JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

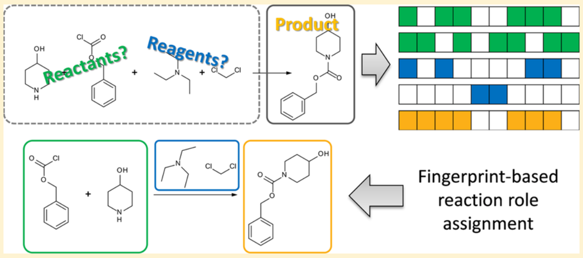
What's What: The (Nearly) Definitive Guide to Reaction Role Assignment

Nadine Schneider,*[†] Nikolaus Stiefl,[†] and Gregory A. Landrum[‡]

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

[‡]T5 Informatics GmbH, Spalenring 11, 4055 Basel, Switzerland

Supporting Information



ABSTRACT: When analyzing chemical reactions it is essential to know which molecules are actively involved in the reaction and which educts will form the product molecules. Assigning reaction roles, like reactant, reagent, or product, to the molecules of a chemical reaction might be a trivial problem for hand-curated reaction schemes but it is more difficult to automate, an essential step when handling large amounts of reaction data. Here, we describe a new fingerprint-based and data-driven approach to assign reaction roles which is also applicable to rather unbalanced and noisy reaction schemes. Given a set of molecules involved and knowing the product(s) of a reaction we assign the most probable reactants and sort out the remaining reagents. Our approach was validated using two different data sets: one hand-curated data set comprising about 680 diverse reactions extracted from patents which span more than 200 different reaction types and include up to 18 different reactants. A second set consists of 50 000 randomly picked reactions from US patents. The results of the second data set were compared to results obtained using two different atom-to-atom mapping algorithms. For both data sets our method assigns the reaction roles correctly for the vast majority of the reactions, achieving an accuracy of 88% and 97% respectively. The median time needed, about 8 ms, indicates that the algorithm is fast enough to be applied to large collections. The new method is available as part of the RDKit toolkit and the data sets and Jupyter notebooks used for evaluation of the new method are available in the Supporting Information of this publication.

INTRODUCTION

Chemical reactions are of great interest in pharmaceutical research. Large commercial databases like SPRESI,[1] Reaxys,[2] or CASREACT[3] have been compiled over the years. Building these sources of reactions has required an enormous effort in extraction, annotation, and curation. This might be one reason why the only publically available data sources that exist are small collections of reactions.[4,5] Recent advancements in text-mining technology allow us to extract chemical structures[6,7] as well as whole reaction schemes automatically from literature or patents.[8,9] This opens up a huge wealth of novel data, which can be used as input for machine-learning models or for other data-driven investigations. The automatically extracted reaction data still needs some kind of cleanup before it can reliably be used in predictive models.[10] This includes basic sanity checks during the text-mining,[8,11] elimination of duplicates, and the assignment of reaction roles (reactants, reagents, and products; see Figure 1 left). The last step is usually done by applying atom-to-atom mapping (AAM) algorithms to the extracted reactions (see Figure 1 right). These algorithms try to find corresponding atoms and bonds between reactants and products. The different approaches to this can be divided in two major subtypes: common substructure-based methods and optimization-based approaches. The former apply algorithms to identify the maximum common substructure (MCS) followed by some post processing steps to correct the remaining atoms, which are not part of the MCS. It is rather challenging for those algorithms to extract the MCS due to the changing atom- and bond-types in a reaction. The optimization-based approaches minimize the number of bonds formed and broken in a reaction. Some more recent methods combine both approaches like the ReactionMap method.[12] A detailed review of these approaches was published recently by Chen and co-workers.[13] Most approaches rely on balanced reaction schemes, meaning all atoms on the reactant side of a reaction are found on the product side. This makes the applicability of these algorithms to "crude" text-mined reactions or other real-world collections of chemical reactions like electronic lab notebooks difficult. For those reactions only the product is identified reliably while the other compounds extracted from the experimental section can often only be approximately assigned as solvent, catalyst, or reactant. For some reactions the materials used for purification might also appear in the reaction scheme. Another peculiarity with patent or literature data is that two- or multistep reactions are often found in one reaction scheme. All these issues and the noisy nature of the text-mined reaction data make it difficult to straightforwardly apply either MCS or optimization-based algorithms to assign reaction roles to the nonproduct molecules in the reaction.

Received: September 19, 2016
Published: November 22, 2016

DOI: 10.1021/acs.jcim.6b00564
J. Chem. Inf. Model. 2016, 56, 2336−2346

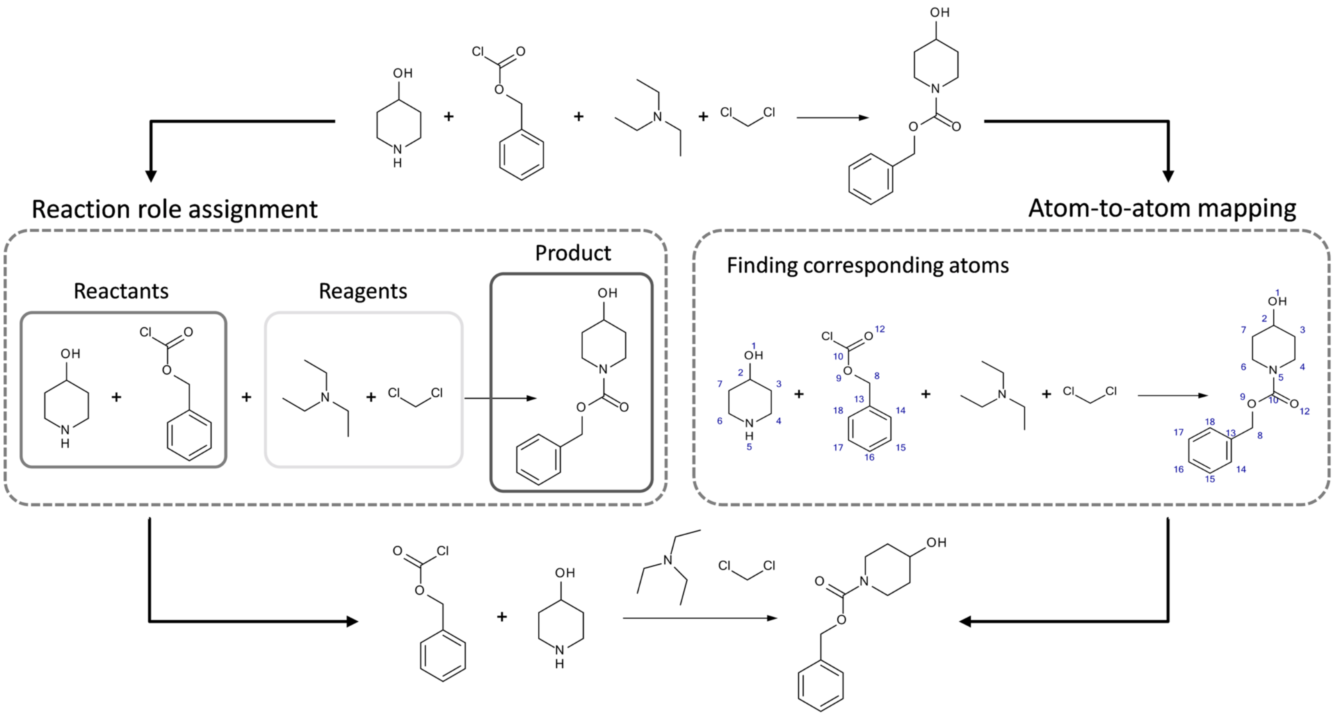


Figure 1. Illustration and comparison of reaction role assignment and atom-to-atom mapping using a simple example (patent ID: US05378712). The result of both is a reaction with correctly assigned reactants (educts) and reagents (e.g., solvent, salts, bases, catalysts, etc.).

