Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

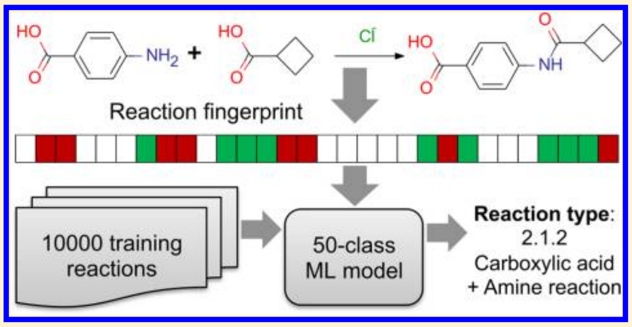
Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity

Nadine Schneider,[†] Daniel M. Lowe,[‡] Roger A. Sayle,[†] and Gregory A. Landrum*,[†]

[†]Novartis Institutes for BioMedical Research, Novartis Campus, 4002 Basel, Switzerland

[‡]NextMove Software, Ltd., Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, United Kingdom

S Supporting Information



ABSTRACT: Fingerprint methods applied to molecules have proven to be useful for similarity determination and as inputs to machine-learning models. Here, we present the development of a new fingerprint for chemical reactions and validate its usefulness in building machine-learning models and in similarity assessment. Our final fingerprint is constructed as the difference of the atom-pair fingerprints of products and reactants and includes agents via calculated physicochemical properties. We validated the fingerprints on a large data set of reactions text-mined from granted United States patents from the last 40 years that have been classified using a substructure-based expert system. We applied machine learning to build a 50-class predictive model for reaction-type classification that correctly predicts 97% of the reactions in an external test set. Impressive accuracies were also observed when applying the classifier to reactions from an in-house electronic laboratory notebook. The performance of the novel fingerprint for assessing reaction similarity was evaluated by a cluster analysis that recovered 48 out of 50 of the reaction classes with a median F-score of 0.63 for the clusters. The data sets used for training and primary validation as well as all python scripts required to reproduce the analysis are provided in the Supporting Information.

■ INTRODUCTION

The classification of reactions is essential in many different applications[1−3] such as systematic indexing of reactions in books or databases, structuring of similar reaction types,[4] analysis of different reaction classes, or retrieval of knowledge for synthesis design. Various approaches exist to classify chemical reactions. These can be divided in two broad classes: model-driven and data-driven.[1] Formerly, reactions were named according to the class of their product, functional groups, or their inventors. The first attempts at systematic classifications were based on the bonds broken or formed or on the type of reaction (elimination, addition, rearrangement, etc.).[5,6] On the basis of these, many more elaborate models were developed for a systematic classification of reaction (for an overview see Kraut et al.[3]). A disadvantage of these model-driven approaches is that concentrating on the reaction center does not allow the derivation of related subclasses.[3] In contrast, data-driven methods, which rely on the computer-based analysis of a set of reactions, try to address this issue. The data-driven approaches mostly require an atom-to-atom mapping of reactants and products[7] to identify the reaction center. The latter is the starting point for different analyses ranging from clustering[8] to similarity-based approaches[9] or for the calculation of hashcodes of the reaction center atoms[3,10,11] that can be used for the reaction classification. A drawback of these methods might be that they rely on the topology of the extended reaction center; they may not recognize reactions that are mechanistically similar but topologically different.[3] These approaches are also dependent on a correct atom-to-atom mapping for the reaction; this can be problematic for automatically generated mappings.[7] An alternative strategy to encode the transformation of a chemical reaction is their representation as reaction fingerprints,[12−15] which does not necessarily demand atom mapping. A class of reaction fingerprints called reaction vectors are difference vectors of descriptors of the reactants and the products. These have been successfully applied for clustering and similarity assessment of metabolic reactions[14] or for *de novo* design of synthetically feasible molecules.[15] They were also used in the QSAR field to represent small similar transformations and analyze the influence of these on the activity of molecules.[16] In this study, we apply difference vectors to train machine-learning methods in order to obtain reaction classifiers based on RXNO ontology.[4]

To build successful models by data-driven approaches, it is essential that a large amount of data be available. There is a very large amount of public data available about small molecules, their interactions, and their properties; examples include PubChem,[17] the Protein Data Bank (PDB),[18] and ChEMBL.[19] In contrast to this, scientists who are interested in studying how those molecules were made often have trouble finding publicly

Received: October 31, 2014
Published: December 26, 2014

DOI: 10.1021/ci5006614

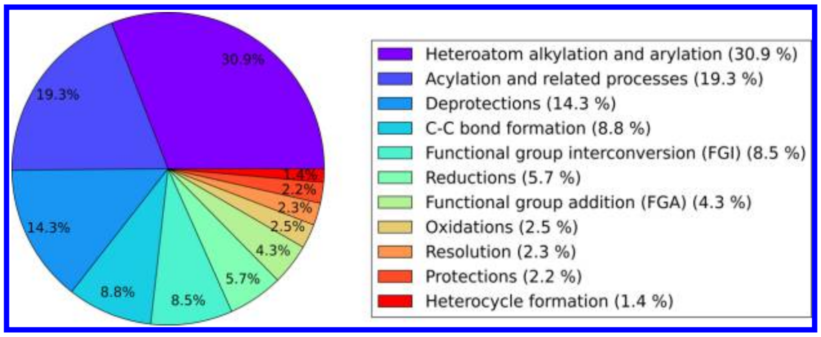


Figure 1. Distribution of reaction superclasses in the patent data set.

available data sources containing reaction schemes in a machine-readable format. A number of closed commercial reaction databases exist, such as CASREACT,[20] Reaxys,[21] or SPRESI,[22] which cover large numbers of chemical reactions but generally do not permit data mining. Some free databases are also available like SyntheticPages[23] or Webreactions;[24] however, these only contain a limited number of reactions that usually cannot be processed automatically. A first attempt at remedying this situation is a collection of more than one million chemical reactions gathered by applying text mining to patent data.[25,26] This new resource opens new areas of investigation in chemical reaction informatics, allowing us to build computational models to, for example, assign classes to the reactions, predict the yield of a new reaction, or find related reactions.

In this study, we investigate the combination of reaction fingerprints and machine learning for classifying chemical reactions to reaction types in the RSC's RXNO ontology.[3] We present the development of a novel transformation fingerprint and evaluate its applicability for model building as well as similarity searches. Further, we analyzed the influence of agents in the prediction of related reaction types. Our new classification models were validated on the patent data as well as on in-house reactions.

■ METHODS AND MATERIALS

Patent Data Set. Chemical reaction data was collected from more than one million United States granted patents from the years between 1976 and 2013 using text-mining technology[25] (the data is freely available on the Internet[26]). This resulted in 1,109,897 chemical reactions, including information about agents, solvents, yields, or quantities if available (for more information, see Lowe[25]). All of these reactions were classified according to a common subset of reaction classes published by Carey et al.[27,28] and the RSC's RXNO ontology[3] using the NameRxn tool (version v2b51) from NextMove software.[29] The classification scheme uses a structure with a maximum depth of three comprising 11 superclasses (e.g., superclass "3" C−C bond formation (RXNO:0000002)), 69 classes/categories (e.g., class "3.1" Suzuki−Miyaura coupling (RXNO:0000140)), and more than 300 named reactions/types (e.g., class "3.1.1" Bromo−Suzuki coupling). Altogether, 318 different reaction types were found in the patent data covering 54% of the extracted reactions. However, for 46% of the chemical reactions, no proper reaction class could be identified. Figure 1 shows an overview of the distribution of reaction superclasses found in the classified patent data.

The largest group of classified reactions in the data set was assigned to the heteroatom alkylation and arylation superclass (30.9%). This class is further split into nine classes, of which eight were proposed by Roughly et al.[28] and one additional was created for miscellaneous heteroatom alkylation and arylation reactions. Seven of these are found in the patent data set.

In addition, a reaction atom mapping based on the reaction classification was generated using NameRxn[29] (version v2b51). For the unclassified reactions, an atom mapping was calculated using the Indigo software[30] (version v1.1.10).

Additional statistics for the patent data set concerning the number of agents, number of partially mapped reactants, ratio of atoms mapped in the reactants, and number of agents per reaction type for the 50 most populated reaction types can be found in Figures S1 and S2 in the Supporting Information.

Experimental Setup for Reaction Classification. The general experimental setup for the classification task was as follows: Using the patent data, training and test subsets were constructed (see below). A reaction difference fingerprint was calculated for each of the reactions. These fingerprints were used to train a multi-class machine-learning (ML) model to predict reaction classes. The model was tested on an external test set calculating recall, precision, and F-score (described in more detail below) as validation measures.

The reaction classification was attempted with different ML methods and fingerprint definitions. The approach was implemented using the open-source cheminformatics toolkit RDKit[31] (version 2014.09 pre) and scikit-learn[32] (version 0.15.1). The scripts to reproduce the work are available as IPython notebooks[33] (version 1.0.0) in the Supporting Information.

Training and Test Data. Several data sets were assembled for this study using the patent data set as well as in-house reaction data for validation of the models.

Model Building Data. To train and select the reaction classification model, we have chosen the 50 largest/most populated reaction types from the patent data, which vary between 44,675 members in the largest class and 2662 members in the smallest class of this subset (Table S1, Supporting Information, for more information). We randomly selected 1000 representatives for each of the 50 reaction types. These were further split into 200 random reactions for training and 800 for testing (test set). The reaction SMILES[34] as well as the assigned reaction type were stored for each of the reactions.

External Test Data. For further validation of our final model, we constructed an external validation set by randomly selecting another 1000 representatives for each of the 50 reaction types from the patent data (external test set A). The reaction SMILES and the assigned reaction type were kept for those.

