JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 4167−4187

pubs.acs.org/jcim

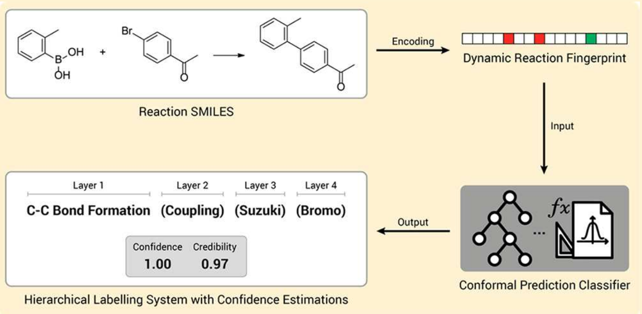
Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature

Gian Marco Ghiandoni,[†] Michael J. Bodkin,[‡] Beining Chen,[§] Dimitar Hristozov,[‡] James E. A. Wallace,[‡] James Webster,[†] and Valerie J. Gillet[*,†]

[†]Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom
[‡]Evotec (U.K.) Ltd., 114 Innovation Drive, Milton Park, Abingdon OX14 4RZ, United Kingdom
[§]Chemistry Department, University of Sheffield, Dainton Building, Brook Hill, Sheffield S3 7HF, United Kingdom

Supporting Information

ABSTRACT: Reaction classification has often been considered an important task for many different applications, and has traditionally been accomplished using hand-coded rule-based approaches. However, the availability of large collections of reactions enables data-driven approaches to be developed. We present the development and validation of a 336-class machine learning-based classification model integrated within a Conformal Prediction (CP) framework to associate reaction class predictions with confidence estimations. We also propose a data-driven approach for "dynamic" reaction fingerprinting to maximize the effectiveness of reaction encoding, as well as developing a novel reaction classification system that organizes labels into four hierarchical levels (SHREC: Sheffield Hierarchical REaction Classification). We show that the performance of the CP augmented model can be improved by defining confidence thresholds to detect predictions that are less likely to be false. For example, the external validation of the model reports 95% of predictions as correct by filtering out less than 15% of the uncertain classifications. The application of the model is demonstrated by classifying two reaction data sets: one extracted from an industrial ELN and the other from the medicinal chemistry literature. We show how confidence estimations and class compositions across different levels of information can be used to gain immediate insights on the nature of reaction collections and hidden relationships between reaction classes.



■ INTRODUCTION

Reaction classification has been a topic of considerable interest for many years.[1] Applications range from efficient indexing of reactions in databases and the management of search output, through to knowledge discovery. In particular, there is a growing interest in applying artificial intelligence (AI) techniques to tasks such as reaction prediction, synthesis planning, and de novo design.[2] Until recently, the availability of large collections of reactions has been restricted to proprietary electronic lab notebooks (ELNs) and commercially available databases such as Reaxys, CASReact, and SPRESI, limiting the possibilities for data mining of reactions. However, large collections of reactions are now available in the public domain, following the work of Lowe and colleagues on the automated extraction of reactions from the U.S. Patent literature. These data are now being used in a number of new approaches to retrosynthesis,[3−5] reaction prediction,[6,7] and reaction classification.[8,9]

Historically reactions were named according to the type of product generated, the functional group or reagent used, or the inventor of the reaction.[1] More systematic approaches to reaction-classification can be divided into model-based and data-driven approaches. The model-based methods are based on predetermined definitions of the reaction center. A variety of models have been generated that vary in the amount of detail that is encoded. For example, Dugundji and Ugi developed the BE-matrices, which describe reactions according to changes in bonds and nonbonded valence electrons.[10] Model-based classification methods do not consider changes beyond the reaction center or subclass.

In data-driven approaches, the classification is generated automatically by the analysis of sets of reactions. For example, InfoChem's CLASSIFY algorithm is an example of a data-driven approach. It makes use of a reaction mapping algorithm, which identifies the reaction center as the atoms and bonds that change in a reaction.[11] The level of specificity can be varied by extending the reaction center in spheres one or two bonds away from the reaction center, the more extended descriptions leading to more specific descriptions. Hashcodes

Received: July 1, 2019
Published: September 17, 2019

DOI: 10.1021/acs.jcim.9b00537

are then calculated for each level of description of the reaction center to produce reaction classification codes. A drawback of this approach is the large number of codes that is generated. A similar approach has been adopted by Christ and co-workers in their analysis of the content of an Electronic Lab Notebook (ELN).[12]

Broughton et al.[13] introduced reaction vectors as a generalization of the earlier Daylight difference fingerprint, whereby a reaction is described by the difference between the Daylight fingerprint of the product molecules and the reactant molecules, assuming a stoichiometric reaction. Different types of descriptors were explored in the reaction vector approach including atom pairs and topological torsions and their relative performance compared in classification tasks. Ridder and Wagener[14] reported a system for predicting potential metabolites in biological reactions based on Sybyl atom types and augmented atom types up to one bond away from the reaction center. Later, Hu et al. have used reaction difference fingerprints to assign EC (Enzyme Commission) numbers to biochemical reactions.[15]

With the availability of large collections of reactions, the application of machine learning approaches has become possible where, given a set of classified training examples, an algorithm can be trained to generate a classification model. Thus, Schneider et al. investigated a range of reaction difference fingerprints for data-driven reaction classification based on 50 reaction classes.[8] They reported their best performing fingerprint as AP3 (atom pair 3) fingerprints of the products and reactants together with a feature-based fingerprint to represent the physicochemical properties of the reaction agents (for example, solvents, catalysts, and reactants with less than 20% atoms mapped). Their method was used to classify reactions based on the NameRxn labeling.

Our interest in classifying reactions is 2-fold. First, an effective reaction classification tool can be very informative for exploring existing collections of reactions, whether these have been extracted from the literature or are historical collections such as those contained in ELNs. The organization of reactions into classes allows more effective knowledge exploitation; for example, monitoring the frequency of occurrence of reaction types or difference in yields can be used to inform decision making associated with synthesis planning, by, for example, identifying under-explored reactions or the success or decline of a particular reaction class over time. Second, the organization of reactions into classes can be used to improve de novo design tools, which aim to simulate the behavior of medicinal chemists by directing the design toward particular reaction classes. We have previously reported a reaction-based approach to de novo design in which reaction transforms are automatically extracted from large collections of reactions and stored as reaction vectors. The reaction vectors can then be applied to previously unseen starting materials to generate de novo product molecules.[16] A reaction classification tool that is based on reaction vectors can be exploited in both fully automated and augmented de novo design to drive the designs to areas of greater synthetic interest.

Here, we describe the development of a data-driven reaction classification tool using machine learning. Our approach is broadly similar to that described by Schneider et al. but with some important differences. First, we extend the approach to classify a much larger set of reaction types than the published method. Second, we explore the use of descriptors that are compatible with our de novo design tool. Third, we employ conformal prediction methods to filter out classifications for which the model has low reliability. We demonstrate the application of the model by comparing the composition of two data sets: reactions extracted from the Evotec corporate ELN and a set of reactions extracted from the medicinal chemistry literature. The use of the classification tool to improve the effectiveness of de novo design will be described in a future publication.

DATA

A large collection of chemical reactions data has been extracted from United States patents and made publicly available. Two data sets were selected for this study: USPD Grants 1976–2016 (referred to here as USPD Grants); and USPD Applications 2001–2016 (referred to here as USPD Apps).[17] USPD Grants and USPD Apps represent reactions extracted from granted patents and patent applications, respectively, and were both released in June 2017. The characteristics of the reactions contained in an earlier version of USPD Grants (i.e., 1976–2015) have already been reported.[18] The USPD Grants contains approximately 1.8 million reactions, and the USPD Apps contains approximately 1.9 million reactions. Classification data generated using NameRxn software (2.0)[19] were obtained from NextMove for the two data sets. NameRxn is a rule-based approach to reaction classification and adopts a nomenclature inspired by the RXNO Ontology developed by the Royal Society of Chemistry[20] and earlier classification system proposals.[21,22] Reactions are named using general descriptions (i.e., such as O-substitution) and specific classes (i.e., such as 1,2-benzoxazole synthesis). NameRxn assigns a reaction to one of over 700 distinct reaction types. These are given a position in a derivative of the hierarchy first published by Carey et al.[21] and later refined by Roughley et al.[22] These positions correspond to either named reactions (e.g., Wittig olefination) or, where the reaction does not have a trivial name, a description of the reaction (e.g., piperidine synthesis). The hierarchy consists of three levels: 11 major reaction classes, 80 reaction subclasses, and more than 700 reaction types. For example, bromo Suzuki coupling is classified as 3.1.1, where 3.1 is any Suzuki coupling and 3 is C–C bond formation. Reaction types are also assigned an identifier in the RXNO reaction ontology. Not all of the reactions were successfully classified, and only those reactions for which classification labels were available were used here.

METHODS

Data Preprocessing. As stated in the Introduction, as well as generating a reaction classification tool that could be used in its own right, which is the subject of this Article, our longer-term aim is to also use the tool to improve the effectiveness of de novo design. It is therefore important that the methods developed are compatible with our de novo design tool, which is based on reaction vectors.[23] The reaction vectors required for effective de novo design should be derived from stoichiometric or balanced reactions, that is, reactions that have the same number of atoms in the reactants as in the products, so that the reaction center is accurately described. Given that reaction data are typically messy, for example, components such as catalysts and reagents may, or may not, be included in the reaction and some components such as byproduct molecules may be missing, it was necessary to "clean" the reactions prior to model building.

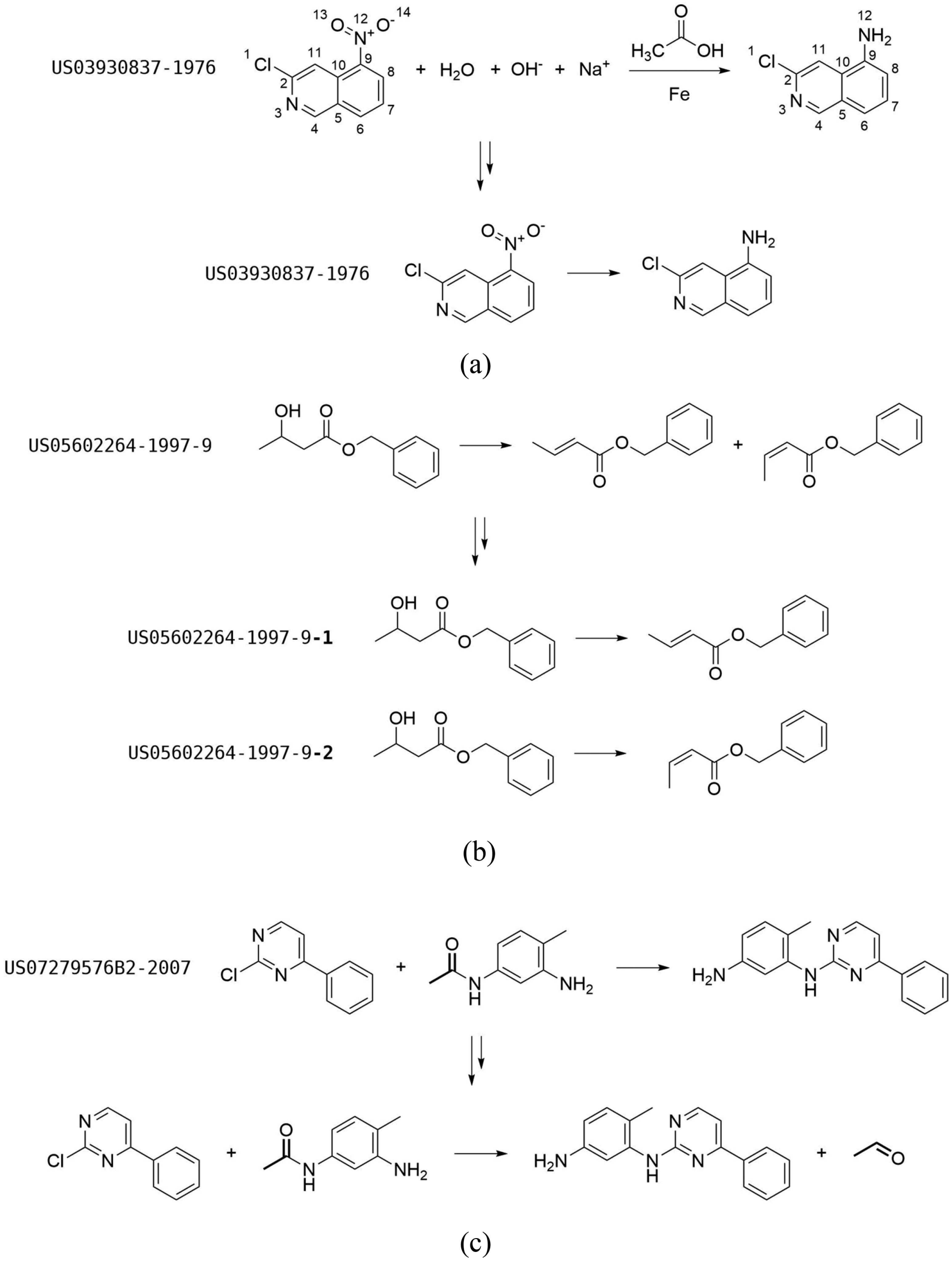


Figure 1. Examples of reaction preprocessing. (a) The reaction is mapped and unmapped components are removed; (b) reactions including multiple products as isomers as separated into distinct reactions; and (c) reaction balancing is used to insert missing components.

Our reaction cleaning workflow consists of a number of steps. First, reaction mapping is carried out using the Indigo Reaction Automapper node in the Indigo Toolkit in KNIME.[24] Note that the USPD data sets already contain mappings generated by Indigo; we recreate them here for convenience only so that we have a single workflow that can be used on data sets regardless of whether or not atom mappings already exist. Also note that the mappings are only used to identify the components that change during a reaction so that, although more recent atom mapping methods have been published, we do not believe that these would impact on the results.[25] Next, components that do not contain any mapped atoms and that are therefore not involved in the reaction center are removed, for example, solvents, catalysts, etc. Multiple reactions that have been assigned the same index are then separated. The final steps attempt to balance the reactions as described by Patel et al.[23] and consist of two procedures. The first is to separate reactions that are unbalanced due to the production of different isomers into separate reaction entries. The second is to handle reactions that are not fully described due to missing components by adding the relevant component(s) to the reaction. Examples of these two categories of reactions are shown in Figure 1. Reactions that cannot be balanced at this stage are rejected, as are reactions with more than three reactants and/or products. The effect on the sizes of the data sets following the data preprocessing is shown in Table 1. Similar processing of the USPD Grants data set was performed by Watson et al.[3] in their use of the data for retrosynthesis.

