ReactionCode: Format for Reaction Searching, Analysis, Classification, Transform, and Encoding/Decoding

Victorien Delannée
National Cancer Institute https://orcid.org/0000-0002-5776-0129
Marc C. Nicklaus (mn1@mail.nih.gov)
National Cancer Institute https://orcid.org/0000-0002-4775-7030

Research article

Keywords: ReactionCode, reaction, encoding, decoding, searching, classification

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-48395/v1

Version of Record: A version of this preprint was published on December 3rd, 2020. See the published version at https://doi.org/10.1186/s13321-020-00476-x.

RESEARCH

ReactionCode: format for reaction searching, analysis, classification, transform, and encoding/decoding

Victorien Delannée* and Marc C. Nicklaus

*Correspondence:
vicorien.delannee@nih.gov
Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, 376 Boyles Street, 21702 Frederic, MD, USA
Full list of author information is available at the end of the article

Abstract

In the past two decades a lot of different formats for molecules and reactions have been created. These formats were mostly developed for the purposes of identifiers, representation, classification, analysis and data exchange. A lot of efforts ha ve been made on molecule formats but only few for reactions where the endeavors have been made mostly by companies leading to proprietary formats. Here, we developed a new open-source format which allows to encode and decode a reaction into multi-layers machine readable code, which aggregates reactants and products into a condensed graph of reaction (CGR). This format is flexible and can be used in a context of reaction similarity searching and classification. It is also designed for database organization, machine learning applications and as a new transform reaction language.

Keywords: ReactionCode; reaction; encoding; decoding; searching; classification

Introduction

Different proprietary and open formats for reactions have been invented over the past 50 years. The first reaction format can probably be attributed to E.J. Corey and W.T. Wipke. They implemented a format based on rules to generate new molecules and integrated it in the first computer-aided organic synthesis program: OCSS (Organic Chemical Simulation of Synthesis). [1] This project split to give birth to LHASA (Logic and Heuristics Applied to Synthetic Analysis) [2, 3] [4] and SECS (Simulation and Evaluation of Chemical Synthesis) [5]. The LHASA team designed the language CHMTRN (CHeMistryTRaNslator), while the SECS group created the ALCHEM (A Language for CHEMistry) language. [6] After their launch, diverse additional reaction transform languages came up along the implementation of programs such as CLASS and IGOR & IGOR2. However, the arrival of SMILES (Simplified Molecular Input Line Entry System) in the late 1980s led to the development of ReactionSMILES and SMIRKS (SMIles ReaKtion Specification). These two formats were largely adopted by the community and are still widely used nowadays. [7–10]

The work around reaction formats has also affected the need for representations and identifiers for data exchange. In the 1990s, Molecular Design Limited (MDL) developed the Chemical Table file (CTfile) format. [11] In this context, the RXNfile and RDfile formats were defined with the objective to store reaction data and quickly

became a reference. RXNfile is used to store the structural information for the reactants and products of a single reaction [11], while RDFile allows one to store a set of RXNs with their associated data. [11] Since then, additional formats have emerged or are under development such as XDfile, MRV, UDM, CMLReact, CDX/CDXML and ReactionSPL. However, none of these formats succeeded in establishing itself widely as the CTfile formats are still much more frequently used. Next to these representations, work on reaction identifiers was also done. The Reaction International Chemical Identifier (RInChI) [12], an application of InChI [13, 14] was recently developed with the objective to offer a unique reaction identifier, which can help to organize and validate reaction databases. [12]

Besides the formats specifically designed to describe reaction transforms and allow easy data exchange, other more versatile formats have been developed in order to try to offer more flexibility and be utilized in different contexts related to reactions. In 1986, Fujita proposed the Imaginary Transition State (ITS) format, which aggregates reactants and products inside a pseudo-molecule in which the bond changes of a reaction are annotated. This pseudo-molecule was created to be used for the purposes of reaction retrieval and design. [15] This format evolved and became known as Condensed Graph of Reaction (CGR). Stored in an SD File, it is mainly employed for machine learning applications, similarity search, and classification. [16, 17] Recently, a SMIRKS-like format for CGR was implemented concomitant with the development of Python-based tools to operate on them (CGRTools) [18]. However, this format cannot be used directly for, e.g., string-based comparisons of reactions. Indeed, all analysis methods using it are based on molecular graph coloration and molecular fragment generated from the CGR [17, 19–21]. Next to the CGR format, three multi-layer formats considering the reaction center and the neighbor atoms have been developed by J.L. Faulon, InfoChem and Elsevier. J.L. Faulon created the reaction signature, where each reactant and product are described as a tree without taking into account the bond type, and calculates the difference between the reactant and products trees. [22] Despite the versatility of this approach, the consideration of only the atom types and their simple connection is a huge limitation. InfoChem developed the reaction ClassCode, which provides a unique identifier (hash) for the reaction center and its two closest atom neighborhood layers. [23] Similarly, Elsevier implemented the BINCODE, which computes, using a pseudo-molecule, a linear string for each layer from the reaction center to the deepest atom neighborhood layers. Each layer contains the atoms that compose it and their connection tables. In addition, the BINCODE also encodes the bond fate and the atom hybridization change. [24] While the ClassCode is limited to a depth of 2 and is strict by its nature as an identifier, the BINCODE appears to offer more flexibility. Indeed, it covers the complete reaction, and its nature as a string allows some modifications for search purposes. However, the BINCODE was made overly generalist by encoding elements into categories (e.g. the halogens Cl, Br, and I have the same encoding). It therefore cannot be used to recover the entire reaction.

To overcome these limitations, we have developed a new format named ReactionCode, which is a multi-layer machine readable code. This open source format is

canonical and designed to be flexible, upgradeable and versatile in order to be applied in a broad range of applications. ReactionCode is particularly useful for reaction similarity searching and classification, but is also conceived for machine learning applications and as a new transform reaction language.

Methodology and Software

ReactionCode format

Structure

The ReactionCode is a multi-layer machine readable code, which is produced from the aggregation of reactants and products into a condensed graph of reaction (CGR) (Figure 1). The ReactionCode is organized into three blocks, which contain their corresponding layers:

1 Reaction center
2 Atoms around the reaction center remaining in the product
3 Leaving atoms around the reaction center (if any)

Each layer is composed of a main sub-layer and up to three optional sub-layers, which describes the stereochemistry, the charges, and the isotope, respectively. A layer starts with a number if it illustrates the reaction center or the remaining group, or a letter if it describes the leaving group. It is always terminated by the symbol '|'.

