not used in training.
303

304 Among the evaluated models, kNN stands out as the best at
305 predicting the experimental solvent, with success rates of 69 to
306 80% using the one-hot encoding of the catalyst identity. The
307 same model using catalyst fingerprints drops in accuracy to 43
308 to 66% across the 5 reaction classes. The decrease in solvent
309 selection accuracy might be attributed to the treatment of the
310 catalyst by the MACCS keys. One aspect is the noise handling,
311 where the kNN fingerprint vector doubles in length, in a sense
312 diluting the quality of the reactant features toward less
313 impactful catalyst features. An increase in catalyst features is
314 not especially helpful for kNN because catalysts typically
315 include metals, which are not particularly well represented in
316 the MACCS keys, where MACCS keys are most useful for
317 main-group compounds. Neural network models also perform
318 reasonably well, reaching accuracies between 48 and 76% using
319 the catalyst identity features. The neural network improves
320 substantially when using the catalyst fingerprints, with
321 improvements over kNN for the Diels–Alder and Wittig
322 reactions by a small amount (<1%), while performing less well
323 on the Friedel–Crafts, Aldol addition, and Claisen con-
324 densation reaction classes. The NN, relative to kNN, is not as
325 influenced by the increased size of the catalyst fingerprint as
326 kNN and intrinsically avoids irrelevant features in the
327 fingerprint. Support vector methods generally underperformed
328 in comparison with the kNN and neural network models.

329 Since more than one solvent may be equally applicable to a
330 given reaction, but databases report only the single solvent
331 used for a particular experiment, testing the computational
332 models for “top N” performance is a natural procedure. Table
333 2 shows the accuracies of the kNN and neural network models

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 %.

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.

