Atom-to-atom mapping: a benchmarking study of popular mapping algorithms and consensus strategies

Arkadii Lin [a], Natalia Dyubankova [b], Timur Madzhidov *,[c], Ramil Nugmanov [c], Assima Rakhimbekova [c], Zarina Ibragimova [c], Tagir Akhmetshin [a, c], Timur R. Gimadiev [d], Rail Suleymanov [e], Jonas Verhoeven [b], Joerg Kurt Wegner [b], Hugo Ceulemans [b] and Alexandre Varnek *, [a]

[a] *Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg*
4, Blaise Pascal str., 67081 Strasbourg, France
**e-mail: varnek@unistra.fr*

[b] *Janssen Pharmaceutica*
30, Turnhoutseweg, 2340 Beerse, Belgium

[c] *Laboratory of Chemoinformatics and Molecular Modelling, A.M. Butlerov Institute of Chemistry, Kazan Federal University*
18, Kremlyovskaya str., 420008 Kazan, Russia
**e-mail: tmadzhidov@gmail.com*

[d] *Institute for Chemical Reaction Design and Discovery, (WPI-ICReDD), Hokkaido University*
Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan

[e] *Arcadia Inc., 28 k2, Bolshoy Sampsonievskiy pr., St. Petersburg, 194044, Russia*

Abstract:

Here, we discuss a reaction standardization protocol followed by a comparison of popular Atom-to-atom mapping (AAM) tools (ChemAxon, Indigo, RDTool, NextMove and RXNMapper) as well as some consensus AAM strategies. For this purpose, a dataset of 1851 manually curated and mapped reactions was prepared (the *Golden dataset*) and used as a reference set. It has been found that RXNMapper possesses the highest accuracy, despite the fact that it has some clear disadvantages. Finally, RXNMapper was selected as the best tool, and it was applied to map the USPTO dataset. The standardization protocol used to prepare the data, as well as the data itself are available in the GitHub repository https://github.com/Laboratoire-de-Chemoinformatique.

.

Keywords: Atom-to-atom mapping, Benchmarking, Chemical reactions, Big Data, Standardization, Reaction data curation

1 Introduction

Performance of any data-driven model largely depends on the data quality. Duplicates and outliers may confuse the model, and non-standard data representation may decrease the accuracy of the model's predictions. For example, the presence of aromatic and Kekule resonance structures can cause differences in descriptor values for chemically identical molecules [1].

Chemical structures standardization became a regular procedure in every Quantitative Structure-Activity/Property Study (QSAR/QSPR). The algorithm of a such cleaning process has been well discussed by D. Fourches et al.[1] The authors explained that the structure standardization process must be done as a sequence of several steps resembling a funnel. At each step of this funnel, chemical structures are discarded, to the end that only unique, chemically valid and fully standardized molecules remain. This workflow ensures that the machine operates with chemically sensible information, and no obvious chemical mistakes will impact the predictions.

Standardization workflows for molecular datasets are now easy to implement and well supported by several frameworks, such as RDKit [2], ChemAxon [3] and recently developed CGRtools toolkit[4]. A classical workflow may include de-aromatization, functional groups standardization, generation of a major tautomer, aromatization, removing salts, handling of radicals and isotopes, and eventually other project specific rules. Here, each subsequent rule works with a new molecular representation produced by the previous rule which makes the protocol order dependant.

The extensive discussion of standardization workflows in the literature[1,5–7] and existence of some published data standardization workflows for molecular datasets[8–10] makes the task of molecular data curation quite straightforward. In contrary, chemical reactions stayed aside this topic. In 2017, the database of US patents was parsed and published by D. Lowe [11]. The application of text-mining techniques revealed more than 3.75 million records. This data set became a very popular source of reaction data, and several projects have already used this database to train models [12–17]. However, little attention was paid so far to the quality of the data representation in this dataset. Although the reactions are provided via canonical SMILES, duplicate entries as well as a non-standard representation of the functional groups may mislead models used for retrosynthetic rule extraction and/or decrease the accuracy of reaction conditions prediction. For instance, in the work of Coley *et al*.[18], data preparation mainly consisted of structure standardization performed separately for reactants and products using RDKit facilities [2]. In our opinion, an more exhaustive data preparation is required. Also, atom-to-atom mapping (AAM) [19] is required for automated reaction templates extraction/application [20], which should be implemented as a part of the standardization pipeline. We believe that this is a serious omission, and, which is more important, it is not clear which tool is the most accurate.

Several benchmarking studies on the performance of different AAM tools have already been carried out [21–24]. However, only few publications reported comparisons of tools using manually mapped dataset as a ground truth[15,23]. The rest were compared using the principle of Minimum Chemical Distance[25] (number of formed and broken chemical bonds in the course of a reaction should be as low as possible).

In this work, we aim to extend the discussion on reaction standardization in the context of chemical reaction structures. Also, we present a benchmarking study to assess the performance of the most popular AAM tools and their underlying consensus. Finally, we provide our reader with the USPTO database standardized and mapped according to the decisions made in this work.

2 Method

2.1 Reactions standardization protocol

It is clear that a structure standardization must be done before any modelling task is run, and a regular standardization pipeline including basic rules such as dearomatization/aromatization, functional groups transformation and generation of the major tautomer. In the context of reaction data, this needs to be extended and adapted to the specifics of chemical reactions. The pipeline that we propose was implemented in Python3[26] language using the CGRtools framework[4]. It includes 11 steps ordered according to the principle “the rule that discards more compounds is posed earlier”:

1) *Functional groups transformation*

First, the workflow tries to read a reaction and attempts to interpret its parts as molecules (in other words, to check if it is a chemical record). It checks if molecular graphs can be reconstructed properly using the given data, and then searches for specific structural patterns in reactants, reagents, and products that represent functional groups. If any pattern is found, the corresponding part of the structure is transformed to its standard representation. Currently, 31 functional groups are included in CGRtools. An example of nitro group standardization is shown in Figure 1. The full list of the available functional groups can be found in the paper of R. Nugmanov et al.[4]

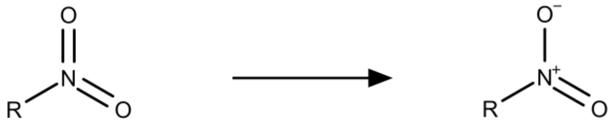

Figure 1. An example of a nitro group standardization.

2) *Dearomatization*

Several algorithms of structure aromatization exist, and at least 4 algorithms can be found in the ChemAxon’s documentation[3], some of which strictly follows the Huckel’s rule[27]. CGRtools has an advanced dearomatization approach that can handle most heterocycles and aromatic groups. However, sometimes (mostly when information on hydrogens position in heterocycles is lost) it is impossible to univocally reproduce molecular structure. Thus, all the chemical structures were transformed to the Kekule form in the beginning to achieve the consistency within the dataset. Molecules (and, hence, reactions where these molecules can be found) that could not be transformed were deleted from the dataset.

3) *Valence checking*

Next, the pipeline checks whether molecules comprising a reaction possess any atom with an incorrect valence. If so, such reactions are removed, and the corresponding record is written to a log file.

4) *Generation of the major tautomer*

Generation of a major tautomer plays an important role in duplicates cleaning since the same molecule might be in different tautomeric forms. To avoid such miscomprehensions, the protocol generates the canonical tautomer for each reactant, reagent and product using ChemAxon's Standardizer tool[3]. Since the latter produces errors when applied to reactions, our wrapper of Standardizer applies tautomerization to molecules in reactions one by one.

5) Removal of explicit hydrogens

Once the major tautomer is selected, the explicit hydrogens are removed.

6) Clear radicals

A reaction containing radicals is usually a consequence of errors arising from text mining techniques (i.e. wrong parsing of a molecule description in a text that leads to incorrect chemical structure recognition). Improving of text parsing techniques might reduce the number of such cases. Till that, we propose to automatically discard all the reactions with radicals to avoid any potential mistakes.