In this paper we present an alternative fingerprint-based and data-driven approach to assign reaction roles to nonproduct molecules of a reaction. The idea for this new approach was inspired by our previous work on reaction fingerprints that do not require atom-mapping information.[14,15] This method was developed to robustly handle unbalanced and noisy reaction schemes. We show that its results are comparable to AAM approaches and that it can be applied as a preprocessing step to speed-up and improve the AAM. The outline of this paper is as follows: First, we describe two data sets for the evaluation of our method followed by a detailed description of the method. Second, we show the overall results of our method on these two data sets and compare the performance to two different AAM approaches and to a baseline model. Along with overall statistics we present detailed results on exemplary reactions and discuss the drawbacks and advantages of our method. As an outlook to the area of application of our method we show how it could be used to enhance the results of an AAM method. Finally, we conclude and give some future perspectives. The data sets and the Jupyter notebooks[16] used for the evaluation of our method are provided in the Supporting Information of this publication. The method itself is available within the RDKit toolkit.[17,18]

METHODS AND MATERIALS

In the following we first report the assembly of data sets to evaluate our new method followed by a description of our new algorithm. The method is implemented in Python and is based on the open-source cheminformatics toolkit RDKit (version 2016.03)[17] but could easily be adapted to any other cheminformatics tool that supports reactions and Morgan-algorithm type fingerprints.[19]

Data Sets. Unfortunately, no public benchmark data sets including unbalanced reaction schemes are available to evaluate atom-to-atom mapping algorithms or similar at the moment. Hence, in order to evaluate our new method we assembled two different data sets. The challenge was to construct reasonably sized, trustworthy, and diverse data sets. One of our goals was that our method should be robust enough to handle the types of unbalanced and noisy reaction schemes, which arise when using text-mining. Therefore, we used reaction schemes text-mined from US patents for both data sets.[8,11,20] The compilation of the reaction data set from patents is described in detail in a former publication.[11] In summary, we extracted about 1.3 million unique reactions from US patents from the past 40 years (1976–2015). Those were prepared using different quality criteria[11] and canonicalized[21] to identify and remove duplicates. We applied two different methods to obtain the atom-to-atom mapping: First, all reactions were atom mapped using the Indigo toolkit.[22] The input reactions to the Indigo mapper originated directly from text mining which provides the preassignment of some reagents (mainly solvents or catalysts which are unambiguous). To improve these mappings further the Indigo mapper was tweaked to ignore both charges and valency and to allow bond order changes.

Second, we used the tool NameRxn (version 2.1.84)[23] to get another atom-to-atom mapping along with the reaction types of the reactions. About 62% of the processed reactions could be assigned to a known reaction type using NameRxn.[11] This software applies a large collection of expert-defined SMIRKS patterns to the reactants of a reaction and compares the result to the product(s) of the reaction. Based on this approach NameRxn is able to provide a reliable atom-to-atom mapping of the reaction along with the reaction type assignment.

The first data set (data set A) comprises 228 diverse reaction types. For each reaction type three reactions were selected to cover different ranges of the number of reactants. This initially resulted in 684 unique reactions out of which one reaction was

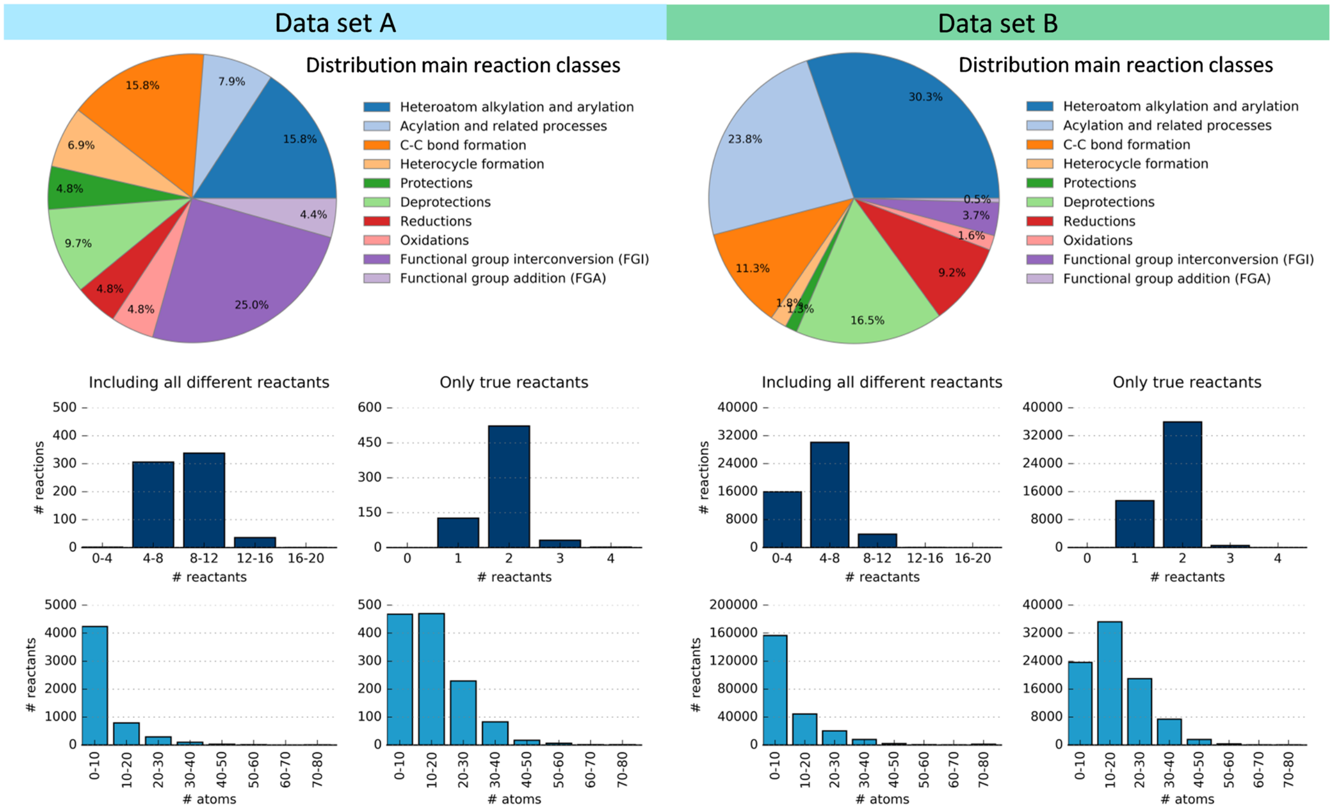


Figure 2. Composition of data sets A and B. (top) Distribution of the major reaction classes within the data sets highlighting the diversity of the reactions. (middle) Distribution of number of possible (left) and true reactants (right) per reaction. (bottom) Distribution of number of atoms per reactant considering all possible reactants (left) and only taking the true reactants into account (right).

discarded due to errors in the reaction scheme to give a final number of 683 reactions. The initial atom-to-atom mapping provided by NameRxn was manually checked and, if necessary, corrected. Figure 2 left shows the distribution of major reaction classes within data set A, the number of possible different reactants (i.e., educts and reagents) compared to the number of true reactants per reaction, and a histogram of the number of atoms in the reactants. This data set was assembled to be highly diverse, to cover edge-cases and to include noisy and rather unbalanced reaction schemes. The second data set (data set B) was constructed to represent the common reaction types in the medicinal chemist's toolkit. For this, two quality criteria were applied to the 1.3 million unique reactions before selecting randomly 50 000 reactions. First, only reactions with an assigned reaction type were considered (about 62% of the data), and second, all product atoms need to be atom mapped by both methods, NameRxn as well as by the Indigo TK. A random selection of 50 000 reactions was made from the 630 000 reactions that satisfied these criteria. Figure 2 right shows some statistics of this data set: the distribution of major reaction classes, the number of different available and true reactants, and the number of atoms per reactants. Considering the major reaction classes we found the same bias as seen for all patent reactions;[11] this set is representative of the most common reaction types used in medicinal chemistry. In both data sets the majority of reactions require two educts (Figure 2 middle diagram) while the set of possible reactants to choose from is between 4 and 12 in data set A and between 4 and 8 in data set B.

