# Reaction Data Curation I: Chemical Structures and Transformations Standardization (Molecular Informatics - 2021)

* Abstract: - the quality of experimental data for chemical reactions is a critical, in this the researchers suggest a 4 steps protocol that includes the curation of individual structures, chemical transformations, reaction conditions, and endpoints.
* Introduction: - Data quality is crucial for the effective storage and exploitation of chemical knowledge.
* So far, two databases have been intensively used in the modelling studies: Reaxys and USPTO reaction dataset.
* The Reaxys dataset is a commercial data set, it has been extensively used for various applications of deep-learning neural networks to retrosynthesis, robochemistry, and prediction of optimal reaction conditions, as well as for analysis of rection network.
* The USPTO is the largest public dataset of chemical reactions extracted from the US patents using text mining techniques.
* The dataset has been extensively used for different applications: analysis of reaction databases, forward and backward-synthesis, reactions classification, atom-to-atom mapping, yield prediction, and compound role assessment.
* Very little effort has been invested to curate reaction data.
* Most of the suggested structures’ standardization workflows, recombine selected steps of the molecules cleaning proposed suggested by Fourches et al.



* A *transformation* term corresponds to a reaction equation that links sets of reactants and products.
* According to IUPAC definition, a *reactant* is “a substance that is consumed in the process of a chemical reaction.”
* Whereas the *product* is “a substance that is formed during the chemical reaction”.
* Atom of reactants that change their connectivity and/or formal charge(s) upon the transformation constitute a *reaction center.*
* To identify a reaction center, a one-to-one correspondence between atoms of reactants and products called *atom-to-atom mapping* (AAM) must be established.
* Reactants are the catalyst, solvents, catalytic, poisons, complexation, agents, redox, agents, detergents, and acid/bases.
* According to IUPAC a reagent is a “substance that is added to the chemical system in order to bring about a reaction” is not precise enough for automatized reagents recognition.
* Schneider et al. attributed a compound to reagents if it wasn’t affected in the course of reactions, whereas Gao et al. considered any compound as a reagent if it did not contribute any carbon atom to the reaction.
* Normally a conditions’ description includes both numeric properties like temperature, pressure, and reaction duration, as well as reagents.
* Because error may occur in each component of a reaction, a curation workflow should consist, at least, of four steps related to the curation of (1) individual chemical structures (reactants, products and reagents), (2) chemical transformation work (3-4) Reaction conditions and endpoints curation.



* The curation protocol may vary as a function of a particular chemoinformatics task.