7) Clear isotopes

Isotope labelling is a useful tool in organic chemistry. It is successfully applied in structures determination and reaction mechanism detection.[28] However, presence of isotopes does not substantially influence reaction property and thus could be ignored. At the same time, some chemoinformatics tools, such as atom-to-atom mappers, ignore this information. This, in turn, might lead to unexpected results. Also, it is unclear how to use the knowledge extracted from such reactions considering that their number is small. In proposed workflow, the user might choose one out of three options: (i) ignore isotopes and leave information as is, (ii) clean them (to remove the isotope label) and (iii) discard a reaction if an isotope was found. In this particular application, we propose to discard these reactions.

8) Split of ions

Reaction data often comes in the form of a reaction SMILES[29]. It is a quite compact format that might contain reactants and products as well as reagents. Different molecules are separated by a dot symbol '.', and different parts of a reaction by an angle bracket '>'. However, dot symbol in reaction SMILES can mean either separate molecules or ions of the same compound. From practical point of view, this is not a proper way to store reactions, and ionic compound should be represented as a single molecule. Unfortunately, some AAM tools reset the standard reaction representation inside applying their own pre-processing that cannot be turned off. This makes the ions regrouping impossible (as it will be shown further, RXNMapper turns a reaction into SMILES and then disorders the ions). Due to that, we concluded that all the ionic compounds must be artificially split to counterions as separate chemical structures if it was not given in the form of a reaction SMILES before. Anyway, the workflow proposed has three options: (i) ignore ion compound and leave ions as is, (ii) merge counterions in the same molecular graph with disconnected components, (iii) split ions in a way that every ion is considered as separate molecule.

9) Aromatization

As it is already mentioned before, there are several aromatization algorithms implemented in different tools. In this work, we consider the aromatization step only as a part of standardization of data representation, and, thus, do not claim the absolute correctness of aromaticity perception in chemical sense. Having that, we follow the Huckel's rule[30] and consider aromatization as an alternation of single and double bonds in a cycle. Hence, cycles with exocyclic double bond cannot be aromatic. If aromatization can cause ambiguity in position of hydrogen (pyrrole-like heterocycles) we keep all the pyrroles in the Kekule form.

10) Removal of unchanged parts

Due to incorrect text mining and/or data storage technique, reagents, solvents and catalysts can be attributed to the reactant part of a reaction. It is used as default setting for the USPTO reaction dataset. It is obviously incorrect and can cause some problems when data are used for retrosynthetic rule extraction, prediction of reaction characteristics, similarity search, etc. However, it is not easy to decide whether particular molecule represents a reagent or reactant. One of the possible solutions applied here is attribute molecule to a reagent/solvent/catalyst if it is in a list of "unwanted chemistry".[31] The disadvantage of this approach is that such list must be prepared beforehand by an expert, and it must be updated permanently depending on a project. For the time being, we rely on an assumption that molecules that present both in reactants and products with no changes are not reactants, and, hence, can be easily moved to the reagents part.

11) Removal of duplicates

The removal of duplicates is a regular data cleaning procedure that is devoted to keeping only unique entities. Depending on the task, it might take into account or ignore stereochemistry as well as considering reaction conditions (reactants/solvents/catalysts if they are not specified in left-hand side of reaction equation). Besides, this step needs some adaptation for identifying duplicates in the dataset since its direct application might still overlook some duplicate reactions. For instance, depending on how duplicate removal is done, two reactions might be recognized as two different ones if the reactants are ordered differently (see an example in Figure 2).

In this workflow, we propose to discard the duplicates basing on the canonical order of the reactants, reagents and products in a reaction SMILES, which is done using CGRtools package. The output of the standardization workflow is stored in an MDL RD file that encompasses structure data in connectivity tables form as well as unlimited meta data (e.g. reaction conditions, reaction ID, etc.).

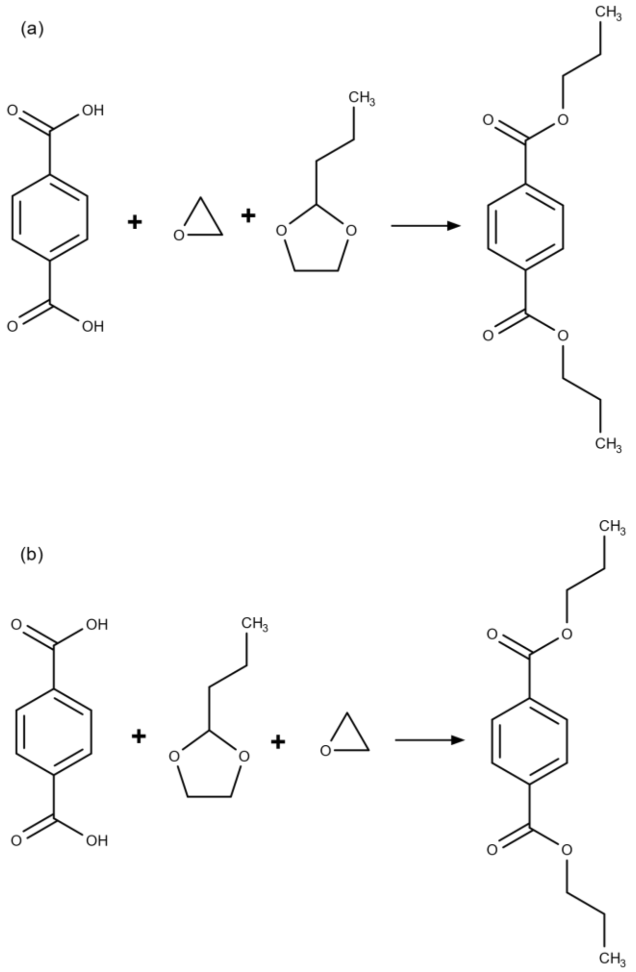

Figure 2. An example of the reaction duplicates. Here, the reaction of esterification is represented where the order of the oxirane and 2-propyl-1,3-dioxolane differs.

2.2 Atom-to-atom mapping: a benchmarking study

Atom-to-atom mapping (AAM) establishes correspondence between atoms of reactants and products. This allows the identification of the transformation occurred in the reaction. AAM is required for reaction substructure[32] and similarity search[33,34], retrosynthetic rules extraction for synthesis planning[35], reaction classification[36], etc. Thus, wrong AAM will lead to errors in downstream tasks, and it will reduce performance of tools related to reaction mining. Thus, we paid special attention to the quality of AAM.

Several public and commercially available tools are performing an AAM exist. However, it is not quantified which one performs better with respect to mapping accuracy. To answer this question, we have decided to perform a benchmarking study of available AAM tools: ChemAxon Standardizer[3], NextMove[37], Indigo[38], RDTool[39], and RXNMapper[15]. RDTool is an open-source software which is based on the internal Maximal Common Substructure (MCS) search. It employs three mapping algorithms: MIN, MAX and MIXTURE (see the details of the algorithms in publication[39]). The tool itself is a consensus of these three algorithms, where the best mapping scheme for the particular reaction is selected according to 8 metrics (minimal energy of bonds formation/breaking, number of structural fragments, etc.; see the full list in paper[39]). In our experiments, it was found that RDTool is slow, and it is not applicable to all chemical reactions. Therefore, the code was revised, bugs were fixed, and some algorithm optimization performed. On top of it we added the functionality to call a particular mapping algorithm. Thus, the initial version of the RDTool will be mentioned further as “old RDTool”, and the new (revised) version of it will be called “new RDTool”. Also, it was decided to add the MIN, MAX and MIXTURE mapping algorithms to the benchmarking study as separate AAM strategies.