In addition to building the evaluation data sets, we used our large collection of reaction data to derive a subset of common reagents. These usually act as solvent, salt, acid, or base in a chemical reaction. To build this set, 1 559 347 unique molecules were extracted from the patent data set along with the number of times they appear in different reactions and in different reaction types. In our set of common reagents, we only included compounds which appear in more than a 1000 reactions and across at least 100 different reaction types. This resulted in a set of 86 different chemicals. Ignoring the compound grouping like, [Br−]·[K+], we were left with a set of 76 different compounds and ions (a list of SMILES is given in the Supporting Information Table S1). This set is used in a preprocessing step of our new method and is described below in detail.

Method. Our new fingerprint-based and data-driven approach to assign reaction roles is outlined in Figure 3. The method consists of four major steps: preprocessing, fingerprint generation, scoring and selection of the best reactant combination, and finally a postprocessing step to add missing reactants.

Preprocessing (Phase 1). First a list of unique molecules from the reactants side as well as one for the products side of the reaction is generated by canonicalizing[21] each of the molecules and checking for identity. This cleanup is necessary for some of the reactions which might include duplicates of solvents or salts as well as other nonreagent molecules. Another cleanup step is checking for molecules which are not modified during the reaction and which appear on both sides of the

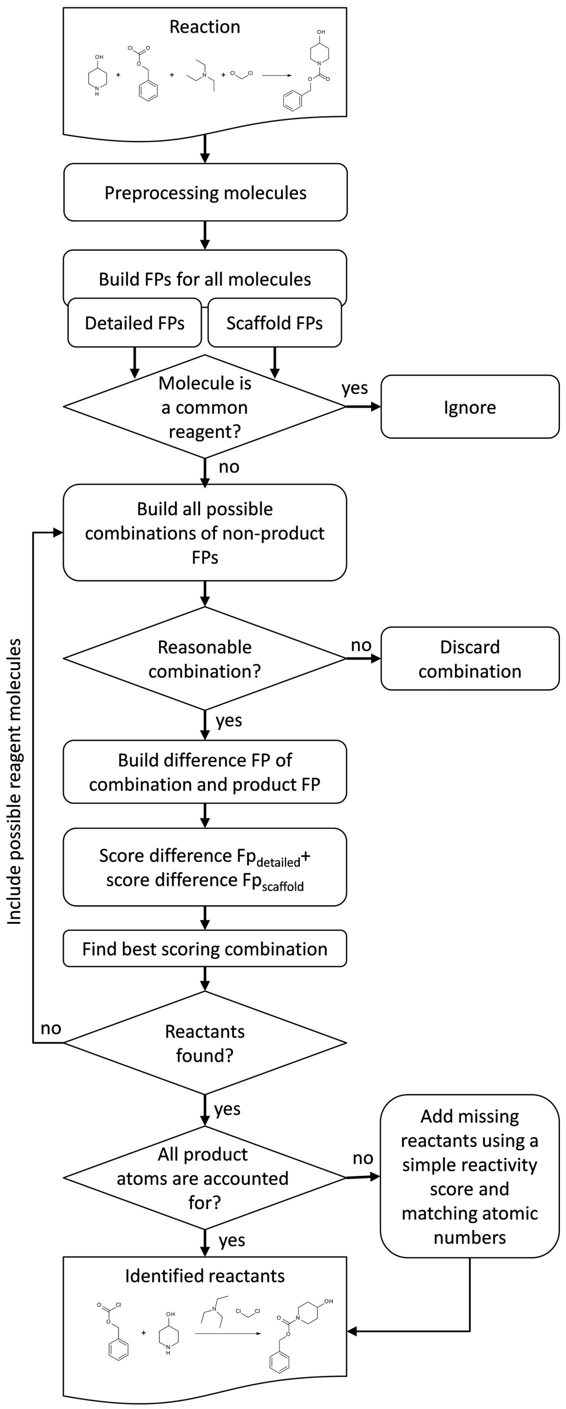

Reaction
Preprocessing molecules
Build FPs for all molecules
Detailed FPs
Scaffold FPs
Molecule is a common reagent?
yes
Ignore
no
Build all possible combinations of non-product FPs
Reasonable combination?
no
Discard combination
yes
Build difference FP of combination and product FP
Score difference Fp$_{detailed}$+ score difference Fp$_{scaffold}$
Find best scoring combination
Include possible reagent molecules
no
Reactants found?
yes
All product atoms are accounted for?
no
Add missing reactants using a simple reactivity score and matching atomic numbers
yes
Identified reactants

Figure 3. Workflow diagram of the reaction role assignment methods. Details to the different steps are given in the text.

reaction. This is not found in many reactions in our data set since almost all of our examples are unbalanced reaction schemes. Product molecules that appear on both sides of the reaction were excluded as products from further steps but were kept as possible reactants or reagents.

Fingerprint Generation (Phase 2). In this step two different kinds of count-based Morgan fingerprints[17,19] are generated for all remaining molecules of the reaction (those not discarded during the preprocessing): a detailed fingerprint and a scaffold fingerprint. The difference between these fingerprints is the selection of the atom invariants (the information used to describe the atoms when generating the fingerprint). In the detailed fingerprint the invariant considers atomic number, degree, total number of hydrogens, aromaticity, bond types, ring membership in general, and special flags for atoms in five- or six-membered rings. The scaffold fingerprint uses simpler invariants: only taking atomic numbers into account. Both fingerprints are generated as an unfolded version to avoid hash collisions. In order to focus on the local atomic environments a radius of one was used: only the direct neighbors of each atom were included in the circular fingerprints.

Selecting the Best FP Combination (Main Phase). The following workflow is done separately for each type of fingerprint (detailed and scaffold FP). The product fingerprints—if more than one product molecule is given—are combined into one fingerprint by summing up their counts. Next, reactants are filtered for common reagents using our common reagent subset derived from the patent reaction data (see above). The fingerprints of common reagents are initially ignored. For the remaining n reactant fingerprints all possible combinations are enumerated. This results in $2^n - 1$ possible combined reactant fingerprints. To ignore irrelevant combinations—those where too few or too many reactant atoms are used to build the product molecule—we discard them using the following heuristic:

$$\begin{aligned}&0.8 \times \text{number of product atoms}\\&\quad \leq \text{number of reactant atoms}\\&\quad \leq 5 \times \text{number of product atoms}\end{aligned}$$

Each of the remaining combinations is transformed into a single reactant fingerprint by summing up the counts of the fingerprints of the considered reactants. Then, the best combination is estimated by building difference reaction fingerprints[14] of the product fingerprint and the different reactant fingerprints and scoring them. The score is calculated using the following equation:

$$\begin{aligned}\text{FPScore} = \max\Bigg(&\left(1 - \frac{\sum \text{ReactionFP}_{\text{positiveCounts}}}{\sum \text{ProductFP}_{\text{allcounts}}}\right)\\&- \left(\frac{\sum \text{ReactionFP}_{\text{negativeCounts}}}{\sum \text{ReactantFP}_{\text{allcounts}}}\right)^2, 0\Bigg)\end{aligned}$$

Here, positive counts are derived by the product molecules while the negative counts result from the reactants. The first term of the score evaluates how well the product is covered by the reactants while the second term considers the size of the leaving group (the reactant atoms not present in the product molecule). Since some types of reactions make use of rather large leaving groups (e.g., the Stille reaction) the second term is weighted less than the first. The score ranges from zero to one with higher scores indicating a better match between the reactant and product fingerprints. The final score of a reactant combination is the sum of the detailed and scaffold FPScores. If no valid reactant combination is found, this phase of the workflow is repeated including the fingerprints of the common reagent molecules.