Main sub-layer The main sub-layer is composed of 4 types of information: the depth, the atom code, the connection table and the atom stoichiometry. This layer starts with the depth followed by ':'. The depth indicates the distance relative to the reaction center. It is expressed in numbers for the reaction center and the remaining group(s) and in letters for the leaving group(s). The atom code is composed of three characters: the first indicates the highest status of the connected bonds encoded using the hexadecimal system (Table S7), the two others encode the atom type (Table S8). Each atom code is followed by a parenthesized connection table, which indicates each bond connected to an atom with a lower index. A bond is encoded by 4 characters: the 1st indicates the bond order in reactants, the 2d encodes the bond order in products (Table S6) and the last two refer to the index of the other atom connected to. The indices are encoded using the hexadecimal system for the atoms to connect that are present in the blocks corresponding to the leaving group (Table S1) and the indices of atoms in the two other blocks are encoded using a lookup table (Table S2). Finally, the square brackets store the atom stoichiometry, i.e. the number of times the same atom is in the products (Example in SI Figure S1).

Optional sub-layers The optional sub-layers qualify the atom and bond in their corresponding layer. Only the sub-layer(s) where a change has to be made are written directly after the end of the main sub-layer. The priority order is: 1) the charge sub-layer (/c), 2) the stereochemistry sub-layer (/s) and 3) the isotope sub-layer (/i) (Figure 3).

1 Charge layer: The charge layer starts with /c and the charge information is contained in a block containing the charged atom index (2 digits) and 2 characters encoding the charge. The first one encodes the state in reactants and the second one the change in products (Table S3). E.g., in /c00HH, "/c" indicates that this layer contains charge information. It having 4 characters means that 1 (4/4) atom has a charge. The only modification is: "00HH". 00 means that the entity at index 00, which is the atom "008", is modified. The third character "H" encodes a negative charge -1, which remains unchanged in products as the fourth character is encoded by the same letter"H".

2 Stereochemistry layer: The stereochemistry layer starts with /s and the relative information is contained in a block containing the atom or bond index (2 digits), which has the corresponding stereochemistry modification and 2 characters encoding the stereochemistry in reactants by the first character and in products by the second one (Tables S4 and S5). E.g., in /s01640364, "/s" indicates that this layer contains stereochemistry information. It having 8 characters means that 2 (8/4) entities (atom(s) and/or bond(s)) have a stereochemistry information. The two modifications are: "0164" and "0346". The first modification is encoded by the first 4 characters "0164". 01 means that the entity at index 01, which is the bond "11GV", is modified. The third character "4" encodes a DOWN bond in reactants, which becomes an UP bond in products indicated by the fourth character "6". The next 4 characters 0364 modify the bond "11GU" from UP to DOWN.

3 Isotope layer:The isotope layer starts with /i and the isotope information is contained in a block containing the isotope atom index (2 digits) and 2 characters encoding the mass difference between the current isotope and the reference. The first one is for the reactants and the second one for products (Table S3). E.g., in /i00JJ02HH "/i" indicates that this layer contains isotope information. It having 8 characters means that 2 (8/4) atoms are isotopes. The two modifications are: "00JJ" and "02HH". 00 means that the entity at index 00, which is the atom "008", is modified. The third character "J" encodes an addition of 2 neutrons to the common isotope. 008 encodes an oxygen with 2 more neutrons, which means that the atom is an ^{18}O. The fourth character "H" is unchanged, which indicates that the atoms in products remains the same isotope.

Encoding/Decoding process

One of the major strengths of ReactionCode is its capacity to be bidirectional: a reaction encoded into ReactionCode can be easily partially or fully decoded to get the reaction back (Figure 4).

In order to generate the ReactionCode, a mapped reaction is necessary. The first step consists in annotating each atom and bond in reactants and products. Three types of annotation are computed:

- atoms and bonds constituting the reaction center
- atoms and bonds present both in reactants and products, which are annotated as the remaining group

- atoms and bonds present in reactants but absent in products (if any), which are annotated as the leaving group

Once the annotation part is finished, reactants and products are aggregated into a CGR. Finally, the ReactionCode is generated from the CGR. Each atom of the CGR is encoded and reverse-ranked by layers. The algorithm starts from the reaction center, reverse-ranks each atom of this layer and makes the connection between them. Then, a Breadth First Search (BFS) algorithm is used to obtain all the surrounding atoms having a depth of 1. These atoms are separated into 2 layers: those belonging to the remaining layer and those that are part of the leaving group. All encoded atoms are reverse-ranked and the connections between each atom with the current and the previous layer are established. The algorithm iterates this procedure until all atoms have been visited (Figure 5).

The decoding process reconstructs the pseudo-molecule from the ReactionCode by transforming each atom code into an atom object and making the bonds between them. This step relies on the chemoinformatics Java libraries contained in CDK (Chemistry Development Kit)[25]. Then, the pseudo-molecule is transformed into reactants and products in order to get the original reaction back. The ReactionCode is set up by default to recover a balanced reaction but the elements present in the leaving group block could be ignored by the user in order to not have them in the products.

ReactionCode software

Java powered by CDK was used to develop the software to generate the ReactionCode, to decode it, to make pseudo-molecules, and to use it as a new transform language. All these functions can be easily used thanks to a CLI (command line interface) and the JAR file can also be directly employed as an API by calling the corresponding class. In addition, this tool has been successfully tested on open source reaction dataset from the USPTO (https://bitbucket.org/dan2097/patent-reaction-extraction/downloads), for which all reactions were successfully encoded.

Encoder The encoder allows one to produce the pseudo-molecules and ReactionCodes. It takes the most common formats as input: SMIRKS (single or a set of SMIRKS in a file), RXN and RDF. The encoder can provide the pseudo-smiles in SDF and in SMILES format and depict them. Finally, the generated ReactionCodes are given in a CSV file.

Decoder The decoder allows one to get the original reaction back. The reactions can be provided as reactionSMILES, SMIRKS, RXN, or RDF. They can also depicted as a PNG file. In addition, a partial reaction can be generated by giving the layers of interest as input (Figure 6).

Transformer Thanks to the structure of ReactionCode, where each layer is only dependent on its previous layers but independent from its subsequent layers, it can be used as a transform language where the ReactionCode is transformed into a pattern applied to a set of reactants (Figure 6). The transformer takes a complete or partial ReactionCode (set of layers) and the reactants as a unique SMILES String or

an SD file. The transformer will generate all unique possible products and outputs them in reactionSMILES, SMIRKS, RXN, or RDF format.

Applications and Results

USPTO reaction data diversity analysis

The USPTO reaction dataset has been used in many machine learning approaches for predicting reactions [26–29]. However, we know of no previous analysis to evaluate the diversity of this dataset. For this purpose, we have used the generated ReactionCodes of each reaction in the USPTO dataset.

To evaluate the diversity, we split the ReactionCodes by incremental layers taking into account a layer and all its previous layers and count the common occurrences. The first part of the analysis consists of extracting all reaction center layers (depth 0) and reverse-sort them as a function of their frequency. In other words, the most frequent reaction center is at the top of the list. Then, the next layers are then processed in the same way until we reach a depth of 9, leading to the generation of 10 CSV files (see supplementary files). Each file starts with the letter 'd' followed by the depth and contains 2 columns: one with the partial ReactionCode and another one with the number of occurrences (number of time this ReactionCode was found in USPTO dataset).