Table 1. Effect on Data Set Sizes Shown for Different Steps of the Reaction Cleaning[a]

		reactions	categories	classes
USPD Grants	original data	1808937	64	753
	classified reactions	1215355	64	753
	six reactants/products filtering	1149212	64	751
	reaction balancing	1114953	64	735
USPD Apps	original data	1939253	65	749
	classified reactions	1374294	64	748
	six reactants/products filtering	1298809	64	745
	reaction balancing	1263602	64	727

[a] The number of distinct reaction classes and categories represented is also reported.

Following the application of reaction cleaning, reaction vectors were calculated for the remaining reactions. Reaction vectors have been described previously and are based on atom pair descriptors.[26] A reaction vector is calculated as follows: a reactant vector is calculated by summing the atom pairs for the reactants; a product vector is calculated by summing the atom pairs for the products; and the reaction vector is then calculated by subtracting the reactant vector from the product vector. Our atom pairs have the following general form:

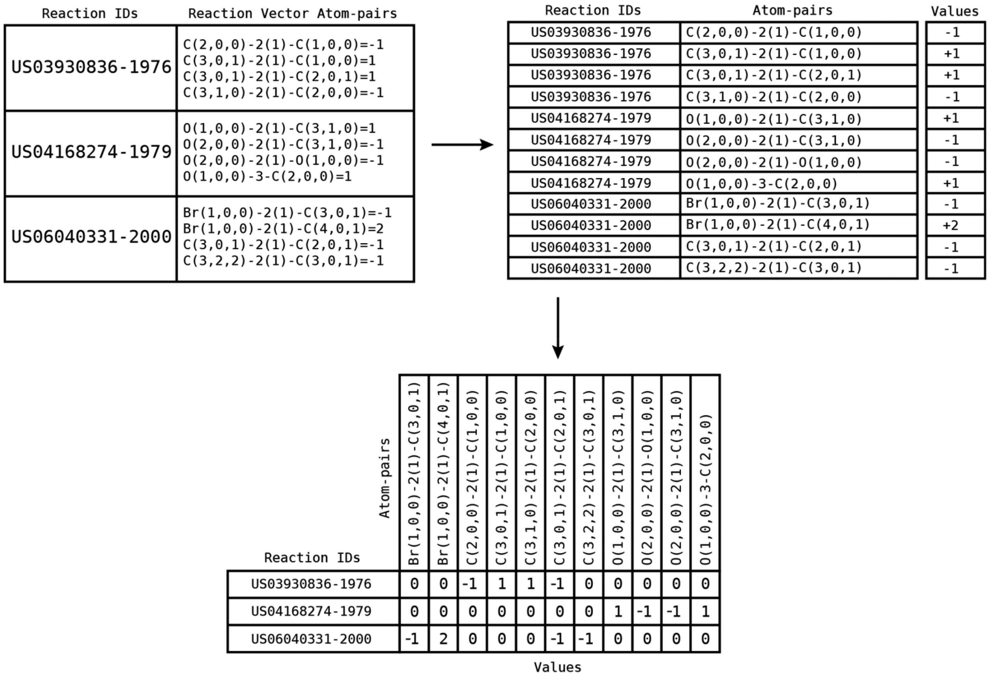
$$\text{AP2: } X1(h,\, p,\, r) - 2(BO) - X2(h,\, p,\, r);$$

$$\text{AP3: } X1(h,\, p,\, r) - 3 - X2(h,\, p,\, r)$$

where $X1$ and $X2$ are the element symbols of the two atoms; h is the number of non-hydrogen connections; p is the number of π bonds incident on the atom; r is the number of rings of which the atom is a member; and BO is the bond order (1 = single bond; 2 = double bond; 3 = triple bond; and 4 = aromatic bond). AP2 atom pairs describe a pair of connected atoms; AP3 atom pairs describe pairs of atoms separated by two bonds; and the reaction vector consists of both sets of atom pairs, that is, AP2+AP3.

Reaction vectors typically consist of a small number of atom pairs, and in their raw form they are represented as lists of strings with an associated count. The number of atom pairs in a reaction vector is variable and depends on the size of the reaction center, and the count can be negative or positive depending on whether the atom pair is lost from the reactants or gained in the products, respectively. For model building, the reaction vector strings were converted into real vectors by first separating the atom pair strings from their values, pivoting the atom pairs into columns, and filling the corresponding cells with the values. The columns were then sorted alphabetically for canonicalization purposes. A set of reaction vectors for a collection of reactions is generated by repeating this process for each reaction vector to build a matrix where each row corresponds to a reaction vector, and each column represents a specific atom pair. New columns are appended to the matrix as new atom pairs are encountered. If an atom pair does not occur in a given reaction vector, the corresponding cell is filled with a zero value. A simplified schematic of the fingerprint generation process is shown in Figure 2.

The reaction vector is referred to as dynamic because the number of columns is data set dependent and represents the minimum number of atom pairs necessary to describe the reactions within it. This means that different data sets will return different numbers and types of atom pairs and will not, therefore, be directly comparable. For supervised machine learning applications, all of the data (including training and external data to which the model is applied) must be represented by vectors consisting of the same number, type, and order of atom pairs. This is achieved by using the training data atom pairs as reference and making adjustments to the



Reaction IDs	Reaction Vector Atom-pairs
US03930836-1976	C(2,0,0)-2(1)-C(1,0,0)=-1 C(3,0,1)-2(1)-C(1,0,0)=1 C(3,0,1)-2(1)-C(2,0,1)=1 C(3,1,0)-2(1)-C(2,0,0)=-1
US04168274-1979	O(1,0,0)-2(1)-C(3,1,0)=1 O(2,0,0)-2(1)-C(3,1,0)=-1 O(2,0,0)-2(1)-O(1,0,0)=-1 O(1,0,0)-3-C(2,0,0)=1
US06040331-2000	Br(1,0,0)-2(1)-C(3,0,1)=-1 Br(1,0,0)-2(1)-C(4,0,1)=2 C(3,0,1)-2(1)-C(2,0,1)=-1 C(3,2,2)-2(1)-C(3,0,1)=-1

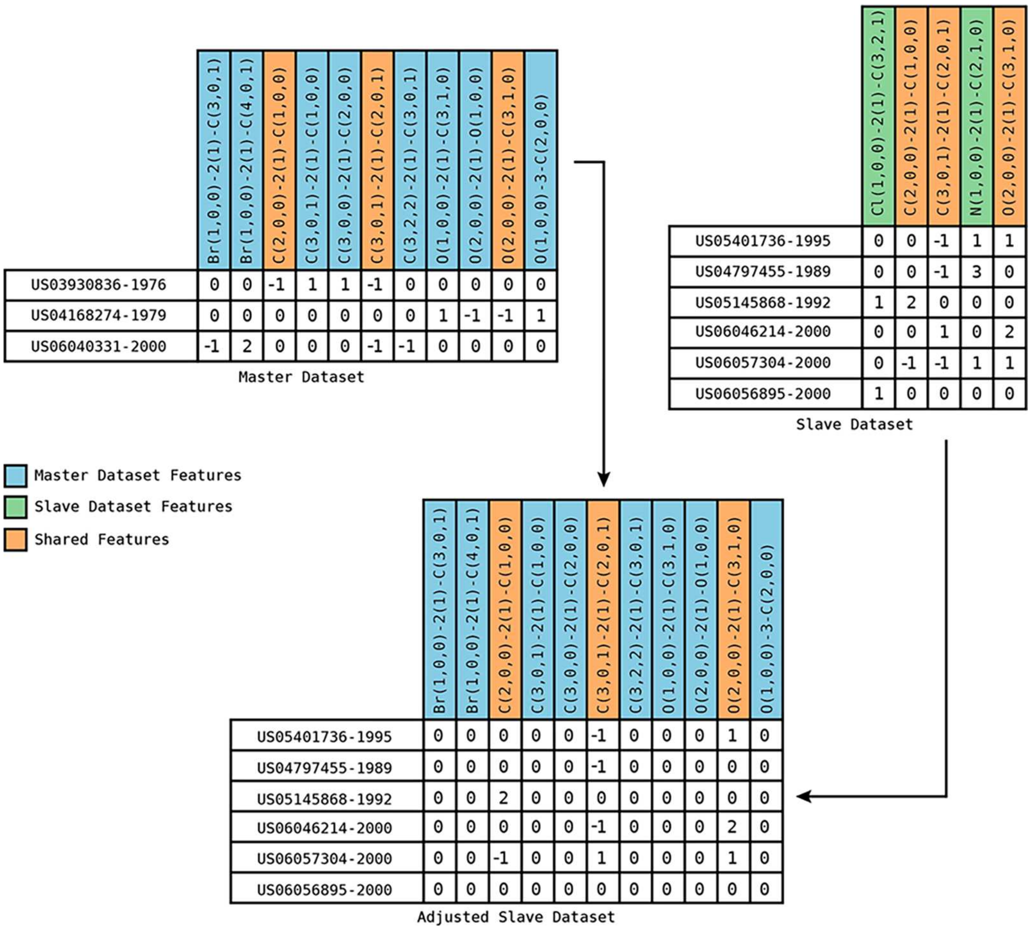
Reaction IDs	Atom-pairs	Values
US03930836-1976	C(2,0,0)-2(1)-C(1,0,0)	-1
US03930836-1976	C(3,0,1)-2(1)-C(1,0,0)	+1
US03930836-1976	C(3,0,1)-2(1)-C(2,0,1)	+1
US03930836-1976	C(3,1,0)-2(1)-C(2,0,0)	-1
US04168274-1979	O(1,0,0)-2(1)-C(3,1,0)	+1
US04168274-1979	O(2,0,0)-2(1)-C(3,1,0)	-1
US04168274-1979	O(2,0,0)-2(1)-O(1,0,0)	-1
US04168274-1979	O(1,0,0)-3-C(2,0,0)	+1
US06040331-2000	Br(1,0,0)-2(1)-C(3,0,1)	-1
US06040331-2000	Br(1,0,0)-2(1)-C(4,0,1)	+2
US06040331-2000	C(3,0,1)-2(1)-C(2,0,1)	-1
US06040331-2000	C(3,2,2)-2(1)-C(3,0,1)	-1

Reaction IDs \ Atom-pairs	Br(1,0,0)-2(1)-C(3,0,1)	Br(1,0,0)-2(1)-C(4,0,1)	C(2,0,0)-2(1)-C(1,0,0)	C(3,0,1)-2(1)-C(1,0,0)	C(3,1,0)-2(1)-C(2,0,0)	C(3,0,1)-2(1)-C(2,0,1)	C(3,2,2)-2(1)-C(3,0,1)	O(1,0,0)-2(1)-C(3,1,0)	O(2,0,0)-2(1)-O(1,0,0)	O(2,0,0)-2(1)-C(3,1,0)	O(1,0,0)-3-C(2,0,0)
US03930836-1976	0	0	-1	1	1	-1	0	0	0	0	0
US04168274-1979	0	0	0	0	0	0	0	1	-1	-1	1
US06040331-2000	-1	2	0	0	0	-1	-1	0	0	0	0

Values

Figure 2. Reaction vectors represented as strings are converted to true vectors. The vector elements are integers with negative values indicating atom pairs that are lost from the reactants; positive values indicating atom pairs that are gained in the products; and zeros indicating atom pairs that are not present in the vector.

DOI: 10.1021/acs.jcim.9b00537



	Br(1,0,0)-2(1)-C(3,0,1)	Br(1,0,0)-2(1)-C(4,0,1)	C(2,0,0)-2(1)-C(1,0,0)	C(3,0,1)-2(1)-C(1,0,0)	C(3,0,0)-2(1)-C(2,0,0)	C(3,0,1)-2(1)-C(2,0,1)	C(3,2,2)-2(1)-C(3,0,1)	O(1,0,0)-2(1)-C(3,1,0)	O(2,0,0)-2(1)-O(1,0,0)	O(2,0,0)-2(1)-C(3,1,0)	O(1,0,0)-3-C(2,0,0)
US03930836-1976	0	0	-1	1	1	-1	0	0	0	0	0
US04168274-1979	0	0	0	0	0	0	0	1	-1	-1	1
US06040331-2000	-1	2	0	0	0	-1	-1	0	0	0	0

Master Dataset

	Cl(1,0,0)-2(1)-C(3,2,1)	C(2,0,0)-2(1)-C(1,0,0)	C(3,0,1)-2(1)-C(2,0,1)	N(1,0,0)-2(1)-C(2,1,0)	O(2,0,0)-2(1)-C(3,1,0)
US05401736-1995	0	0	-1	1	1
US04797455-1989	0	0	-1	3	0
US05145868-1992	1	2	0	0	0
US06046214-2000	0	0	1	0	2
US06057304-2000	0	-1	-1	1	1
US06056895-2000	1	0	0	0	0

Slave Dataset

🟦 Master Dataset Features

🟩 Slave Dataset Features

🟧 Shared Features

	Br(1,0,0)-2(1)-C(3,0,1)	Br(1,0,0)-2(1)-C(4,0,1)	C(2,0,0)-2(1)-C(1,0,0)	C(3,0,1)-2(1)-C(1,0,0)	C(3,0,0)-2(1)-C(2,0,0)	C(3,0,1)-2(1)-C(2,0,1)	C(3,2,2)-2(1)-C(3,0,1)	O(1,0,0)-2(1)-C(3,1,0)	O(2,0,0)-2(1)-O(1,0,0)	O(2,0,0)-2(1)-C(3,1,0)	O(1,0,0)-3-C(2,0,0)
US05401736-1995	0	0	0	0	0	-1	0	0	0	1	0
US04797455-1989	0	0	0	0	0	-1	0	0	0	0	0
US05145868-1992	0	0	2	0	0	0	0	0	0	0	0
US06046214-2000	0	0	0	0	0	-1	0	0	0	2	0
US06057304-2000	0	0	-1	0	0	1	0	0	0	1	0
US06056895-2000	0	0	0	0	0	0	0	0	0	0	0

Adjusted Slave Dataset

Figure 3. The "Master" data set is the training data, and the "Slave" data set is the test/external data. The atom pairs that are unique to the Master are shown in blue, those unique to the test data are in green, and those common to both data sets are in orange. The green columns are removed from the test data and blue columns are added. All of the entries in the blue columns are set to zero because these atom pairs are not present in the test data.

test/external data. Atom pairs present in the test data but not in the training data are removed because they are not accounted for by the model; and training data atom pairs not included in the test data are simply added as columns to the test data, and all of the new cells are filled with zeros. An example of reaction vector data set adjustment is shown in Figure 3 where the reference and processed data sets are indicated as "Master" and "Slave" data sets, respectively.