The second external test set (external test set B) we collected from the patent data was a set of 50,000 randomly chosen reactions that could not be classified by the NameRxn tool before. For these only, the reaction SMILES were available together with an atom mapping generated by the Indigo software. Another test set (external test set C) was constructed by selecting reactions that were deposited in our in-house electronic laboratory notebook (ELN) between January and August of 2014. For these, the reaction type assignment and the atom mapping were also calculated using the NameRxn software. All of the reactions that could be assigned to one of our 50 reaction types were included in the external test set C. This results in a set of 38,326 reactions that were stored as reaction SMILES with reaction type.

Reaction Fingerprints. Our goal was to develop a reaction fingerprint that can be used for both model building and similarity evaluation. In general, the reaction fingerprints tried were calculated using the difference of the chemical fingerprints of reactants and products. Reaction agents were treated separately in the fingerprint calculation. In addition to solvents and known catalysts that were already assigned to agents by the text-mining software,[25,29] we considered all reactants with less than 20% atoms mapped as an agent. These can then be included in the difference fingerprint as part of either reactants or products by varying the weight/sign of the agents in the formula below

$$\text{reactionFP} = w_{\text{nonAgent}}\Big(\sum_{\text{products } i} \text{productFP}_i - \sum_{\text{reactants } i} \text{reactantFP}_i\Big) + w_{\text{agent}} \sum_{\text{agents } i} \text{agentFP}_i$$

To test and control for the influence of the agents, a weighting for reactants and products ($w_{\text{nonAgents}}$) as well as for agents (w_{agents}) can be chosen. An alternative approach to including agents was to generate different types of fingerprints/descriptors for these separately and to concatenate reaction and agent fingerprints. In the reaction classification analysis, we investigated different types of chemical fingerprints (AtomPairs[35] (AP), Morgan2[31] (equally to ECFP4[36]), and TopologicalTorsions[37] (TT)), different sizes (max. path length for AP as well as bit size), and different weighting schemes. For the agents, we additionally calculated a feature fingerprint (MW, NumAtoms, NumRings, LogP, NumRadicalElectrons, TPSA, NumHeteroAtoms, NumHAcceptors, NumHDonors) and a dictionary-based fingerprint. For the latter case, we analyzed all of the agents in our data set and selected those that occurred at least 200 times, resulting in 72 different agents (Table S2, Figures S3 and S4, Supporting Information). Canonical SMILES of these were used as keys in our dictionary-based fingerprint. For each reaction in the data set, the agents were recorded using the dictionary-based fingerprint.

All of the fingerprints/descriptors were generated using the open-source cheminformatics toolkit RDKit[31] as count vectors. Note that negative counts are possible for these difference fingerprints. Folded (128 bit, 256 bit, 512 bit, 1024 bit, 2048 bit, and 4096 bit) and unfolded versions of the reaction fingerprints were tested.

Machine-Learning Methods. In this study, five different ML methods were applied for the classification of reactions: Random Forest[38] (RF), Naïve Bayes (NB), K-Means, Logistic Regression (LR), and k-Nearest Neighbors (kNN). All classifiers as well as the K-Means clustering were calculated using the open-source machine-learning toolkit scikit-learn[32] (version 0.15.1).

Random Forest. The RF classifier was trained setting the number of trees to 200 and using the Gini impurity as selection criterion. The model was tested with a maximum tree depth of 15 and 25 to account for the large number of possible features in the fingerprints. Further, we used a preselected seed for the RF in order to be able to obtain consistent results across several runs. For the other parameters, we have taken the default values (min_samples_split = 2, min_samples_leaf = 1).

Naïve Bayes. The multinomial NB classifier was selected because the fingerprints are count vectors. The alpha parameter of the NB classifier was set to 0.0001 to reduce the weighting of nonpresent features. Because negative counts are not possible in the NB classifier, the reaction fingerprints have to be adapted. Therefore, we doubled the length of the fingerprint and constructed it with the first half representing the positive counts, while the second half covers the absolute values of the originally negative counts.

K-Means. As a baseline model, a K-Means clustering ($k = 3$) was performed with the training set reaction fingerprints of each reaction class separately. The test set fingerprints were classified by determining the cluster/reaction class with the minimum distance (Euclidean distance).

Logistic Regression. For the LR classifier, all of the default parameters were used from scikit-learn. In an additional trial, the LR model was trained and tested with an unfolded version of the reaction fingerprints using the DictVectorizer functionality from scikit-learn.

k-Nearest Neighbors. The kNN classifier was trained with two different values for k ($k = 3$ and $k = 30$). For all other parameters, we used the default values of scikit-learn. As metric, the Euclidean distance was employed. The kNN model was trained with folded versions of the reaction fingerprints.

Reaction Fingerprint Clustering. To assess the similarity of the fingerprints, the reaction fingerprints were clustered using the Butina clustering algorithm[39] implemented in the RDKit. The clustering was calculated using Dice[40] as the similarity metric and 0.5 as the threshold for similar fingerprints. On the basis of these parameters, the clustering algorithm first generates centroids by constructing a distance matrix and calculating the number of neighbors for each fingerprint by applying the similarity threshold. Later, these were sorted in descending order depending on their number of neighbors. Starting from the first fingerprint in this list, the pairwise similarity is calculated to all other fingerprints. Those with a similarity equal or larger than the given threshold were added to that cluster and simultaneously excluded from any further comparison (for more details, see Butina[39]). Before the reaction fingerprints could be used in the clustering, they have to be converted to their "positive" form as described for the NB classifier above. This has to be done to allow a reliable calculation of the Dice similarity which is not defined for negative counts.

Validation Measures. To evaluate the performance of the models, recall, precision, and F-score were calculated. All of these were based on number of true positives (TP), false positives (FP), and false negatives (FN). The recall, also called sensitivity, is estimated as fraction of correct predictions (recall = TP/(TP+FN)). The precision is the ratio of correct predictions compared to all positive predictions (prec = TP/(TP+FP)). The F-score is derived from the recall and precision (F = 2 (recall × prec)/(recall + prec)).

Table 1. RF Results with Different Reaction Fingerprints (FP)[a]

fingerprint		RF (max depth 15)			RF (max depth 25)		
type	parameters[b]	recall	prec	F-score	recall	prec	F-score
reactionFP 2048 bit (AP)	10_1_wA	0.79	0.85	0.82	0.89	0.89	0.89
	10_−1_wA	0.78	0.82	0.8	0.86	0.88	0.87
	1_1_wA	0.79	0.85	0.82	0.89	0.89	0.89
	w/oA	0.92	0.92	0.92	0.94	0.94	0.94
reactionFP 4096 bit (AP)	10_1_wA	0.8	0.85	0.82	0.89	0.9	0.9
	10_−1_wA	0.79	0.83	0.81	0.87	0.88	0.87
	1_1_wA	0.8	0.85	0.82	0.89	0.9	0.9
	w/oA	0.92	0.92	0.92	0.94	0.94	0.94
reactionFP 2048 bit + featureFP (agents)		0.95	0.95	0.95	0.97	0.97	0.97
reactionFP 2048 bit + dictionary-based FP (agents)		0.95	0.95	0.95	0.96	0.96	0.96
reactionFP 2048 bit + Morgan2 FP (agents)		0.95	0.95	0.95	0.97	0.97	0.97

[a] Classification results for 40,000 test set reactions equally distributed among 50 different reaction classes. [b] Signature: $\text{weight}_{\text{nonAgents}}$_$\text{weight}_{\text{agents}}$_includeAgents (wA = included, w/oA = not included).

■ RESULTS AND DISCUSSION

In the first part of this section, we present the development of a novel reaction fingerprint. We explored multiple different fingerprints and strategies for the inclusion of agents. The fingerprints were used in several ML algorithms to train models for classifying reactions and were evaluated on the patent data test set. Finally, the most promising model was tested on external data sets compiled from in-house ELN data as well as from unclassified reactions of the patent data. In the second part, we apply cluster analysis to evaluate the performance of the novel fingerprint for use in similarity assessment.

Chemical Reaction Class Assignment. The 50 largest classes from the patent data set were chosen to build a 50-class predictive model. We conducted a series of experiments to find the most effective classification model. In the first setup, we tested the reaction fingerprints for classification in general as well as the influence and handling of agents. First, reaction fingerprints were generated with and without agents and using different weighting schemes for the agents. For instance, in the first two tests, the weighting of the agents was scaled down by increasing the reactants' and products' weights by 10. Agents were either included as products (Table 1 first row: $\text{agent}_{\text{weight}} = 1$) or reactants (Table 1 second row: $\text{agent}_{\text{weight}} = -1$). As the underlying molecular fingerprint, a folded version (2048 bit and 4096 bit) of AP with RDKit default settings (max. path length = 30) was chosen. The resulting fingerprints were used to train RF classifiers using two different maximum tree depths (15 and 25). Furthermore, special fingerprints were created where agents were not directly included in the difference reaction fingerprints but concatenated with them so as to not mix the bits from the reaction components and the agents. We used feature fingerprints for the agents to cover a very general description of the physicochemical properties of the agents, Morgan2 fingerprints as a substructure description of the agents and finally a dictionary-based fingerprint representing the most common agents found in the patent reactions. All results of this first experiment are in Table 1. In addition, Figure 2 shows confusion matrices for the 50 classes; other confusion matrices can be found in Figure S5 of the Supporting Information.

Directly including the agents in the difference reactions fingerprint led to moderate performance of the classification model with an F-score varying between 0.8 and 0.82 on our test set (Table 1, Figure 2, top) independent of the weighting of the agents. Ignoring agents increases the performance significantly to an F-Score of 0.92 (Table 1). Allowing the model to be more flexible by using a maximum tree depth of 25 instead of 15 in the RF classifier further improved the classification to an F-score of 0.94 and simultaneously reduced the number of errors (Table 1). We also tried a maximum tree depth of 30, which did not improve the accuracy of the model. Using a larger bit size for the reaction fingerprint (4096 bit) also did not enhance model performance (Table 1).