In addition, it was decided to add some consensus strategies to the benchmarking study as well (besides the consensus that is done within the RDTool mapper). Comparing several mapping schemes for the same reaction we select the best one which is returned (we called it “consensus mapping”[40]). In RDTool consensus selection of best mapping requires calculation of 8 different metrics. Instead, we used a concept of the Chemical Distance (CD). Here, the CD is the number of dynamic bonds (formed, broken or with changed bond order) for a particular reaction. It is computed from a Condensed Graph of a Reaction (CGR)[41,42] that represents the entire reaction as a single pseudo-molecule (see an example in Figure 3). Briefly speaking, the latter condenses the reactants and products into a new pseudo-molecule connecting the atoms via dynamic (formation of a single bond, reduction of a double bond to a single one, etc.) as well as static (single, double, triple, aromatic) bonds. Once a reaction is represented as a CGR, the corresponding reaction center can be extracted, and its CD can be computed.

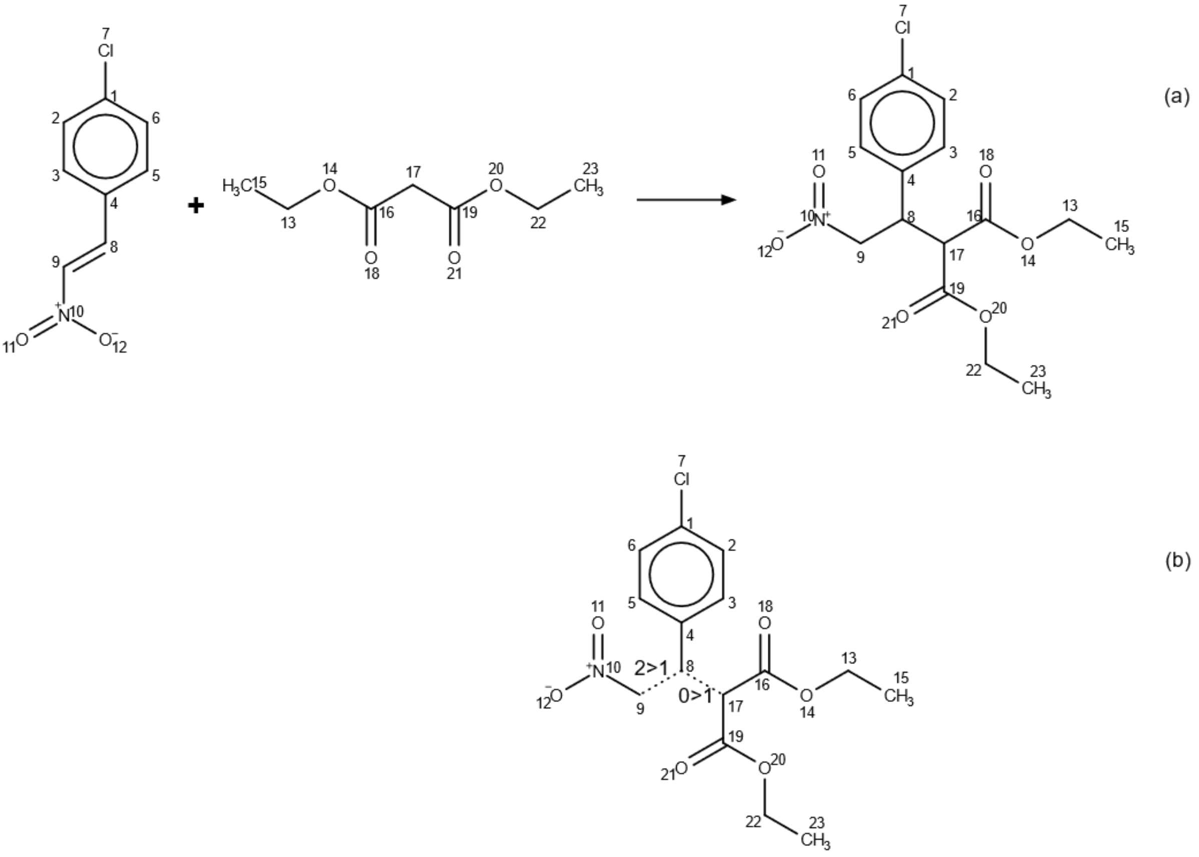

Figure 3. An example of the Condensed Graph of a Reaction (CGR). Here, a reaction of addition of a malonic ester to 4-chloro-beta-nitrostyrene is mapped (a) and then transformed to a CGR (b) where the dynamic bond “2>1” means the reduction of a double bond between two carbon atoms 8 and 9 to a single bond, and “0>1” means the formation of a single bond between two carbon atoms 8 and 17.

Within a consensus strategy, the best mapping scheme is selected according to the minimal CD following the principle “lower the CD – less transformations are in the reaction – better the mapping”. Five consensus strategies were tried out:

1) *Consensus 1* – consensus of RXNMapper & MAX algorithm & MIXTURE algorithm & MIN algorithm & ChemAxon & Indigo & NextMove;

2) *Consensus 2* – consensus of RXNMapper & MAX algorithm & MIXTURE algorithm & MIN algorithm & ChemAxon & Indigo;

3) *Consensus 3* – consensus of RXNMapper & MAX algorithm & MIXTURE algorithm & MIN algorithm & ChemAxon;

4) *Consensus 4* – consensus of RXNMapper & MAX algorithm & ChemAxon;

5) *Consensus 5* – consensus of RXNMapper & ChemAxon.

Here, the order of the tools corresponds to their mapping accuracy described in the Results and Discussion section. Also, an additional consensus approach operating just with MAX, MIN and MIXTURE algorithms was tested (named further as “new RDTool CD consensus”) in order to compare the CD concept to the 8 metrics computed in the RDTool software.

The benchmarking study was performed evaluating the performance of each AAM strategy using two metrics: speed of calculations and mapping accuracy (percentage of correctly mapped reactions out of the given reaction data set). To decide whether the mapping is correct or not, the reactions were first transformed to CGRs, and then the AAM provided by automatic tool was compared to the experts’ AAM (see the details in the Data section). The machine-produced mapping is considered to be correct one if two CGRs based on the automatic and manual mappings, respectively, coincide totally. Comparison of CGRs and calculation of CD is performed using CGRtools library[4].

2.4 Data

In this study, the data set of 1405 organic reactions published by Jaworski et al.[23] and mapped by experts was taken as a reference data set (further mentioned as the reference data set). In the original publication some atoms were added to reactions after mapping procedure and in mapped reactions some SMARTS symbols were used in reaction SMILES, in mapped reaction we’ve left only atoms that were present in the initial reaction (prior to AAM generation). Otherwise, it caused some inconsistencies while reading, appearance of radical moieties, etc. Dataset was transformed to the RDF format and extended with new 469 USPTO reactions of types common in medicinal chemistry which were mapped manually as well. This data set was used to test the standardization workflow and to compare the AAM strategies. The designed workflow was then applied to the database of US patents containing 3.75M chemical records.[11]

3 Results and Discussion

3.1 Reference dataset curation

The standardization workflow described in the Method section was implemented and applied to pre-process the reference data set.

Errors in atom-to-atom mapping were corrected by expert organic chemist to ensure that the corresponding reaction mechanism is reflected. Ambiguous reactions and reactions with unclear mechanism were excluded from the dataset. In the process of Jaworski et al.[23] data curation, 5 records were not interpreted as chemical reactions, 10 reactions were identified as duplicated (prior to addition of 469 USPTO reactions), 30 reactions were discarded since they were comprised by the molecules with invalid valences at some atoms (see an example in Figure 4), and 350 reactions contained radicals.

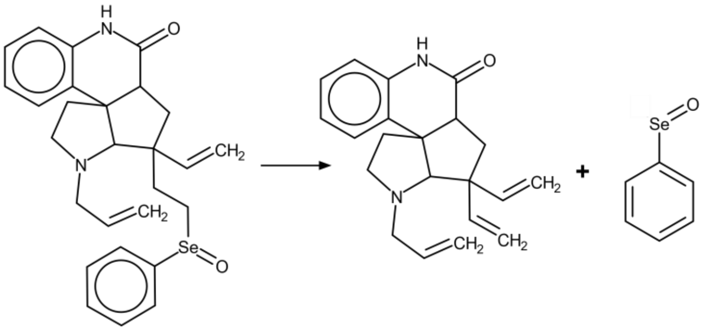

Figure 4. An example of a reaction with invalid valence. Here, Se atom in the product belongs to unsubstituted selenoxyde which is considered likely impossible.