Postprocessing (Phase 4). During the scoring process we track how many and which product atoms are not accounted for by the bits of the scaffold fingerprint. Unmapped atoms are

DOI: 10.1021/acs.jcim.6b00564
J. Chem. Inf. Model. 2016, 56, 2336−2346

usually found in the products of functional group additions (halogenations, aminations, etc.) where sometimes only one atom is attached. Due to the bias toward atom economy introduced by the second term of our scoring function, the method selects one reactant instead of two in cases when the leaving group is too large. In a postprocessing step we check the remaining reagents for small and reactive molecules that could provide these missing atom(s). The reactivity calculation is based on simple criteria like the number of heteroatoms, the number of bonds between heteroatoms, or ring membership. Details of the reactivity estimation can be found in the Supporting Information (Table S2) and in the Python implementation of the method.[18]

Finally, after these four phases the algorithm proposes a set of the most probable reactants and assigns the remaining molecules as reagents. If the algorithm detects a tie (= equally scoring reactant combinations) all possible solutions are provided.

Baseline Model. In addition to our new algorithm we built a baseline model to compare to. In this model all possible reactant combinations are enumerated and the sum of the number of reactant atoms is compared to the number of product atoms. The best matching combination is chosen. If a tie occurs, the smallest combination (minimum number of reactants) is selected. In cases where more than one combination fulfills this criterion one of the possible combinations is randomly picked. We implemented this simple model to be able to compare our new algorithm to a naive approach.

Evaluation Criteria. In the evaluation of our method we compare the reactants assigned by an expert or an AAM approach to the assignment based on our method. For this we define all nonproduct molecules with at least one mapped atom as reactants and those without any mapped atoms as reagents. We calculate the number of differences between the set of reactants obtained from the experts or from the AAM approaches and the set of reactants our method proposes by computing the symmetric difference between the sets.

RESULTS AND DISCUSSION

In this section we show the results of the method described above on our validation data sets and compare to other mapping algorithms and our baseline model. Along with the overall performance we present some edge cases and examples in detail.

Performance on Data Set A. Our hand-curated data set was designed to be challenging: it includes a broad range of diverse reaction types and a large number of possible reactants per reaction (compare Figure 2 left). Figure 4 shows the overall result: in 88% of the reactions no difference is found between experts and our automatically assigned reactants; 3% of the reactions have a single reactant assigned differently from the expert selection; 8% have a difference of two reactants and only 1% of the reactions have three different reactants. The performance of the baseline model on test set A is dramatically worse: only 18% of the reactions are correctly assigned, and 53% of the reactions have one incorrectly assigned reactant (see Figure S1).

We have further broken down the results into the major reaction classes to see if there are significant interclass performance differences. Figure 5 shows the results of this analysis. Most incorrect assignments can be found in oxidations, protections, and functional group interconversions.

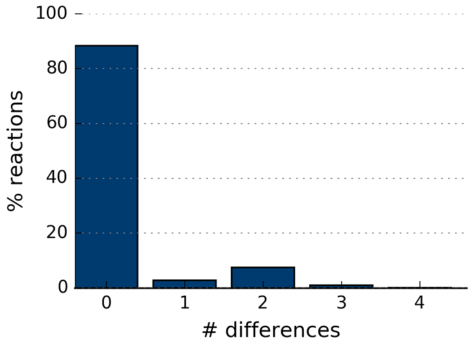


Figure 4. Result reactant assignment on data set A. The number of differences is given by the size of the symmetric intersection of the set of true reactants and the set of assigned/selected reactants. For 88% of the reactions the reactants were correctly assigned.

The FP-based assignment performs almost perfectly for deprotections, reductions, and functional group additions (Figure 5). For the remaining four classes (heteroatom alkylation and arylation, acylation and related process, C−C bond formation, and heterocycle formation) our new method correctly assigns reactants in more than 85% of the cases. We also compared this result to the baseline model (see Figure S2) and in all of the reaction classes the FP-based assignment performs significantly better. The simple baseline model is only able to assign a reasonable number of reactants correctly for reductions (88%), functional group additions (70%), and deprotections (64%). The common feature of these classes is that the product is usually produced by a small modification of the reactant. These circumstances allow a method based only on the number of atoms in the reactants and products to have a reasonable chance of picking the correct reactant. Some deprotections will fail if the protecting group is too large, leading to a large difference in the number of atoms in the reactants and products.

Figure 6 shows two of the incorrectly assigned oxidation reactions. Two typical reasons for incorrect assignment are a wrong additional reactant is selected and a wrong/nonexisting reaction type is chosen. The first type of failure is the most common type throughout all major reaction classes. The challenge here is most often when the second reactant only transfers a small number of atoms (e.g., a functional group) to the major reactant. In these cases our method might select the second reactant by the smallest possible leaving group and/or the "most reactive" molecule. In Figure 6 top the FP-based assignment method chooses the smaller thiosulfate reagent as the oxidizing agent instead of selecting 3-chloroperbenzoic acid. This leads to two differences using our evaluation criteria. In Figure 6 bottom a wrong and nonexistent reaction type is selected by the FP-based method (resulting in four differences): the transfer of an oxygen atom from the bromination agent (*N*-bromosuccinimide (NBS)) to the 5-methylthiophene-2-carbonitrile. This reaction scheme is an example of a typical unbalanced, two-step reaction that is frequently found in the reaction data extracted from patents. In this scheme reactants 1, 3, 5, and 8 belong to the first step, a bromination reaction, and the remaining five reactants are involved in the oxidation of the bromine to an aldehyde.

Another exemplary result is shown in Figure 7. In this case the fingerprint-based assignment provides two possible selections for a C−C bond formation: reactant 2 and 4 or reactant 2 and 5. Both solutions scored equally using our

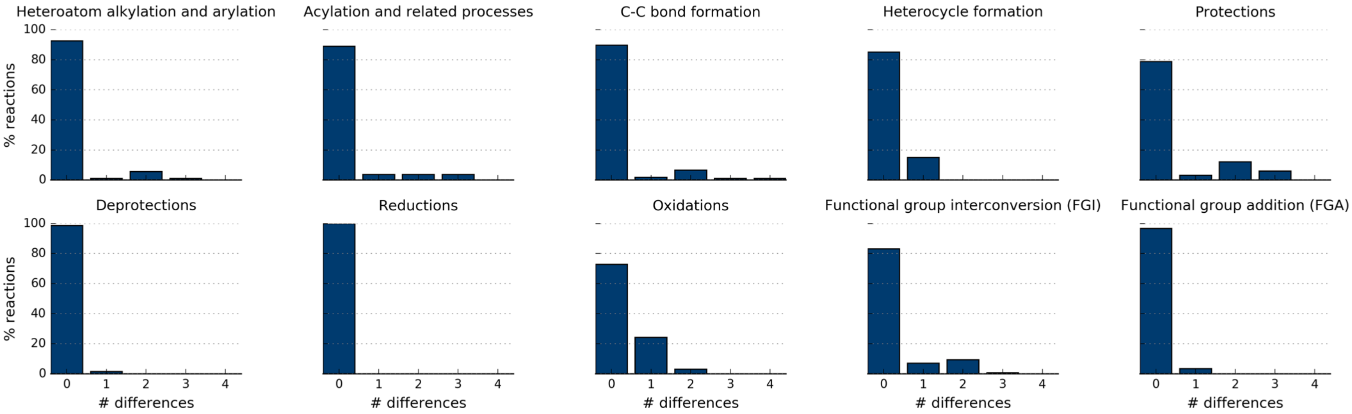


Figure 5. Result reactant assignment on data set A broken down to major reaction classes. The number of differences is given by the size of the symmetric intersection of the set of true reactants and the set of assigned/selected reactants.