Figure 2 shows that most of the classification errors occur between related classes, for example, Suzuki coupling and Suzuki-type coupling or Mitsunobu aryl ether synthesis and Williamson ether synthesis. Because some reaction types differ only in the agents used, better discrimination should be possible by incorporating agents in the fingerprints. Directly including agents in the difference reaction fingerprint resulted in a sizable decrease in the performance; therefore, we tried several approaches to consider agents separately. Concatenating the reaction fingerprints without agents with feature, dictionary-based, or Morgan2 fingerprints of the agents all led to improvements in our classification performance (Table 1). The best two models we generated in this first setup (reaction fingerprint combined with feature or Morgan2 fingerprint of the agents, RF max. tree depth of 25) obtained an excellent F-score of 0.97 on our test set (Table 1, bottom). Comparing the results in Figure 2 (top) and Figure 2 (bottom) clearly shows that some of the confusion between related reaction classes is remedied by properly including the agents, for example, Bromo−Suzuki coupling (3.1.1) and Bromo−Suzuki-type coupling (3.1.5) or Fischer−Speier esterification (2.6.1) and methyl esterification (1.7.6). To summarize the first experiment, we found that reaction fingerprints can be used to build a classification model and that agents can be helpful to distinguish related reaction classes if properly incorporated. Further, we found that a 2K fingerprint size is sufficient and that a larger tree depth in the RF reduces the number of errors.

In the second experiment, we tested if more local molecular fingerprints can also capture details of the transformation of the chemical reaction. In the preceding section, we applied an AP fingerprint with a maximum path length of 30 bonds, allowing the atoms in the pairs to span most if not all of the molecular structures. This will lead to many bits in the difference fingerprint, which are very specific for one particular reaction but which may introduce noise to the general transformation characteristic for a reaction type. To explore this, we generated three reaction fingerprint types that are more local in character:

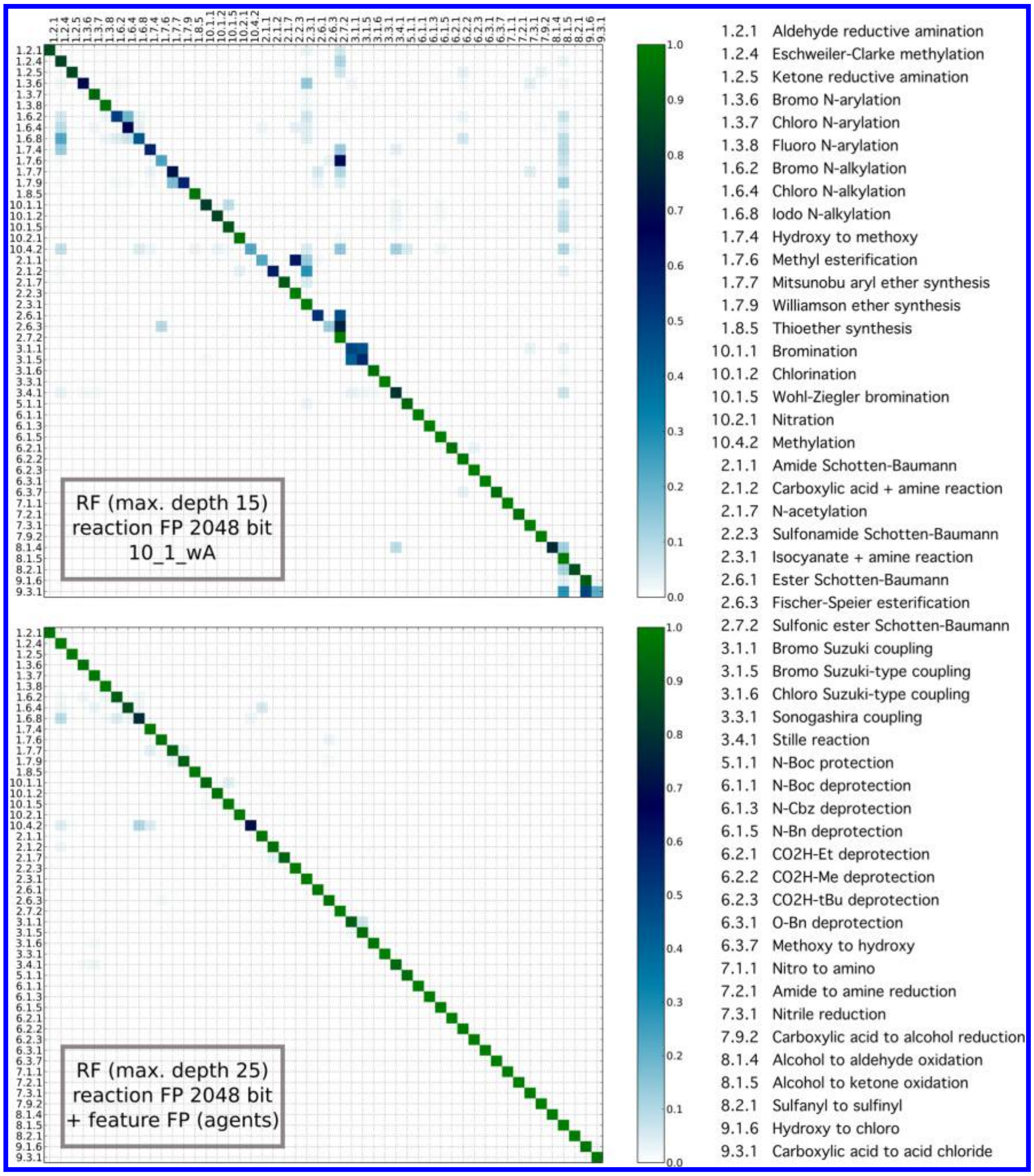


Figure 2. Confusion matrices of two different RF models: Top: Using reaction fingerprints (FP) with weighted agents. Bottom: Using reaction fingerprints (FP) combined with agent feature fingerprint. Classification results are shown for 40,000 test set reactions equally distributed among 50 different reaction classes. The color bar is on a percentage scale. Results for a certain cell are only shown if at least 1% of the reactions were classified in that cell, otherwise it is left white. On the *x*-axis, the true reaction type is plotted, while on the *y*-axis, the predicted reaction type can be found.

Table 2. RF Results with Different Transformation Fingerprints (FP)[a]

fingerprint		RF (max depth 15)			RF (max depth 25)		
type	parameters	recall	prec	F-score	recall	prec	F-score
transformationFP 2048 bit	AP3	0.93	0.93	0.93	0.95	0.95	0.95
	Morgan2	0.93	0.93	0.93	0.94	0.94	0.94
	TT	0.77	0.82	0.79	0.82	0.84	0.83
transformationFP 4096 bit	AP3	0.93	0.93	0.93	0.95	0.95	0.95
	Morgan2	0.92	0.92	0.92	0.94	0.94	0.94
	TT	0.78	0.81	0.80	0.83	0.86	0.84

[a] Classification results for 40,000 test set reactions equally distributed among 50 different reaction classes.

Table 3. Results for Different ML Models and Different Transformation Fingerprints (FP)[a]

fingerprint		RF (max depth 25)			K-Means (k = 3)		
type	parameters	recall	prec	F-score	recall	prec	F-score
transformationFP 2048 bit	AP3	0.95	0.95	0.95	0.88	0.88	0.88
	Morgan2	0.94	0.94	0.94	0.89	0.89	0.89
	TT	0.82	0.85	0.84	0.76	0.79	0.78
transformationFP 4096 bit w/o negative counts	AP3	0.95	0.95	0.95	–	–	–
	Morgan2	0.94	0.94	0.94	–	–	–
	TT	0.83	0.85	0.84	–	–	–
		multinomial NB (α = 0.0001)			LR		
transformationFP 2048 bit	AP3	–	–	–	0.95	0.95	0.95
	Morgan2	–	–	–	0.94	0.94	0.94
	TT	–	–	–	0.91	0.91	0.91
transformationFP 4096 bit w/o negative counts	AP3	0.93	0.93	0.93	–	–	–
	Morgan2	0.89	0.9	0.89	–	–	–
	TT	0.87	0.87	0.87	–	–	–
		kNN (k = 3)			kNN (k = 30)		
transformationFP 2048 bit	AP3	0.92	0.92	0.92	0.86	0.87	0.87
	Morgan2	0.81	0.89	0.85	0.79	0.83	0.81
	TT	0.80	0.84	0.82	0.64	0.77	0.70

[a] Classification results for 40,000 test set reactions equally distributed among 50 different reaction classes.

an AP fingerprint with maximum path length of three (AP3), a Morgan2 fingerprint, and a TT fingerprint. These are referred to as transformation fingerprint in the following. Agents were not included in these fingerprints. We again investigated two bit lengths of the folded fingerprints (2048 bit and 4096 bit) as well as two maximum tree depths for the RF classifier. All results are summarized in Table 2. Figure S6 of the Supporting Information shows the confusion matrices of the best results of this second setup.

In this experiment, the best classification model, which resulted in an F-score of 0.95, is obtained using the AP3 fingerprint and a maximum tree depth of 25 of the RF classifier. The 4K fingerprint did not improve the classification as also discovered in the first experiment (Table 2). The performance using the Morgan2 fingerprint was not significantly inferior to the AP3 fingerprint, producing an F-score of 0.94 on our test set. In contrast, the TT fingerprint was markedly worse, showing an F-score of 0.84 (Table 2, Figure S6, Supporting Information). We conclude that transformation fingerprints are sufficient for the classification of chemical reactions, even performing slightly better than reaction fingerprints.

After analyzing several fingerprint types for reaction classification, our next round of experiments probed the influence of various ML methods. Due to the excellent performance of the RF classifier, the question arises if simpler models are also capable of building a 50-class predictive model. In this setup, we also applied the three different transformation fingerprints (AP3, Morgan2, and TT) described before and compared the performance of five ML methods: RF (max. tree depth = 25), K-Means clustering (k = 3), multinomial NB (α = 0.0001), LR, and kNN (k = 3 and k = 30). We used a folded 2K bit version of the fingerprints except for the NB classifier, which cannot handle the negative counts of the difference transformation fingerprint. Hence, we provided a special version of the fingerprints to represent both positive and negative counts; a more detailed description is given in the Methods and Materials section. This special fingerprint was also tested with the RF classifier to allow comparison. In this experiment, the agents were neglected. The results of this setup are summarized in Table 3. The confusion matrices are depicted in Figure S7 of the Supporting Information.