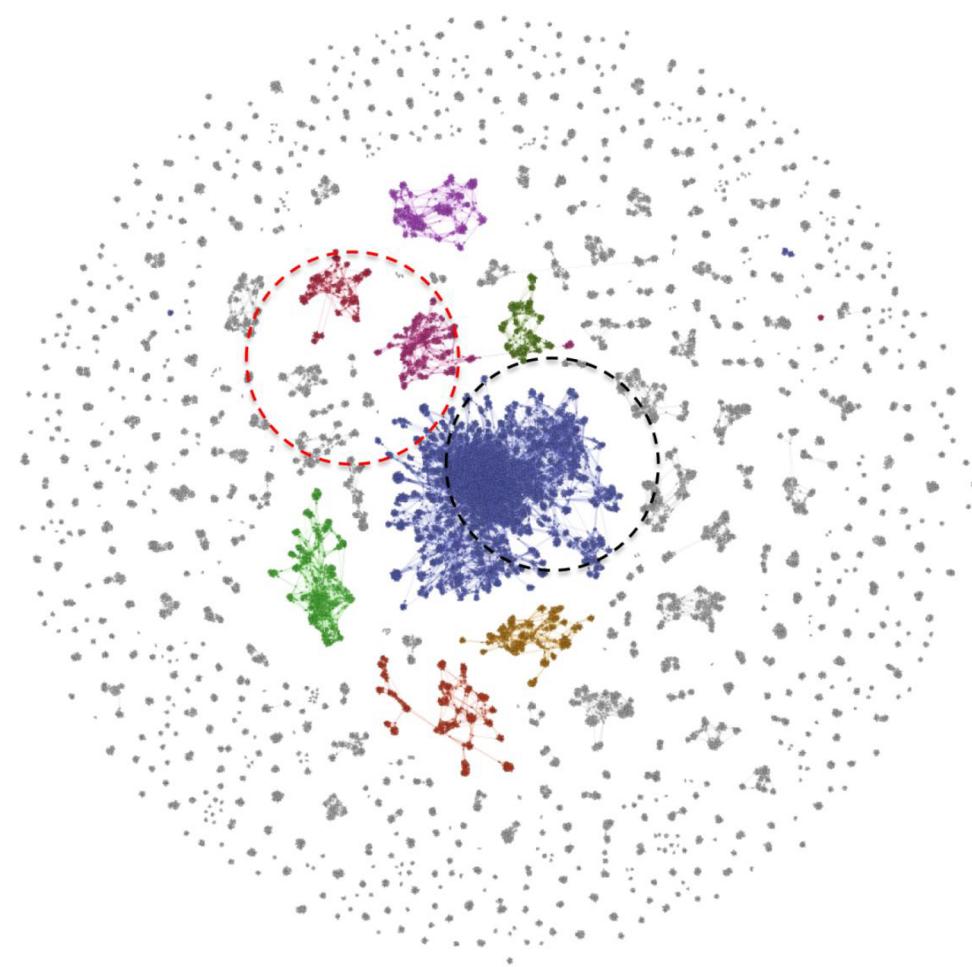


Figure 2. Friedel–Crafts reaction network. The network is chromatically labeled by a frequent catalyst with the scheme provided in the legend of Figure 3(a).

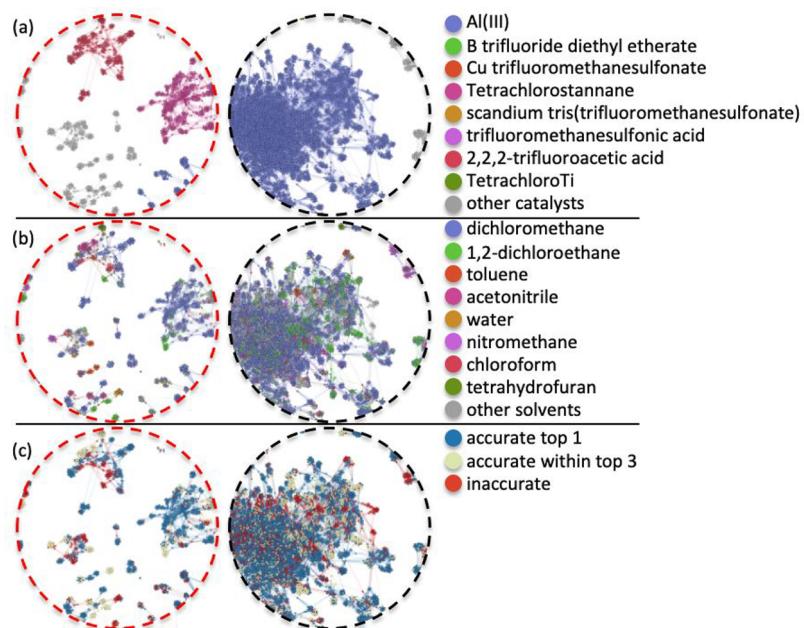


Figure 3. Two network cutouts chromatically labeled by property. One cutout is red, and the other cutout is black: (a) frequent catalysts chromatic label, (b) frequent solvents label, and (c) solvent prediction accuracy label.

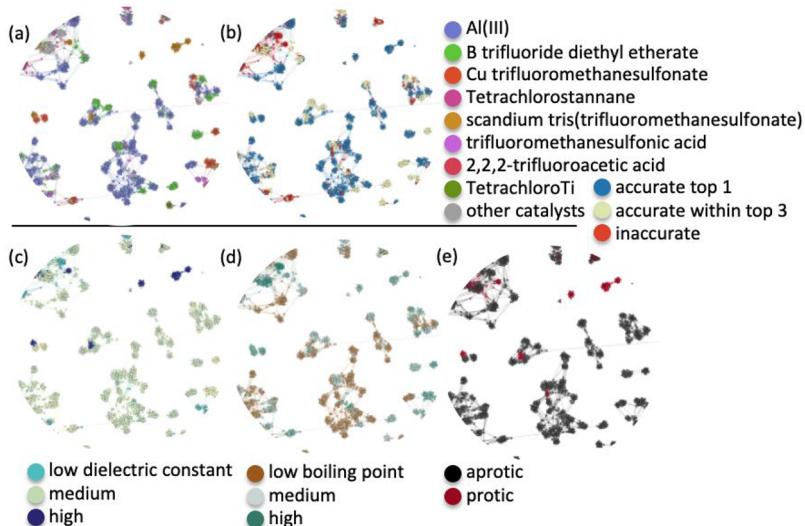


Figure 4. Frequent solvents cutout of the Friedel–Crafts network. The network is able to cluster reactions by solvent without any prior knowledge of solvent. Beyond solvent identity, solvent properties reveal clusters of consistency at least as large as solvent identity: (a) solvents, (b) solvent prediction accuracy, (c) dielectric constant, (d) boiling point, and (e) protic or aprotic.