The USPTO is formed of 479,035 reactions. Among these reactions, 9,532 different reaction centers were identified by ReactionCode, i.e. our approach determines that the UPSTO dataset contains 9532 reaction types. The 10 most-represented reaction types in this dataset are found in 203776 (42.5%) of the reactions. 90% of the USPTO dataset is covered by only 400 reaction types, which corresponds to 4.2% of all reaction types identified in this dataset (Figure 7. We finally note that 4,607 reaction types (48.3%) are only represented by one single reaction in the USPTO dataset.

Other applications of ReactionCode

ReactionCode is a format that can be used for multiple purposes. We describe a few of them here.

Reaction balancing correction

Unbalanced reactions (typically, one or more molecules are missing in products) are not uncommon in reaction databases. This can complicate or entirely throw off analyses and work-up of reactions. As ReactionCode aggregates both reactants and products, it can be used to restore the balance of a reaction by encoding and then re-decoding the flawed reaction.

Searching for similar reactions

The ReactionCode is perfectly suited to search for similar reactions in a database as it is in string format. In addition, a wild card can replace each figure or letter. This

can be employed in order to match with any atom, or to ignore the bond order, or any property desired by the user. The syntax of the ReactionCode thus provides the user with a broad flexibility.

Reaction transform language

The ReactionCode is also designed as a new reaction transform language. One or multiple layers can be used to match a set of reactants in order to generate all the possible products and get all possible reactions. This can be easily done by using our software. Note, however, that this approach does not incorporate any knowledge about the actual synthetic accessibility of the proposed reaction (in contrast to CHMTRN/PATRAN [30]) but operates strictly on the basis of pattern matching.

Classification

The structure by layers of the ReactionCode allows one to classify the reaction in order to make statistical analyses, study the diversity or just to have an idea of the contents of a database. A clusterization of reaction data can also be useful in the context of machine learning, for trying to build the best possible training, testing, and validation sets.

Machine Learning

The ReactionCode could be useful for machine learning applications as descriptors or directly for reaction prediction by predicting one or multiple layers. The ReactionCode describes the reaction center and its neighboring environment, which provides additional descriptors compared to current methods.

Compression

In the context of graph databases, the ReactionCode could be used as a tree structure where a node corresponds to a layer. This structure could improve the searching process but also help save disk usage because only the unique layers are stored. This structure permits one to retrieve and regenerate each reaction. Such a tree structure could be used to develop a reaction encoding process. Each layer could be transformed into a bit vector similarly to fingerprints used for molecules, which could allow one to compress a reaction and speed up the reaction comparison process.

Discussion

USPTO analysis

The diversity analysis of the USPTO dataset showed that this database is covered in the vast majority by only about 400 reaction types while conversely 48.3% of the dataset consists of reactions that do not share a common reaction center with any other reaction in the dataset. This analysis shows that the USPTO has an unbalanced diversity with some significantly over-represented reaction types, which may explain the good accuracy of the models predicting reactions. However, as 48.3% of the dataset consists of unique reactions, it may be wise to define an appropriate strategy during training, testing and the validation of a predictive model. The

unique reactions cannot be learned by ML (if they are in the validation dataset it will decrease the score, if they are in the training set, they cannot be validated). In other words, if one does not apply cross-validation, you cannot trust such models. We therefore hope that this first diversity analysis of the USPTO dataset via ReactionCodes be helpful for better sampling during model-building and for future implementations using the USPTO dataset.

Future developement

The ReactionCode was designed to be an upgradeable format. This format is open to the community, which can submit a new version. For instance, the aromatic bonds are encoded with the same character "9", which can fail to encode some tautomeric reactions. In a lot of mapped reactions, the correspondence of the Kekule versions in reactants and products is wrong, which will be considered as a change of the molecule and integrated into the reaction center. To avoid this problem, it is safest in most of the cases to adopt the aromatic annotation. However, if the user is sure of his/her mapping, this parameter can be easily changed and the bond will be encoded as single or double. Besides, it can be of interest to have the number of hydrogens in the ReactionCode for tautomeric reaction studies. This can be easily integrated into the reaction code by modifying one single parameter in the code.

Conclusion

ReactionCode has been implemented as a new, open source, versatile reaction format that avoids the drawbacks of others. The field of its possible applications is large and we believe that it can be profitable for the community working on reactions. Freely available and open source software has been developed to generate the ReactionCode from a reaction format, to convert the ReactionCode to a reaction format, and to use it as a reaction transform language. This program and the source code are available at https://cactus.nci.nih.gov/reactioncode.

Declarations

Availability of data and materials

The software is available at https://cactus.nci.nih.gov/reactioncode and the source code at https://github.com/victoriendelannee/reactioncode. Future updates will be available at the same URLs. The generated data for USPTO diversity analysis is attached to the manuscript as 11 supplementary files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, Center for Cancer Research, National Cancer Institute.

Author's contributions

Victorien Delannée developed the idea of ReactionCode and wrote all the code. Marc Nicklaus has been leading the project.

Acknowledgements

We thank Scott Hutton, Matthew Clark, and Hans Kraut for useful discussions. This work was supported by the Intramural Research Program of the National Institutes of Health, Center for Cancer Research, National Cancer Institute. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