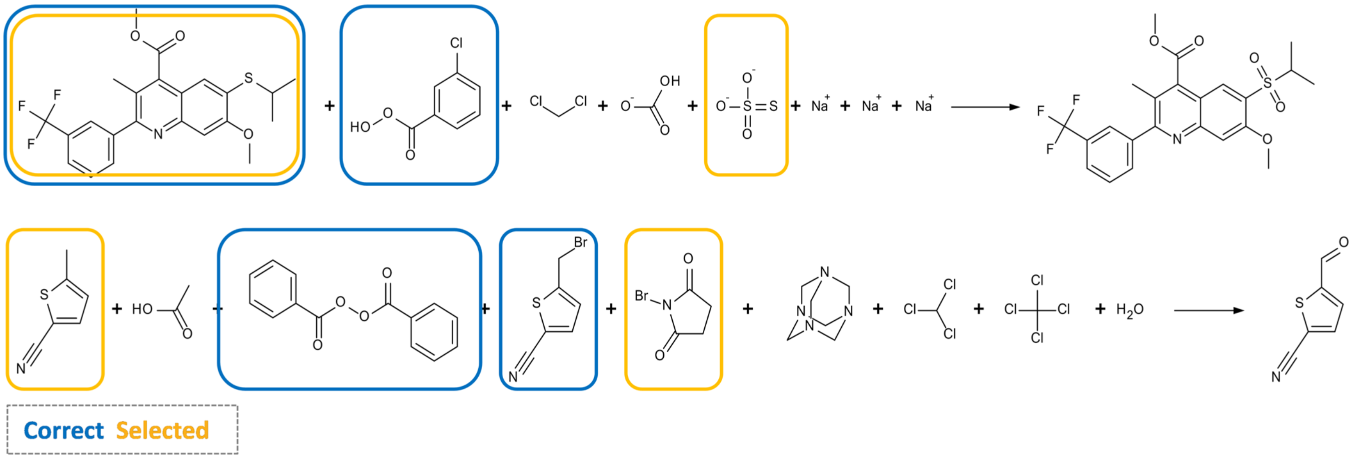


Figure 6. Examples of reactions with incorrectly assigned reactants. In blue the true reactants are marked and in yellow the ones selected by the fingerprint-based method. (top) Example of an oxidation reaction (patent ID: US08658636B2) where the FP-based method selects the smaller thiosulfate reagent as the oxidizing agent instead of selecting the correct reactant (3-chloroperbenzoic acid). (bottom) Another example of an oxidation reaction (patent ID: US05411982). In this case the FP-based method picks the wrong/nonexisting reaction type due to the multistep reaction in one scheme.

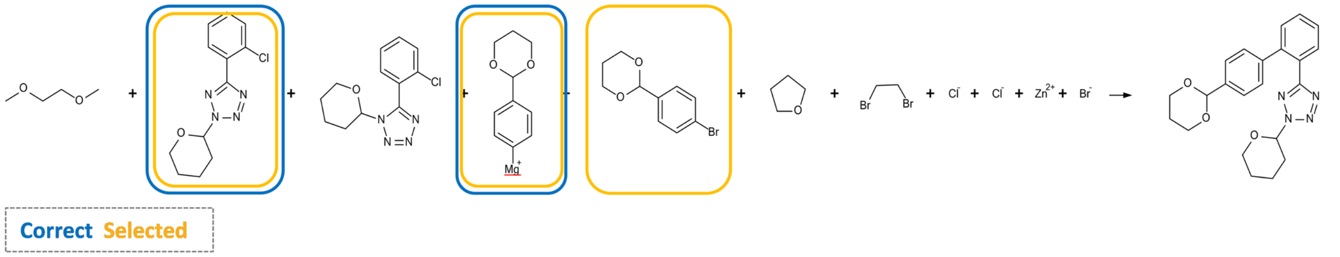


Figure 7. Example of a reaction with multiple solutions of assigned reactants (C−C bond formation reaction, patent ID: US20100152458A1). In blue the true reactants are marked and in yellow the ones selected by the fingerprint-based method. In this case the FP-based method provides two equally scoring reactant combinations: reactants 2 and 4 and reactants 2 and 5.

algorithm since the possible second set of reactants (2-(4-bromo-phenyl)-[1,3]dioxane or 4-([1,3]dioxan-2-yl)-phenylmagnesium bromide) only differ in the leaving atom (bromide or magnesium). In this scheme again a two-step reaction is shown which comprises the preparation of the 4-([1,3]dioxan-2-yl)phenylmagnesium bromide and the C−C bond formation of the latter with 5-(2-chlorophenyl)-2-(tetrahydropyran-2-yl)-2*H*-tetrazole to yield 5-(4′-[1,3]dioxan-2-yl-biphenyl-2-yl)-2-(tetrahydro-pyran-2-yl)-2*H*-tetrazole. For the statistics we counted this example as a success since our method was able to propose the correct solution. For data set A two solutions were provided in only 1.5% of the cases (see Figure S3); in 89% of those the correct solution was one of the two. No more than two solutions were proposed for any of the reactions.

Finally, we present three challenging reactions from data set A where the FP-based method performs perfectly in Figure 8. The first example in Figure 8 top shows a Bromo Suzuki-type coupling reaction with two correctly assigned reactants. Although a toluene is present as a solvent, the FP-based

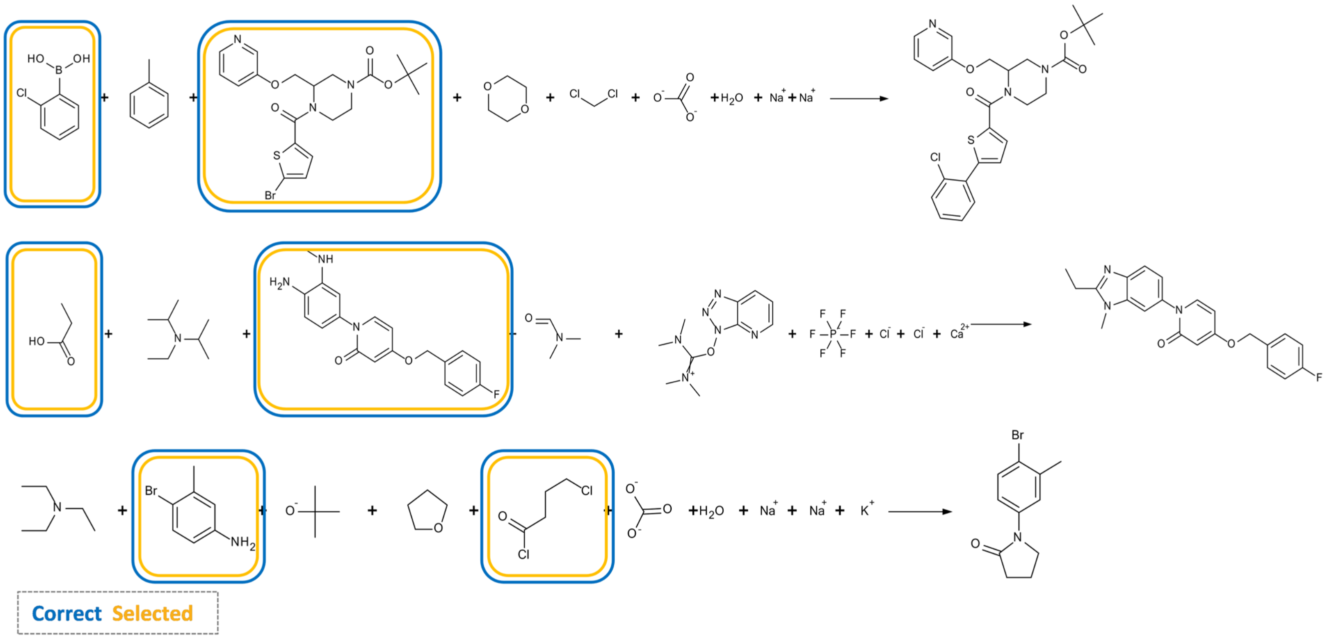


Figure 8. Examples of reactions with successfully assigned reactants. In blue the true reactants are marked, and in yellow are the ones selected by the FP-based method. (top) Example of a C−C bond formation reaction (patent ID: US08822472B2). (middle and bottom) Two examples of more complicated heterocycle formations (patent IDs: US20150018363A1 and US06156783).