The goal of this test was to find a simpler ML method and define a baseline model for the reaction classification task. Our K-Means-based classifier that generates a K-Means clustering for each reaction type separately and finally employs the minimum distance to a cluster for the classification (see Methods and Materials section for more details) even shows a reasonable performance (best F-score = 0.89) when compared to the other more elaborate methods (Table 3). This result provides strong evidence, to be revisited below, that these fingerprints will also be useful for similarity searching and related tasks. With K-Means, the TT fingerprints provided a markedly worse performance, as was found when using them in the RF classifier (Tables 2 and 3). As a second baseline, we generated kNN models with two different values for k. The performance of the model was significantly better when considering three nearest neighbors instead of choosing the 30 nearest neighbors (Table 3). The TT fingerprints were once again inferior compared to the AP3 fingerprints. The very good performance when using the AP3 fingerprints in the kNN (k = 3) classifier indicates their suitability for similarity-based approaches such as clustering or similarity search (see also Clustering of Reaction Fingerprints section). The multinomial NB model performs comparably to the RF classifier with the AP3 fingerprint but is worse with the Morgan2 fingerprint. In contrast, it is more effective than RF and K-Means when using the TT fingerprint (Table 3). The best model in this comparison proved to be the LR classifier, which performs very well for all three different fingerprint types, achieving F-scores between 0.91 and 0.95 (Table 3). The most effective fingerprint type in this setup is once again AP3. On the basis of these results, the combination of AP3 fingerprint and LR produced the most robust classification models.

As a final step to identify the most effective approach for reaction classification, we explored the bit size of the transformation fingerprint and the inclusion of agent fingerprints. Additionally, we tested an unfolded version of the AP3 transformation fingerprint using the DictVectorizer functionality from scikit-learn. This allowed us to determine the real number of different bits found in our data sets (training as well as test set) and assess the influence of bit collisions due to the folding. The results of this analysis are outlined in Table 4 and are shown in Figure S8 of the Supporting Information.

Table 4. LR Results with Different Transformation Fingerprint (FP) Bit Sizes[a]

fingerprint		LR		
type	parameters	recall	prec	F-score
transformationFP AP3 (folded)	4096 bit	0.95	0.95	0.95
	2048 bit	0.95	0.95	0.95
	1024 bit	0.95	0.95	0.95
	512 bit	0.95	0.95	0.95
	256 bit	0.95	0.95	0.95
	128 bit	0.93	0.93	0.93
transformationFP AP3 (unfolded)	(1232 bit)	0.95	0.95	0.95
transformationFP AP3 (folded) + featureFP (agents, unfolded)	256 bit +9 bit	0.97	0.97	0.97
transformationFP AP3 (folded) + Morgan2 FP (agents, folded)	256 bit +256 bit	0.97	0.97	0.97
transformationFP AP3 (unfolded) + featureFP (agents, unfolded)	(1241 bit)	0.98	0.98	0.98
transformationFP AP3 (unfolded) + Morgan2 FP (agents, unfolded)	(13828 bit)	0.98	0.98	0.98

[a] Classification results for 40,000 test set reactions equally distributed among 50 different reaction classes. The nine features of the agent FeatureFP are MW, NumAtoms, NumRings, LogP, NumRadicalElectrons, TPSA, NumHeteroAtoms, NumHAcceptors, and NumHDonors.

We discovered that the bit size of the transformation fingerprint could be reduced to 256 bits without decreasing the performance of the classification model. Further reduction of the size to 128 bits led to a small loss of 2% (F-score = 0.93) (Table 4). Comparing this result to the unfolded version of the transformation fingerprint showed that no gain is obtained by using all of the 1232 unique bits found in our data set. Although decreasing the bit size introduced a lot of collisions (a detailed analysis of this can be found in Figure S9 of the Supporting Information), this did not strongly influence the performance of the model. Either the fingerprint contains a large amount of redundant information or the bits containing critical information are not affected by the folding process. The first explanation seems more plausible. Combining the 256 bit transformation fingerprint with either the agent feature or Morgan2 fingerprint further improved the results to an excellent F-score of 0.97 (Table 4). We also tested the unfolded version of this fingerprint combination that did not result in a significant improvement, however demanding 976 bits more in case of combining with the agent feature fingerprint or 13,316 bits more using the agent Morgan2 fingerprint (Table 4).

As the final model for reaction classification, we chose the combination of the AP3 transformation fingerprint (256 bit) and the feature fingerprint for the agents trained using the LR classifier. A Y-scrambling was performed for the final model to eliminate the correlation by chance of our fingerprint and the reaction types. This tests results in an F-score of 0.02 providing very strong support for the argument that the excellent performance of our models is not due to chance/overfitting. Finally, we applied our final model on an external set (external test set A) to additionally check for overfitting of the model. On this data set that is composed of another 1000 reactions for each of the 50 chosen reaction types, we achieved the same convincing results as before (F-score = 0.97, recall = 0.97, precision = 0.97) refusing any assumption of overfitting.

Furthermore, we analyzed some of the incorrectly predicted reactions from our test set. We selected the nine reaction types where more than 5% of the reactions were incorrectly predicted. From these, we randomly chose three examples each and looked into the details of the reactions. We found different reasons for the incorrect classifications. Some could be attributed to incorrect atom mapping leading to erroneous reaction role assignment (Figure 3). For example, a Williamson ether synthesis was classified as a Bromination reaction. In this reaction, a phenol and a 2,6-dibromopyridine were fused to 2-bromo-6-phenoxypyridine. However, in the reaction SMILES, both reactants were not mapped, but a third reactant 2-methanol-6-phenoxypyridine was included in the atom mapping (Figure 3, top). Because unmapped reactants were included in the agents, the difference reaction fingerprint captures the removal of the methanol bits and the addition of the bromine bits. A similar case was identified for an N-Boc protection, which was predicted as a Williamson ether synthesis. Here also, the atom mapping was incorrect, and on the basis of the mapping, the classification model gave the correct reaction type (Figure 3, bottom).

Another reason for incorrectly classified reactions is the existence of ambiguous classifications of reaction types. For example, we found a methylation that was classified as an Iodo N-alkylation (Figure 4). In this reaction, both reaction types are correct because a methyl is added to a nitrogen atom using an iodomethane as catalyst. Other classification errors also occurred between related classes, for example, Bromo−Suzuki and Bromo−Suzuki-type coupling, or related reaction mechanisms, for example, N-acetylation and carboxylic acid + amine reaction.

Reaction Class and Superclass Prediction Models. The hierarchical assignment of reaction types according to the name reaction ontology (RNXO) allowed creation of models for reaction class and superclass prediction. For this task, the training and test set fingerprints (AP3 transformation fingerprint folded (256 bit) + featureFP (agents)) were reassembled with regard to their reaction class and superclass membership. This results in 28 different reaction classes and 10 superclasses within the 50 selected reaction types. We used 20% of the fingerprints as training data for the LR classifier and the remaining 80% as test set. Both new models show the same excellent performance as the reaction type prediction model obtaining an F-score of 0.97 (recall = 0.97 and precision = 0.98). The confusion matrices of both models are shown in Figure 5.

Most of the confusion can be found between related transformations, for example, 2% of the reactions in class "O-substitution" (1.7) were predicted as "O-acetylation to ester" (class 2.6) or "NH protections" (class 5.1) where 3% of the reactions were classified as "N-acetylation to amide" (class 2.1) (Figure 5, top). These two models were also externally validated on test set A, which gave the same excellent results as on our test data before (F-score = 0.98, recall = 0.97, and precision = 0.98. This experiment indicated that the trans-