References

1. Corey, E.J., Wipke, W.T.: Computer-assisted design of complex organic syntheses. Science **166**(3902), 178–192 (1969)
2. Corey, E.J., Cramer, R.D., Howe, W.J.: Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. Journal of the American Chemical Society **94**(2), 440–459 (1972). doi:10.1021/ja00757a022. https://doi.org/10.1021/ja00757a022
3. Corey, E.J., Wipke, W.T., Cramer, R.D., Howe, W.J.: Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. Journal of the American Chemical Society **94**(2), 421–430 (1972). doi:10.1021/ja00757a020. https://doi.org/10.1021/ja00757a020
4. Pensak, D.A., Corey, E.J.: 1. LHASA—Logic and Heuristics Applied to Synthetic Analysis, vol. 61, pp. 1–32. ACS Symposium Series, USA (1977). https://pubs.acs.org/doi/pdf/10.1021/bk-1977-0061.ch001. doi:10.1021/bk-1977-0061.ch001. https://pubs.acs.org/doi/abs/10.1021/bk-1977-0061.ch001
5. Wipke, W.T., Ouchi, G.I., Krishnan, S.: Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. Artificial Intelligence **11**(1), 173–193 (1978). doi:10.1016/0004-3702(78)90016-4. Applications to the Sciences and Medicine
6. Yanaka, M., Nakamura, K., Kurumisawa, A., Wipke, W.T.: Automatic knowledge base building for the organic synthesis design program (secs). Tetrahedron Computer Methodology **3**(6, Part A), 359–375 (1990). doi:10.1016/0898-5529(90)90062-D
7. Hunter, R.S., Culver, F.D., A., F.: SMILES User Manual. A Simplified Molecular Input Line Entry System. Includes Extended SMILES for defining fragments. Review Draft, Internal Report, Montana State University, Institute for Biological and Chemical Process Control (IPA), Bozeman, MT (1987)
8. Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (1988). doi:10.1021/ci00057a005. https://pubs.acs.org/doi/pdf/10.1021/ci00057a005
9. Weininger, D., Weininger, A., Weininger, J.L.: SMILES. 2. Algorithm for generation of unique smiles notation. Journal of Chemical Information and Computer Sciences **29**(2), 97–101 (1989). doi:10.1021/ci00062a008. https://pubs.acs.org/doi/pdf/10.1021/ci00062a008
10. Anderson, E., Veith, G.D., D., W.: SMILES: A line notation and computerized interpreter for chemical structures. Report No EPA/600/M-87/021 US Environmental Protection Agency (1990). Report No. EPA/600/M-87/021. U.S. Environmental Protection Agency, Environmental Research Laboratory-Duluth, Duluth, MN 55804
11. Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A., Laufer, J.: Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of Chemical Information and Computer Sciences **32**(3), 244–255 (1992). doi:10.1021/ci00007a012. https://pubs.acs.org/doi/pdf/10.1021/ci00007a012
12. Grethe, G., Goodman, J.M., Allen, C.H.: Internationa Chemical Identifier for Reactions (RInChI). J Cheminform **5**(1), 45 (2013)
13. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I.: InChI - the worldwide chemical structure identifier standard. J Cheminform **5**(1), 7 (2013)
14. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC International Chemical Identifier. J Cheminform **7**, 23 (2015)
15. Fujita, S.: Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. Journal of Chemical Information and Computer Sciences **26**(4), 205–212 (1986). doi:10.1021/ci00052a009. https://pubs.acs.org/doi/pdf/10.1021/ci00052a009
16. Hoonakker, F., Lachiche, N., Varnek, A., Wagner, A.: Condensed Graph of Reaction: Considering a Chemical Reaction As One Single Pseudo Molecule. Springer (2009)
17. de Luca, A., Horvath, D., Marcou, G., Solov'ev, V., Varnek, A.: Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. Journal of Chemical Information and Modeling **52**(9), 2325–2338 (2012). doi:10.1021/ci300149n. PMID: 22894688. https://doi.org/10.1021/ci300149n
18. Nugmanov, R.I., Mukhametgaleev, R.N., Akhmetshin, T., Gimadiev, T.R., Afonina, V.A., Madzhidov, T.I., Varnek, A.: CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. J Chem Inf Model **59**(6), 2516–2521 (2019)
19. Ruggiu, F., Marcou, G., Varnek, A., Horvath, D.: ISIDA Property-Labelled Fragment Descriptors. Molecular Informatics **29**(12), 855–868 (2010). doi:10.1002/minf.201000099. https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201000099
20. Muller, C., Marcou, G., Horvath, D., Aires-de-Sousa, J., Varnek, A.: Models for Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by Automated Algorithms. Journal of Chemical Information and Modeling **52**(12), 3116–3122 (2012). doi:10.1021/ci300418q. PMID: 23167287. https://doi.org/10.1021/ci300418q
21. Glavatskikh, M., Madzhidov, T., Horvath, D., Nugmanov, R., Gimadiev, T., Malakhova, D., Marcou, G., Varnek, A.: Predictive Models for Kinetic Parameters of Cycloaddition Reactions. Molecular Informatics **38**(1-2), 1800077 (2019). doi:10.1002/minf.201800077. https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201800077
22. Faulon, J.-L., Visco, D.P., Pophale, R.S.: The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. Journal of Chemical Information and Computer Sciences **43**(3), 707–720 (2003). doi:10.1021/ci020345w. PMID: 12767129. https://doi.org/10.1021/ci020345w
23. Kraut, H., Eiblmaier, J., Grethe, G., Löw, P., Matuszczyk, H., Saller, H.: Algorithm for Reaction Classification. Journal of Chemical Information and Modeling **53**(11), 2884–2895 (2013). doi:10.1021/ci400442f. PMID: 24102490. https://doi.org/10.1021/ci400442f
24. Elsevier: BinCoder. https://www.elsevier.com/solutions/reaxys. Accessed: 2020-04-06 (n.d.). https://www.elsevier.com/solutions/reaxys

25. Willighagen, E.L., Mayfield, J.W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Cherto, M., Spjuth, O., Torrance, G., Evelo, C.T., Guha, R., Steinbeck, C.: The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform **9**(1), 33 (2017)
26. Coley, C.W., Barzilay, R., Jaakkola, T.S., Green, W.H., Jensen, K.F.: Prediction of organic reaction outcomes using machine learning. ACS Central Science **3**(5), 434–443 (2017). doi:10.1021/acscentsci.7b00064. PMID: 28573205. https://doi.org/10.1021/acscentsci.7b00064
27. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C., Laino, T.: "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem. Sci. **9**, 6091–6098 (2018). doi:10.1039/C8SC02339E
28. Baylon, J.L., Cilfone, N.A., Gulcher, J.R., Chittenden, T.W.: Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. Journal of Chemical Information and Modeling **59**(2), 673–688 (2019). doi:10.1021/acs.jcim.8b00801. PMID: 30642173. https://doi.org/10.1021/acs.jcim.8b00801
29. Bai, R., Zhang, C., Wang, L., Yao, C., Ge, J., Duan, H.: Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level. Molecules **25**(10) (2020)
30. Judson, P., Ihlenfeldt, W.D., Patel, H., Delannée, V., Tarasova, N., Nicklaus, M.C.: Adapting CHMTRN (CHeMistry TRaNslator) for a new use. Journal of Chemical Information and Modeling **0**(ja), (0). doi:10.1021/acs.jcim.0c00448. PMID: 32539385. https://doi.org/10.1021/acs.jcim.0c00448