The discarded reactions were analyzed manually and corrected if possible. Namely, 17 reactions represented chemically wrong were fixed (e.g. a reaction contained an abbreviation of the acetic functional group as a part of a structure is depicted in Figure 5), and around 350 reactions were restored: radicals were removed (we assume that it was an attempt of Jaworski and others to balance the reactions using single atoms in the form of radicals).

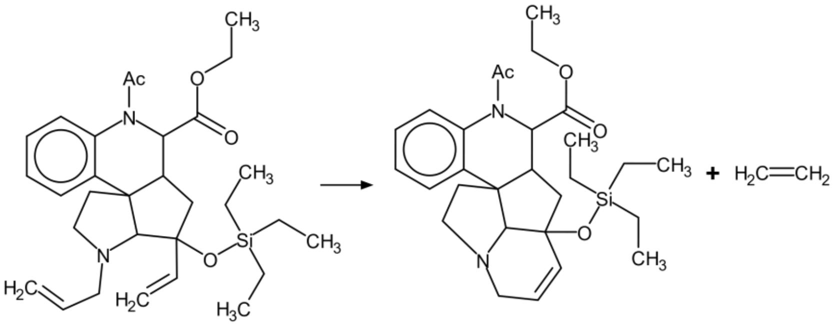

Figure 5. A reaction stored in an inappropriate way. Here, the "Ac" atom is the acetyl group abbreviation that should be written explicitely.

For all reactions in our golden dataset atom-to-atom mapping is curated by chemistry expert. It was found that 8 reactions from the dataset published by Jaworski *et al.* contain obvious errors (e.g. one carbon atom in a reactant corresponded to both an oxygen and a carbon atoms in a product simultaneously), 22 reactions had non-obvious mapping which may depend on conditions used, and 3 reactions had unclear reaction mechanism or lost reactants (see Figure 6). The former 8 reactions were fixed and the latter 25 were removed.

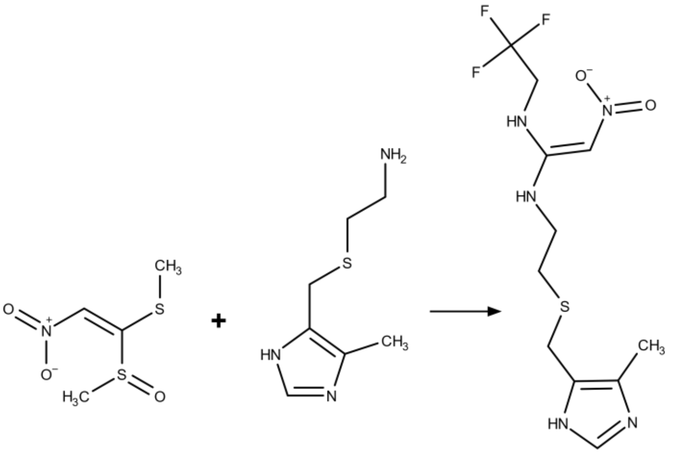

Figure 6. A reaction with lost reactants. This is a two step reaction in which the second amine reagent (2,2,2-trifluoroethylamine) is missing among the reagents.

The final “golden” data set comprises 1851 reactions. The number and the nature of the found mistakes demonstrate that the standardization protocol is efficient, and it detects the errors that are not obvious and may compromise the model’s quality. Next, the dataset was mapped by 15 mapping strategies described in the Method section. The results of timing of the AAM tools (excluding the consensus) are shown in Figure 7.

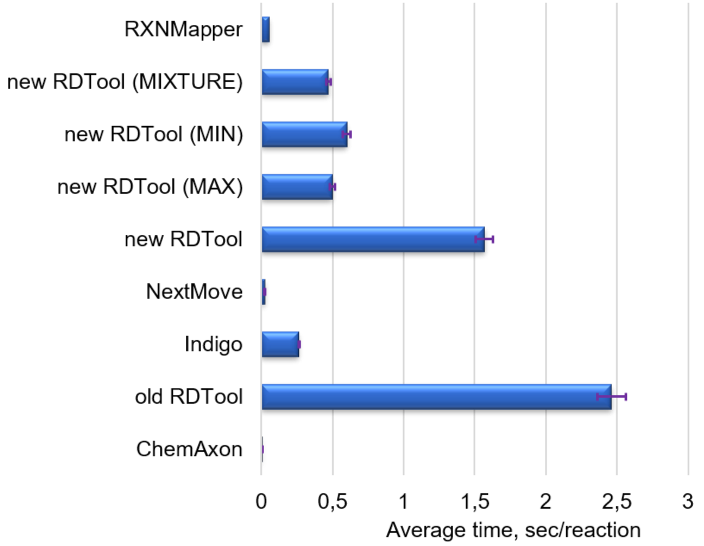

Figure 7. Mapping tools timing.

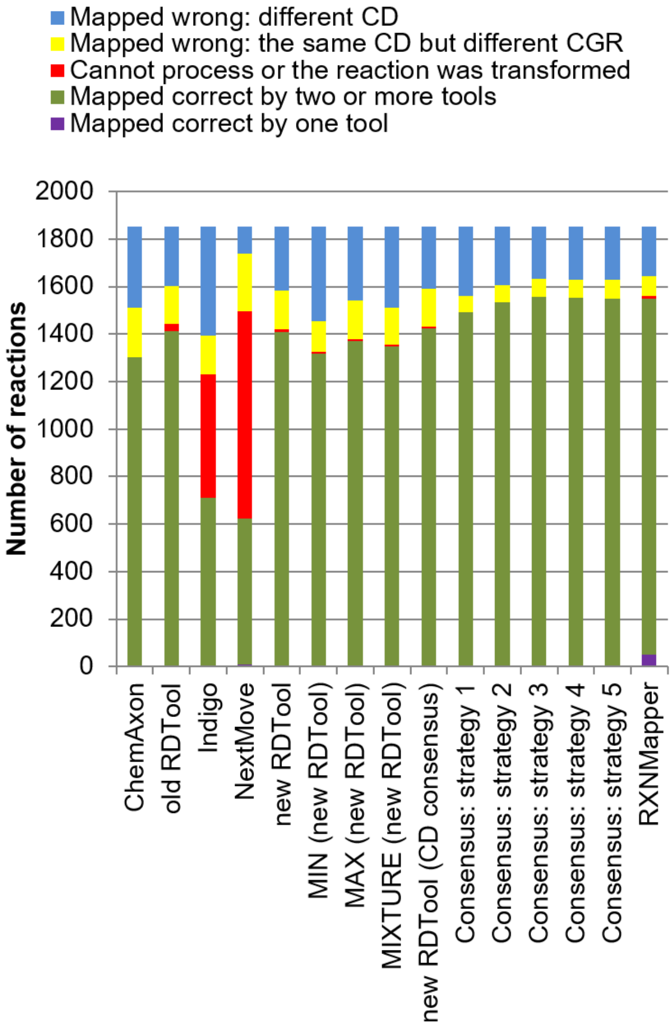

Figure 8. Comparison of AAM accuracies. Here, the consensus strategies correspond to the ones described in the Method section.

As it can be seen from the Figure 7, ChemAxon is the fastest tool, and it needs approximately 0.003 second to map one reaction. NextMove and RXNMapper follow ChemAxon, and they share the second and the third places spending 0.02 and 0.05 seconds per reaction, respectively. The old version of RDTool is the slowest – it takes around 2.46 seconds to map one reaction. The reason for that is the fact that it runs three mapping algorithms (MIN, MAX and MIXTURE). Note that the timing shown for MIN, MAX and MIXTURE was measured for the revised version of RDTool already. Despite that it does not change the conclusion. It can also be seen that the code revision impacted significantly: the new RDTool spends about 1.57 seconds per reaction which is about 50% faster but still quite slow.