approach chooses the correct chlorobenzene boronic acid as the secondary reactant. This is achieved by prefiltering the reaction for common reagents. The other two examples are both heterocycle formations, which are challenging reaction types for AAM algorithms, especially for substructure-based methods. In our case the combination of the detailed and scaffold fingerprints allows identification of the correct reactants in most of the cases for this type of reaction. In Figure 8 middle the two amino substituents of the main reactant (1-(4-amino-3-(methylamino)phenyl)-4-((4-fluorobenzyl)oxy)pyridin-2(1*H*)-one) form a benzimidazole ring with propionic acid to finally yield 1-(2-ethyl-1-methyl-1*H*-benzimidazol-6-yl)-4-((4-fluorobenzyl)oxy)pyridin-2(1*H*)-one. The correct secondary reactant (propionic acid) is selected by the FP-based approach, which does not pick atoms from the coupling reagent HATU (1-[bis(dimethylamino)methylene]-1*H*-1,2,3-triazolo[4,5-*b*]pyridinium 3-oxid hexafluorophosphate) like the substructure-based AAM algorithm of the Indigo toolkit does. In Figure 8 bottom a 4-bromo-3-methylaniline and 4-chlorobutyryl chloride form a pyrrolidin-2-one ring in the product (1-(4-bromo-3-methylphenyl)-pyrrolidin-2-one). Here again the correct secondary reagent (4-chlorobutyryl chloride) is chosen since the tetrahydrofuran is correctly identified as a frequent reagent (solvent) using our common reagent subset.

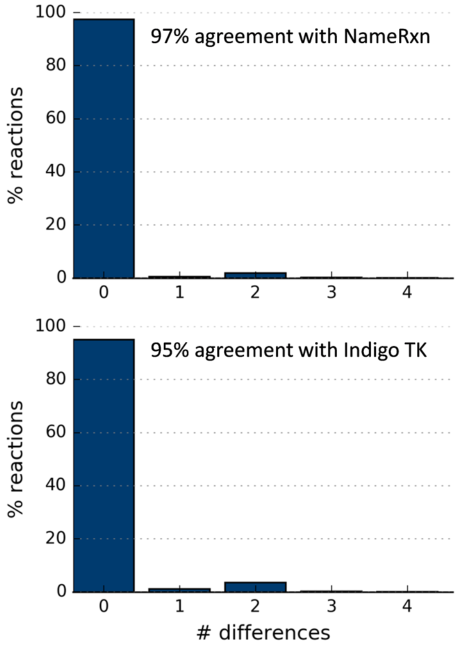


Figure 9. Result reactant assignment on data set B. The number of differences is given by the size of the symmetric intersection of the set of assigned reactants (using AAM) and the set of assigned reactants using the FP-based method. (top) Compared to reactant assignment using the NameRxn mapping (97% agreement). (bottom) Compared to reactant assignment using the Indigo TK mapping (95% agreement).

Performance on Data Set B. Data set B was constructed to represent the common reaction types found in the medicinal chemist's toolkit. These reactions are most likely encountered when assigning reaction roles for reactions extracted from patents or electronic laboratory notebooks (ELN). In this section we compare the assignment of reaction roles of the FP-based method to the assignment provided by two AAM algorithms: NameRxn and the Indigo toolkit. Figure 9 shows the overall results of this comparison. In 97% of the reactions (48 709) the FP-based methods agrees with the reaction role assignment using the AAM provided by NameRxn. Comparing the assigned reactants using the AAM of the Indigo toolkit to the FP-based method we found the same assignment for 95% of the reactions (47 501). The detailed results broken down to the ten major reaction classes can be found in Figure S4. We also applied our baseline method to this data set and compared the results to reactant assignment based on the NameRxn AAM.

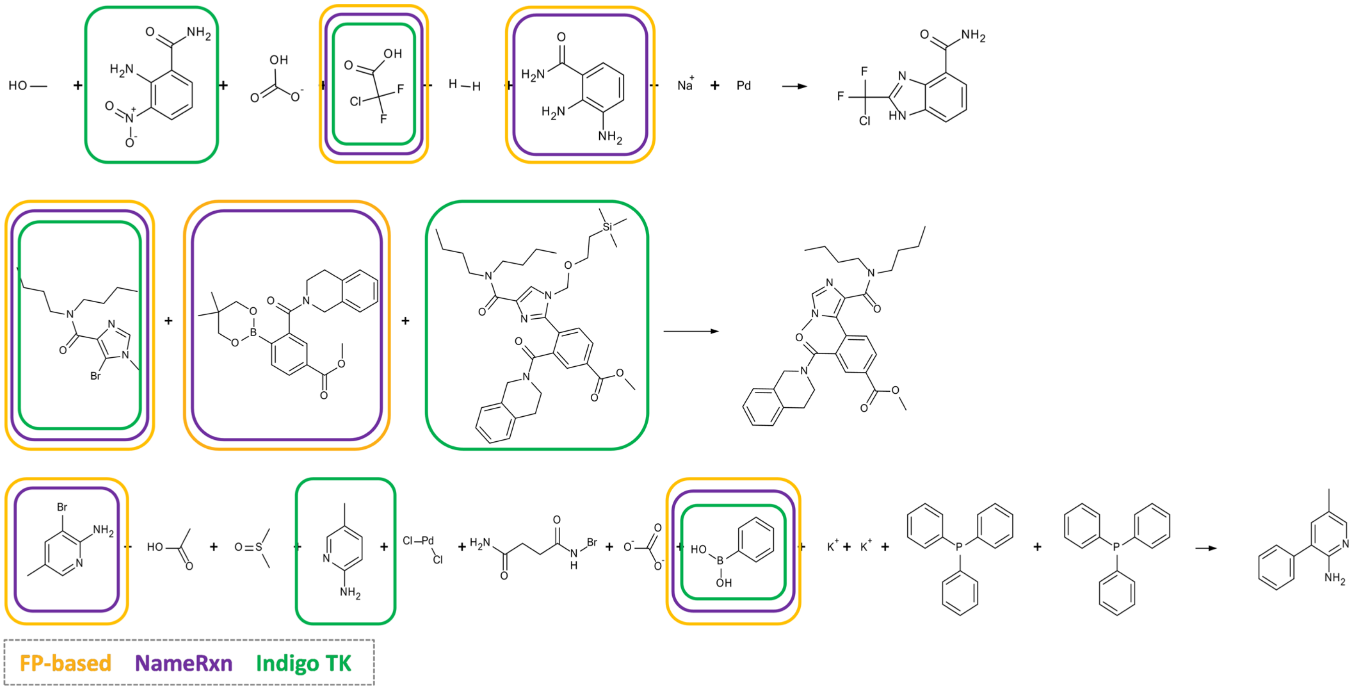


Figure 10. Examples of reactions with successfully assigned reactants using the FP-based method. In yellow reactants are marked which were assigned using the FP-based method, in violet the ones are highlighted selected by the NameRxn mapping and in green the selection done using the Indigo TK mapping is shown. (top) Example of a heterocycle formation (patent ID: US20110218103A1). (middle) Examples of C−C bond formation (patent ID: US20140135318A1). (bottom) Another example of a C−C bond formation (patent ID: US07572791B2).