Figures

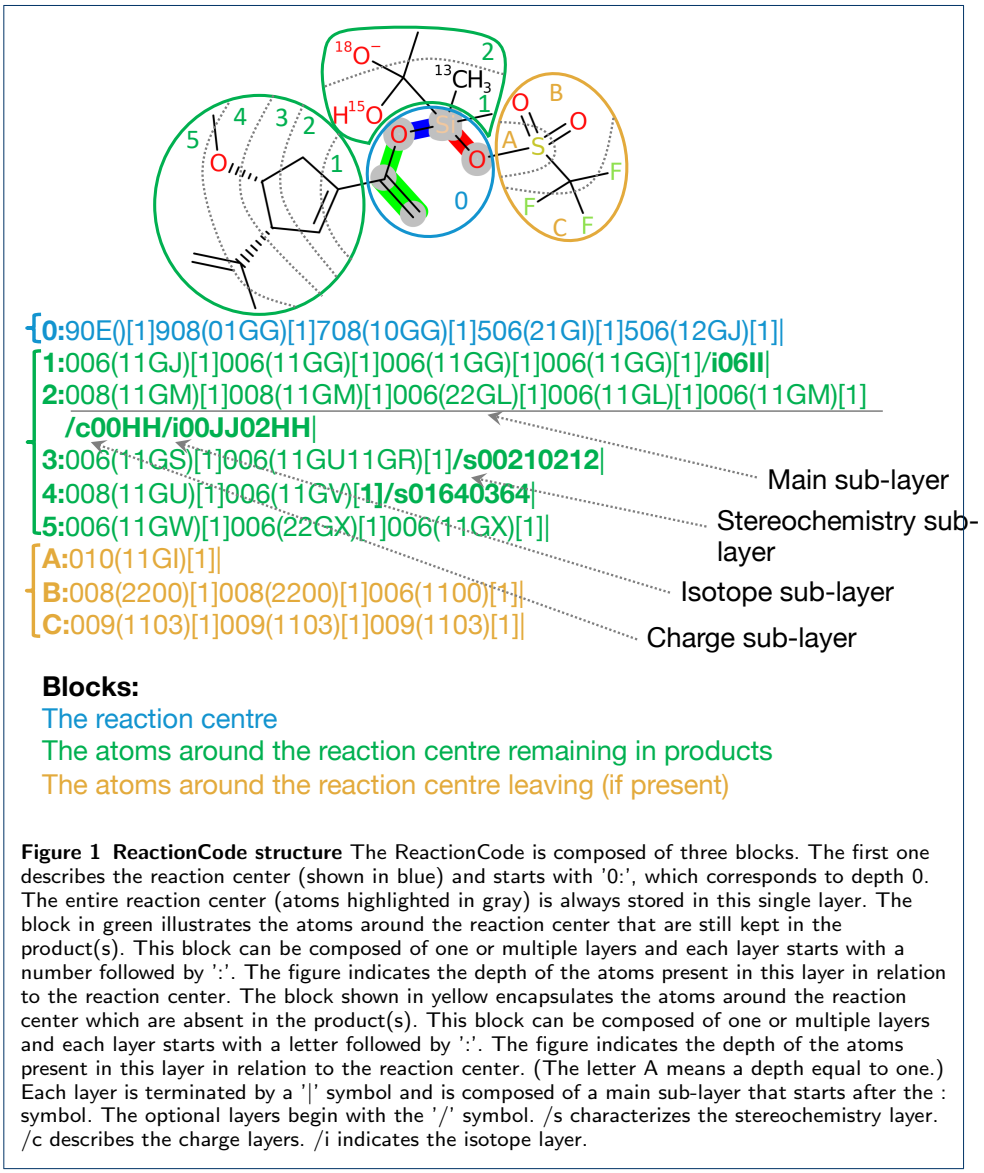


Figure 1 ReactionCode structure The ReactionCode is composed of three blocks. The first one describes the reaction center (shown in blue) and starts with '0:', which corresponds to depth 0. The entire reaction center (atoms highlighted in gray) is always stored in this single layer. The block in green illustrates the atoms around the reaction center that are still kept in the product(s). This block can be composed of one or multiple layers and each layer starts with a number followed by ':'. The figure indicates the depth of the atoms present in this layer in relation to the reaction center. The block shown in yellow encapsulates the atoms around the reaction center which are absent in the product(s). This block can be composed of one or multiple layers and each layer starts with a letter followed by ':'. The figure indicates the depth of the atoms present in this layer in relation to the reaction center. (The letter A means a depth equal to one.) Each layer is terminated by a '|' symbol and is composed of a main sub-layer that starts after the : symbol. The optional layers begin with the '/' symbol. /s characterizes the stereochemistry layer. /c describes the charge layers. /i indicates the isotope layer.

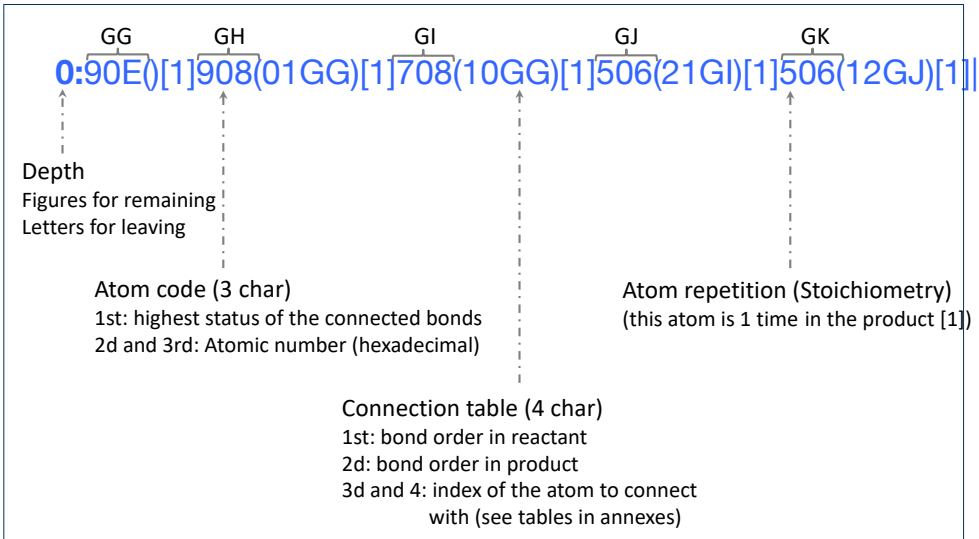


Figure 2 Main sub-layer composition A layer starts with its depth, which is the distance of its atoms from the reaction center (which has depth = 0). The next three characters characterize the atom. The first character indicates the highest bond status among all bonds connected to this atom. (Table S7. For instance, the Si atom (90E) is connected by 4 single bonds: 2 are not part of the reaction centre and are encoded by 0; 1 bond is broken, encoded by 7; and another one is made, encoded by 9, which is the highest bond status. The two other characters encode the atom symbol (Table S8). E.g., 0E stands for Si. The connection table is contained between the brackets. Each bond in the connection table is encoded by 4 characters. E.g., the connection table (11GU11GS) encodes 2 bonds. The first two characters indicate the bond order in reactants for the first one and in products for the second one, respectively (Table S6). 01GG means a bond is made with the atom at index GG. The last two characters represent the index of the other atom to connect to (see Table S2 for the reaction center and the remaining group and Table S1 for the leaving group). The square brackets store the atom stoichiometry, i.e. the number of times a same atom is in the products (Example in SI Figure S1).