Next, the mapping accuracy was checked. The histogram depicting the results is shown in Figure 8, and the detailed results are given in Table 1. AAM is considered wrong if CGR of automatically mapped reaction does not coincide with reference expert curated one. However, for wrong AAM we specially identified cases when CD for the right and wrong AAM is different and when CDs coincide. In the latter case, error might be caused by selection of wrong AAM out of two having same CD by the mapping tool. We also found cases when mapping software returns reactions with fully or partially incomplete mapping, or changes reactants or products (adds, loses atoms, changes product, etc). And finally, we identified reactions for which AAM was done correctly solely by a given tool, while the all other mappers were wrong.

Table 1. AAM benchmarking results.

AAM strategy	# of reactions mapped correct	Accuracy, %	# of reactions mapped correct only by one tool	# of corrupted reactions[a]	# of reactions with different CD[b]	# of reactions with the same CD but different CGR
ChemAxon	1304	70.45	5	0	339	208
Old RDTool	1411	76.23	0	30	250	160
Indigo	712	38.47	4	518	457	164
NextMove	622	33.60	9	873	113	243
New RDTool	1410	76.18	0	9	269	163
MIN	1316	71.10	0	8	395	132
MAX	1369	73.96	0	8	311	163
MIXTURE	1347	72.77	0	8	339	157
New RDTool (CD consensus)	1422	76.82	0	8	262	159
RXNMapper	1550	83.74	49	9	208	84
Consensus: strategy 1[c]	1494	80.71	-[d]	0	289	68
Consensus: strategy 2	1532	82.77	-	0	246	73
Consensus: strategy 3	1556	84.06	-	0	220	75
Consensus: strategy 4	1554	83.95	-	0	222	75
Consensus: strategy 5	1549	83.68	-	0	221	81

[a] A corrupted reaction is the reaction that caused an error of the mapping tool or was transformed by the mapper in a way such this reaction could not be compared to the initial one.

[b] Chemical Distance is the number of dynamic bonds (formed, broken or transformed) counted for a particular reaction.

[c] See the description of the consensus strategies in the Method section.

[d] No statistics on the reaction mapped correct by a single tool was prepared for the consensus strategies since the latter is already a combination of several approaches.

The accuracy of ChemAxon's AAM (70.45%) is comparable to the one that was observed in previous publications.[15,23] The analysis of its mistakes revealed reactions split into two almost same size groups: (1) reactions with different CD (obvious discrepancy; 339 reactions) and (2) reactions with the same CD but different CGRs (208 reactions). The Prilezhaev epoxidation shown as an example in Figure 9a represents the first group. Instead of using the peroxide oxygen [O:32], ChemAxon maps the oxygen atom [O:31] in the epoxide reaction product. This decision leads to breaking of two single bonds instead of one as it is shown in Figures 9b and 9c. Thus, the CD for the correct mapping is 4, and the CD for the wrong one is 5. Such kind of mistakes are easy to detect, whereas the reactions in the second group are not so obvious. For instance, the reactions of esters hydrolysis (Figure 10) confuse most of the mappers and usually they are mapped wrong.

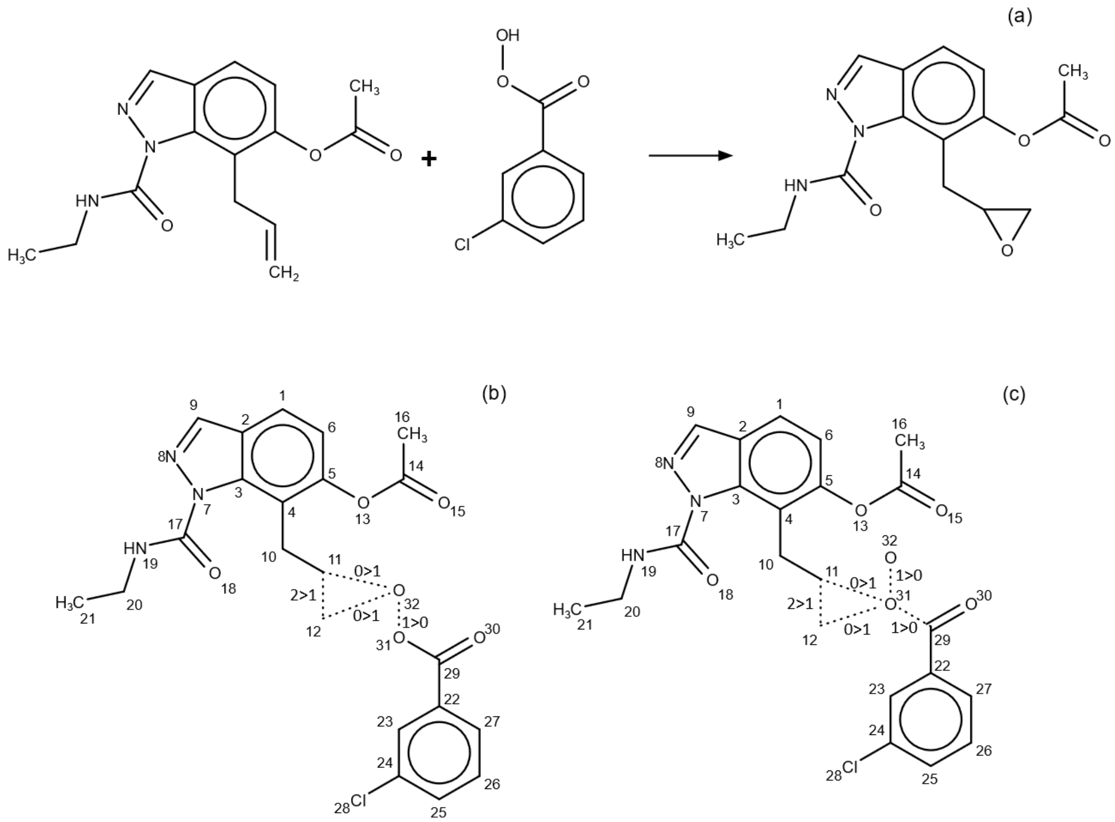

Figure 9. An example of a correct (b) and wrong (c) mappings produced for the Prilezhaev reaction of epoxidation (a). Here, the corresponding condensed graphs are depicted to demonstrate the difference in the correct and wrong mapping schemes. The annotations like “1>0” mean the bond transformation type, here single bond breaking. “0>1” and “1>0” mean formed and broken single bond according to convention in CGR depiction[4].

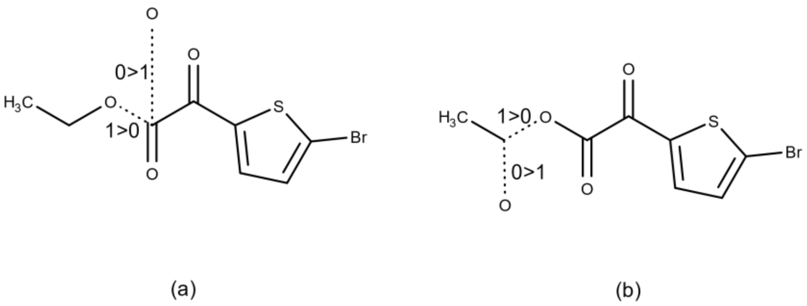

Figure 10. Correct (a) and wrong (b) condensed graphs for the reaction of ester hydrolysis. Here, “0>1” and “1>0” mean formed and broken single bond according to convention in CGR depiction[4].

Even though the Indigo mapper worked well for more than 700 reactions, it still fails on reactions of halogenations. Namely, it prefers to map the atom of a halogen in a product (e.g. bromine) as it comes from unspecified reactant, even if halogen molecule present in reactant part (i.e. in form of Br_2). In addition, in case of symmetric molecules, Indigo makes one-to-many correspondence of one reactant atom to many product atoms. Since it breaks AAM convention that AAM is essentially one-to-one correspondence, it causes problems with such mapping usage. For example, generation of CGR is impossible (each atom in reactants/products should possess the unique number). Such reactions were classified as “corrupted” and depicted in Figure 8 in red.