Model Training. The USPD Grants data set was used for model training and internal validation, and the USPD Apps data set was used for external validation.

50-Class Models. We initially developed a 50-class model to determine the best choice of input descriptors and machine learning method. This also allowed us to compare results with Schneider et al. because their published work is based on a 50-class model.[8] Although we had previously established that AP2+AP3 reaction vectors were most effective for de novo design, it is not necessarily the case that these descriptors will be most effective for reaction classification. Therefore, we investigated the effectiveness of different versions of the atom pair descriptors. We considered AP2, AP3, and AP4 atom pair descriptors (AP4 atom pairs describe atom pairs separated by three bonds) to examine the effect of increasing the environment encoded along with the reaction center itself (AP2). We also used the combined AP2+AP3 descriptors, which are the descriptors used by our structure generation method.

For each descriptor type, the clean reaction data were converted to the appropriate descriptors, reaction vectors were calculated, and duplicates were removed. The "no. of atom pairs" column in Table 2 indicates the number of unique atom pairs required to describe the data. As expected, the use of AP2 descriptors alone leads to the greatest reduction in the number of unique reaction vectors because only the immediate reaction center is encoded. There is also a small reduction in the number of reaction classes encoded. AP3 and AP4 descriptors represent atom pairs separated by two and three bonds, respectively, and therefore capture more of the environment of

Table 2. Numbers of Unique Reaction Vectors and the Reaction Classes Covered for Different Types of Atom Pair Descriptors

data set (descriptor)	unique reaction vectors	no. of atom pairs	classes
USPD Grants (AP2)	41726	1592	715
USPD Grants (AP3)	113975	2613	726
USPD Grants (AP4)	112119	2898	726
USPD Grants (AP2+AP3)	115602	4205	727
USPD Apps (AP2+AP3)	110802	4046	718

the reaction center. AP2+AP3 descriptors represent the combination of AP2 and AP3 descriptors. Each of the data sets in Table 2 was preprocessed as follows. First, the 50 most populated reaction classes were retained and then randomly sampled so that they were equally sized according to the smallest class size, illustrated in Figure 4. Down-sampling was used to reduce any bias toward the most populated classes during training. Second, columns containing only zeros were removed.

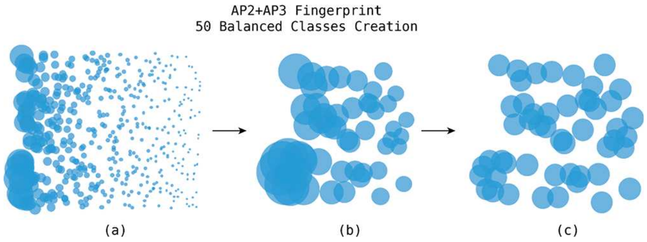


Figure 4. Creation of a balanced collection of 50 reaction classes from the USPD Grants data set: (a) 727 unbalanced classes; (b) selection of 50 most populated classes; and (c) balanced down-sampling according to the minority class.

Table 3 shows the number of reaction vectors and atom pairs for each descriptor-type where it can be seen that the data set contents vary in the number of reaction vectors in each class, and therefore the total number of examples. The data sets also vary in the coverage of reaction classes, as shown in Figure 5. The AP2 fingerprint data set is characterized by a significantly lower number of unique reaction vectors as compared to the other data sets. This is due to the descriptor encoding the reaction center only, so that it is much less discriminating than the descriptors that encode more of the reaction environment, and a much smaller number of unique reaction vectors are produced. Also, the coverage of reaction classes is different from the other descriptors. The extended reaction vector descriptors cover the same number of reaction classes; however, six of the classes represented by these descriptors are omitted for the AP2 descriptors, and replaced by other reaction classes, due to AP2's focusing on the reaction center only. For example, the reaction class "ketone to alcohol reduction", which describes a C=O group reduced to a CH–OH group, is represented by a small number of unique reaction vectors when only the reaction center itself is encoded and it does not appear in the top 50 populated classes.

Table 3. Total Number of Unique Reaction Vectors (Number of Rows in the Input Data) Shown for the Different Types of Reaction Vector Descriptors, along with the Number of Unique Atom Pairs (Number of Columns in the Input Data), and the Number of Unique Reaction Vectors in Each Class, for the Fifty Most Populated Reaction Classes

descriptor	total number of reaction vectors	retained atom pairs	reaction vectors per class
AP2	10000	1167	200
AP3	25650	2103	513
AP4	25500	2292	510
AP2+AP3	25700	3146	514

The four data sets were then used to train and validate the models as follows. Each data set was partitioned into a training set (40%) and a test set (60%) using stratified sampling on the reaction classes to preserve the distribution of examples across the classes. For the AP2+AP3 data set, the training set was arbitrarily fixed at 10 000 reactions (∼40%) to reproduce conditions similar to those reported by Schneider et al.[8] The results of the partitioning process are shown in Table 4. The training sets then formed the input to the classifiers, and the resulting models were used to infer the NameRxn reaction classes for the entries in the corresponding test sets.

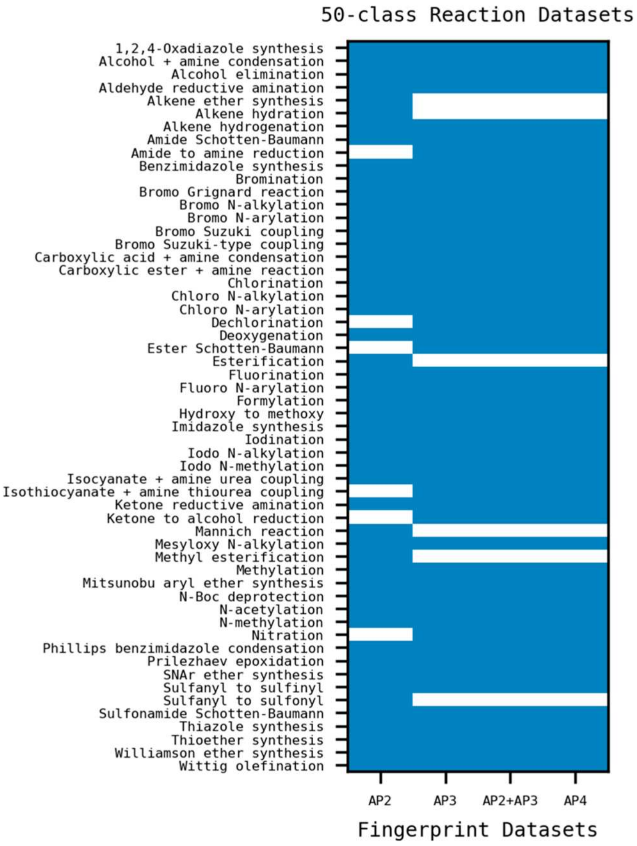


Figure 5. Reaction class representation. Represented classes are represented in blue, and missing classes are in white.

Table 4. Training Set and Test Set Sizes Shown for the Different Types of Reaction Vector Descriptors, along with the Number of Unique Reaction Vectors in Each Class, for the Fifty Most Populated Reaction Classes

USPD Grants	training set	reaction vectors per class	test set	reaction vectors per class
AP2	4000	80	6000	120
AP3	10660	205	14990	308
AP4	10200	204	15300	306
AP2+AP3	10000	200	15700	314

The USPD Apps data were used as an external test set for the models built using AP2+AP3 descriptors and were prepared by retaining only those classes contained in the USPD Grants AP2+AP3 training data set. AP2+AP3 descriptors were calculated, and the atom pairs were adjusted to be compatible with those in the training data. Reaction vectors, which were already described in the USPD Grants data set, then were excluded so that there was no overlap between the external test set and the training set. The final USPD Apps data consisted of 15 193 reaction vectors. Note also that classes were not down sampled due to the low number of examples in the minority classes in the USPD Apps.

Hierarchical Reaction Classification System. Our reaction vector approach is not fully compatible with NameRxn. For example, some reactions that fall into different categories in NameRxn are indistinguishable using reaction vectors, such as reactions that vary according to the reagent used or stereochemical effects. Furthermore, although NameRxn adopts a three-level classification system, it is based on traditional nomenclature, which is not optimal for browsing because subclasses and reaction types are often described by the names of the scientists who discovered the reactions. In addition, NameRxn labels are not optimal for alphabetic sorting because they often contain redundant information. For example, Heck, Negishi, Stille, Suzuki, and Sonogashira couplings are all cross-coupling reactions that involve the formation of a carbon−carbon bond, and, although they are all placed into the major class "C−C bond formation", they are then assigned to separate subclasses such that the relationship between them, that is, that they are all couplings, is lost.

The NameRxn ontology was, therefore, replaced by a novel, manually curated, four-level Hierarchical Reaction Classification System, which we call SHREC (Sheffield hierarchical reaction classification system) and which is more consistent with the reaction vector de novo design framework and which is a true hierarchy. For each NameRxn class, multiple examples of reactions were evaluated to identify the general cores of their transformations and thus produce a new set of reaction classes. Reaction classes that describe transformations that could not be processed using reaction vectors (e.g., stereochemistry inversions, resolutions, etc.) were not considered. The procedure condensed the NameRxn specific classes into 598 new classes. The SHREC System is distributed across four levels ranging from general reaction categories to increasingly more specific subclasses and allows the most specific reaction classes to be merged into more generic categories (e.g., "C−C bond formation (condensation)" and "C−C bond formation (coupling)" can be merged into "C−C bond formation" by moving up a level in the hierarchy and vice versa). The different levels in the hierarchy are shown by the use of parentheses. The hierarchical arrangement enables the classification algorithm to be run once only while allowing the results to be investigated across different levels of generalization according to the selected level. The first level in the hierarchy describes the transformation according to some basic chemistry definitions (e.g., C−C bond formation, functional conversion, protection, etc.); the second level describes the type of the transformation (e.g., coupling), or, in some cases, a specific substrate involved in the reaction (e.g., alcohol to alkene). The third and fourth levels contain additional information on the substrates/products (e.g., isocyanate + amine), reaction inventors (e.g., Suzuki), or functionalities (e.g., Bromo). Examples are given in Table 5 for the C−C bond formation reaction described above.

Note that the four-level hierarchical labeling in SHREC is not exhaustive in terms of nomenclature due to its bias toward the USPD and NameRxn. A table showing the mapping of the original NameRxn labels to the four-level SHREC is shown in the Supporting Information.

336-Class Classification Models. Following validation on 50 reaction classes, the approach was extended to include a much larger range of reaction classes. The cleaned USPD Grants data set was converted to unique AP2+AP3 reaction vectors as before. The reactions were then mapped to the SHREC labeling system. This time, all reaction classes containing at least 30 examples were retained. This resulted in 336 classes being represented with a median of 129.5 examples per class; see Table 6. The data were then partitioned into 40% training and 60% test data using stratified sampling to preserve the distribution of examples across the classes. This resulted in 44 792 unique reaction vectors in the training set with a median number of 52 examples per class; and 67 189 unique reaction vectors in the test set with a median of 77.5 examples per class. Note that the training data now consist of unbalanced classes, unlike for the 50 class model.

Table 5. Mapping of NameRxn Labels to SHREC Labels for a Set of C−C Bond Formation Reactions

	SHREC			
NameRxn class	level-1	level-2	level-3	level-4
bromo Heck reaction	C−C bond formation	coupling	Heck	bromo
Negishi coupling	C−C bond formation	coupling	Negishi	
chloro Stille reaction	C−C bond formation	coupling	Stille	chloro
iodo Sonogashira coupling	C−C bond formation	coupling	Sonogashira	iodo
iodo Suzuki coupling	C−C bond formation	coupling	Suzuki	iodo

Table 6. Total Number of Unique AP2+AP3 Reaction Vectors (Number of Rows in the Input Data) Shown for the Different Types of Reaction Vector Descriptors, along with the Number of Reaction Classes, the Median Number of Reaction Vectors Per Reaction Class, and the Number of Atom Pair Descriptors for the USPD Grants and USPD Apps Data Sets

classification system	data set	number of reaction vectors	number of reaction classes	median number of examples per class	number of descriptors
SHREC	USPD Grants	111981	336	129.5	4119
	USPD Apps	25026	335	29	4119

The cleaned USPD Apps data set was also preprocessed to produce an external test compatible with the extended model. The reaction classes in the USPD Apps were mapped to the SHREC labels, and reaction vectors, which were already described in the USPD Grants data set, were excluded. One class present in USPD Grants was missing in USPD Apps, the "C−C bond formation (methylation) (Blanc chloromethylation)" class, which was not therefore evaluated externally. The characteristics of the two data sets are shown in Table 6.

Evaluation Measures. The performance of the models was assessed using recall, precision, and the F1-score, all of which can be derived from the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN); see Table 7. Macro-averages of recall, precision, and the F1-score were calculated by first calculating the values for each class independently and then taking the unweighted means. Macro-averages are appropriate for balanced classes because all classes are treated equally. Micro-averages were calculated

DOI: 10.1021/acs.jcim.9b00537
J. Chem. Inf. Model. 2019, 59, 4167−4187

Table 7. Performance Measures Used To Evaluate the Models

performance measure
$\text{recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}$
precision = TPTP + FP
$\text{F1} = 2 \times \dfrac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$

from the global counts of TPs, FNs, and FPs and are more appropriate when the classes are unbalanced because they give more emphasis to the majority classes. Weighted macro-averages were also calculated to account for unbalanced classes by weighting the individual class values according to the relative number of examples available in the class.

Confidence in Predictions.

The Random Forests (RF) classifier implemented in scikit-learn infers class probabilities using the soft-voting method, which averages the probabilities associated with each class and assigns the class with highest average probability. To assess the level of confidence in the predictions, the best model identified during training was retrained using the entire USPD Grants data, and predictions were made on the USPD Apps data. For each reaction class, the ratios of true and false predictions were calculated for increasing probability scores to determine a confidence level for a given prediction.

The RF probability scores are the direct outputs of the model and reflect the variability of the model itself. We also investigated the use of conformal prediction (CP) to assign reliability scores.[27] Conformal predictors are built on top of machine learning algorithms and make use of calibration data, which are used to determine nonconformity scores for each class, for example, class probabilities given as the percentage of trees that assign the correct class. When applied to bioactivity prediction of compounds consisting of two classes, active and inactive, the usual approach to determining which class to assign is to calculate a p-value for each class as the number of nonconformity scores with lower values than the compound to be predicted, divided by the total number of calibration compounds in the class. To be assigned to a particular class, the p-value should be greater than a user-defined significance level.[28−30] Thus, a new compound can be predicted as belonging to just one class, both classes, or neither of the classes.