This results in only 36% agreement (see Figure S5). On that data set the baseline method achieves a reasonable performance of more than 90% correctly assigned reactants for reduction reactions but considering the other nine classes we found more than 50% agreement with NameRxn only for deprotections and oxidations (see Figure S6).

In Figure 10 we show three exemplary reactions from data set B: one heterocycle formation and two C−C bond formations. In all three examples the FP-based method and NameRxn assign the correct reactants, only the Indigo toolkit MCS-based approach shows deviations due to the noisy reaction data. The first example (Figure 10 top) is once again a two-step reaction: in the first step 2-amino-3-nitrobenzamide is reduced to 2,3-diaminobenzamide before the latter reacts with chlorodifluoroacetic acid to form the final product (2-[chloro(difluoro)methyl]-1*H*-benzimidazole-4-carboxamide). The second example shows another type of noise/problem sometimes occurring in the patent reactions. In the preparation description of the reaction to yield methyl-4-(4-(dibutylcarbamoyl)-1-methyl-1*H*-imidazol-5-yl)-3-(1,2,3,4-tetrahydroisoquinoline-2-carbonyl)benzoate an analogue of the product molecule was mentioned (third reactant Figure 10 middle) which is not part of the current reaction. This leads to confusion of the MCS-based method. The third reaction in Figure 10 bottom is another example of a two-step reaction where 2-amino-5-picoline (fourth reactant) is first brominated to 2-amino-3-bromo-5-methylpyridine and then reacts with phenylboronic acid to form 2-amino-3-phenyl-5-methylpyridine. This example shows that the FP-based method is able to identify subtle differences between the reactants (bromine atom) and chooses the correct one.

On data set B we also measured the time required to run the FP-based methods to assign reaction roles. The median computing time for a reaction in this data set was 8 ms (mean 8 ms; min 3 ms; max 12 s; machine Intel Xeon 64 bit, 3.6 GHz, 8 core processor). This is rather fast compared to the MCS-based approaches like the one implemented in the Indigo toolkit (mean 3.6 s, min 0.06 s, max 60 s; using 100 randomly selected reactions of data set B and measuring the performance in KNIME[24]) or ReactionMapper (compare ref 12).

Application as a Preprocessing Step to AAM. Finally, we show the application of our FP-based algorithm as a preprocessing step to an AAM algorithm to speed-up the calculation and to improve the results by providing the assignment of reaction roles. For this we applied the Indigo AutoMapper KNIME (version 3.1.3) node to data set B. The difference here to our former Indigo AAM is that now all nonproduct molecules are on the reactant-side of the reaction, no solvents or catalysts were preassigned due to the text mining. We again tweaked the Indigo AAM by selecting to ignore charges and valency in the configuration of the node. This resulted in a lower agreement of 87% between our FP-based reaction role assignment and the one based on Indigo's AAM as well as between the reactant assignment using NameRxn and the Indigo AutoMapper (Figure S7). Most of the differences were obtained in oxidation, reduction, and C−C bond formation reactions (Figure S8). Furthermore, the product molecules of 1735 reactions were not completely mapped which indicates an incorrect AAM. These reactions were preprocessed using our new FP-based method and subsequently atom mapped using the Indigo AutoMapper in KNIME. As criteria to measure the improvement of the AAM we calculated the percentage of unmapped atoms in the assigned reactants of a reaction; this should be small since it indicates the size of the leaving groups. Another criterion is the percentage of mapped atoms in the products; in a correct AAM this should be 100%. The results of this test are shown in Figure 11. Both criteria were improved by preprocessing the

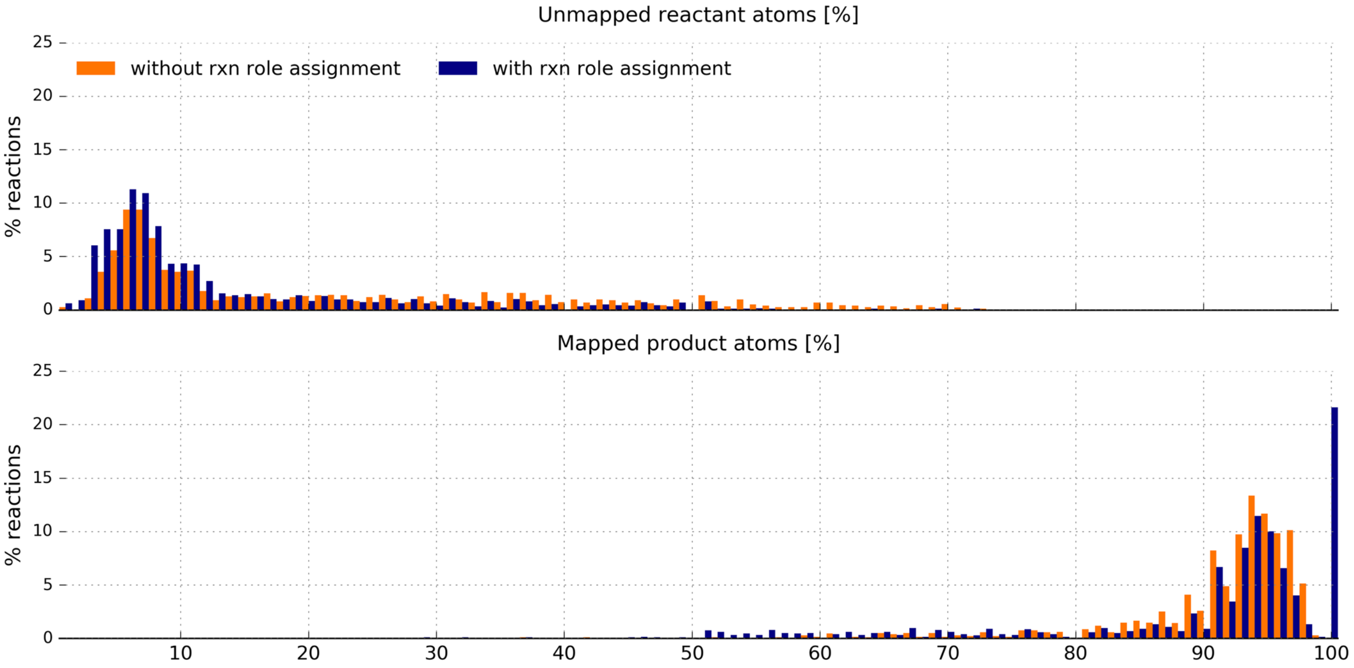


Figure 11. Percentage of unmapped reactant atoms and mapped product atoms for a set of 1735 reactions having an incomplete AAM for the products (<100% mapped product atoms). The results are binned into 1% bins The AAM was obtained using the Indigo AutoMapper in KNIME. (orange) Results without the reaction role assignment preprocessing. (blue) Results of the AAM after preprocessing the reactions using the FP-based method.