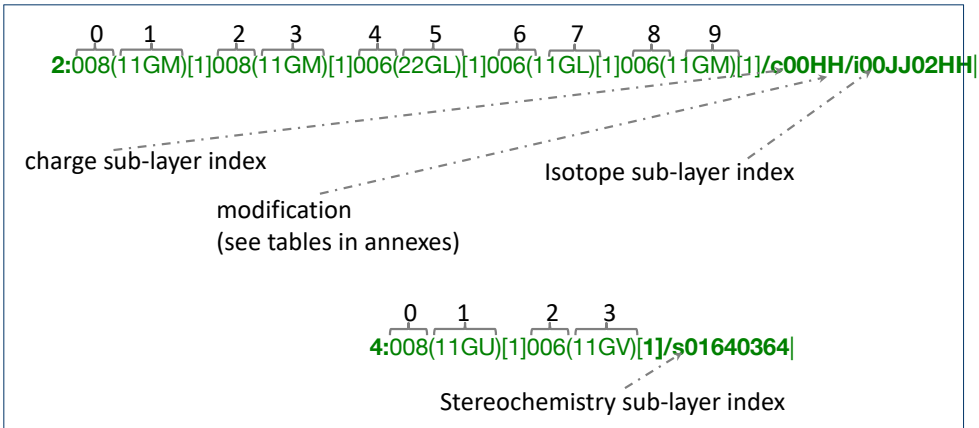


Figure 3 Optional sub-layer composition An optional sub-layer is directly located after the main sub-layer. Such a code is composed by the symbol '/' followed by a letter and indicates the modification (/c for charge, /i for isotope and /s for stereochemistry). Each block qualifying an atom or a bond is composed of four characters. The first two characters are a number (2 characters) indicating the index (decimal system) of the entity (atom or bond) in the current layer which has to be modified. The next 2 characters are encoding the change to apply to reactants (first character) and products (second character) (for isotope and charge see table S3; for atom stereochemistry see table S4; and for bond stereochemistry, see table S5).

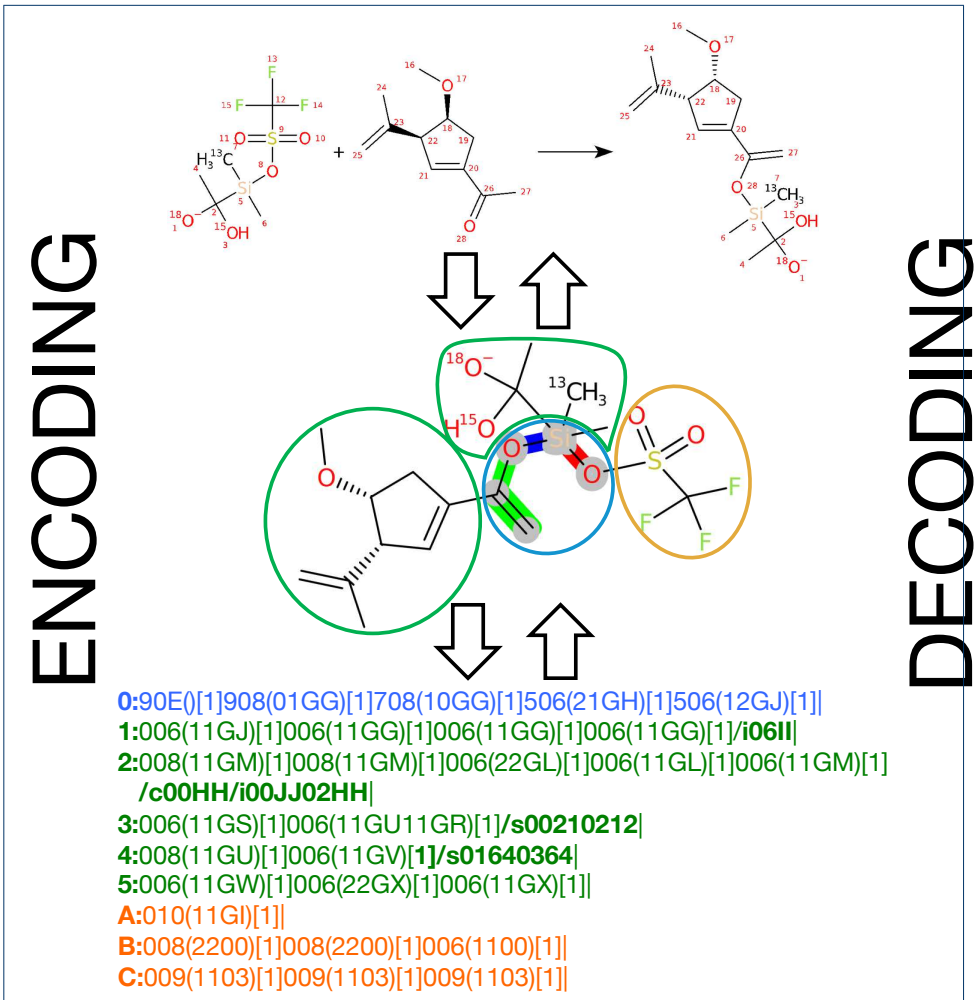


Figure 4 Encoding and decoding process Encoding: The reactants and products of the mapped reaction are aggregated into a pseudo-molecule. The bond changes are annotated: green for a bond order change, blue for a bond made, and red for a bond broken. All atoms and bonds are annotated as part of the reaction center (inside blue circle), remaining in the final product (green circles), or leaving the final product (yellow circle). Finally, all atoms and bonds are encoded into ReactionCode by layers starting from the reaction center to the outermost layer. **Decoding:** The ReactionCode is transformed into a pseudo-molecule, which allows one to recover the initial reaction.