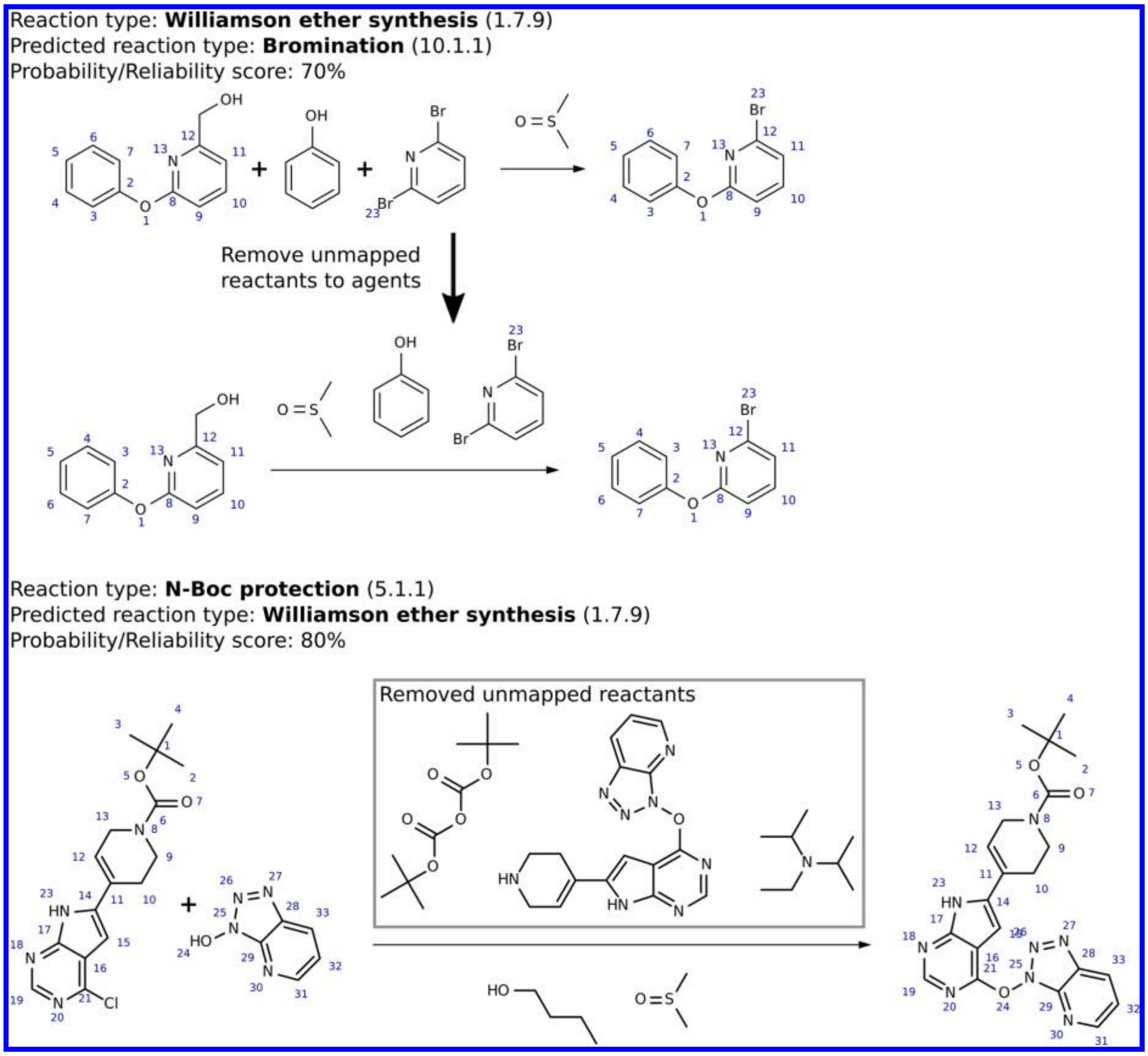

Reaction type: Williamson ether synthesis (1.7.9)
Predicted reaction type: Bromination (10.1.1)
Probability/Reliability score: 70%
Remove unmapped reactants to agents
Reaction type: N-Boc protection (5.1.1)
Predicted reaction type: Williamson ether synthesis (1.7.9)
Probability/Reliability score: 80%
Removed unmapped reactants

Figure 3. Two examples of incorrect classified test reactions: Top: Williamson ether synthesis that was classified as bromination. Bottom: N-Boc protection reaction that was predicted as a Williamson ether synthesis. Both reactions were incorrectly classified due to incorrect atom mapping.

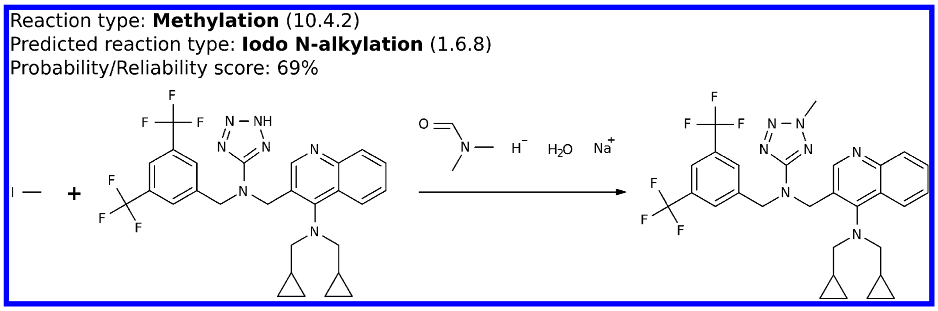

Reaction type: Methylation (10.4.2)
Predicted reaction type: Iodo N-alkylation (1.6.8)
Probability/Reliability score: 69%

Figure 4. Methylation that was classified as an Iodo N-alkylation reaction. This is an example for an ambiguous classification of a reaction type.

formation fingerprint could also be used to build models for classifying related chemical transformations.

Classification of In-House Reactions. We applied the reaction type, class, and superclass prediction model to a set of in-house reactions (external test set C). The transformation fingerprints combined with the agent feature fingerprints were calculated for the reactions. Using the reaction type classification, we obtained a slightly worse performance compared to the application on the patent data (F-score = 0.89, recall = 0.93, precision = 0.86). Most of the errors can be found between related classes ("Amide−Schotten−Baumann" (2.1.1), "Carboxylic acid + amine reaction" (2.1.2), or "N-acetylation" (2.1.7)) or similar chemical transformations ("Iodo N-Alkylation" (1.6.8) and "Methylation" (10.4.2). The same trend can be found for the other two models. The class prediction model achieved an F-score of 0.88 (recall = 0.92, precision = 0.85), while the superclass prediction model obtained an F-score of 0.89 (recall = 0.86, precision = 0.91).

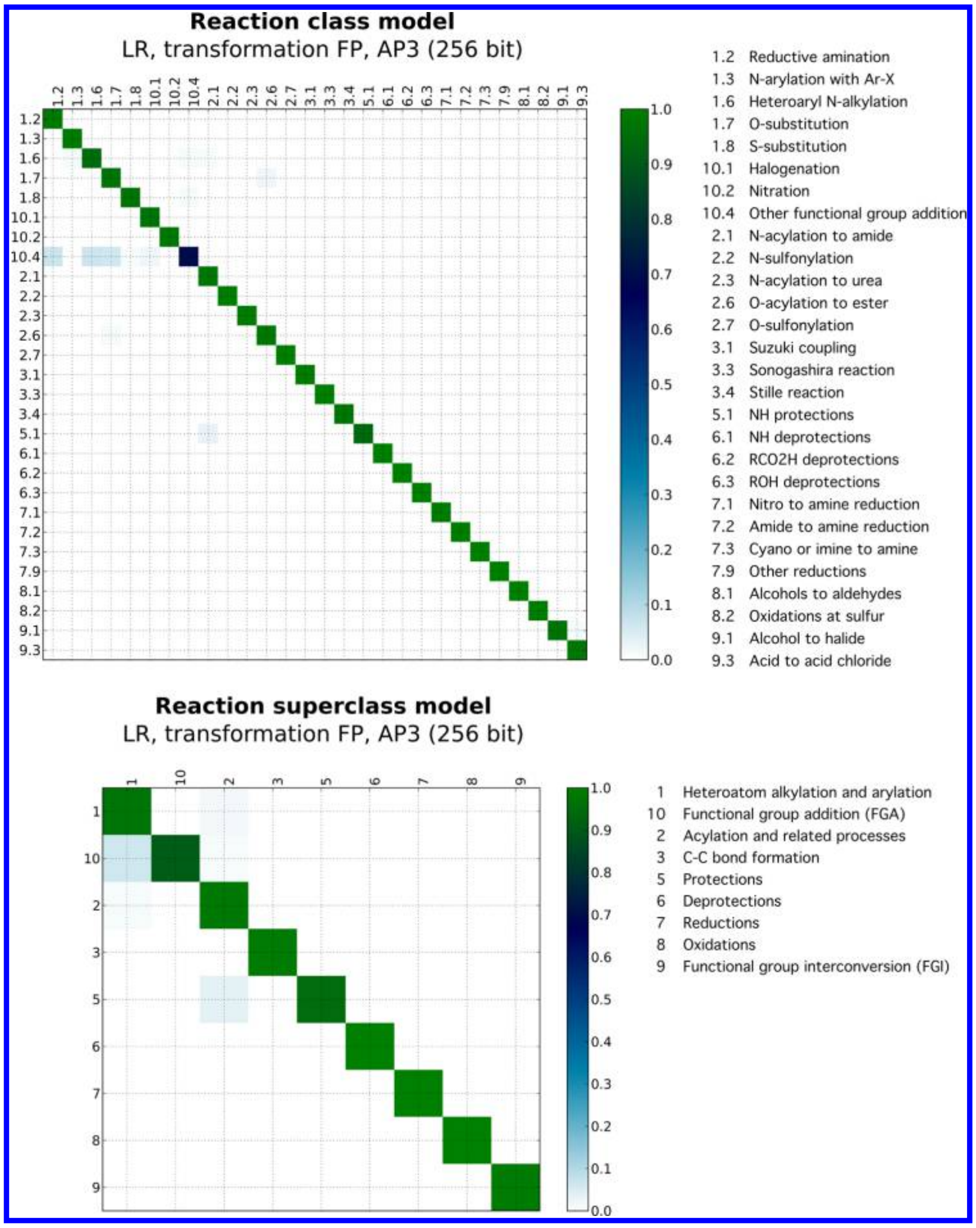


Figure 5. Confusion matrices of class (top) and superclass (bottom) prediction models (LR classifier, AP3 transformation fingerprint folded (256 bit) + featureFP (agents)). Classification results for 40,000 test set reactions. The color bar is on a percentage scale. Results for a cell are only shown if at least 1% of the reactions were classified in that cell, otherwise it is left white. On the *x*-axis, the true reaction type is plotted, while on the *y*-axis the predicted reaction type can be found.

We believe that the decrease in model performance is due to differences in representation/standardization of the reactions, particularly the specification of agents/reactants. Addressing that would require a more sophisticated scheme for assigning reaction roles when processing the ELN data.

Classification of Unclassified Reactions. A further application of our classification models was to a randomly selected subset of the unclassified reactions in the patent data (50,000 reactions) to check if some of those could be recovered. First, the transformation and agent feature fingerprint combination for all of the sample reactions were generated. Subsequently, these were classified using the reaction type prediction model. The results of this classification can be found in Figure 6, top.