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

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

408 In addition to the graphical analysis, Figure 5 shows that
 409 Zipf's law²¹ approximates the distribution of catalysts in all five
 410 reaction data sets. The fact that Zipf's law generally applies to

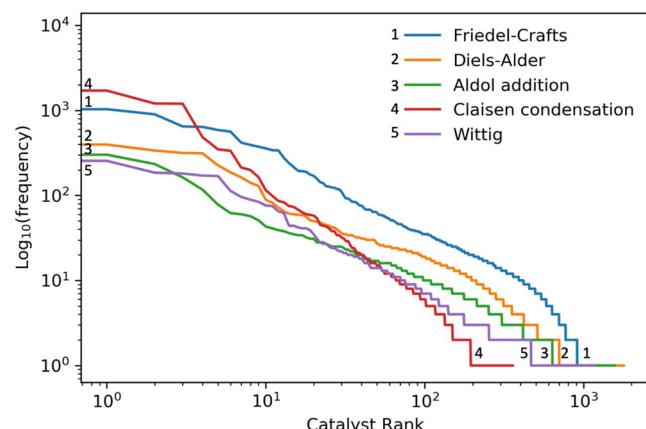


Figure 5. Zipf's law describes catalyst distributions. For all 5 reaction data sets, catalysts appear with a frequency described by a power law known as Zipf's law.

the catalyst distribution of all 5 data sets suggests a significant 411 trend: there are a large number of catalyst identities which 412 appear only once, forming a non-negligible slice of the data 413 where predictions cannot be easily made. In the interpretable 414 kNN solvent prediction algorithm, single-catalyst reactions 415 have no neighbors. Not only does this greatly limit the ability 416 of kNN to make predictions, it implies that there is not enough 417 data to train machine learning algorithms in general for 418 these important “outlier” cases. For the 5 reaction classes, 419 0.76% to 7.30% of the data points are single-catalyst. 420

3.3. Human Chemist Focus Group Trials. To provide 421 feedback on the algorithm, a small group of chemists was 422 assembled for evaluation and trials. The purpose of gathering 423 these chemists was 3-fold: 1. to provide comparisons between 424 computer and expert solvent predictions, 2. to evaluate 425 whether computer solvent predictions were within reason on 426 unlabeled data points, and 3. to give a general discussion of 427 strengths and weaknesses of the algorithm. See the **Supporting 428 Information** for a complete description of the focus group 429 procedure (Section VIII). 430