The old RDTool succeeded with more than 1400 reactions which is 76.23% of the reference data set. Despite such high efficiency, it fails on 30 reactions since its MCSs search time exceeded timeout considerations. Based on its speed, it is not the most promising tool. Our code revision improved the speed (see Figure 7) and gave us an opportunity to call MIN, MAX and MIXTURE algorithms separately. Although the revised new RDTool is faster and it fails on fewer number of reactions (9 reactions), its accuracy is still comparable with the old version. Comparing the results of individual mapping approaches to their consensus (named “new RDTool” in Figure 8), we found that the MAX algorithm already provides a 73.96% accuracy, which corresponds to 97.1% of the total number of reactions mapped correctly by new RDTool. Therefore, application of only the MAX algorithm allows us to achieve almost the same accuracy as it for the whole RDTool but the result might be obtained 5 times faster (comparing to the old version).

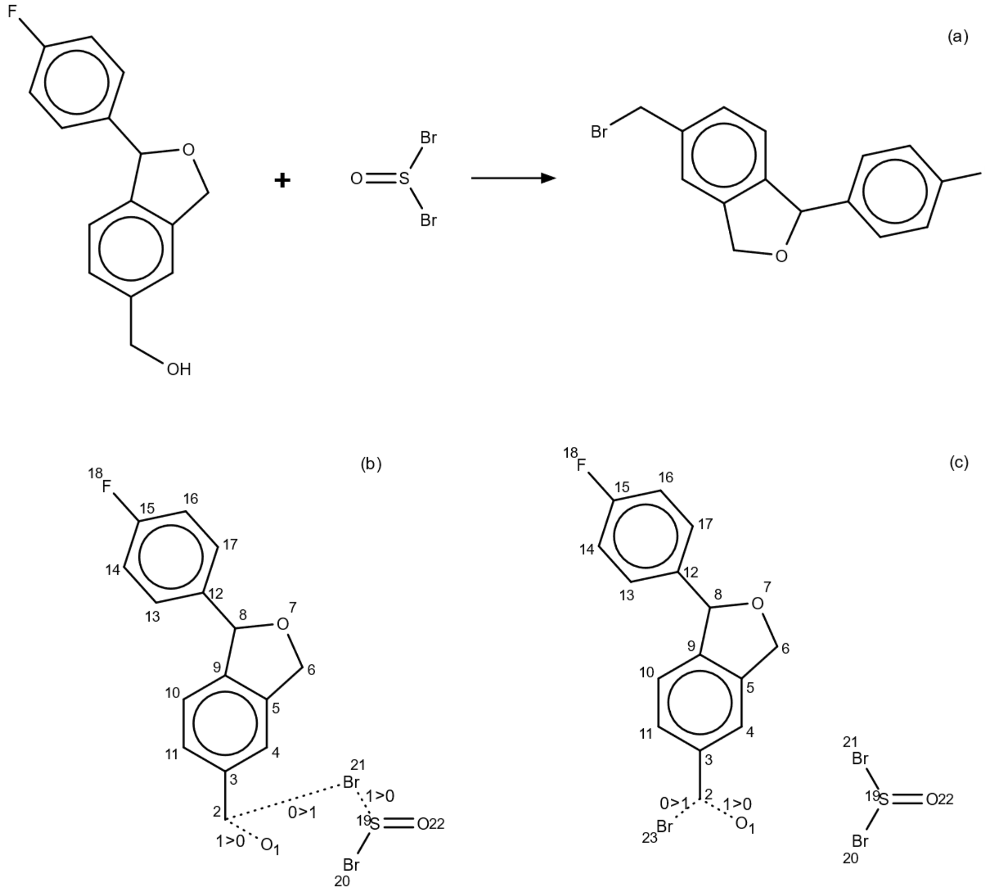

Figure 11. An example of (a) a bromination reaction, (b) CGR corresponding to correct AAM and (c) CGR based on wrong AAM made by NextMove tool. On CGR depiction, ‘0>1’ corresponds to single bond formation, and ‘1>0’ corresponds to single bond breaking. Notice that incorrect AAM has also lower number of dynamic (broken and formed) bonds.

NextMove software demonstrated the lowest number of correct atom-to-atom mapping results comparing to the other tools with accuracy 33.6%. It cannot map 873 reactions. The results are explained by the fact that AAM is done based on a large set of rules, *i.e.* NameRXN[43] reaction classification system. If no rule exists for a given reaction, no mapping can be obtained for that entry. Since NameRXN software is based on handcrafted chemical rules collected by chemists, it is surprising to see that expert system AAM mistakes (see blue and yellow bars in Figure 8). The problems are often caused by ignoring present reactants as it can be seen on the bromination reaction example depicted in Figure 11. Similar scenario was met in 113 reactions where CD differed from the one computed for the manual mapping, and 243 reactions had the same CD but different transformations (different CGRs). This demonstrates that the specificity of the rules is insufficient, and some improvement to the code/rules is needed.

In comparison to the classical mapping approaches, deep-learning used in the RXNMapper tools has demonstrated its power mapping correctly more 1550 reactions (83.74%). This is the best result among considered reaction mappers. In addition to that, the latter tool consistently generates a correct AAM for esterification and esters hydrolysis reactions, which were found to be challenging for most of the mappers. Unfortunately, this approach has several drawbacks:

1) It is applicable only to reaction SMILES, which resets the standardization operations done before;

2) The tool requires larger computation resources in terms of CPU/GPU as well as computational memory (RAM);

3) The RDKit pre- and, perhaps, postprocessing used by the tool cannot be skipped, which makes the tools less flexible and leads to unexpected behaviour (see Figure 12);

4) The tool reorders the reactants and the products, which makes the ions regrouping impossible.

5) And finally, we found some reaction types for which RXNMapper never makes AAM errors on “golden dataset”, however when applied to entire USPTO dataset errors appear. For example, out of 541 reactions of reduction of carboxyl groups to alcohols found in USPTO only 58 were mapped correctly, while 483 had wrong AAM. For reactions of methyl ester formation using methanol 493 reactions were correct but 225 were found to be incorrectly mapped by RXNMapper. Similar situation with other esterification reactions: error rate was 10-30%, despite no errors of this kind were observed for “golden” dataset. Reactions with correct and incorrect AAM were identified using reaction signatures implemented in CGRtools[4]. Thus, despite RXNMapper was reported [15] to avoid some types of AAM errors, it still makes them on large reaction dataset.

Because of first four listed drawbacks, it was decided to repeat the standardization procedure for the reference data set mapped by each AAM tool.

Analysing the consensus strategies, we see that the consensus done for the RDTool mapping algorithms using the CD concept is slightly more efficient in comparison to the initial one (the accuracy is 76.82% vs 76.18%). However, the consensus of several AAM tools almost did not improve the accuracy of the best tools. For instance, the consensus strategy 1 which includes all the mapping algorithms considered in this project shows higher accuracy than ChemAxon and RDTool individually, but it is still worse than RXNMapper. Removal of Indigo and NextMove (consensus strategy 3) improved the accuracy (now it is 84.06%) and this is the best solution if timing could be ignored. Unfortunately, this approach is 30 times slower than individual RXNMapper. Once we exclude MAX, MIN and MIXTURE algorithms (consensus strategy 5), the accuracy becomes slightly lower than it is for the RXNMapper.