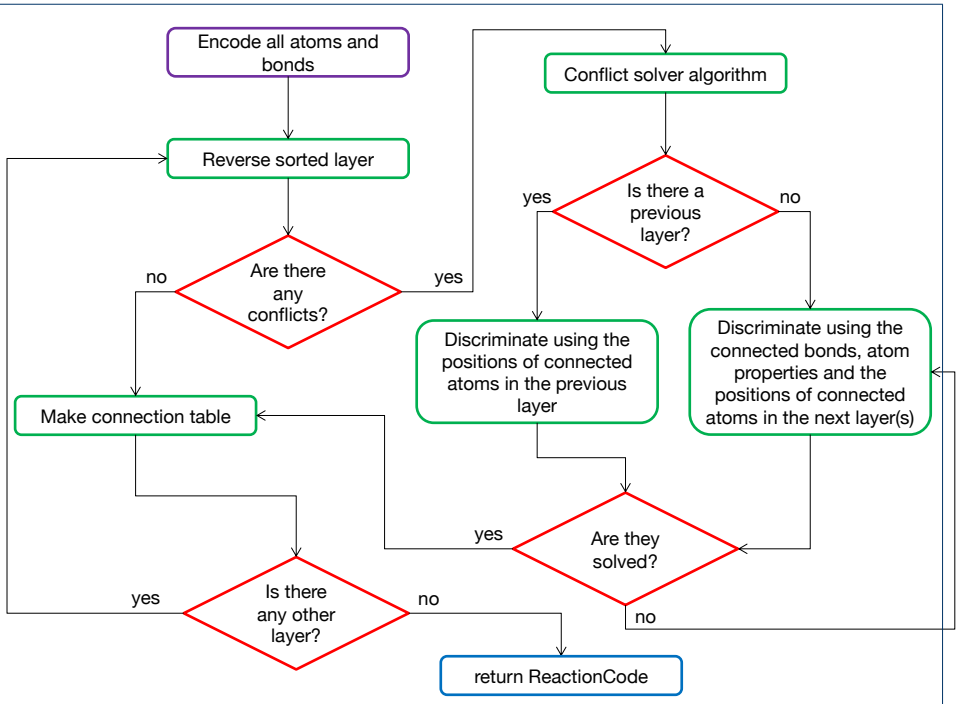


Figure 5 ReactionCode encoding algorithm Once all atoms and bonds are encoded, all encoded atoms present in the reaction center (depth = 0) are reverse sorted. If two or more atoms have the same code, the conflict solver algorithm is started. The conflict algorithm will make a tree of the connected atom (see Figure S2 in Supplementary Information) and compare the atom codes and the bond codes of the atom present in the next layer using a BFS algorithm. The algorithm iterates until the conflicts are solved and a proper order can be set up. Then, the connections between the atoms in the current layer are made. Then, all the atoms present in the depth n+1 (if any), which will be all atoms connected with those in the reaction center in this situation, are reverse sorted. If there are any conflicts, they are first solved by comparing the position of the connected atoms in the previous layer, then using the connected bonds, the atom properties (stereochemistry, charge and isotopy) and finally in the next layer(s) if the conflict cannot be solved. We iterate this process until all layers are processed.

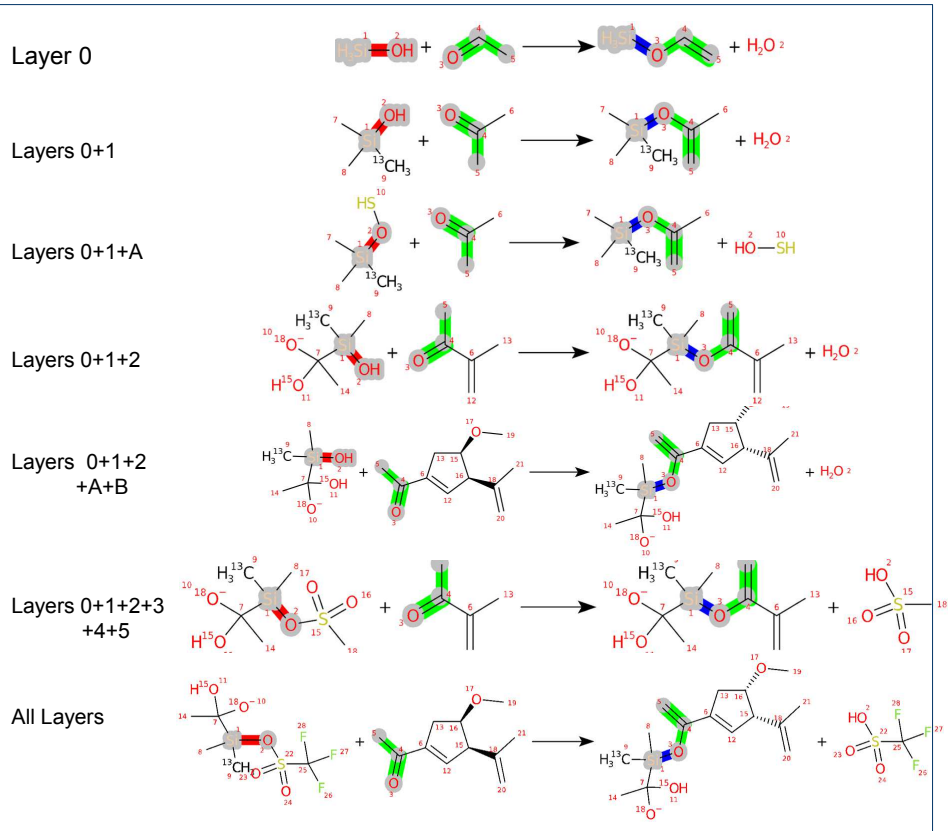
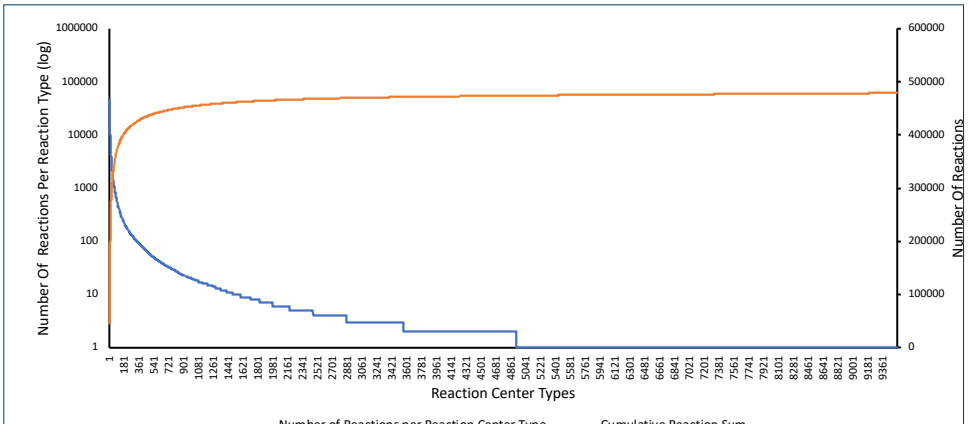


Figure 6 ReactionCode partial and complete decoding The ReactionCode can be decoded by taking all or some layers. The layer 0 corresponds to the decoding of the reaction center only. Layers 0+1 illustrates the decoding of the reaction center and all surrounding atoms in the remaining group having a depth equal to 1, while layers 0+1+A incorporates all surrounding atoms (remaining and leaving groups) with a depth of 1. Layers 0+1+2 and 0+1+2+A+B are considering all surrounding atoms present in the depth lower or equal to 2. Layers 0+1+2+3+4+5 is an example where the reaction is decoded in its entirety but without the leaving group. Finally, "All Layers" represents the decoding of the complete reaction.



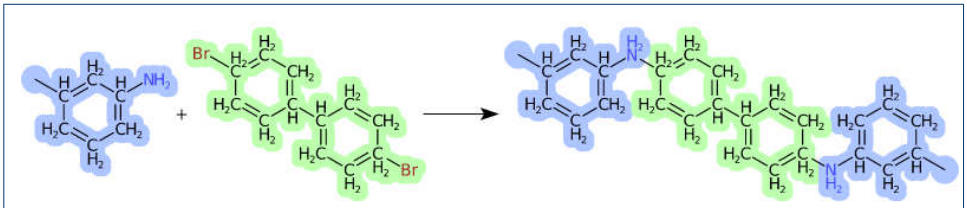
SMIRKS	[N:1].[C:2][Cl:3]>>[N:1][C:2].[Cl:3]	[N:1].[C:2][O:3]>>[N:1][C:2].[O:3]	[O:1][C:2]>>[O:1].[C:2]	[N:1][C:2]>>[N:1].[C:2]	[N:1].[C:2]([O:3])[O:4]>>[N:1][C:2]=[O:4].[O:3]	[N:1].[C:2][Br:3]>>[N:1][C:2].[Br:3]	[O:1].[C:2][Cl:3]>>[O:1][C:2].[Cl:3]	[O:1]=[N+:3][O-:2]>>[O:1].[O-:2].[N:3]	[N:1].[C:2]=[O:3]>>[N:1][C:2].[O:3]	[O:1]([C:2])[C:4]=[O:3]>>[O:1]=[C:4][O:3].[C:2]
ReactionCode	0:907()[1]906(01GG)[1]711(10GH)[1]|	0:907()[1]906(01GG)[1]708(10GH)[1]|	0:708()[1]706(10GG)[1]|	0:707()[1]706(10GG)[1]|	0:907()[1]906(01GG)[1]708(20GH)[1]508(12GH)[1]|	0:907()[1]906(01GG)[1]723(10GH)[1]|	0:908()[1]906(01GG)[1]711(10GH)[1]|	0:708()[1]708()[1]707[20GG10GH][1]/c01HH02I0|	0:907()[1]906(01GG)[1]708(20GH)[1]|	0:708()[1]706(10GG)[1]508()[1]506(21GG)[12GG][1]|
Number of Reactions	45534	29834	28081	19265	18842	17153	12330	11833	11055	9849
Cumulative Reaction Sum	45534	75368	103449	122714	141556	158709	171039	182872	193927	203776