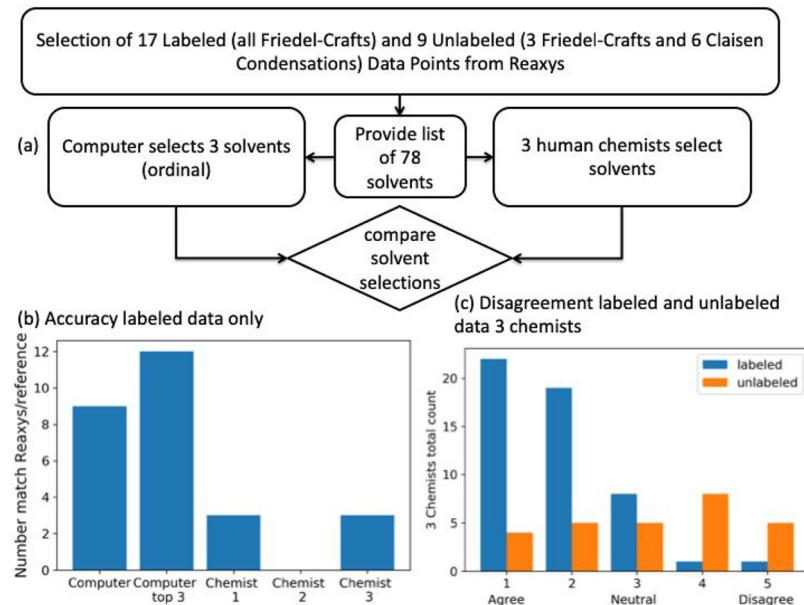


Figure 6. Human chemist focus group trials of computer solvent selection. (a) The computer ranks its top 3 solvents, and human chemists select solvents from a reaction test set. (b) The number of matches between the computer and the human chemists is counted for solvent-labeled reactions. (c) The spectrum of agreement to disagreement of the human chemists to the computer is totaled.

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

Scheme 1. Focus Group Observations

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.

507 by the kNN algorithm was acceptable to this group of experts.
 508 The algorithm “thought chemically” up to a point because the
 509 functional groups of the fingerprints were driving the similarity
 510 measures, and therefore the neighbor choices and ultimately
 511 the solvent predictions were seen as made through acceptable
 512 chemical reasoning.

4. CONCLUSIONS

513 Networks of named organic reactions show clusters of
 514 consistency that are predictive of experimental conditions.
 515 This technique was shown to accurately select solvents for
 516 these reactions, an essential component of experimental
 517 conditions. The kNN metric, a Tanimoto similarity with a
 518 catalyst label, was shown to be not only particularly effective in
 519 solvent classification but also rich in relevant conceptual
 520 chemical information. The raw molecular structure informa-
 521 tion was insufficient to accurately predict solvent, but the
 522 inclusion of a catalyst label provided the necessary chemical
 523 information to fill in this gap. Additionally, the solvent choices
 524 through this metric mimicked those of a human chemist and
 525 were visually interpretable, allowing an expert panel of
 526 chemists to view the algorithm favorably in critical testing.
 527 While the kNN method was tested on five common named
 528 reactions, it may be expanded to less-well-known classes of
 529 reactions based on the results of the present work. For
 530 example, a more general prediction algorithm could classify
 531 reactions by SMIRKS,⁵⁶ the reaction analogy to SMILES.
 532 SMIRKS would allow systematic classification of reactions by
 533 their mechanisms and the chemical substructure involved in
 534 those mechanisms, allowing treatment of essentially any class
 535 of reaction where the reactant/product pairs are known.

536 In addition to having strong interpretive value, the kNN
 537 algorithm was the most accurate technique on 4 of the 5
 538 reaction data sets, with accuracies of 91.0 to 98.0% on top-3
 539 predictions. The highly popular technique, a deep neural
 540 network, was found to be numerically useful as well and
 541 slightly outperformed kNN on 1 of the 5 data sets (98.7% vs
 542 98.0% accuracy), performed similarly on 3 of the 5, and
 543 performed somewhat worse on the most challenging data set,
 544 the Friedel–Crafts reaction (89.0% vs 92.8%), cf. Table 2. In
 545 total, kNN was found to be a strong technique, having
 546 interpretive value as well as statistical accuracy, making it
 547 outcompete neural networks from our point of view.
 548 Continued studies using kNN are likely to provide a route
 549 for machine-expert interfaces, where feedback from the
 550 machine learning technique allows the expert chemist to
 551 learn deep insights from large data sets.

■ ASSOCIATED CONTENT**● 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)

■ AUTHOR INFORMATION**Corresponding Author**

*Phone: (734) 615-0191. E-mail: paulzim@umich.edu.

ORCID 

Joshua Kammeraad: 0000-0003-0386-7198

Paul M. Zimmerman: 0000-0002-7444-1314

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.

Notes

The authors declare no competing financial interest.