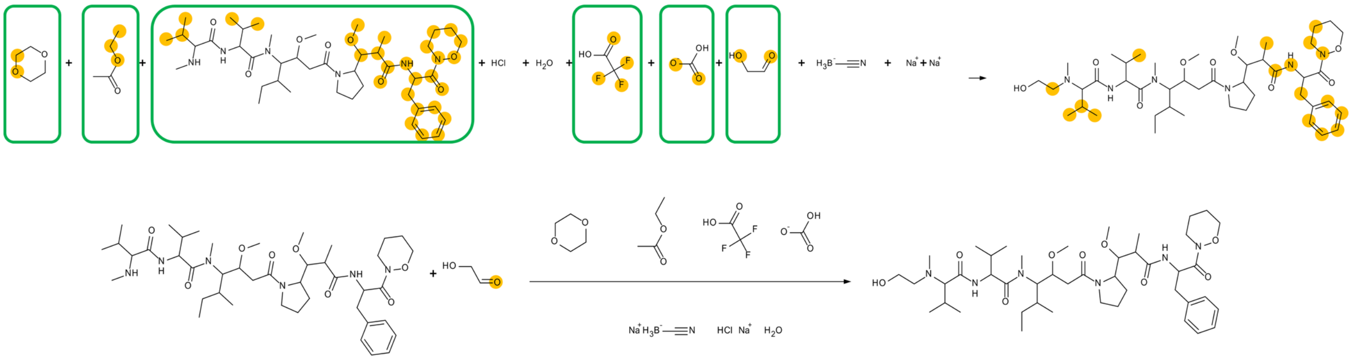


Figure 12. Exemplary aldehyde reductive amination reaction (patent ID: US08722629B2). The green boxes show reactants with mapped atoms, atoms without a mapping number are highlighted in yellow. (top) Result AAM without reaction role preprocessing. (bottom) AAM result after the preprocessing was applied.

reactions using our FP-based method (compare Figure 11 top and bottom). The median percentage of unmapped reactant atoms decreases from 13.3% to 7.7%. Although the median percentage of mapped product atoms stays constant at 93% we could shift the histogram to a more accurate mapping: 375 reactions exhibit a percentage of 100% product atoms mapped now compared to 1 before (compare blue peak Figure 11 bottom left).

Figure 12 shows a detailed example for the AAM improvement of a reaction resulting from the reaction role assignment preprocessing step. In this aldehyde reductive amination reaction the AAM is confused by the presence of acid, solvent and salts in the reaction scheme (see Figure 12 top). This leads to the selection of atoms from six different compounds, leaving a large number of atoms unmapped in reactants and products. Applying our FP-based method to this reaction the correct two reactants were assigned and the other nonproduct molecules were annotated as reagents. This drastically improves the Indigo AAM for the reaction: unmapped product atoms are no longer present.

In addition to an improvement in the quality of the AAM we also achieved a 2 orders of magnitude improvement in runtime. On the 1735 reactions tested the mean computation time for the AAM was only 0.08 s (min 4 ms; max 2.6 s) compared to 3.6 s before (see above). Our FP-based reaction role assignment is a useful preprocessing step to improve the AAM of a reaction.

■ CONCLUSION

We presented a new FP-based approach to assign reaction roles to the nonproduct molecules of a reaction. Our method can be used as an alternative to the more complicated and time-consuming atom-to-atom mapping algorithms when only the reaction roles need to be assigned. Evaluating the method on different data sets we could show that it is efficient, reliable and robust to unbalanced and noisy reaction data extracted from patents using text-mining. It could also be applied as a preselection step before an AAM algorithm is used in order to lower the error-rate in noisy reaction schemes and to save time due to the smaller number of potential reactants.

As another possible application we also have used the new method to cleanup the unclassified reactions in our data set. This allowed us to shed some light on these reactions by building reaction fingerprints, clustering them and investigating those clusters for novel reaction types. In this way we have identified almost 50 different heterocycle formations. We also found a substantial number of multistep reactions as shown in some of the examples above. In the next version of our approach we intend to include the detection of multisite reactions even if the stoichiometry of the given reaction is not correct. Further we plan to implement a new AAM method using the presented approach combined with reaction fingerprints to identify the reaction center.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00564.

Further plots and details discussed in this study (PDF)
Data sets used in this study (ZIP)
Jupyter notebooks to evaluate the new method described in this study (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: nadine-1.schneider@novartis.com.

ORCID

Nadine Schneider: 0000-0001-5824-2764
Gregory A. Landrum: 0000-0001-6279-4481

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Daniel Lowe and Roger Sayle of NextMove Software for providing the patent reaction data set and for valuable and inspiring discussions. The authors thank Nikolas Fechner for many critical discussions. The authors also wanted to thank Michael Tarselli for his help with the manual curation of data set A and Anna Pellicciolі and Brian Kelley for critical proofreading of the manuscript. N. Schneider thanks the Novartis Institutes for BioMedical Research education office for a Presidential Postdoctoral Fellowship.

■ ABBREVIATIONS

FP, fingerprint; MCS, maximum common substructure; AAM, atom-to-atom mapping; TK, toolkit

■ REFERENCES

(1) SPRESI Database. http://infochem.de/products/databases/spresi.shtml (accessed July 17, 2016).

(2) Reaxys Database. http://www.elsevier.com/online-tools/reaxys (accessed July 17, 2016).

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

(4) ChemSpider SyntheticPages Database. https://cssp.chemspider.com/ (accessed July 17, 2016).

(5) Webreactions Database. http://www.openmolecules.org/webreactions/index.html (accessed July 17, 2016).

(6) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: a Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220−D1228.

(7) IBM Contributes Data to the National Institutes of Health to Speed Drug Discovery and Cancer Research Innovation. http://www.prnewswire.com/news-releases/ibm-contributes-data-to-the-national-institutes-of-health-to-speed-drug-discovery-and-cancer-research-innovation-135275888.html, 2011 [accessed July 17, 2016].

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

(9) Lowe, D. M.; Sayle, R. A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition. *J. Cheminf.* **2015**, *7* (Suppl 1), S5.

(10) Sayle, R. A.; Lowe, D. Standardized Representations of ELN Reactions for Categorization and Duplicate/Variation Identification. Presented at the 247th ACS National Meeting, Dallas, TX, March 16−20, 2014; Paper CINF-1. http://www.slideshare.net/NextMoveSoftware/standardized-representations-of-eln-reactions-for-categorization-and-duplicatevariation-identification (accessed August 2, 2016).

(11) 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* (9), 4385−4402.

(12) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **2013**, *53*, 2812−2819.

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

(14) 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* (1), 39−53.

(15) Sayle, R. A.; Griffen, E.; Kogej, T.; Drake, D. Efficient Searching and Similarity of Unmapped Reactions: Application to ELN Analysis. Presented at the 243rd ACS National Meeting, San Diego, CA, March 25−29, 2012; Paper CINF-1; http://www.slideshare.net/NextMoveSoftware/filbert-19143105 (accessed August 2, 2016).

(16) Pérez, F.; Granger, B. E. IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* **2007**, *9* (3), 21−29.

(17) Landrum, G. A. RDKit: Open-Source Cheminformatics Software [Online], version 2016.03. *Zenodo* **2016**, DOI: 10.5281/zenodo.58441.

(18) Reaction Role Assignment Code on GitHub. https://github.com/rdkit/rdkit/tree/master/Contrib/RxnRoleAssignment (accessed November 11, 2016).

(19) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(20) Patent data. http://nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/, https://bitbucket.org/dan2097/patent-reaction-extraction/downloads (accessed July 17, 2016).

(21) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55* (10), 2111−2120.

(22) Epam Life Sciences. Indigo Toolkit [Online], version 1.2.2beta-r13. http://lifescience.opensource.epam.com/indigo (accessed July 17, 2016).

(23) NameRxn, version 2.1.84. NextMove Software Limited, 2015. https://www.nextmovesoftware.com/namerxn.html (accessed on July 17, 2016).

(24) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The

Konstanz Information Miner. *In Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer: New York, 2007.