Here, the problem is a multitask classification where the aim is to assign a reaction to a single reaction class, which is one of many possible reaction classes. We assign the reaction class as the one with the highest p-value and assess the reliability of the prediction using the highest p-value as a confidence score; and the difference between the two highest p-values as a credibility score. Thus, the credibility score indicates the separation between the class associated with the highest p-value and the class associated with the second highest p-value. The ideal case would be when, for a given instance, the resulting confidence value is high (i.e., the prediction is close to the likely observations) and the credibility score is also high (i.e., the second highest p-value is very low, and the separation between the two highest p-values tends to 1).

The USPD Grants data set was split into 90% for training and 10% for the calibration set using a stratification algorithm on the reaction class column. Although a higher percentage of the training set is usually recommended when using CP for QSAR prediction, for example, 30%,[28] increasing the accuracy of the conformal predictor by decreasing the accuracy of the underlying algorithm was not thought to be desirable. (It will become evident in the Results that >80% of the training data were required to achieve a good model.) The training data in this case consisted of 100 782 unique reaction vectors with a median number of 116.5 reaction vectors per class. The calibration set consisted of 11 199 unique reaction vectors and a median of 13.0 per class.

Implementation.

The machine learning classifiers from the scikit-learn package were used: Random Forests (RF), K-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and Gradient Boosted Tree (GB). Default parameters were used for the 50-class models as shown in Table S1. The hyperparameters of the RF classifier were optimized (results not shown) for the 336-class model and are reported in Table S2. A Python implementation of the Inductive Conformal Prediction (ICP) framework (https://github.com/donlnz/nonconformist) was integrated with the optimized Random Forests (RF) classifier for the CP-augmented classification algorithm (RF-CP).

■ RESULTS

Data Characteristics.

From Tables 1 and 2, it can be seen that around 30% of the reactions in each data set did not have classification labels. When the classified reactions were transformed to descriptors and only the unique reaction vectors retained, there was an approximately 90% reduction in the number of data points (for example, considering the USPD Grants data, 1 114 953 cleaned and classified reactions result in 115 692 unique AP2+AP3 reaction vectors). This indicates that both data sets (Grants and Applications) contain a high degree of redundancy in terms of unique reaction centers. The high redundancy reflects the nature of pharmaceutical patents, which are aimed at covering specific regions of the chemical space exhaustively, often by combining very similar molecules with similar reagents. The mapping of reactions to unique reaction vectors is also highly skewed as shown in Figure 6 with a small number of reaction vectors associated with thousands of reactions and a long tail, where the majority of the reaction vectors are associated with fewer than 10 reactions

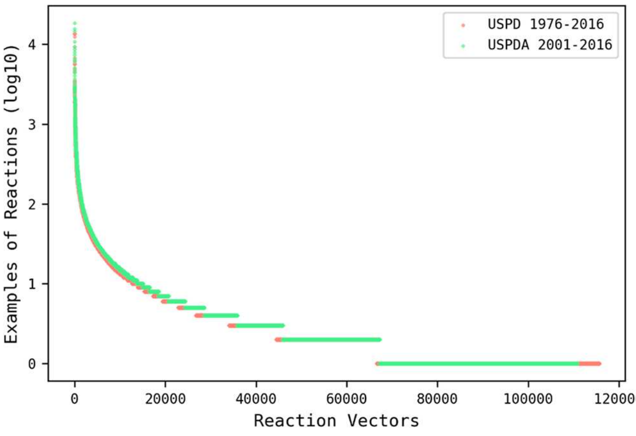


Figure 6. Number of reactions (log 10 format) represented by each AP2+AP3 reaction vector in the USPD Grants and USPD Apps data sets, respectively.

DOI: 10.1021/acs.jcim.9b00537

and 40% of the reaction vectors are associated with a single reaction only.

The distribution of unique reaction vectors across the reaction classes is also high skewed (Figure 7). Fewer than 5% of the reaction classes are associated with more than a thousand unique reaction vectors, while ∼20% of the reaction classes contain 5 or fewer examples, thus evidencing the presence of very unbalanced data.

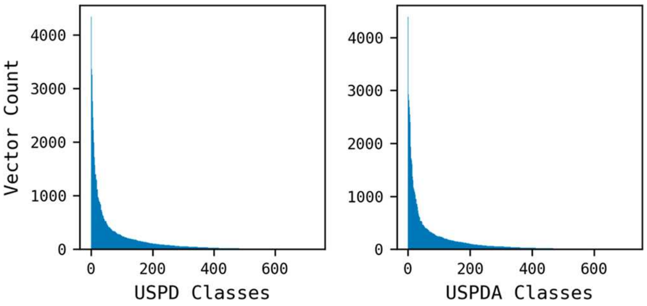


Figure 7. Reaction vector class distributions for the filtered USPD Grants and USPD Apps sets. Classes are sorted in descending order according to their numbers of vectors.

The overlap between the reaction vectors for the USPD Grants and USPD Apps data sets is shown in Figure 8.

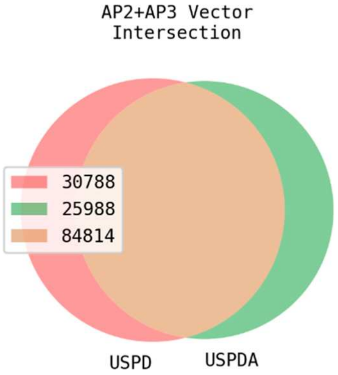


Figure 8. Overlap between unique AP2+AP3 reaction vectors in the USDP Grants and USPD Apps data sets.

50-Class Models. Table 8 presents the performance of the 50-class classification models on the USPD Grants test set based on the four different descriptors and the four different machine learning methods. The Random Forests (RF) classifier performed slightly better than the other models in all cases except the AP4 descriptor data set. The Gradient Boosted Trees (GB) and Support Vector Machine (SVM) also performed well in most cases. The best of the single descriptors was AP3, indicating that it is necessary to encode some of the environment of the reaction to improve discrimination between reactions, as would be expected for reaction classes that differ by features that are external to the reaction center. The reduced performance of the AP4 descriptors can be rationalized as an increase in noise with the extended environment not being relevant for class discrimination. The AP2+AP3 combined descriptors performed better than any of the single descriptors, and the performance is comparable to that of Schneider et al.[8]

The normalized confusion matrix for the RF and AP2+AP3 is shown in Figure 9. The lowest scores are reported for classes that cannot be distinguished effectively using reaction vectors. Examples include “bromo Suzuki coupling” and “bromo Suzuki-type coupling”; and “*N*-methylation” and “iodo *N*-methylation”. The first pair differs only in the reaction conditions through which the reactions occur, which are not encoded by reaction vectors, whereas the second pair represents the same reaction class in a generic and in a more specific form. These findings led to the introduction of the SHREC hierarchical classification scheme as described in the Methods.

Classes with small reaction centers such as “methylation” or “alcohol + amine condensation”, where extended environments are characterized by significantly different atom pair features,

Table 8. Macro Averages of Precision, Recall, and the F1-Score for Different Descriptors and Different Machine Learning Methods for the Fifty Class Models[a]

descriptor	classifier	precision	recall	F1-score
AP2	RF	0.80	0.80	0.80
	k-NN	0.61	0.59	0.59
	SVM	0.77	0.76	0.76
	GB	0.80	0.78	0.79
AP3	RF	0.87	0.87	0.87
	k-NN	0.76	0.75	0.75
	SVM	0.87	0.87	0.87
	GB	0.86	0.85	0.85
AP4	RF	0.80	0.80	0.79
	k-NN	0.67	0.65	0.65
	SVM	0.81	0.81	0.81
	GB	0.77	0.76	0.76
AP2+AP3	RF	0.90	0.90	0.90
	k-NN	0.80	0.79	0.79
	SVM	0.89	0.89	0.89
	GB	0.90	0.89	0.90

[a] Performance is shown on the internal validation data (i.e., the test data set extracted from USPD Grants as shown in Tables 3 and 4).

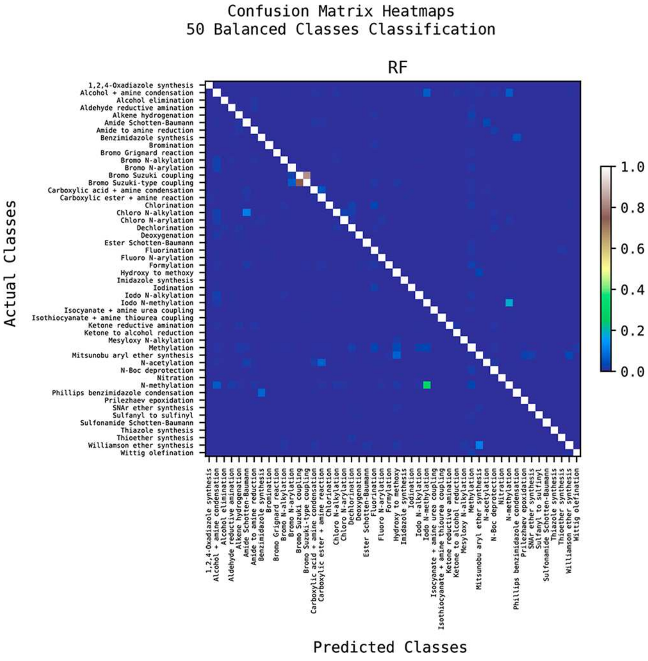


Figure 9. Normalized confusion matrix for the RF and the 50 reaction classes.

also contribute negatively to the model performance. In these cases, the large difference between examples can lead the classifier toward a misclassification of the unseen examples.

An external validation was carried out using the 15 193 USPD Apps data set to confirm the selection of AP2+AP3 as the default descriptor-type for reaction classification. The distribution of examples in this data set is unbalanced, and these data are therefore a more realistic reflection of the distributions of reactions in real data sets to which the model might be applied. The prediction performances were evaluated using micro-average and macro-weighted recall, precision, and F1-scores. Near identical results were found for both the micro- and the macro-weighted scores, and only micro-average results are shown in Table 9.

Table 9. External Validation Set Consisting of 15 193 Reaction Vectors; "Micro" Classification Report

classifier	micro precision	micro recall	micro F1-score	external validation F1-score
RF	0.86	0.86	0.86	0.90
k-NN	0.69	0.69	0.69	0.79
SVM	0.85	0.85	0.85	0.89
GB	0.85	0.85	0.85	0.90

Similar trends were seen as for the internal validation. k-NN was the lowest performing classifier (although no attempt was made to optimize the performance of the k-NN), whereas the scores reported for RF, SVM, and GB are almost comparable to those found in the internal validation. These results support the selection of the AP2+AP3 as an appropriate descriptor for classification and RF as the best choice of machine learning method, taking both effectiveness and efficiency into account.

336-Class Classification Models. Having established the suitability of the AP2+AP3 descriptors for reaction classification, the approach was extended to a much large data set consisting of a much large number of reaction classes, as shown in Table 6. The NameRxn reaction classes were converted to the SHREC classes described in the Methods. Note that the hyperparameters of the RF were optimized with the final parameters shown in the Supporting Information. Finally, the best performing model was further explored using the built-in probability estimation method in RF and CP.

Models were trained using the RF and AP2+AP3 descriptors, and the performance based on macro-weighted F1-scores is shown in Table 10, for both the internal validation (USPD Grants test set) and the external validation (USPD Apps). The reaction vector descriptors are also compared to 4096 bit RDKit reaction fingerprints, which are conceptually similar to reaction vectors, although they are fixed-length hashed fingerprints. The reaction vector descriptors performed slightly better than the RDKit fingerprints.

Table 10. Performance of the RF Model Trained on 336 Reaction Classes Reported As Macro-Weighted F1-Scores[a]

data set	RDKit 4096	Sheffield AP2+AP3
USPD Grants test set	0.87	0.90
USPD Apps	0.82	0.85

[a] The test set is the internal validation, whereas the USPD Apps represents an external validation consisting of reaction vectors that are not present in the training data.

The relationship between model performance and number of training examples was investigated by examining the F1-scores for the individual classes for both the internal and the external validations. The recall, precision, and F1-scores plotted against training class size are shown in Figure 10. As expected, the scatter plots show slightly improved performance in the internal validation as compared to the external validation, and they confirm that the three metrics are highly correlated. Each plot shows a very broad variance in performance when the number of training examples in a class is lower than 100, which is the case for the majority of the classes, because the median number of examples is 52.

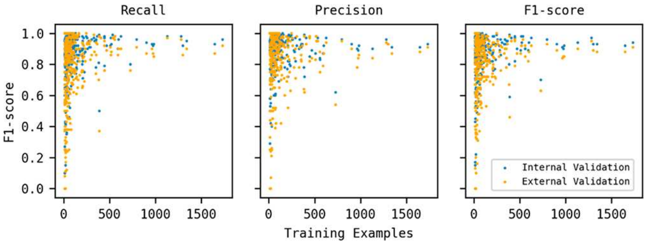


Figure 10. Distribution of model performance across the reaction classes is shown for both the internal validation (USPD Grants test data, blue) and the external validation (USPD Apps, orange). The x-axes represent the number of examples in each class in the training data.

This variation can be explained by the intrinsic nature of each reaction class and the varieties of reaction centers it potentially describes. For example, although the "synthesis (1,2,4-triazole)" class contains only 13 examples in the training data, the F1-score for 19 unseen examples in the internal validation set is 0.97. This class performed well because the variety of its extended reaction centers is generally very narrow (Figure 11). Conversely, an F1-score of 0.7 is reported for the "C−C bond formation (methylation)" class for 1094 example reactions in the internal validation set, even though its corresponding training set contains 729 examples. This is because the reaction center itself, a simple methylation, is small; however, it occurs in many different extended environments and is, therefore, not an easy class to match using the current implementation of reaction vectors. All classes that are described by a small number of AP2 descriptors and a high variety of AP3 descriptors are affected by this issue.

The scatter plots also highlight that the variance in performance is strongly reduced when the number of training examples increases. A minimum threshold of 150 examples per class returns a lowest F1-score equal to 0.59 for the "C−N bond formation (amination)" class (internal validation), which, as for methylation, consists of a small reaction center presented in a wide variety of extended environments. A threshold of 250 examples per class returns lowest and median F1-scores equal to 0.70 and 0.93, respectively. Therefore, the use of a bigger and more curated source of training data is expected to yield better performing models in the future.