The LR classifier also provides the possibility to obtain prediction probabilities per reaction type. We used these probabilities to select the most likely reaction type per reaction/fingerprint and stored it together with the probability itself. Because the unclassified reactions may include some errors, unusual characteristics or belong to rare or yet unspecified reaction types, we only kept those with a high probability for one of the 50 reaction types of our model. As a threshold, we chose a probability of greater than 95% for one reaction type, which excluded 48,586 of the reactions (Figure 6, bottom) and decreased the number of reaction types from 50 to 48. To obtain a more reliable prediction of the reaction type for the remaining 1414 reactions, we classified these using the class and subsequently the superclass prediction model. This allowed confirmation of consistent reaction type, class, and superclass and reduced the subset of reactions to 1167. From these, we selected the five most probable (if available) predictions per reaction type for manual validation, which

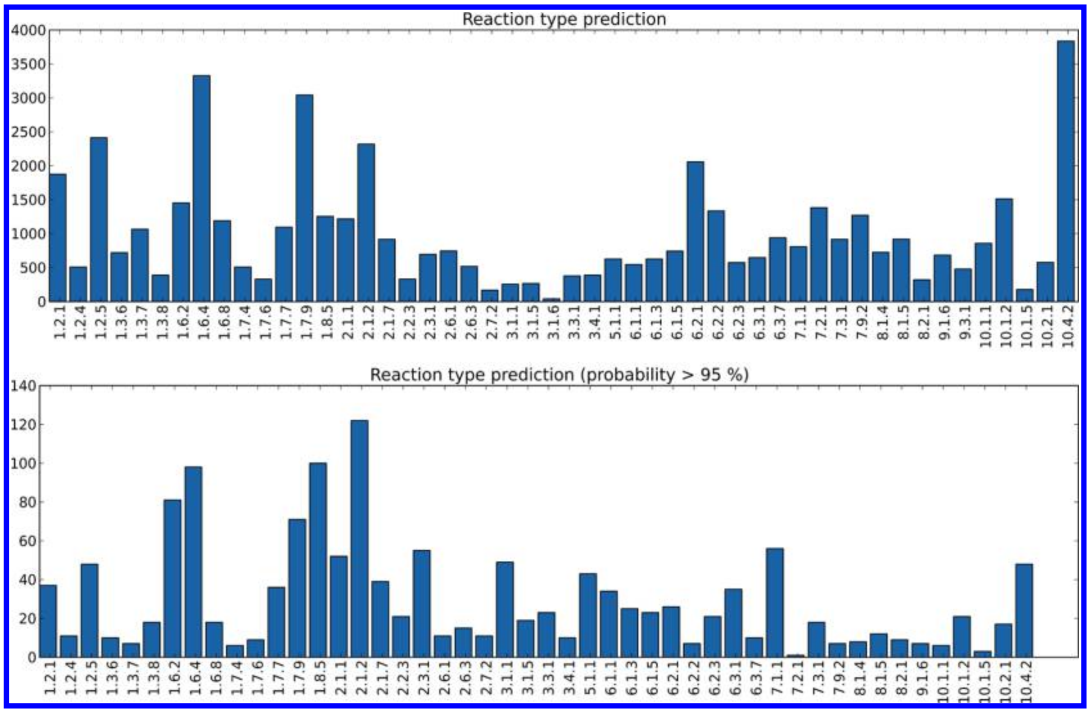
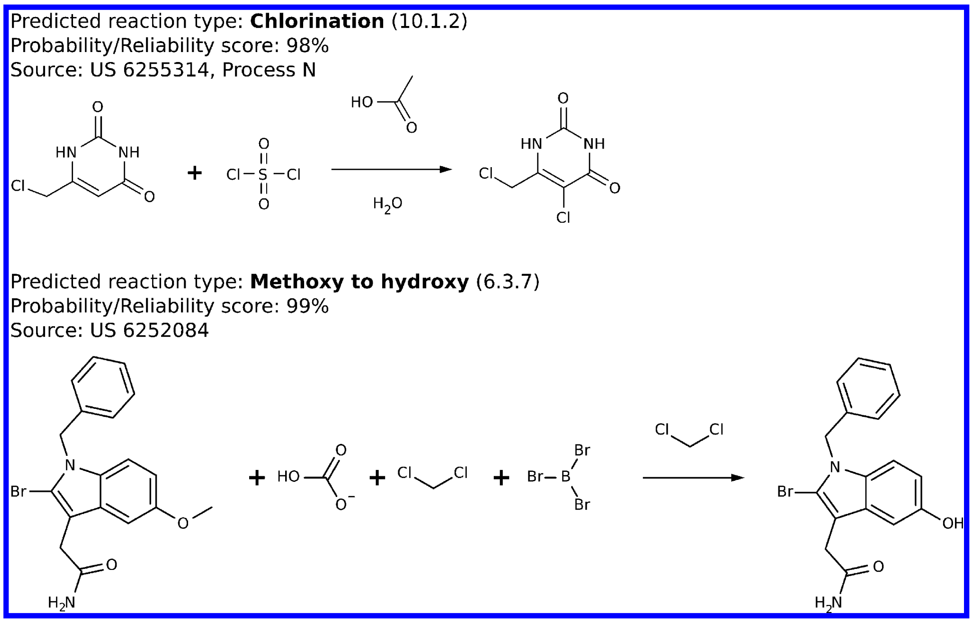


Figure 6. Distribution of predicted reaction types for 50,000 unclassified reactions: Top: Reaction type assignment based on the LR classifier. Bottom: Reaction type assignment using a probability threshold >0.95 for one reaction type.

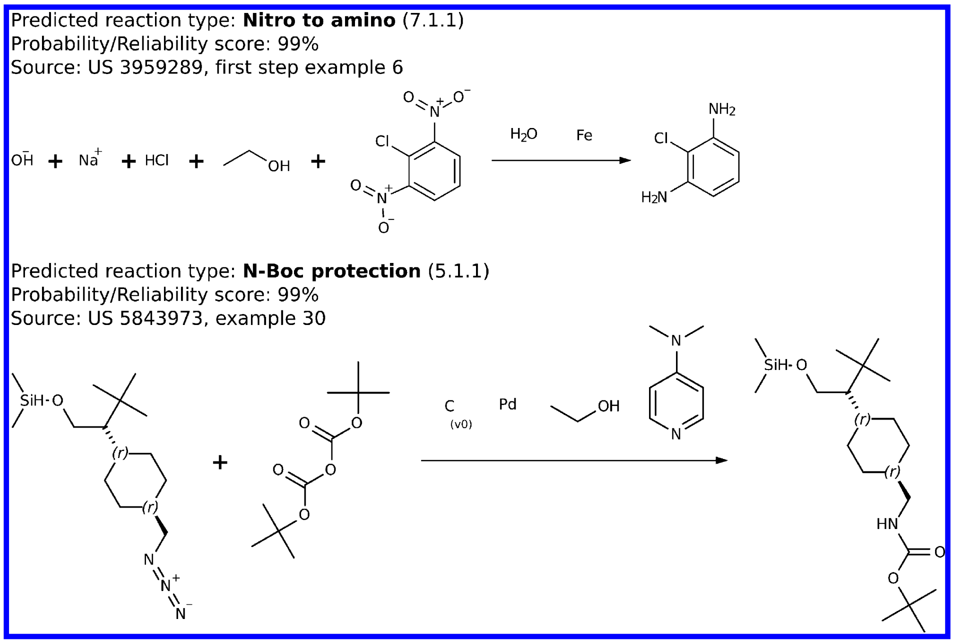


Predicted reaction type: **Chlorination** (10.1.2)
Probability/Reliability score: 98%
Source: US 6255314, Process N

Predicted reaction type: **Methoxy to hydroxy** (6.3.7)
Probability/Reliability score: 99%
Source: US 6252084

Figure 7. Two examples of unclassified reactions from the patent data. Top: Chlorination reaction. Bottom: Methoxy to hydroxy reaction. Both reactions were correctly classified using the reaction type prediction model.

resulted in a representative subset of 226 different reactions. More than two-thirds of these reactions (69%) could actually be classified correctly. Some of these reactions are complicated multi-step reactions or "multi-site" reactions, for example, double deprotections. However, using the classification model, some bits in the fingerprint seem to be sufficient to accurately predict at least one step of the reaction. Other reaction SMILES we had found in this set showed some errors, for example, split reactants, product with missing or interchanged substituents, or some special or even missing agents. These errors could be attributed to the contents of the patents themselves or to text-mining problems. This analysis proves our model to be very robust against small errors in the data. In some cases, the reaction is classified as a related reaction because the true reaction type was not contained in our model, for example, Iodo−Suzuki coupling was classified as Bromo−Suzuki coupling. Most of these incorrectly classified reactions could be explained by the functional groups that undergo a particular transformation and were captured in our fingerprint. In the following, we show three examples of recovered reactions from the unclassified data in more detail. All classifications of these 226 reactions with reaction SMILES and annotation can be



Predicted reaction type: **Nitro to amino** (7.1.1)
Probability/Reliability score: 99%
Source: US 3959289, first step example 6

Predicted reaction type: **N-Boc protection** (5.1.1)
Probability/Reliability score: 99%
Source: US 5843973, example 30

Figure 8. Two examples of unclassified reactions from the patent data. Top: Nitro to amine multi-site reaction. Bottom: Orthogonal protection reaction and azide deprotection to N-Boc protection. Both reactions were correctly classified using the reaction type prediction model.