■ 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, M. A.; Azencott, C. A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209–2222.
- (3) Chen, J. H.; 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.
- (4) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.
- (5) Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. Enabling Technologies for the Future of Chemical Synthesis. *ACS Cent. Sci.* **2016**, *2*, 131–138.
- (6) Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. M. Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry. *Angew. Chem., Int. Ed.* **2012**, *51*, 7928–7932.
- (7) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.
- (8) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73–76.
- (9) Gómez-Bombarelli, R.; Duvenaud, D.; Hernández-Lobato, J. M.; Aguilera-Iparragirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

- 613 (10) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep
614 Neural Nets as a Method for Quantitative Structure-Activity
615 Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.
 616 (11) Lin, A. I.; Madzhidov, T. I.; Klimchuk, O.; Nugmanov, R. I.;
617 Antipin, I. S.; Varnek, A. Automated Assessment of Protective
618 Group Reactivity: A Step Toward Big Reaction Data Analysis. *J.
619 Chem. Inf. Model.* **2016**, *56*, 2140–2148.
 620 (12) Reaxys, Elsevier.: <https://www.reaxys.com> (accessed January
621 2018).
 622 (13) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.;
623 Jensen, K. F. Prediction of Organic Reaction Outcomes Using
624 Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.
 625 (14) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of
626 Complex Organic Syntheses. *Science* **1969**, *166*, 178–192.
 627 (15) Corey, E. J. General Methods for the Construction of Complex
628 Molecules. *Pure Appl. Chem.* **1967**, *14*, 19–38.
 629 (16) Pensak, D. A.; Corey, E. J. Computer-Assisted Organic
630 Synthesis. *ACS Symp. Ser.* **1977**, *61*, 1–32.
 631 (17) Sello, G. Reaction Prediction—the Suggestions of the Beppe
632 Program. *J. Chem. Inf. Model.* **1992**, *32*, 713–717.
 633 (18) Jorgensen, W. L.; Laird, E. R.; Gushurt, A. J.; Fleischer, J. M.;
634 Gothe, S. A.; Nelson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO – A
635 Program for the Logical Prediction of the Products of Organic
636 Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921–1932.
 637 (19) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine
638 Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.*
639 **2017**, *23*, 5966–5971.
 640 (20) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning
641 to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, *23*, 6118–6128.
 642 (21) Zipf, G. K. *Human Behavior and the Principle of Least Effort: An
643 Introduction to Human Ecology*; Addison-Wesley Press: Cambridge,
644 MA, 1949.
 645 (22) Grzybowski, B. A.; Szymkuc, S.; Gazewska, E. P.; Molga, K.;
646 Dittwald, P.; Wolos, A.; Kluczniak, T. Chematica: A Story of Computer
647 Code That Started to Think like a Chemist. *Chem.* **2018**, *4*, 390–398.
 648 (23) Szymkuc, S.; Gajewska, E. P.; Kluczniak, T.; Molga, K.; Dittwald,
649 P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted
650 Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.*
651 **2016**, *55*, 5904–5937.
 652 (24) Fefilatyev, S.; Shreve, M.; Kramer, K.; Hall, L.; Goldgof, D.;
653 Kasturi, R.; Daly, K.; Remsen, A.; Bunke, H. Label-Noise Reduction
654 with Support Vector Machines. Proceedings of the 21st International
655 Conference on Pattern Recognition (ICPR2012).
 656 (25) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.;
657 Jensen, K. F. Using Machine Learning to Predict Suitable Conditions
658 for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465–1476.
 659 (26) Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.;
660 Haycock, P.; Piccione, P. M. Computer-Aided Molecular Design of
661 Solvents for Accelerated Reaction Kinetics. *Nat. Chem.* **2013**, *5*, 952–
662 957.
 663 (27) Chemical Identifier Resolver. National Cancer Institute,
664 National Institutes of Health. [https://cactus.nci.nih.gov/chemical/