The F1-score represents a global measure of model performance, which can mask the negative effect caused by a very small number of poor performing classes, in particular for those models that are assessed using large test sets with a strongly unbalanced distribution of examples per class. As shown above, in general the performance improves with the number of training examples within a class. Therefore, the use

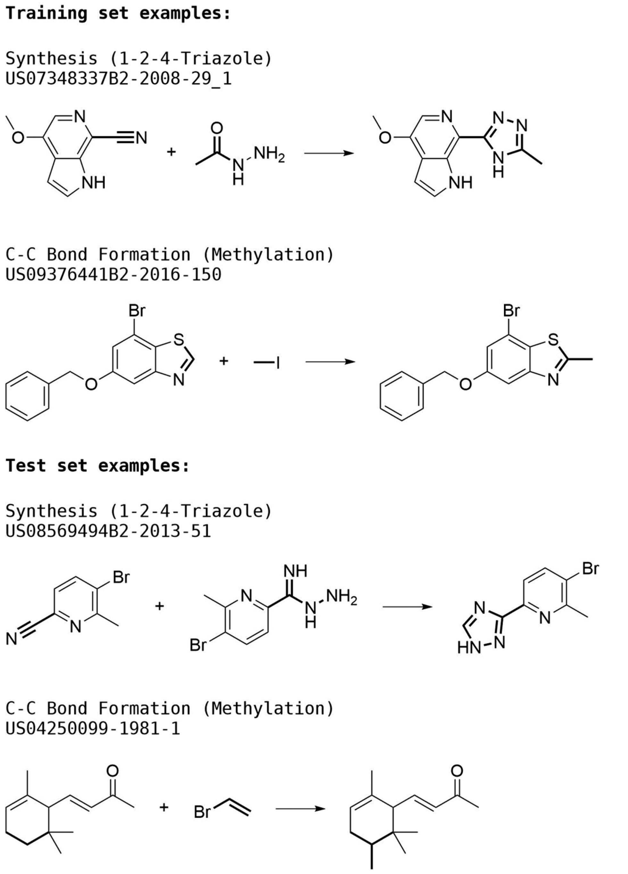


Figure 11. Examples of classes involving more ("C−C bond formation (methylation)") or less ("synthesis (1,2,4-triazole)") variable reaction centers.

of class weights was explored to investigate if biasing the classifier toward the less populated classes would result in improved overall performance. Three sets of weights were tested: default settings of 1.0 (i.e., all classes weighted equally) as a control; balanced weights; and empirical weights. Balanced weights were calculated in scikit-learn according to the heuristic shown in eq 1 inspired by King et al.,[31] where the weight (w_y) associated with a given class y is calculated by dividing the total number of examples (n_samples) by the product of the total number of classes (n_classes) and the number of examples in the class y (y_samples):

$$w_y = \frac{n\_\text{samples}}{(n\_\text{classes} \times y\_\text{samples})} \qquad (1)$$

Empirical weights were determined by inspecting the internal validation results to identify classes that produced a high number of false positives. This was achieved as follows. The classes were sorted in descending order of the number of false positives. The 10 classes with the highest number of false positives are shown in Table 11 with their corresponding F1-scores. Three classes with relatively large numbers of false positives and low F1-scores were selected for the manual weight tuning: "C−N bond formation (methylation)", "C−N bond formation (amination)", and "C−N bond formation (N-arylation) (bromo)".

Results are shown in Table 12. All weighting schemes resulted in the same global performance except the balanced weights, which showed slightly worse results as compared to the default (all classes weighted equally). This may be due to the unbalanced nature of the validation sets: a classifier trained with some bias toward the most populated classes might actually perform better than an unbiased classifier in these contexts. Although the manually assigned weights did not affect the global performance of the model, some reduction was seen in the numbers of false positives for the classes where these were most prevalent and the false positives were more evenly spread across classes instead of being concentrated in one or two classes, Figure 12. For this reason, weighting scheme 4 was chosen for the final model.

Table 11. Ten Classes with the Highest Number of False Positives in the Internal Validation Data Set

reaction class	false positives	F1-score
C−C bond formation (methylation)	*538*	*0.68*
C−C bond formation (condensation) (carboxylic acid + amine)	210	0.93
C−N bond formation (*N*-alkylation) (bromo)	203	0.94
functional conversion (hydrogenation) (alkene to alkane)	190	0.94
C−N bond formation (*N*-methylation)	173	0.90
C−N bond formation (*N*-arylation) (chloro)	172	0.90
C−N bond formation (amide formation) (Schotten−Baumann)	166	0.93
C−N bond formation (amination)	*141*	*0.56*
C−N bond formation (N-arylation) (bromo)	*138*	*0.78*
C−O bond formation (etherification) (Williamson)	136	0.91

Table 12. Performance of Models Trained Using Different Weighting Schemes[a]

classifier	weights	validation set weighted F1-score	external data weighted F1-score
1	*none*	*0.90*	*0.85*
2	balanced	0.87	0.82
3	*C−N bond formation (N-arylation) (bromo): 0.8* *C−C bond formation (methylation): 0.3* *C−N bond formation (amination): 0.6*	*0.90*	*0.85*
4	C−N bond formation (*N*-arylation) (bromo): 0.6 C′−C bond formation (methylation): 0.1 ′C−N bond formation (amination): 0.4	0.90	0.85
5	*C−N bond formation (N-arylation) (bromo): 0.8* *C′−C bond formation (methylation): 0.1* *C′−N bond formation (amination): 0.8*	*0.90*	*0.85*

[a] Note that the micro F1-scores were identical to the weighted scores.

Effect of Training Data Size. The optimized RF classifier was trained with increasing proportions of the USPD Grants data set and performance reported on the USPD Apps. The training set size was varied from 2% to 100% in 2% intervals using stratified sampling with three data sets produced for each size by varying the seed in the stratification algorithm. "Weighted" and "micro" F1-scores on the external data set are reported in Figure 13.

DOI: 10.1021/acs.jcim.9b00537

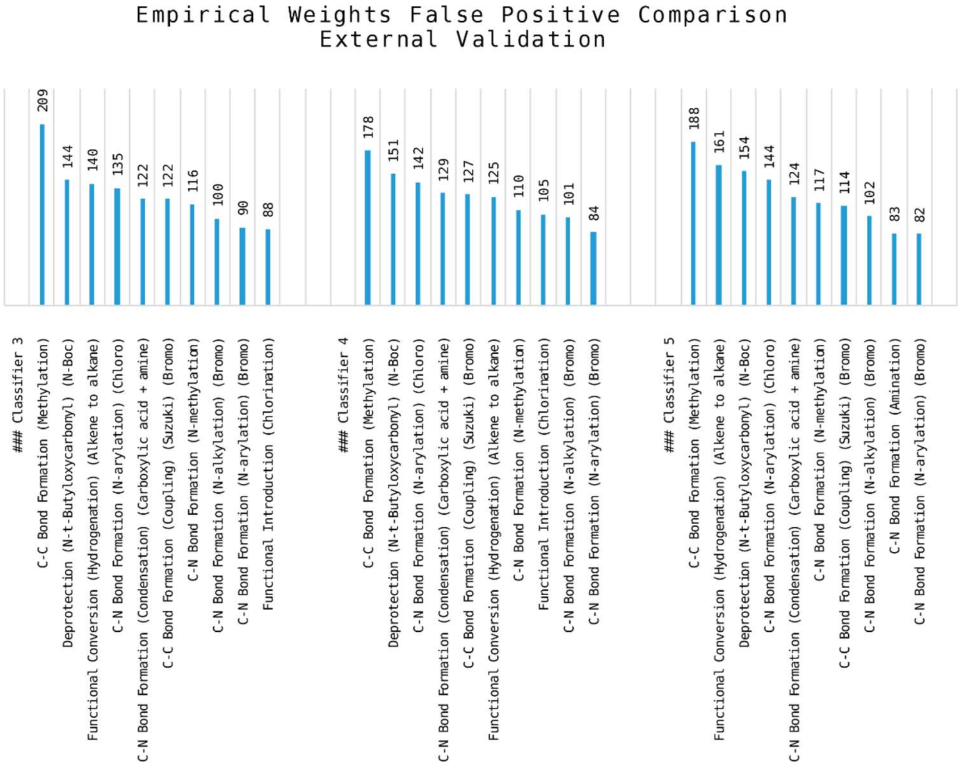


Figure 12. Frequency distribution of false positives across the reaction classes in the external validation set.

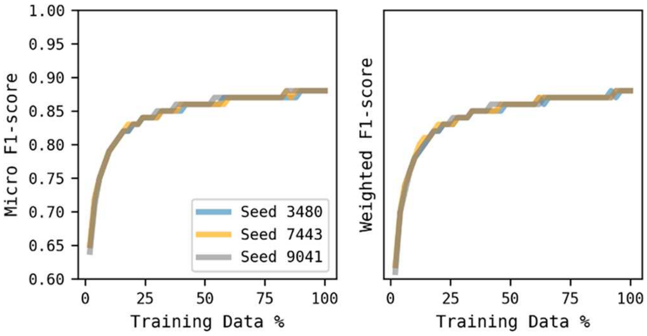


Figure 13. "Micro" (left) and "weighted" (right) F1-scores trends at increasing amounts of training data on the prediction of the external data set.

Both plots show consistent trends, demonstrating that the combination of RF and reaction vectors produced efficient models at almost any percentage of the training data, even close to zero percent. "Micro" and "weighted" F1-score trends are closely comparable, except for very low amounts of training data (i.e., lower than 10%) where the "weighted" scores are slightly worse than the "micro" scores. The best "micro" F1-scores were found using a percentage of training data higher than 86%, whereas the best "weighted" F1-scores were found with a percentage of training data higher than 92%. The general performance trends show that after a steep increase in performance between 0% and 20%, the curve reaches a plateau beyond which there are diminishing gains.

Confidence in Predictions. The model was trained using the entire USPD Grants data, and predictions were made on the USPD Apps data. The confidence levels associated with true and false predictions were then evaluated for each reaction class as follows. The data were sorted on ascending probability score and then binned into 98 bins ranging from 0.03 to 1.00. The absolute numbers and ratios of true and false predictions were calculated for each bin/probability level and are shown in Figure 14. The left graph shows that the number of correct predictions increases steadily as the probability scores increase; however, it does not show how the false predictions change due to their lower absolute numbers as compared to the true predictions. The right graph shows the ratios of true predictions to false predictions where it can be seen that a probability of 0.22 results in 49% true and 51% false predictions.

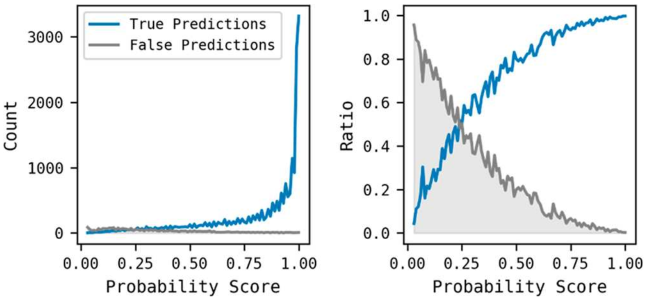


Figure 14. Absolute numbers (left) and ratios (right) of true and false predictions associated with each level of probability.

Table 13 shows how the classification performance improves by removing entries with low probability values. When the model was trained on the entire USPD Grants set (~111 K examples), the "weighted" F1-score was 0.88 even without applying any confidence score filtering, which can be already be considered good performance for the classification of an external data set. The performance of the model increases as the probability cut-off is increased, by sacrificing an increasing

Table 13. Variations in Performance (Left) and Percentage of Filtered Reactions (Right) Associated with Different Probability Cut-Off Levels

probability cut-off	weighted F1-score	percentage of filtered reactions
0.0	0.88	0.0
0.15	0.90	3.65
0.25	0.93	7.81
0.35	0.94	13.37
0.45	0.96	17.05
0.60	0.97	25.26
0.80	0.99	39.76

percentage of reactions for which predictions are not reported. The performance improves even for low cut-off values, ranging from 0.15 to 0.35, where the percentage of filtered reactions is under 15%. The results provide insights on how to set numerical cut-offs to enhance the reliability of the model, for example, by only assigning classes to reactions that have a high chance of being correctly predicted. It should be noted, however, that these specific values are not directly transferable to other data sets because they will vary according to the composition of the test set.

For the CP, the classifier, referred to as RF-CP, was trained and calibrated using 100 782 and 11 199 unique reaction vectors, respectively, and confidence and credibility scores were inferred on the entries of USPD Apps. Two separate binning processes were carried out: the confidence scores were binned into nine bins ranging from the values 0.92 to 1.00; and the credibility scores were binned into 93 bins ranging from 0.08 to 1.00. The absolute numbers and ratios of true and false predictions associated with each confidence and credibility level are plotted on the left of Figure 15; ratios are plotted on the right.

The top-left chart shows a trend similar to that in Figure 14, although in this case the range of scores is significantly smaller. The top-right plot shows how a satisfactory separation between true and false predictions is achieved only when the confidence score tends to the value 1. These results are supported by the theory of conformal prediction: the highest *p*-value indicates how close the observed prediction is to the typical distribution of results for a given class, but it does not provide information on the presence of other high *p*-values associated with other classes. This effect was verified by plotting the credibility scores, which show how far the predicted class is from the rest of the possible class predictions. The bottom-left and the bottom-right plots report a much broader separation between true and false predictions. Both plots show that the percentage of wrong predictions remains very low for a credibility score higher than 0.3.

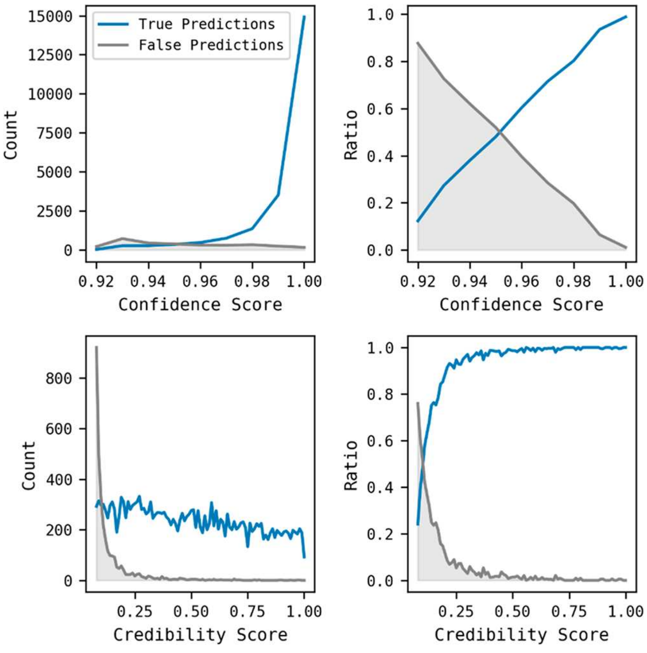


Figure 15. Absolute numbers (left) and ratios (right) of true and false predictions associated with each level of confidence (top) and credibility (bottom).