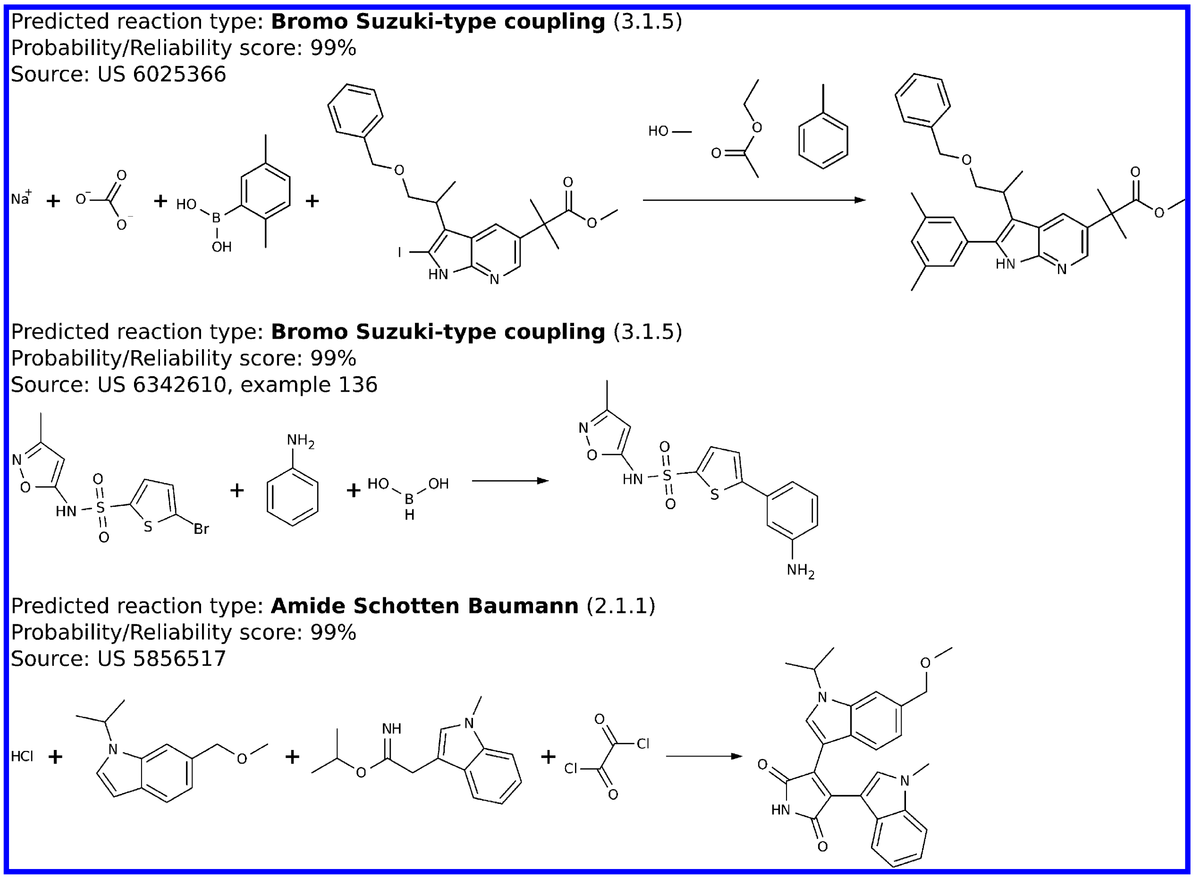
found in the Supporting Information. Figure 7 shows two examples of correctly classified reactions, a chlorination and a methoxy to hydroxyl reaction. In both reactions, the reactants as well as the products are correctly text-mined from the patents. We are investigating why these reactions were not classified by NameRxn.

Figure 8 illustrates two other examples of reactions that could not be classified using NameRxn. Using our model, we correctly identified the nitro to amino reaction, which is a multi-site reaction with two nitro groups being modified to two amino groups at the phenyl ring. The second reaction in this example is a 2-step reaction, an orthogonal protection, which first includes an azide deprotection followed by a N-Boc protection. We correctly classified the N-Boc protection, which is the final step of this orthogonal protection.

The last examples (Figure 9) reveal different problematic aspects of either our model, the text-mining, or the reaction itself. Figure 9 (top) shows a Iodo−Suzuki-type coupling predicted as Bromo−Suzuki-type coupling. This misclassification was caused by the limited number of reaction types we had included in the model. Hence, this reaction was predicted as a related reaction type. Additionally, this reaction exhibited the problem of interchanged methyl substituents at the 2,5-dimethylphenylboronic acid, which resulted in m-xylene in the product. This type of error can be attributed to either the text-mining of the patents or the content of the patents themselves. Figure 9 (middle) illustrates a Bromo−Suzuki-type coupling that was correctly classified despite the fact that one of the reactants, the boronic acid, was disrupted. This could be triggered by the notation in the patent where an unusual spelling of 3-aminophenylboronic acid was used (US 6342610 "3-amino benzene boronic acid") and led to an error in the text-mining. Figure 10 (bottom) shows a multi-step ring-forming reaction that was incorrectly classified as an Amide−Schotten−Baumann reaction. Here, the reaction itself is very specific and rare, complicating the assignment to a certain reaction type.

In this section, we used the probability vector of the LR classifier for the 50 reaction types as a kind of reliability score and combined it with hierarchical prediction of reaction class and superclass as a surrogate measure of the applicability domain of our model. In several examples, we showed that this procedure was successful in the prediction of the correct reaction type. One could argue that predicting one aspect of a multi-step reaction or recovering erroneous reactions is not useful (or is impractical). However, we believe that it shows the robustness of the model; these "almost right" predictions could be used as a hint for a chemist who immediately will identify the correct reaction. Nevertheless, the correct prediction of all steps of a multi-step reaction is definitely outside the domain of our single-step reaction-type prediction model.

Clustering of Reaction Fingerprints. To evaluate the applicability of our new reaction fingerprints (AP3 difference reaction fingerprint) for similarity-based analyses, we clustered the 10,000 fingerprints of our training data set using the Butina clustering algorithm[39] implemented in RDKit. As a similarity measure, the Dice[40] coefficient was used with a threshold of 0.5 for similar reaction fingerprints. The threshold was chosen after calculating the Dice similarity for random pairs of reaction fingerprints, which resulted in a mean similarity of 0.08. The clustering produced 655 clusters of which 115 clusters had more than 10 members, covering 8702 reaction fingerprints (87% of the data). For each of these, we evaluated the homogeneity of the cluster considering the reaction type, class, and superclass. The purity of the cluster is calculated by determining the main reaction type, class, and superclass and normalizing this quantity by the size of the cluster. On the basis of this, 48 reaction types out of 50 could be recovered by clustering the reaction fingerprints. The two missing reaction types, "N-Cbz deprotection" (6.1.3) and "Bromo−Suzuki-type coupling" (3.1.5), were contained in other clusters. All 200 instances of the N-Cbz deprotection type were found in the "O-Bn deprotection" (6.3.1) cluster. This cluster, composed of 644 members, contains all reaction instances of both of these two types, 89% of the reaction type "N-Bn deprotection" (6.1.5), and also one-third of the "N-Boc deprotection" (6.1.1.) reaction type. For the other missing reaction type, Bromo−


Predicted reaction type: **Bromo Suzuki-type coupling** (3.1.5)
Probability/Reliability score: 99%
Source: US 6025366

Predicted reaction type: **Bromo Suzuki-type coupling** (3.1.5)
Probability/Reliability score: 99%
Source: US 6342610, example 136

Predicted reaction type: **Amide Schotten Baumann** (2.1.1)
Probability/Reliability score: 99%
Source: US 5856517

Figure 9. Three examples of unclassified reactions from the patent data. Top: Iodo−Suzuki-type coupling predicted as Bromo−Suzuki-type coupling. Middle: Correctly classified Bromo−Suzuki-type coupling, the boronic acid is disrupted. Bottom: Multi-step ring-forming reaction incorrectly classified as an Amide−Schotten−Baumann reaction.

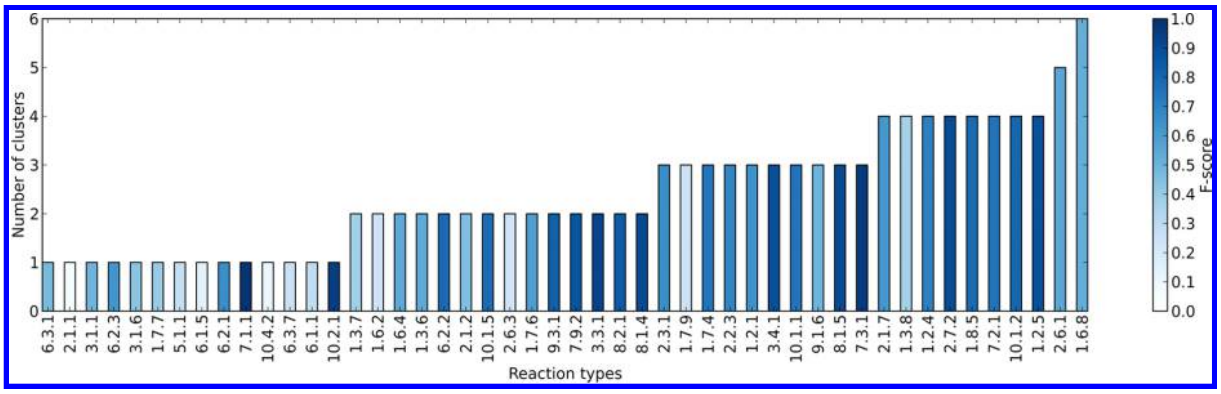


Figure 10. Results reaction fingerprint clustering. Clusters evaluated by reaction type. Bars are sorted by the number of the cluster the reaction type was split in. Bars are colored by the mean F-score of the clustering per reaction type.

Suzuki-type coupling, two-thirds were assigned to the "Bromo−Suzuki coupling" (3.1.1) cluster, and another 25% could be found in the "Chloro−Suzuki-type coupling" (3.1.6) cluster. Some of the 48 reactions types could be reassembled from various clusters (Figure 11), for example, "Ketone reductive amination" (1.2.5), which was split into four clusters with a mean purity of 100%. However, of the 200 reaction instances, only 159 were recovered. On the contrary, the "Nitro to amino" reaction type (7.1.1) was found in one cluster with a purity of 99% and 197 members, implying that these reaction fingerprints cover the inherent chemical transformation of this reaction type. The mean purity per reaction type can be regarded as a kind of precision value. We were also interested in the recall, the number of reactions per reaction type that could be actually recovered within the clusters. Hence, we calculated the recall by the number of recovered reactions divided by number of reactions of that type contained in the data. Using the recall and precision, we were able to calculate the F-score of the clustering for each reaction type. This resulted in a mean F-score of 0.6 and a median F-score of 0.63 (min = 0.06, max = 0.99). In general, we found for many of the reaction types a good precision (mean precision = 0.78, min = 0.31, max = 1.0); however, the recall for some of the reaction types was quite low (mean recall = 0.58, min = 0.03, max = 1.0). Figure 10 shows the results of the clustering for the different reaction types.