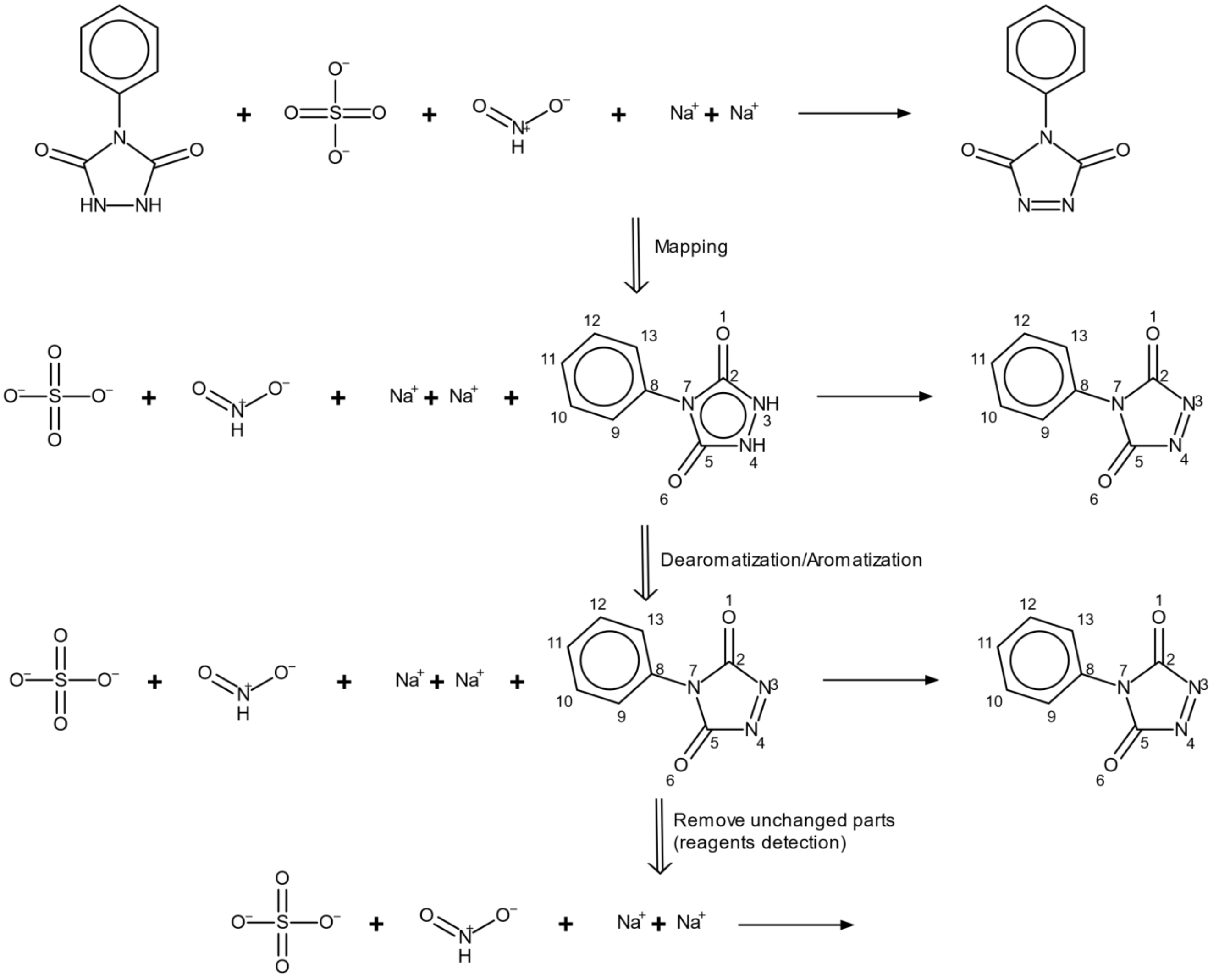

Figure 12. An example of a reaction passed through the RXNMapper tool. Here, the urazole fragment was aromatized by RDKit inside the RXNMapper that violets the valence. However, CGRtools made an attempt to dearomatize it keeping the valence of the atoms of nitrogen to be 3 which let to appearance of the double bond. This, in turn, let to removal of the 4-phenyl-1,2,4-triazolidine-3,5-dione molecule recognized as 4-phenyl-1,2,4-triazole-3,5-dione together with the product considering it as unchanged parts (reagents).

The cause of such behaviour lies in the methodology of the consensus approach. As it was mentioned, the “best” mapping scheme is selected according to the lowest chemical distance (CD) assuming that lower CD leads to the lower number of transformations, which, in turn, corresponds to better AAM. However, this is not always the case which is illustrated in the example of the Diels-Alder reaction in Figure 13. In this reaction, wrong AAM possesses CD of 5, and the correct one possesses the CD of 6. In this case, the consensus approach will prefer the wrong AAM instead of correct one, and the accuracy will decrease. To solve this problem, mapping correction can be done by application of remapping rules.[4] However, this is out of the scope of the current work.

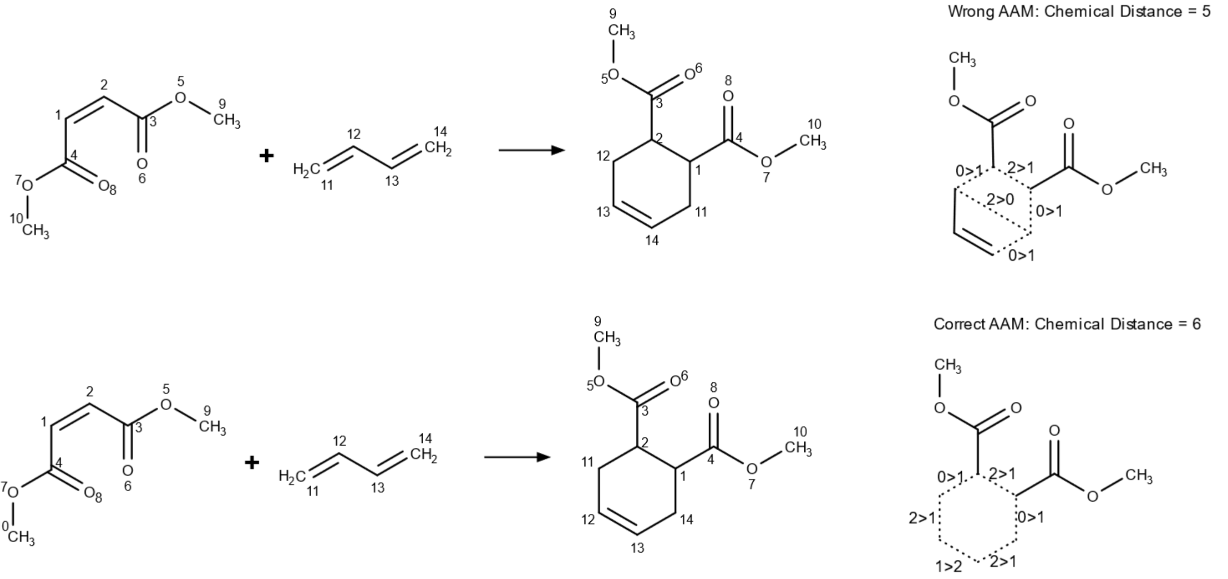
Wrong AAM: Chemical Distance = 5
Correct AAM: Chemical Distance = 6
0>1
2>1
2>0
0>1
0>1
0>1
2>1
2>1
0>1
1>2
2>1

Figure 13. An example of a reaction of Diels-Alder with correct (a) and wrong (b) atom-to-atom mappings. Here, the reaction itself is on the left side, and the corresponding CGR is on the right side. Symbol "*n*>*m*" mean change of bond type in reaction from bond order *n* to *m*[4].

Generally, the best solution from accuracy-performance point of view is application of the RXNMapper. Taking into account its specifics, standardization is required to repeat after mapping procedure. Thus, strategy 'standardize-map-standardize again' using RXNMapper and described above standardization workflow was applied to clean and to map the USPTO database. The statistics of its errors are shown in Table 2. Due to the reactions lost during the second standardization (an example of such reaction is given in Figure 12), the statistics on errors increased, and the results shown in Table 2 already cumulate both first and second standardizations.

Table 2. The statistical results of the USPTO cleaning procedure[a].