Different credibility scores were then used as threshold values to filter the predictions. Table 14 shows the trade-off between F1 score and number of entries filtered out as the credibility cut-off increases. The trends obtained using CP are comparable to those seen using the probability scores in RF, with the performance improving notably even for low cut-off values ranging from 0.09 to 0.12, where the percentage of filtered entries remains under 15%. Although these results are also dependent on the composition of the test data set as for the RF probability scores, the statistical basis of CP is such that this is our preferred approach to assessing prediction reliability.

Table 14. Variations of Performance (Left) and Percentage of Filtered Reactions (Right) Associated with Different Credibility Cut-Off Levels

credibility cut-off	“weighted” F1-score	percentage of filtered reactions
0	0.88	0
0.09	0.91	4.73
0.10	0.93	8.04
0.12	0.95	12.43
0.15	0.96	17.72
0.20	0.98	24.74
0.25	0.99	36.94

Applications. This section reports on the application of the reaction classification model on two unseen data sets for which classification data were not available, a subset of the Evotec ELN and a collection of reactions extracted from the medicinal chemistry literature. In general, these data sets are not curated, and therefore the first step was to prepare them using the same protocol as used when training the model. The reactions were then classified using the RF-CP classifier with credibility scores used to enhance the reliability of the predictions, and the composition of each data set was examined and compared.

The reactions extracted from the medicinal chemistry literature are expected to be more diverse as compared to in-house pharmaceutical data, and to consist of a greater variety of syntheses with no necessary prerequirement for robustness. Syntheses reported in the literature usually involve the formation of new scaffolds, which are relevant for drug discovery use, thus describing novel reaction environments that can be used to evaluate the classification model flexibility.

Evotec Electronic Laboratory Notebook (ELN). The 170 770 reactions deposited between September 9, 2009 and February 27, 2018 were extracted from the Evotec (UK) ELN server. The reactions were described by reactants, reagents, products, yields, and time information. Entries were then cleaned and balanced using the reaction standardization protocol: reagents were filtered out as were entries with more than six reactants or six products, which left 168 375 entries; the reactions were mapped using the Indigo Reaction Automapping tool, and reaction components, which did not contain any atom mapping information, were removed to preserve only the components involved in the reaction center, and the reactions were balanced. 136 240 entries were retained after this process, and a total of 144 330 single-step reactions were generated. Entries identified with the label "test" were filtered out because they did not represent real experiments. The final set of reactions consisted of 144 014 entries.

Reaction vectors were then generated with 144 008 entries processed successfully, yielding a table described by 3305 atom pairs. For comparison, the numbers of atom pairs necessary to describe the original USPD Grants (115 602 entries) and USPD Apps (110 802 entries) data sets is 4205 and 4046, respectively. Thus, although the ELN data set contains almost 25% more entries as compared to the USPD Grants data set, 21% fewer atom pairs are required to represent it using reaction vectors. This indicates that the ELN data are less diverse than the USPD Grants data, which is to be expected. The reaction vectors were adjusted according to the USPD Grants data set by removing atom pairs present in the ELN set but missing from the training data, and adding columns filled with zeros to the ELN set for atom pairs present in the training data but not in the ELN data. The procedure yielded a table of 144 008 entries and 4119 features. The RF-CP classifier was then used to classify the ELN entries including assigning confidence and credibility scores to the predictions. The distributions of scores are plotted in Figure 16.

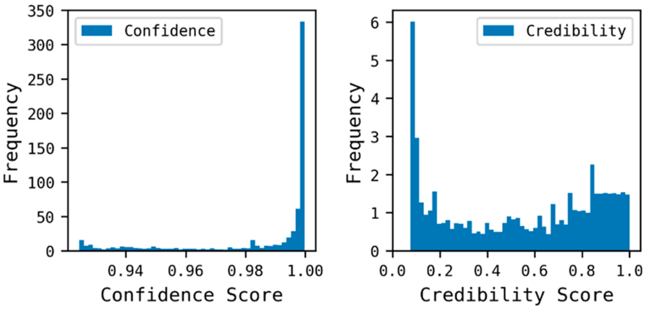


Figure 16. Confidence (left) and credibility (right) scores of the Evotec ELN data reaction classification.

The plots are similar to those reported on the USPD Apps data. The confidence scores fall within a narrow range of values (0.924−1.000) and are mostly concentrated between 0.98 and 1.0. This indicates that the model identifies most of the examples as being very similar to those used in the calibration set. The credibility scores fall within a larger range of values (0.075−1.000) with an intense peak on the lower bound. This suggests that some examples have high *p*-values for more than one reaction class. Different credibility thresholds were then applied to determine the absolute numbers and percentages of entries filtered out at each cut-off level. Results are reported in Table 15.

Table 15. Credibility Score Threshold Filtering Tests Applied to the Evotec ELN Data

credibility threshold	absolute number (percentage) of retained entries	absolute number (percentage) of filtered entries
0	144008 (100%)	0 (0%)
0.09	129679 (90.05%)	14329 (9.95%)
0.10	124103 (86.18%)	19905 (13.82%)
0.12	118754 (82.46%)	25254 (17.54%)
0.15	114120 (79.25%)	29888 (20.75%)
0.20	105680 (73.38%)	38328 (26.62%)
0.25	100569 (69.84%)	43439 (30.16%)

A minimum credibility threshold of 0.12 was applied to remove the entries with very low chances of being correct predictions, in this case, 17.5% of the reactions in the ELN. This value was chosen on the basis of the analysis of the USPD Apps data where the same credibility threshold resulted in 12.4% of the entries being removed while the F1-score for the remaining reactions increased to 0.95.

The classification data were then analyzed at different levels of the classification hierarchy. Level-1 labels (e.g., "C−C bond formation") were grouped to produce a pie chart for comparison with the statistics on reaction superclasses identified in the USPD data[18] (Figure 17). Level-2 labels (e.g., "C−C bond formation (coupling)") and level-4 labels (e.g., "C−C bond formation (coupling) (Suzuki) (bromo)") were grouped to examine the most frequent reaction classes, Tables 16 and 17, respectively. Level-3 labels were ignored because they produced statistics very similar to the level-4 labels.

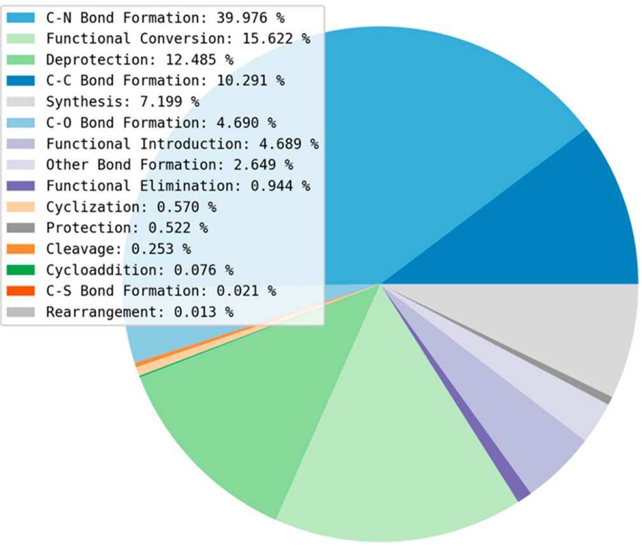


Figure 17. Level-1 classification of the Evotec ELN data.

The level-1 classification provides a general description of the ELN composition. C−N, C−C, and C−O bond formations constitute almost 55% of the total composition of the data set. This result is in accord with expectations because medicinal chemistry synthetic strategies are usually bottom-up, that is, start with small fragments, which are "grown" into drug-like molecules. Functional conversions describe almost 16% of data set. This percentage is comparable to the sum of the reductions, functional group interconversions (FGI), and oxidations percentages (17.3%) found in the U.S. patent literature (these classes are all grouped into a single class in the hierarchical classification system).

The proportion of functional introductions (~4.7%) is also similar to that reported for the USPD literature (3.4%). The

Table 16. Top Fifteen Reaction Classes in the ELN Data According to the Level-2 Labeling

level-2 classification	count
C−N bond formation (condensation)	15995
C−N bond formation (*N*-arylation)	12667
C−C bond formation (coupling)	10198
deprotection (*N*-*t*-butyloxycarbonyl)	6293
C−N bond formation (*N*-alkylation)	6024
C−O bond formation (etherification)	4401
C−N bond formation (amide formation)	4013
C−N bond formation (amination)	3947
functional conversion (reduction)	3276
other bond formation (sulfonamide formation)	3106
deprotection (COO-methyl)	2984
functional introduction (bromination)	2359
functional conversion (nitro to amino)	2133
C−N bond formation (carboxylic ester + amine)	1985
deprotection (COO-ethyl)	1796

Table 17. Top Fifteen Reaction Classes in the Evotec Data According to the Level-4 Labeling

level-4 classification	count
C−N bond formation (condensation) (carboxylic acid + amine)	14211
C−N bond formation (*N*-arylation) (chloro)	8220
deprotection (*N*-*t*-butyloxycarbonyl) (*N*-Boc)	6293
C−C bond formation (coupling) (Suzuki) (bromo)	4820
C−N bond formation (amide formation) (Schotten−Baumann)	3874
C−N bond formation (*N*-alkylation) (bromo)	3229
other bond formation (sulfonamide formation) (Schotten−Baumann)	3106
deprotection (COO-methyl) (COO-Me)	2984
C−O bond formation (etherification) (Williamson)	2937
C−N bond formation (*N*-arylation) (bromo)	2429
functional introduction (bromination)	2359
functional conversion (nitro to amino)	2133
C−N bond formation (carboxylic ester + amine)	1985
C−N bond formation (*N*-alkylation) (chloro)	1828
deprotection (COO-ethyl) (COO-Et)	1796

high percentage of functional group interconversions and additions can be explained by their use in both molecule construction and molecule optimization phases. Deprotections (∼12.5%) are more frequent as compared to protections (∼0.5%), suggesting the use of protected building blocks as starting materials for the syntheses. A similar result is also reported for the U.S. patent data. "Synthesis" (∼7.2%) is another frequent class, and describes reactions related to the preparation of particular scaffolds such as thioethers, imidazoles, pyrazolamines, thiazoles, and similar heterocycles. This class can be compared to the "heterocycle formation" class in the U.S. patent analysis, which represents a smaller percentage (1.4%) of that data. This suggests the frequent use of smaller building blocks and robust reactions for the preparation of larger scaffolds, as an alternative to the use of commercially available functionalized building blocks. These statistics are also supported by the analysis of the number of reactants in the data set: 63.2% of the entries were described by two reactants (i.e., C−C, C−N, C−O bond formations, and scaffold syntheses), 35.8% by only one reactant (i.e., functional introductions, conversions, and deprotections), and the remaining 1% of reactions were split between 3-, 4-, and 5-reactant reactions.

Other classes report lower percentages because of their minor efficacy in the synthesis of compounds of pharmaceutical interest (e.g., other bond formation), because of their unsuitable involvement in molecule construction (e.g., cleavage or functional elimination), or because of the use of already functionalized reagents that allowed those classes to be skipped (e.g., cyclization and C−S bond formation).

The level-2 ranking drills down to show how the broader classes are distributed across more specific examples. The C−N bond formation class is strongly supported by the subclasses "condensation" and "*N*-arylation", which consist of almost 29 000 reaction examples (∼24% of the total ELN set). This means that one in four reactions in the data set is a "C−N bond formation (condensation)" or a "C−N bond formation (*N*-arylation)". The remaining classes ("C−N bond formation (*N*-alkylation)", "C−N bond formation (amide formation)", "C−N bond formation (amination)", and "C−N bond formation (carboxylic ester + amine)") represent an additional almost 16 000 examples confirming that the creation of C−N bonds is a typical strategy in medicinal chemistry due to the general robustness and versatility of these reactions in the construction of pharmaceutically relevant structures. It is important to point out that the class "C−N bond formation (amination)" is not considered as a functional introduction in the SHREC because reaction vectors do not encode chemical environments outside the reaction center; thus reactions that involve building blocks containing an amine group are often indistinguishable from secondary or tertiary amine group introductions. The "C−C bond formation (coupling)" is represented by more than 10 000 examples, indicating the high efficiency of this reaction class as well. A large number of "C−O bond formation (etherification)" examples also indicate the relevance of structures linked as ethers (i.e., R_1−O−R_2, where R is a hydrocarbon group) as an alternative to the "C−N" and "C−C" bond formations.

Although the other bond formation class is not included among the majority classes in the level-1 classification, the specific "other bond formation (sulfonamide formation)" class is represented by more than 3100 examples of reactions, indicating its particular efficacy in the creation of S−N bonds between amines and sulphones. Despite its relatively high frequency in the level-1 classification, the "functional conversion" is represented by only one class in the top 15 of the level-2 classification ranking, which is the "functional conversion (nitro to amino)" class with approximately 2100 examples. This suggests the presence of many different functional conversions that contribute to the broader class, but that, with the exception of "nitro to amino", there are no particular preferred subclasses. In fact, functional conversions are commonly used to make small modifications to molecules to prepare them for bond formation reactions. The opposite effect is seen for the "functional introduction" level-1 class, which is not very frequent as compared to the other level-1 classes even though the "functional introduction (bromination)" subclass is represented by more than 2300 examples in the level-2 classification.

Deprotections are dominated by three specific examples with the protective agents of *t*-butyloxycarbonyl (BOC), COO-methyl, and COO-ethyl groups used in more than 11 000 examples. The high number of deprotections suggests the use of protected building blocks to enforce selective reactivity or to avoid catalyst poisoning as suggested in the U.S. patent analysis.