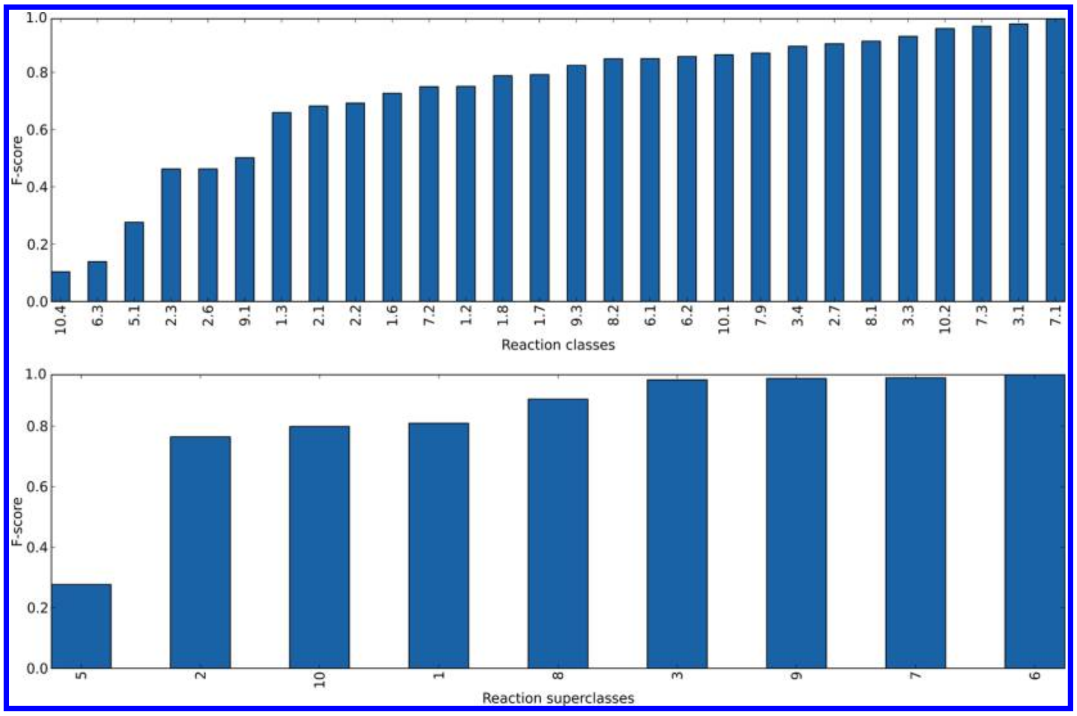


Figure 11. F-score results of reaction fingerprint clustering for reaction class and superclasses. Bars were sorted according to the mean F-score of class/superclass. Top: Clusters evaluated by reaction class. Bottom: Clusters evaluated by reaction superclass.

The two clusters with the lowest F-score were "Amide−Schotten−Baumann" (2.1.1, F-score = 0.06) and "Methylation" (10.4.2, F-score = 0.1). Almost half of the Amide−Schotten−Baumann reaction instances were assigned to the "Isocyanate amine reaction" (2.3.1) cluster, while the rest were spread over different clusters. Although these two reaction types are not related by mechanism, this probably arises because their reaction fingerprints both contain amide or amide-like bits. The majority, 68%, of the methylation instances were found as singletons or in small clusters (less than 10 members). This is caused by the broad spectrum of chemical variation for this reaction type.

We have also calculated the F-score for the reaction classes and superclasses (Figure 11). We obtained a mean F-score of 0.73 and a median F-score of 0.81 for the reaction classes (min = 0.1, max = 0.99). Here also, the precision was much higher (mean precision = 0.89, min = 0.66, max = 1.0) than the recall (mean recall = 0.67, min = 0.06, max = 1.0). This observation was also confirmed in the reaction superclasses, achieving a mean F-score of 0.82 (mean precision = 0.93, mean recall = 0.76).

These results demonstrate that intratype similarities vary widely between types. At one end of the spectrum is the methylation reaction type, where a broad chemical variety is observed. At the other end, one finds examples like the hydroxy to methoxy (1.7.4) transformation and the methyl esterification (1.7.6) or the Fischer−Speier esterification (2.6.3)—all of which involve adding a carbon atom to an OH, where both intra- and intertype variability are low.

■ CONCLUSIONS

In this study, we presented the development of a novel reaction transformation fingerprint that can be used for model building as well as similarity search. We showed that applying the transformation fingerprint in a multi-class prediction model allows classification of reactions. Furthermore, the results indicate that the inclusion of agents provides a better discrimination of related reaction classes. However, agents have to be treated separately and cannot be directly integrated in the difference reaction fingerprint. Finally, creating a rather simple classification model (LR + 265 bit fingerprint) demonstrates the robustness of our transformation and agent feature fingerprint combination. Evaluating the model on different external data sets from patents as well as in-house reactions resulted in excellent performance of predicting the different reaction types. We also applied the model to recover a set of unclassified patent reactions. Despite some of these reactions apparently containing errors or encoding complicated multi-step reactions, our classification model was able to determine the correct reaction type in many cases or at least helped to identify one step of the multi-step reaction. To assess the usability of the novel transformation fingerprint in similarity searching and related tasks we conducted a clustering experiment. This showed that the fingerprint was able to re-extract 48 out of our 50 reaction types, proving its applicability for searching similar reactions. In a separate exercise, not described here, we have used this approach to find clusters of similar reactions within the unclassified reactions of the patent data allowing us to detect new or not yet registered reaction types. One drawback of the current approach to generate the fingerprint is its dependence on the atom mapping to identify the agents or more generally on the correct annotation of agents and nonagents. Improper handling of agents has been shown to introduce a reasonable amount of noise into the fingerprint.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Further figures and results of this study are summarized in the PDF. An Excel sheet contains the 226 unclassified patent reactions that were manually checked. Training and tests sets constructed from the patent data set are in zipped Python pickle files. Also inlcuded are all IPython notebooks to reproduce the results shown in the paper. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: gregory.landrum@novartis.com.

Notes

The authors declare the following competing financial interest(s): R.S. and D.L. are employees of NextMove Software that markets the NameRxn tool used in this contribution.

■ ACKNOWLEDGMENTS

N.S. thanks the Novartis Institutes for BioMedical Research education office for a Presidential Postdoctoral Fellowship. The authors thank Nikolas Fechner for his help with scikit-learn and many insightful discussions on machine-learning and fingerprints. The authors also wanted to thank Mike Tarselli for his help with manually classifying the set of unclassified patent data.

■ ABBREVIATIONS

AP atom pairs; ELN electronic laboratory notebook; FP fingerprint; kNN k-neareast neighbors; LR logistic regression; ML machine learning; NB naïve Bayes; RF random forest; TT topological torsions

■ REFERENCES

(1) Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 348−390.

(2) Warr, W. A.; Short, A. Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Informatics* **2014**, *33* (6−7), 469−476.

(3) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **2013**, *53* (11), 2884−2895.

(4) RSC's RXNO Ontology: http://www.rsc.org/ontologies/RXNO/index.asp (accessed January 2015).

(5) Weygand, C. *Organische-Chemische Experimentierkunst*; Barth: Leipzig, Germany, 1938.

(6) *Theilheimer's Synthetic Methods of Organic Chemistry*; Tozer-Hotchkiss, G., Ed.; Karger: Basel, Switzerland.

(7) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3* (6), 560−593.

(8) Gelernter, H.; Rose, J. R.; Chen, C. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (4), 492−504.

(9) Sello, G.; Termini, M. Classification of organic reactions using similarity. *Tetrahedron* **1997**, *53* (41), 14085−14106.

(10) Grethe, G. Analysis of Reaction Information. In *Handbook of Chemoinformatics*; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 1407−1427.

(11) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* **2012**, *52* (7), 1745−1756.

(12) Daylight Reaction Fingerprint. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed October 17, 2014).

(13) Broughton, H. B.; Hunt, P. A.; MacKey, M. D. Methods for Classifying and Searching Chemical Reactions. U.S. Patent 2003/0182094 A1, 2003.

(14) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem.* **2008**, *3*, 821−832.

(15) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de NovoDesign Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163−1184.

(16) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180−192.

(17) Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Vol. 4; American Chemical Society: Washington, DC, 2008.

(18) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535.

(19) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(20) Blake, J. E.; Dana, R. C. CASREACT: More than a million reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (4), 394−399.

(21) Reaxys Database. http://www.elsevier.com/online-tools/reaxys (accessed October 17, 2014).

(22) SPRESI Database. http://infochem.de/products/databases/spresi.shtml (accessed October 17, 2014).

(23) ChemSpider SyntheticPages Database. https://cssp.chemspider.com/ (accessed October 17, 2014).

(24) Webreactions Database. http://www.openmolecules.org/webreactions/index.html (accessed October 17, 2014).

(25) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Ph.D. Thesis. University of Cambridge: Cambridge, U.K., 2012.

(26) Patent Data: http://nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/,https://bitbucket.org/dan2097/patent-reaction-extraction/downloads (accessed on October 17, 2014).

(27) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the Reactions Used for the Preparation of Drug Candidate Molecules. *Org. Biomol. Chem.* **2006**, *4*, 2337−2347.

(28) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451−3479.

(29) NextMove Software. www.nextmovesoftware.com (accessed October 17, 2014).

(30) Indigo Software. http://ggasoftware.com/opensource/indigo/ (accessed October 17, 2014).

(31) RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed January 2015).

(32) 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.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(33) Pérez, F.; Granger, B. E. *IPython: A System for Interactive Scientific Computing, Comp. Sci. Eng.* **2007**, *9* (3), 21−29. URL: http://ipython.org (accessed January 2015).

(34) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.

(35) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs As Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(36) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(37) Nilakantan, R.; Baumann, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR

Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(38) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5−32.

(39) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.

(40) Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297−302.