665 structure](https://cactus.nci.nih.gov/chemical/structure) (accessed January 2018).
 666 (28) Weininger, D. SMILES, A Chemical Language and Information
667 System. 1. Introduction to Methodology and Encoding Rules. *J. Chem.
668 Inf. Model.* **1988**, *28*, 31.
 669 (29) Weininger, D. SMILES. 2. Algorithm for Generation of Unique
670 SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97.
 671 (30) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.;
672 Vandermeersch, T.; Hutchinson, G. R. Open Babel: An Open
673 Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.
 674 (31) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*;
675 Springer: Dordrecht, The Netherlands, 2007; DOI: [10.1007/978-1-4614-0200-9](https://doi.org/10.1007/978-1-4614-0200-9).
 676 (32) Fernández-de Gortari, E.; García-Jacas, C. R.; Martínez-
677 Mayorga, K.; Medina-Fraco, J. L. Database Fingerprint (DFP): An
678 Approach to Represent Molecular Databases. *J. Cheminf.* **2017**, *9*, 9.
 679 (33) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
 680 (34) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115–1118.
 681 (35) Bajusz, D.; Racz, A.; Heberger, K. Why is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7*, 20.
 682 (36) Suykens, J. A. K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Proc. Lett.* **1999**, *9*, 293–300.
 683 (37) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55* (6), 1136–1147.
 684 (38) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: 2016. <http://www.deeplearningbook.org> (accessed Aug 9, 2019).
 685 (39) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
 686 (40) Ralaivola, L.; Swamidas, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18*, 1093–1110.
 687 (41) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148*, 241718.
 688 (42) Hinton, G. E. Connectionist Learning Procedures. *Artif. Intell.* **1989**, *40*, 185–234.
 689 (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.
 690 (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.
 691 (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.
 692 (46) Hutaït, 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.
 693 (47) Yeh, M.-C. P.; 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.
 694 (48) Xu, X.-L.; Li, Z. Catalytic Electrophilic Alkylation of p-Quinones through a Redox Chain Reaction. *Angew. Chem.* **2017**, *129*, 8308–8312.
 695 (49) Meshram, H. M.; Kumar, D. A.; Goud, P. R.; Reddy, B. C. Triton B-Assisted, Efficient, and Convenient Synthesis of 3-Indolyl-3-Hydroxy Oxindoles in Aqueous Medium. *Synth. Commun.* **2009**, *40*, 39–45.
 696 (50) Sobral, A. J. F. N.; Rebanda, N. G. C. L.; Da Silva, M.; Lampreia, S. H.; Ramos, S.; Matos, B.; Paixao, J. A.; Rocha Gonsalves, Antonio M. D’A. One-Step Synthesis of Dipyrrromethanes in Water. *Tetrahedron Lett.* **2003**, *44*, 3971–3973.
 697 (51) Jiang, H.; Zhang, J.; Xie, J.; Liu, P.; Xue, M. Water-Soluble (salicylidimine)₂Cu Complex as an Efficient and Renewable Catalyst for Michael Addition of Indoles to Nitroolefins in Water. *Synth. Commun.* **2017**, *47*, 211–216.
 698 (52) Chatterjee, P. N.; Roy, S. Allylic Activation Across an Ir-Sn Heterobimetallic Catalyst: Nucleophilic Substitution and Disproportionation of Allylic Alcohol. *Tetrahedron* **2012**, *68*, 3776–3785.
 699 (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.

- 748 (54) Xu, B.; Guo, Z.-L.; Jin, W.-Y.; Wang, Z.-P.; Peng, Y.-G.; Guo,
749 Q.-X. Multistep One-Pot Synthesis of Enantioenriched Polysubsti-
750 tuted Cyclopenta[b] Indoles. *Angew. Chem., Int. Ed.* **2012**, *51*, 1059–
751 1062.
- 752 (55) Morkved, E. H.; Andreassen, T.; Froehlich, R.; Mo, F.;
753 Gonzalez, S. V. Thiophen-2-Yl and Bithienyl Substituted Pyrazine-2,3-
754 Dicarbonitriles as Precursors for Tetrasubstituted Zinc Azaphthalo-
755 cyanines. *Polyhedron* **2013**, *54*, 201–210.
- 756 (56) SMIRKS – A Reaction Transformation Language. Daylight
757 Chemical Information Systems, Inc. [https://www.daylight.com/
758 dayhtml/doc/theory/theory.smirks.html](https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html) (accessed January 2018).