Table 18. Number of Reactions per Year in the ELN

year	2010	2011	2012	2013	2014	2015	2016	2017
number of reactions	7082	9760	14695	16075	20407	19879	15839	12041

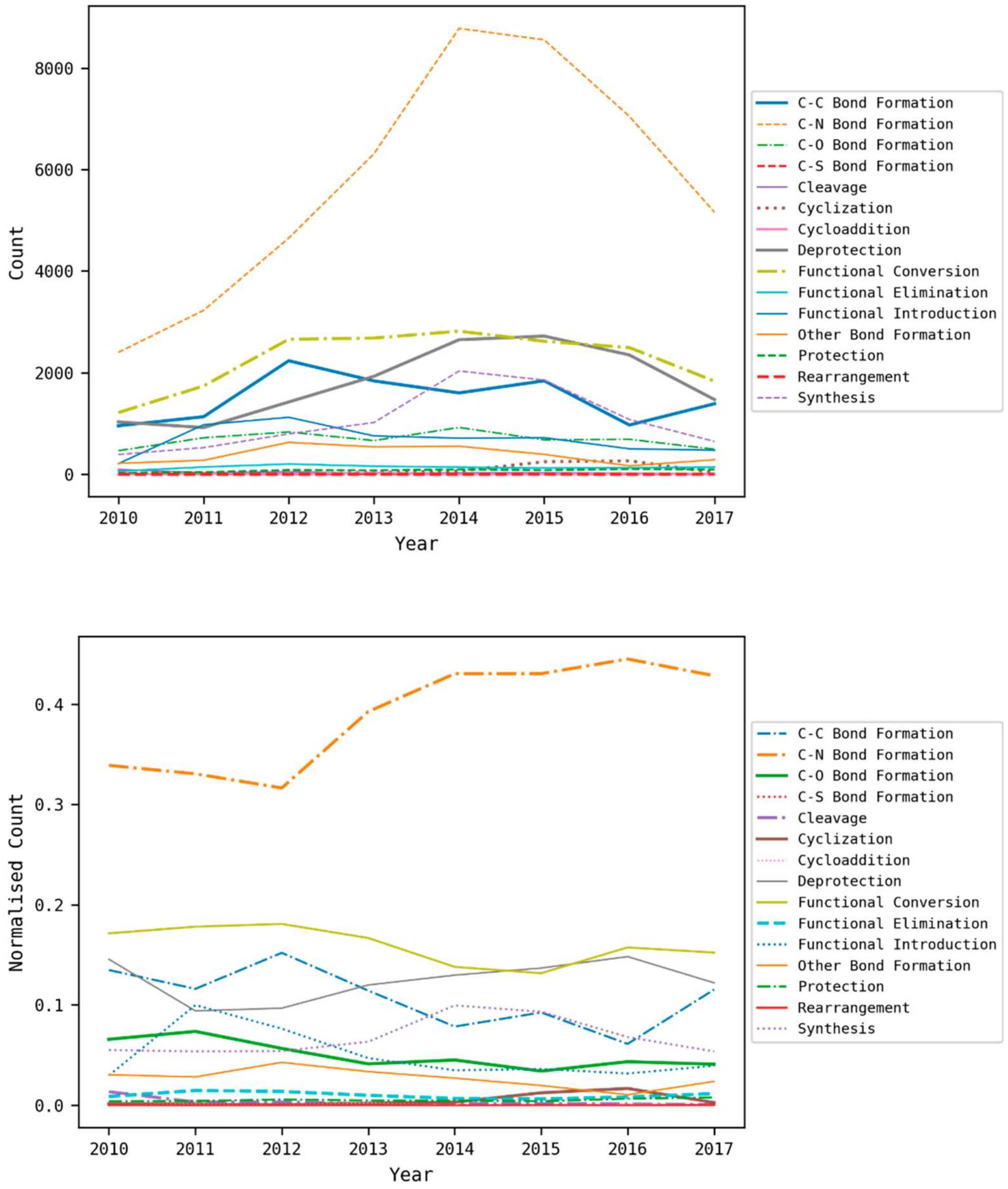


Figure 18. (a) Absolute and (b) normalized count time series of the level-1 class labels.

The most frequent classes based on the most detailed (level-4) classification, shown in Table 17, more or less preserve the same order as compared to Table 16. However, some reaction classes, such as "C−C bond formation (coupling)" and "C−N bond formation (amination)", lose their positions due to being split into smaller subclasses; for example, "C−N bond formation (amination)" is split into four subclasses, none of which are present in Table 17.

On the other hand, classes such as "C−N bond formation (amide formation)" do not drop in numbers significantly after adding subclass information because most of the examples belong to a single subclass. The "C−N bond formation (*N*-arylation)" is split into more specific subclasses such as "C−N bond formation (*N*-arylation) (bromo)" and "C−N bond formation (*N*-arylation) (chloro)", with the latter at position two in level-4 table. Similarly, "C−N bond formation (*N*-alkylation)" is split into "C−N bond formation (*N*-alkylation) (bromo)" and "C−N bond formation (*N*-alkylation) (chloro)".

The addition of more detailed classification levels does not affect several class counts at all for two reasons: first, some classes such as "functional conversion (nitro to amino)" or "functional introduction (bromination)" are not discriminated further by passing from level-2 to level-4 in the hierarchy, so they preserve the same labels and counts; and, second, the "other bond formation (sulfonamide formation)" is transformed into "other bond formation (sulfonamide formation) (Schotten−Baumann)" and preserves the same count because it is the only sulfonamide formation class in the data set.

Table 18 and Figure 18 show the results following a time series analysis, which can be useful if, for example, focused on

DOI: 10.1021/acs.jcim.9b00537

the correlation between classes and financial (e.g., company profits or project lengths) or scientific (e.g., molecule activities or successful properties) parameters. In particular, this type of analysis can be used to remove the human bias on certain reaction classes, and identify ones that are more successful. The analysis presented here is restricted to a correlation study between classes due to a lack of accessibility to the company financial data. The time series analysis considered counts and yields associated with each reaction class. The results are not supposed to be exhaustive; rather they are intended to provide some hints on how reaction classification can bring useful information for decision making in drug discovery. Level-1 classification labels were selected due to their more generalized nature and the lower number of classes. Values (i.e., counts or yields) were split by year, then retained only for the years between 2010 and 2017, inclusive (years 2008 and 2018 were excluded due to their partial contents). A total of 115 778 reactions were retained, and class counts were normalized by total counts per year.

Table 18 shows a steady increase in the total number of reactions since the introduction of the corporate ELN. The growth reaches a peak in 2014 and then gradually drops by the end of 2017. This behavior can be explained by the introduction of client ELNs, which are private notebooks that cannot be accessed internally. The use of private databases could have affected the composition of the classes as well, although this hypothesis was not tested. Time series plots of absolute and normalized counts are reported in Figure 18, for the 15 classes identified by level-1 labels.

The overwhelming presence of “C−N bond formation” compresses slightly the other classes, although many of their trends are still clearly visible. The absolute counts plot shows an increasing trend for almost every class with most peaking in 2014−2015, followed by a rapid decrease. Exceptions are “C−C bond formation”, “functional introduction”, and “other bond formation”, which are characterized by earlier peaks (i.e., 2011−2012), and increasing trends in 2017. The normalized data provide a different perspective of the same scenario: the “C−N bond formation”, “deprotection”, and “synthesis” classes show an increase moving from early (2010 to 2012) to late years (2014 to 2016 excluding 2017). This general increase is obtained at the expense of the other classes such as “C−C bond formation”, “C−O bond formation”, and “other bond formation”, which regain some positions only in 2017. As was already reported in the literature,[32−34] this result indicates a higher propensity toward the use of C−N bond formations due to their simplicity and robustness.

The correlation between normalized class counts was inspected by calculating the Pearson correlation coefficient (R) for each pair of classes, shown graphically in Figure 19. In general, the molecular growth classes such as “C−C bond formation”, “C−O bond formation”, and “other bond formation” show positive correlations with “cleavage” and all of the functional-related class, whereas they are negatively correlated with “C−N bond formation”, “cyclization”, “deprotection”, and “synthesis”. Conversely, “C−N bond formation” shows opposite trends, suggesting that the substrates involved in these reactions do not need to be prefunctionalized in situ (i.e., functional introduction or conversion) to react correctly with each other. This can produce a decrease in the number of steps required to obtain the final products, thus explaining the growing success of this class over time. This hypothesis could be further tested by comparing the average number of steps in routes with and without “C−N bond formation”, although this was not done here. Furthermore, “C−N bond formation” and “deprotection” show a positive correlation with each other, suggesting the deprotection of the products after the union of two building blocks through the formation of a C−N bond. “Synthesis” shows a positive correlation with “deprotection” probably for the same reason. “Functional elimination” and “deprotection” show a strongly negative correlation. Deprotections are comparable to functional eliminations that remove protective groups from the molecules; thus it would be unlikely to observe an increasing occurrence of these two classes at the same time. Interestingly, “C−S bond formation”, “cycloaddition”, “protection”, and “rearrangement” do not show relevant relationships with the other classes. This can be a consequence of their lower popularity in the data set.

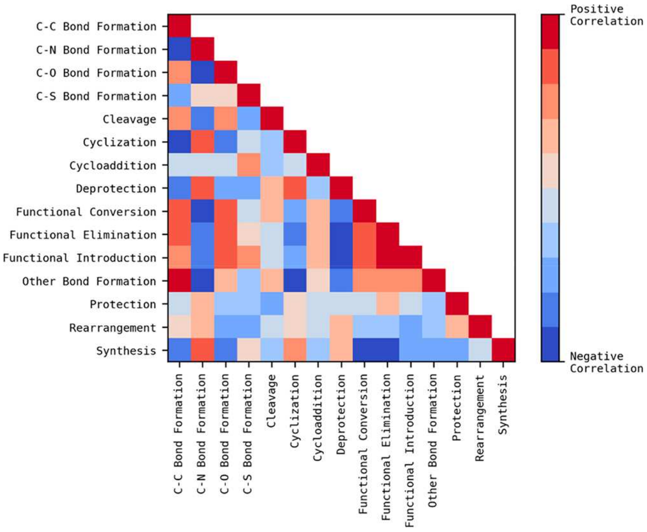


Figure 19. Heatmap that describes the lower triangular pairwise matrix of the level-1 class correlation coefficients.

Figure 20 shows how the yields of the reactions vary over time. The data were processed as follows. When multiple yields were reported for a single reaction, they were averaged and reactions for which no yield was reported were filtered out. The yields were then averaged to produce a mean yield for each reaction class for each year. Reaction classes described by fewer than 250 entries in the years between 2010 and 2017 were not analyzed, leaving a total number of 83 343 entries. The plot shows three different trends: increasing, decreasing, and stable yields. The yields of “deprotection” and “C−C bond formation” reactions increase over time, whereas they decrease for “functional elimination” and “functional introduction” reactions. The remaining classes show stable yields characterized by either low variance (i.e., “functional conversion”, “synthesis”, and “C−N bond formation”) or high variance (i.e., “cyclization”, “other bond formation”, and “C−O bond formation”). This type of analysis could be readily implemented in the ELN framework to monitor how each different class performs over time with the aim of maintaining a high global efficiency. For example, it could be used to assess the performance of the medicinal chemists in a specific time range, or to highlight differences in yield due to the impurity of the reagents, after the introduction of a new chemical supplier.

Medicinal Chemistry Literature Reactions. The medicinal chemistry data set consists of reactions from the *Journal of*

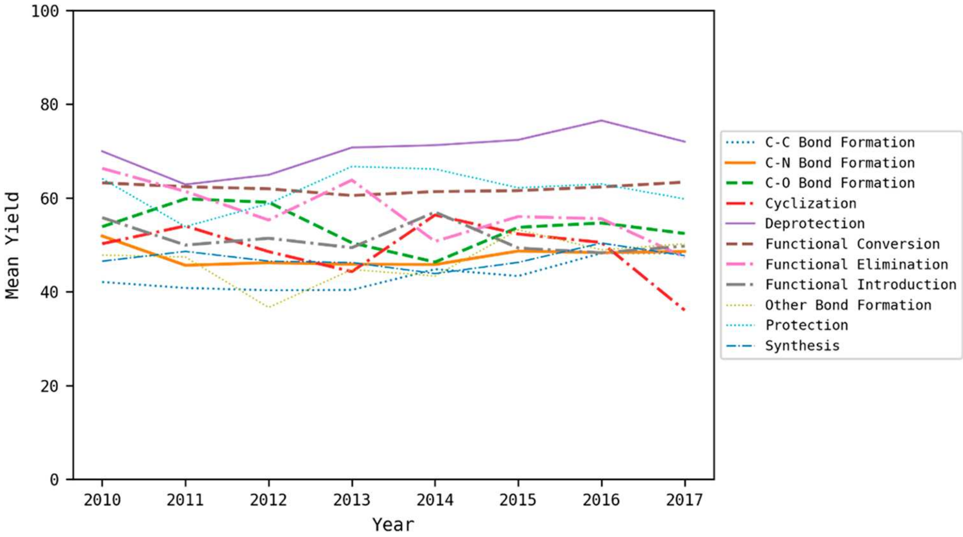


Figure 20. Variation in yields over time.

Medicinal Chemistry from the year 2008 and was originally prepared to test the performance of our de novo design tool.[16] Around 24K reactions were compiled by first collecting all reactions with yields of 100%, 75%, and 50% and excluding those consisting of solid-phase chemistry. The data set was reduced to 19 209 single-step reactions after cleaning, etc., using the same procedure as reported for the ELN. The data set is referred to here as JMC. When converted to reaction vectors, these were described by a total of 12 242 reaction vectors and 5331 atom pairs. This relatively high number of atom pairs for a relatively small data set already suggests that a large variety of reaction centers are described within it: although it is represented by 90% fewer unique reaction vectors as compared to the original USPD Grants data set, it requires almost 27% more atom pairs to be fully described. This preliminary result suggests that the data are more diverse than the patent data, which is perhaps not surprising given that the patent literature is aimed at capturing local regions of chemical space, whereas the medicinal chemistry literature is more likely to consist of a greater variety of syntheses with no necessary prerequirement for robustness or coverage of particular regions of chemical space.

The JMC reaction vectors were adjusted to be compatible with the reaction classification model, and the RF-CP classifier was used to classify the entries and to assign confidence and credibility scores, which are plotted in Figure 21. The confidence scores fall into a narrow range of values (0.924—1.000) that is similar to the ELN data and are similarly characterized by a peak on the left, but they are more spread. This indicates that the classifier identified the majority of the JMC reactions as very similar to the reactions contained in the calibration set, although they presented lower similarities as compared to the ELN reactions. The JMC credibility scores show a range of values identical to that found for the ELN data (0.075—1.000); however, the majority of the reactions are associated with lower scores. This means that the JMC data generally consist of examples with higher ambiguity as compared to the ELN distribution, causing a decrease in distance between the first and second best *p*-values computed by the CP. Different minimum thresholds on the credibility score were applied to determine the absolute numbers and percentages of filtered entries at each level as reported in Table 19. The threshold of 0.12 results in 49.09 of the JMC reactions being filtered out, as compared to only 17.54 of the ELN reactions. A manual inspection of the filtered entries confirmed that most were not classified correctly. Two conclusions were drawn from these results. First, data from the scientific literature tend to be more difficult to classify due to their higher diversity in terms of (extended) reaction centers. Second, the use of the credibility score thresholds in a more difficult classification problem highlights the practical advantages of integrating the classification model within a CP framework to improve model reliability.

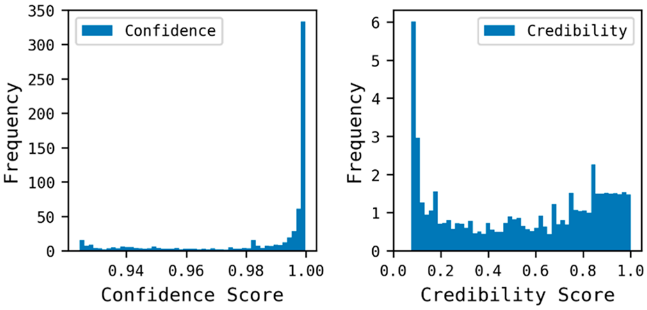


Figure 21. Confidence (left) and credibility (right) scores of the JMC data set reaction classification.