* Reaction Curation Workflow: - the most popular non-commercial libraries RDKit, CDK, Indigo, and OpenBabel can be applied to standardize only individual molecules extracted from reaction equation.
* The absence of ready-to-use tools for reaction curation motivated the researchers to develop a curation workflow that combines several selected standardization steps.
* Chemical Structures Curation: - (Specific Feature of Structures Curation) - some steps of the structures curation protocol, like functional groups standardization, aromatization and valences checking are similar with those recommended for medchem databases.
* Organometallic compounds represent a particular problem because they are poorly represented using conventional bond types.
* Normally a metal atom is linked to its neighbors via single covalent bonds or is represented as an isolated cation.
* A special coordination bond type between metal and its environment could be a solution of this problem.
* Structure Curation Steps (Functional Groups Standardization) – big data collections gathered from different sources often suffer from inconsistent representation of certain functional groups.
* Aromatization: - although aromatization solves the problem of standard representation of arenes and the vast majority of heterocycles, it may cause ambiguity in structure of heterocycles containing both pyrrolic and pyridinic nitrogen atoms, like, imidazole.
* In order to achieve the aromatization consistency within a given dataset, the researchers suggest, first, to transform all chemical structure to the kekulé form and then re-aromatize them again at the end of the “Structures curation” part using selected aromatization algorithm.
* The *“partial”* option, which keeps heterocycles with pyrrolic nitrogen in the kekulé form looks rather convenient.
* The researchers believe that the dearomatization/re-aromatization procedure described can also be recommended for data standardization in any database of individual compounds, including medchem databases.
* Valence Checking (major tautomer generation): - the generation of the major tautomer is needed because a given molecule can be recorded in a database in different tautomeric forms.
* It creates a problem with duplicates detection, substructural or similarity searches and molecular descriptors generation.
* The RDKit and ChemAxon tools used for identification of a major tautomer cannot be directly applied to chemical reactions: RDKit gives an error message, whereas ChemAxon corrupts the data by reassigning each compound to reactants.
* Although many rules for ring-chain tautomers generation have earlier been proposed, they were not implemented into any standardization workflow.
* Removal of Explicit Hydrogens: - once the major tautomer is selected, explicit hydrogens should be removed unless they are required in a particular chemoinformatics task.
* (Discarding Reactions Containing Radicals): - in USPTO, reactions containing radicals frequently result from erroneous recognition of chemical structures by text mining tool or bad reaction balancing. On the other hand the Reaxys dataset contains some 100K reactions where radicals, like (2,2,6,6-Tetramethylpiperidin-1-yl)oxyl, are used as a reactant.
* There are no algorithms able to distinguish erroneously parsed reactions from truly radical reactions. Improvement of the text parsing and reaction balancing techniques may help to solve this problem.
* Cleaning Isotopes: - atom-to-atom mappers ignore this information and isotopes do not almost influence reaction characteristics. Therefore, to avoid any confusion, the researchers propose to clean up the isotope labels.
* Split of lons: - this option helps to detect ions exchange reaction, which proceed without any transformation of convalent bonds.
* These reactions have no reaction center. They are out of the chemoiniformatics, however, formal representation as entire salts may significantly complicate their identification.
* Clear Stereo: - the stereochemistry labels of bonds and atoms should be discarded. Thus, reactions involving stereoisomers of some compounds will be later considered as duplicates.
* Transformations Curation: - transformation curation aims to perform the following operations: (i) atom-to-atom mapping, (ii) reaction role assignment, (iii) reaction duplicates removal, and (iv) reaction equation balancing.
* A CGR (Condensed Graphs of Reactions) represents an ensemble of reactants and products as a single molecular graph resulting from superposition of related atoms in reactants and products.
* Dynamic bonds describe a chemical transformation: breaking or formation of chemical bonds, transformation of a single bond to a double bond etc.
* Unlike reaction SMILES that encode a set of individual reactions and products, CGR and related CGR/SMILES account for both structure of all molecular species and information about reaction center issued from atom-to-atom mapping.
* SMILES-based operations don’t need any mapping and, therefore, are recommended as a relatively fast procedure to perform reaction role assignment and reaction duplicates removal.
* Atom-to-atom Mapping: - atom-to-atom mapping (AAM) is a one-to-one correspondence between an atom in a reactant and an atom in a product.
* AAM is needed to identify the reaction center, many AAM algorithms and related software tools are known, but none of them is perfect
* From many AMM tools (ChemAxon Automapper, Indifo etc.) only the ChemAxon Automapper produces one-to-one correspondence between reactant and product atoms, does not leave unmapped atoms and does not modify molecular representation.
* Reaction Role Assignment: - it is one of the most delicate operations of the reaction data curation. The erroneous assignment of reactants to reagents or vice versa can be introduced by incorrect text parsing, data storage technique, or special rules in particular databases.
* The presence of reagents can also slow down AAM assignments as mapping complexity increases with the number of atoms.
* Therefore, the researchers suggest the before the AAM step we should identify the reagents as molecules present in both reactants and products ad remove them from the reaction equations.
* In order to assign a reaction role automatically, one can follow the definition from Schneider et al. which considered compounds unaffected in the course of a reaction as reagents.
* Such compounds can be automatically detected in AAM as reactants that are not mapped into products.
* Duplicates Detection: - duplicates detection and removal are a regular data cleaning procedure that aims to keep only unique entities.
* Two types of duplicates were considered: (i) transformations with the same list of reactants and products and (ii) transformations that contain a different number of reactants and products but corresponding to the same CGR.
* The researchers recommend duplicate detection using canonical reaction SMILES strings before the AAM step. Its helps to decrease the number of reactions considered in time consuming mapping procedures.
* As soon as AAM is established, reactions can be converted into CGR which, in turn, can be used to identify remaining duplicates.
* Reaction Balancing: - some chemoinformatics tasks, like the preparation of reaction vectors require fully balanced reactions.
* CGR allows to balance a given reaction upon its reversible decomposition. Application of some additional heuristics can be used to balance redox processes or for compounds that are used many times in a reaction, e.g., reagents for exhaustive alkylation.
* Missing components in reactions can be identified using deep networks. It should be noted that balancing can only be performed if the list of reactants and/or products can only be performed if the list of reactants CGR.
* Curation of Popular Reaction Databases: - the simplified curation workflow described in the paper has been applied to clean three popular reaction databases: Reaxys, USPTO and Pistachio.
* This simplified workflow did not include the AAM step and, therefore, some other curation procedures that require mapped reaction equations were not considered.
* Statistics on reactions discarded at each step of the protocol are given below table. Here, the difference in initial numbers of reactions and transformations in the USPTO and Pistachio databases is explained by the fact that some transformations proceed under different reaction conditions.



* A large portion of the transformations with invalid valences in all three databases largely stems from problems with the encoding of organometallic stems from problems with the encoding of organometallic molecule chemical bonds in organometallic compounds.
* Almost 7.2 M reactions (26% of the initial one-step reactions dataset) with no reactants or products were found in Reaxys.
* Reactions without either reactants or products are much rarer in USPTO and Pistachio and are often caused by improper text mining.
* Transformations without reactants or products are very rare given their absence in the initial reaction equation.
* In some 200K transformations from Reaxys, 2 K from USPTO and 3.8 K from Pistachio, reactants and products coincide, and, therefore, the curation protocol excludes them from reaction equation and moves to reagents.
* Conclusions: - chemical reactions are complex processes that involve reactants, products, reagents. Their yield and thermodynamic or kinetic parameters depend on reaction conditions.
* The protocol proposed in this paper/work for reactions complexity and consists of four steps related to the curation of (1) individual chemical structures, (2) chemical transformations, (3) reaction conditions and (4) endpoints (only first two points are related in this paper/work).
* Only few software tools can be applied to reactions directly, the others can be applied only to molecular structures extracted from the reaction equation.
* The curation of organometallic compounds poses a particular problem because it cannot be performed with existing tools.
* Transformation curation includes (i) atom-to-atom mapping, (ii) reaction role assignment, (iii) reaction duplicates removal, and (iv) reaction equation balancing.
* First two steps can be done via SMILES and Condensed Graph of Reaction (CGR). The CGR can only prepared for atom-to-atom mapped reaction equations.
* Therefore, less time-consuming SMILES-based operations are recommended as the first step of transformation curation because it may (i) significantly reduce the size of a database due to removal of reaction duplicates, and (ii) simplify reaction equation by reassigning some reactants to reagents.
* The removal of reaction duplicates, transformations without reactants and/or products as well as records containing structures with invalid valences led to the 2-3-fold size reductions.

Timur R. Gimadiev, Arkadii Lin, Valentina A. Afonina, Dinar Batyrshin, Ramil I. Nugmanov, Tagir Akhmetshin, Pavel Sidorov, Natalia Duybakova, Jonas Verhoeven, Joerg Wehner, Hugo Ceulemans, Andrey Gedich, Timur I. Madzhidov, and Alexandre Varnek