	Statistics[a]
Initial number of USPTO records	3.75M
Number of unique reaction SMILES before reactions standardization	1.48M
Number of reaction SMILES that could not be recognised	620
Number of reactions failed on dearomatization	75
Number of reactions with invalid valences	55.4K
Number of reactions containing isotopes	45
Number of reactions without reactants	1.1K
Number of reactions without products	10.3K
Number of reactions without reactants and products[b]	1.9K
Number of reactions that could not be mapped	220
Number of duplicates	86.5K
Final number of reactions	1.316M

[a] This statistics includes errors obtained before and after RXNMapper and counted by the standardization protocol.

[b] Once the unchanged parts are removed, it may happen that no reactants and products left in the reaction. In this case, an exception of an empty reaction is raised.

The final version of the USPTO database contains 1.316M reactions.

3 Conclusions

In this work, two aspects of chemical reaction standardization were considered: *(i)* curation of structural information and *(ii)* reaction mapping. For the first part, the workflow for standardization of structural representation was proposed. It consists of regular molecular standardization, ions cleaning, identification of the same molecules in different sides of a reaction equation, duplicates removal, etc. Second, an extensive comparison of reaction mappers has been performed. For this task, we collected a dataset of 1851 reactions carefully curated and mapped by our expert chemist. The dataset is an extension of the previously published manually mapped reaction dataset [23] that was also augmented with 469 USPTO records for better representativity of industrially important reactions. The prepared "*Golden*" reaction dataset was used to compare 15 mapping strategies (including 5 consensus strategies). The results show that the consensus strategy 3 is the most accurate one (84.06% of correct mappings), whereas single RXNMapper possesses similar accuracy (83.74%) and it is much faster (0.1 seconds per reaction instead of 1.7 seconds used by the consensus strategy 3). At the same time, RXNMapper resets the standardization efforts made before which leads to unexpected behavior. To solve this issue, a new round of standardization procedure was found to be essential. Such trick makes the tool less attractive, whereas retraining the model and refusing of RDKit will certainly solve the found problems. Despite that, we still recommend RXNMapper since wrongly standardized reactions can be fixed or removed, but wrong atom-to-atom mapping cannot be even detected automatically.

The developed workflow was tested on USPTO reaction dataset comprising 3.75M records. The resulted USPTO dataset contains 1.316M unique reactions. The standardized USPTO dataset, manually curated "golden" reaction dataset, standardization workflow and optimized RDTool are available on GitHub: https://github.com/Laboratoire-de-Chemoinformatique. We encourage the society to collaborate on further their extension and improvement.

Funding

This project was sponsored by the VLAIO HBC.2018.2287 MaDeSMArt (Machine Design of Small Molecules by AI). TM, AR, and RN are grateful to Russian Science Foundation for support of reaction standardization workflow development (project No 19-73-10137).

Reference

[1] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

[2] "RDKit: Open-source cheminformatics," can be found under http://www.rdkit.org, **2019**.

[3] ChemAxon, "JChem," can be found under https://chemaxon.com/products/jchem-engines, **2020**.

[4] R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov, A. Varnek, *J. Chem. Inf. Model.* **2019**, *59*, 2516–2521.

[5] D. Fourches, E. Muratov, A. Tropsha, *J Chem Inf Model* **2016**, *56*, 1243–1252.

[6] A. Tropsha, *Mol. Inform.* **2010**, *29*, 476–488.

[7] D. Fourches, E. Muratov, A. Tropsha, *Nat Chem Biol* **2015**, *11*, 535.

[8] V. D. Hähnke, S. Kim, E. E. Bolton, *J. Cheminform.* **2018**, *10*, 36.

[9] K. Karapetyan, C. Batchelor, D. Sharpe, V. Tkachenko, A. J. Williams, *J. Cheminform.* **2015**, *7*, 30.

[10] D. Gadaleta, A. Lombardo, C. Toma, E. Benfenati, *J. Cheminform.* **2018**, *10*, 60.

[11] D. Lowe, **2017**, DOI 10.6084/m9.figshare.5104873.v1.

[12] I. A. Watson, J. Wang, C. A. Nicolaou, *J. Cheminform.* **2019**, *11*, 1.

[13] A. Thakkar, N. Selmi, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, **2020**.

[14] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *Science (80-.).* **2019**, *365*, eaax1566.

[15] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, **2020**, DOI 10.26434/chemrxiv.12298559.v1.

[16] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Chem. Sci.* **2020**, *11*, 154–168.

[17] G. M. Ghiandoni, M. J. Bodkin, B. Chen, D. Hristozov, J. E. A. Wallace, J. Webster, V. J. Gillet, *J. Chem. Inf. Model.* **2019**, *59*, 4167–4187.

[18] C. W. Coley, W. H. Green, K. F. Jensen, **2019**, DOI 10.26434/chemrxiv.7949024.v1.

[19] T. E. Moock, J. G. Nourse, D. Grier, W. D. Hounshell, in (Ed.: W.A. Warr), Springer Berlin Heidelberg, Berlin, Heidelberg, **1988**, pp. 303–313.

[20] P. P. Plehiers, G. B. Marin, C. V Stevens, K. M. Van Geem, *J. Cheminform.* **2018**, *10*, 11.

[21] E. E. Litsa, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett, L. E. Kavraki, *J. Chem. Inf. Model.* **2019**, *59*, 1121–1135.

[22] G. A. Preciat Gonzalez, L. R. P. El Assal, A. Noronha, I. Thiele, H. S. Haraldsdóttir, R. M. T. Fleming, *J. Cheminform.* **2017**, *9*, 39.

[23] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, B. A. Grzybowski, *Nat. Commun.* **2019**, *10*, 1434.

[24] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, **2020**, DOI 10.26434/chemrxiv.12298559.v1.

[25] C. Jochum, J. Gasteiger, I. Ugi, *Angew. Chemie Int. Ed.* **1980**, *19*, 495–505.

[26] T. E. Oliphant, *Comput. Sci. Eng.* **2007**, *9*, 10–20.

[27] E. Hückel, *Zeitschrift für Phys.* **1932**, *76*, 628–648.

[28] D. A. Semenow, J. D. Roberts, *J. Chem. Educ.* **1956**, *33*, 2.

[29] Daylight Chemical Information Systems Inc., “Daylight Theory Manual,” can be found under http://www.daylight.com/dayhtml/doc/theory/index.html, **2011**.

[30] S. Kikuchi, *J. Chem. Educ.* **1997**, *74*, 194.

[31] N. Schneider, N. Stiefl, G. A. Landrum, *J. Chem. Inf. Model.* **2016**, *56*, 2336–2346.

[32] L. Chen, J. G. Nourse, B. D. Christie, B. A. Leland, D. L. Grier, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1296–1310.

[33] A. I. Lin, T. I. Madzhidov, O. Klimchuk, R. I. Nugmanov, I. S. Antipin, A. Varnek, *J. Chem. Inf. Model.* **2016**, *56*, 2140–2148.

[34] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov’ev, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 693–703.

[35] M. H. S. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604.

[36] **2009**.

[37] NextMove Software, “NameRXN,” can be found under www.nextmovesoftware.com, **2020**.

[38] A. Savelev, I. Puzanov, V. Samoilov, V. Karnaukhov, **2019**.

[39] S. A. Rahman, G. Torrance, L. Baldacci, S. Martínez Cuesta, F. Fenninger, N. Gopal, S. Choudhary, J. W. May, G. L. Holliday, C. Steinbeck, J. M. Thornton, *Bioinformatics* **2016**, *32*, 2065–2066.

[40] T. I. Madzhidov, R. I. Nugmanov, A. I. Lin, I. S. Antipin, T. R. Gimadiev, A. Varnek, *Butlerov Commun.* **2015**, *44*, 170–176.

[41] F. Hoonakker, N. Lachiche, A. Varnek, A. Wagner, *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

[42] T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, A. A. Varnek, I. S. Antipin, *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

[43] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, G. A. Landrum, *J. Med. Chem.* **2016**, *59*, 4385–4402.