Table 19. Credibility Score Threshold Filtering Levels Applied on the JMC Data Set

credibility threshold	absolute number (percentage) of retained entries	absolute number (percentage) of filtered entries
0	19209 (100%)	0 (0%)
0.09	13335 (69.42%)	5874 (30.58%)
0.10	11632 (60.55%)	7577 (39.45%)
0.12	9779 (50.91%)	9430 (49.09%)
0.15	8339 (43.41%)	10870 (56.59%)
0.20	6994 (36.41%)	12215 (63.59%)
0.25	6308 (32.84%)	12901 (67.16%)

The 9779 reactions (50.9%) retained at the 0.12 credibility level were analyzed as for the ELN data. Results are reported in

Figure 22 and Tables 20 and 21. The level-1 classification shows different trends as compared to the Evotec ELN data.

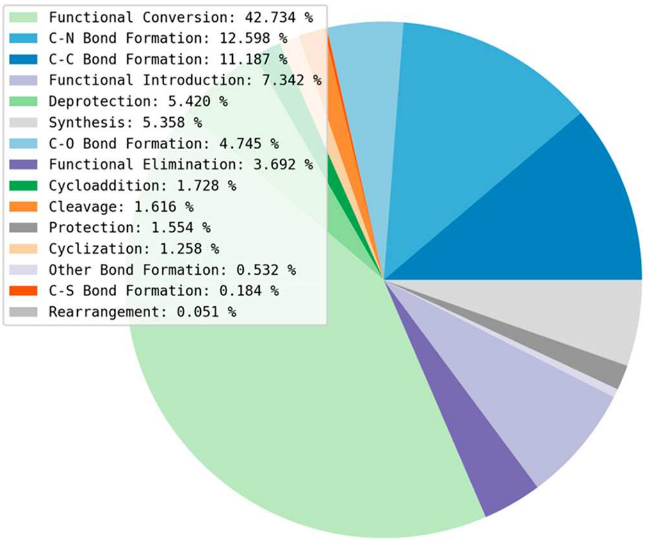


Figure 22. Level-1 classification of the JMC data set.

Table 20. Top Fifteen Reaction Classes in the JMC Data Set According to the Level-2 Labeling System

level-2 classification	count
functional conversion (hydrogenation)	1034
functional conversion (reduction)	776
C−C bond formation (coupling)	466
functional conversion (alkene to epoxide)	307
functional conversion (cyano to carboxy)	293
functional conversion (oxidation)	293
synthesis (thioether)	293
functional conversion (nitro to amino)	277
C−N bond formation (condensation)	250
functional introduction (hydroxylation)	244
functional conversion (alcohol to alkene)	227
C−O bond formation (esterification)	225
C−O bond formation (etherification)	218
C−N bond formation (*N*-alkylation)	203
functional introduction (bromination)	164

Table 21. Top Fifteen Reaction Classes in the JMC Data Set According to the Level-4 Labeling System

level-4 classification	count
functional conversion (hydrogenation) (alkene to alkane)	909
functional conversion (reduction) (aldehyde/ketone to alcohol)	528
functional conversion (alkene to epoxide) (Prilezhaev)	307
functional conversion (cyano to carboxy)	293
synthesis (thioether)	293
functional conversion (nitro to amino)	277
functional conversion (alcohol to alkene)	227
functional conversion (oxidation) (alcohol to aldehyde/ketone)	208
functional introduction (hydroxylation) (alkene hydration)	205
functional introduction (bromination)	164
cycloaddition (diene + dienophile) (Diels−Alder)	160
C−O bond formation (esterification)	155
functional elimination (deoxygenation)	155
functional conversion (sulfanyl to sulfinyl)	147
C−N bond formation (condensation) (carboxylic acid + amine)	141

The functional conversion class dominates all other classes and represents almost 43% of the entire data set, as compared to 15.4% of the ELN data. This suggests that these reactions were focused on scaffold modifications more than C−N, C−C, and C−O bond formations, which constitute 28.5% of the total classification. "Functional introduction" (7.3%) and "synthesis" (5.4%) also describe a significant number of examples in the data, indicating their persistent roles in medicinal chemistry. Deprotections constitute only 5.4% of the total classification in comparison to 12.5% reported for the Evotec ELN, supporting the existence of a positive correlation between C−N bond formations and deprotections.

The higher percentages of the minority classes such as "functional elimination" (3.7%) and "cleavage" (1.6%) as compared to the ELN can be explained by these reaction classes being generally avoided in industrial pharmaceutical chemistry where the objective is to construct the final products from the minimum number of building blocks. Conversely, the academic literature is usually more concerned with the presentation of new scaffolds with particular properties, with limited regard for the number of steps used to obtain such molecules. The "cycloaddition" (1.7%) (22.7 times higher) and "cyclization" (1.3%) (2.2 times higher) classes are also more prevalent as compared to the ELN data analysis.

The ranked frequencies of reactions using the level-2 classification are reported in Table 20. Subclasses of the functional conversions class occupy seven of the top 15 positions with five subclasses ("hydrogenation", "reduction", "alkene to epoxide", "cyano to carboxy", and "oxidation") together representing more than 2700 reactions, which corresponds to 28% of the data set. From an organic chemistry point of view, these reactions tend to preserve the total number of heavy atoms in a given structure; thus they are used for structural activation or functionalization. The relative frequency of the particular scaffold synthesis class "synthesis (thioether)" shows a strong focus on a particular motif, which can be typical of a data set covering a short period of time. This is also supported by the presence of "functional conversion (alkene to epoxide)" as the fourth most frequent class in the top 15. This class indicates a particular interest toward the transformation of alkenes into their corresponding epoxides, which is not a typical transformation observed in the preparation of molecules of pharmaceutical relevance. The highest ranking bond formation subclasses include "C−C bond formation (coupling)", "C−N bond formation (condensation)", "C−N bond formation (*N*-alkylation)", "C−O bond formation (esterification)", and "C−O bond formation (etherification)". It is also worth noting that the C−C bond formation class has almost twice as many examples as compared to the most popular C−N bond formation class. This result is consistent with the analysis carried out by Schneider et al.[18] where they highlighted increasing attention on C−C bond formations in recent years.

The level-4 classification ranking reported in Table 21 almost preserves the same order of the level-2 ranking except for a few classes. The "C−C bond formation (coupling)" and "C−N bond formation (*N*-alkylation)" are split into multiple classes, among which no one subclass is sufficiently populated to appear in the top 15 classes. However, "cycloaddition (diene + dienophile) (Diels−Alder)", "functional conversion (sulfanyl to sulfinyl)", and "functional elimination (deoxygenation)" appear in the top 15 positions, highlighting that the JMC data set composition is more related to particular transformations, which are perhaps aimed at producing novel scaffolds. The presence of specific functional conversions and, in particular, of a functional elimination class among the top 15 classes describes a trend diametrically opposed to the statistics found for the Evotec ELN data set and the U.S. patent reactions.

CONCLUSIONS

Reaction classification is a complex task that has traditionally been accomplished using hand-coded rule-based approaches; however, the availability of large collections of reactions enables data-driven approaches to be developed. Building on the work of Schneider et al.,[8] we have used machine learning to develop a model capable of predicting over 300 organic reaction classes. The classification task is configured as a multitask classification problem and is trained using reactions extracted from U.S. patents, with random forests (RF) chosen as the best performing method. We have extended the previous work in a number of ways. First, we have increased the number of reaction classes that can be predicted from 50 to 336. This scaling up of the approach enables a more complete analysis of data sets to be carried out. Second, and unlike the previous approach, our workflow involves cleaning and balancing the reaction data prior to model building including reducing each reaction to only those components that change during the reaction. We believe that use of “clean” data may well have impacted positively on the scaling-up by minimizing noise in the training data. Third, we remove duplicate fingerprints from our training and test data and also ensure that there is no overlap between training and test data; thus we believe that we have created a more difficult modeling task, yet we still obtain impressive statistics. Fourth, the classifier uses a dynamic reaction fingerprint to reduce feature noise in the classification task by accounting only for the features that are described in the training data set. Finally, we also introduce a novel hierarchical reaction classification system, SHREC, which distributes the label information across four hierarchical levels and allows data sets to be browsed at different classification levels. We first demonstrated performance comparable to that seen in the literature for a smaller set of 50 reaction classes and then extended the approach to the much larger task of over 300 reaction classes.

Prediction confidence is evaluated by integrating a conformal prediction module on top of the classification model. Two confidence estimations are associated with each prediction: a confidence value that is related to the variance in the prediction; and a credibility score that is related to the separation in confidence value between the two highest scoring classes. A systematic evaluation has been carried out on the separation between true and false predictions for different credibility thresholds to enhance the performance and reliability of the model.

The classification model was used to compare two reaction data sets, one obtained from industry (the Evotec ELN) and the other from the medicinal chemistry literature (JMC), respectively. Results showed that reaction classification can be used to gain immediate insights on the nature of data sets by analyzing their confidence estimations and general class compositions, as well as providing detailed information for data analysis purposes. In particular, the analysis of the classification data revealed that the industrial data set was more focused on typical synthetic routes for molecular growing using commercial fragments, while the literature collection was more related to particular functionalizations and scaffold syntheses.

A limitation of our approach is the composition of the training data, which was derived from pharmaceutical reactions in patent and is not expected to cover organic reaction space exhaustively. This was demonstrated by the lower percentage of reactions that could be predicted reliably in the medicinal chemistry literature (around 50%) as compared to the ELN (around 85%). The training data may also have introduced some bias in our classification system because it was restricted to evaluation of the labels in the original patent set. A further potential issue is the unbalanced distribution of reaction classes in the patent set, and the use of more representative and curated training data would be expected to result in a better performing model with wider coverage. Other limitations relate to the information encoded within the reaction vector, which is used in model training. For example, the reaction vector takes no account of stereochemistry or of catalysts; thus, reactions where these characteristics are important cannot be distinguished, as highlighted in the Methods.

The reaction classifier has been developed to be fully compatible with our reaction-based de novo design tool, and we are currently exploring two applications in this context, both of which are aimed at controlling the combinatorial explosion that is inherent in de novo design. First is simply to allow the user to select preferred reaction classes during augmented de novo, for example, from a drop down list. The second approach is the use of a Reaction Class Recommender, which is able to automatically suggest preferred reaction classes based on the characteristics of a starting material. The development of the Reaction Class Recommender is the subject of a forthcoming paper.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00537.

Parameters used to build the models and the mapping between the NameRxn and the Sheffield Hierarchical Reaction Classification system (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: v.gillet@sheffield.ac.uk.

ORCID

Valerie J. Gillet: 0000-0002-8403-3111

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge funding provided by the Engineering and Physical Sciences Research Council (EPSRC) and Evotec in the form of a studentship for G.M.G and by the Biotechnology and Biological Sciences Research Council (BBSRC) and Evotec in the form of a studentship for J.W., grant no. BB/R505821/1. We also gratefully acknowledge NextMove for making the U.S. patent data available to the community and for providing us with the reaction classification data.

REFERENCES

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

(2) Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23*, 1203−1218.

(3) Watson, I. A.; Wang, J.; Nicolaou, C. A. A Retrosynthetic Analysis Algorithm Implementation. *J. Cheminf.* **2019**, *11*, 1.

(4) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **2019**, *59*, 673−688.

(5) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(6) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(7) Coley, Connor W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370−377.

(8) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39−53.

(9) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9*, 6091−6098.

(10) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs In *Computers in Chemistry. Fortschritte Der Chemischen Forschung*; Springer Berlin, Heidelberg, 1973; Vol. *39*, pp 19−64.

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

(12) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745−1756.

(13) Broughton, H.; Hunt, P.; MacKey, M. Methods for Classifying and Searching Chemical Reactions. US20030182094A1, 2003.

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

(15) Hu, Q.-N.; Zhu, H.; Li, X.; Zhang, M.; Deng, Z.; Yang, X.; Deng, Z. Assignment of Ec Numbers to Enzymatic Reactions with Reaction Difference Fingerprints. *PLoS One* **2012**, *7*, e52901.

(16) Patel, H. *Knowledge-Based De Novo Design Using Reaction Vectors*; 2009.

(17) Lowe, D. Chemical Reactions from US Patents (1976-Sep2016); https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (August 29, 2019).

(18) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists Bread and Butter. *J. Med. Chem.* **2016**, *59*, 4385−4402.

(19) Nextmove Software. NameRxn; https://www.nextmovesoftware.com/namerxn.html (August 29, 2019).

(20) Royal Society of Chemistry. RXNO: Reaction Ontologies; https://github.com/rsc-ontologies/rxno/ (August 29, 2019).

(21) 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*, 2336−2337.

(22) 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*, 3451−3479.

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

(24) Epam. Indigo Toolkit; http://lifescience.opensource.epam.com/indigo/ (August 29, 2019).

(25) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic Mapping of Atoms across Both Simple and Complex Chemical Reactions. *Nat. Commun.* **2019**, *10*, 1434.

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

(27) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, 2005; pp 323−324.

(28) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. The Application of Conformal Prediction to the Drug Discovery Process. *Annals of Mathematics and Artificial Intelligence* **2015**, *74*, 117−132.

(29) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from Toxcast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29*, 1003−1010.

(30) Ahlberg, E.; Hammar, O.; Bendtsen, C.; Carlsson, L. Current Application of Conformal Prediction in Drug Discovery. *Annals of Mathematics and Artificial Intelligence* **2017**, *81*, 145−154.

(31) King, G.; Langche Zeng, G. H. E.; Alt, J.; Freeman, J.; Gleditsch, K.; Imbens, G.; Manski, C.; McCullagh, P.; Mebane, W.; Nagler, J.; Russett, B.; Scheve, K.; Schrodt, P.; Tanner, M.; Tucker, R.; Bennett, S.; Huth, P.; Zeng, L. *Logistic Regression in Rare Events Data*; 2001.

(32) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443−4458.

(33) Bostrom, J.; Brown, D. G.; Young, R. J.; Keseru, G. M. Expanding the Medicinal Chemistry Synthetic Toolbox. *Nat. Rev. Drug Discovery* **2018**, *17*, 709−727.

(34) Campbell, I. B.; Macdonald, S. J. F.; Procopiou, P. A. Medicinal Chemistry in Drug Discovery in Big Pharma: Past, Present and Future. *Drug Discovery Today* **2018**, *23*, 219−234.