Figure 7 Reaction center diversity analysis The UPSTO contains 9532 different reaction centers (by extension 9532 reaction types). The graph illustrates the diversity of the reaction types of UPSTO dataset. The x-axis corresponds to the reaction types, where 1 indicates the reaction center, which is the most frequent in the dataset and 9432 the one, which is the less frequent. For instance, the reaction center is common to 45534 reactions. The blue line shows the number of reaction center types in the logarithmic scale (the reaction center 1 is present in 45534 reactions, and the reaction center 2 in 29834 reactions). The orange line depicts the cumulative sum of reactions. For example, the cumulative sum for the reaction center 2 is 75368 as the reactions center 1 is identified in 45534 reactions, and the reaction center 2 in 29834 (45534 + 29834 = 75368). The table details the 10 most frequent reaction centers.

Tables

Additional Files

Supplementary Figure S1 — Stoichiometry management



0:907()[2]906(01GG)[1]906(01GG)[1]723(10GH)[1]723(10GI)[1]|
1:006(11GG)[2]006(99GH)[1]006(99GH)[1]006(99GI)[1]006(99GI)[1]|
2:006(99GN)[1]006(99GP)[1]006(99GL)[2]006(99GL)[2]006(99GM)[1]006(99GO)[1]|
3:006(99GU99GQ)[1]006(99GW99GV99GR)[1]006(99GT)[2]006(99GS)[2]|
4:006(99GZ99GY)[2]006(11GY)[2]|

Figure S1 Stoichiometry management The green molecule has 2 Br atoms, which both react with the amine function of the blue molecule. The stoichiometry is 2 blue molecules for and 1 green molecule. Consequently, each atom of the green molecule has to be represented 2 times in the product, which is encoded by [2].

Supplementary Figure S2 — Conflict solver algorithm

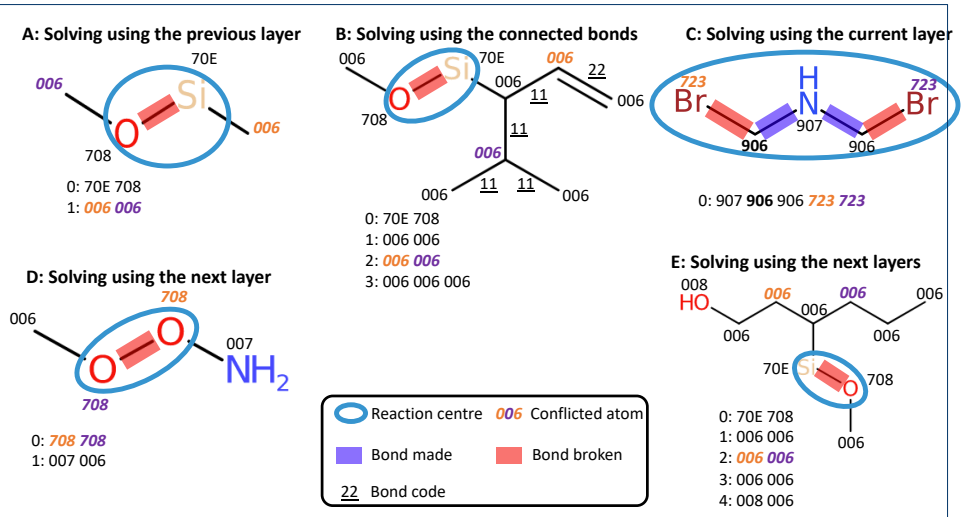


Figure S2 Conflict solver algorithm After ranking each atom according to their codes, a conflict algorithm solves the conflicted atoms (i.e. those sharing the same code). This algorithm discriminates first by looking at atoms in the previous layers. If the conflict can not be solved, the bond type is used. If the two previous processes have failed, the priorities of the connected atoms in the next layer(s) are analyzed. **A:** The carbon atom with an orange atom code has a higher priority because its connected atom (Si atom with atom code 70E) in the previous layer is ranked first, while the carbon atom in purple is connected to the atom at position 2 (O atom with atom code 708). **B:** As the first method failed to solve the conflict, the bonds are taking into account. It takes into account the encoded bond types and the number of connected bonds with both layers. All bond codes related to the connected bond to the atom are reverse sorted and merged into a string. The merged code for carbon atom with an atom code in orange is 2211 and the one for carbon atom with an atom code in purple is 111111. The carbon atom with an atom code in orange is ranked first because the priority of its merged code is higher ($2211 > 111111$) by using the string comparison method: compareTo). **C:** As both previous methods failed to solve the conflict, the atom priorities are analyzed in the current layer. In this example, the two carbons (906) were ranked by looking at the next layers. The Br atom with an atom code in orange is connected to a carbon (906 in bold), which has a higher rank than the Br in purple connected to a different carbon (906). (The carbon 906 in black is positioned 2d, while the other 906 carbon is 3rd). **D:** If the current layer method can't discriminate the conflicts, the atom priorities are analyzed in the next layer. The oxygen atom with an atom code in orange has a higher priority because it is connected to a nitrogen atom having a higher priority than the neighbor (carbon atom) of the oxygen atom with an atom code in purple ($007 > 006$). **E:** In this example, the algorithm iterates over the next layers until being able to solve the conflict. The carbon atom with an atom code in orange has a higher priority because the oxygen atom in layer n+2 has a higher rank than the n+2 carbon atom ($008 > 006$). In the case of symmetry, where both path are equivalent, a unique position is attributed to each atom and is used as a reference for solving the next conflicts (ie sub-figure C, where both 906 carbon are ranked and used to discriminate the ranking of both bromines (723)).

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
2	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
3	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63
4	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
5	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
6	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111
7	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127
8	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
9	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
a	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
b	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
c	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
d	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
e	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
f	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

Table S1 Atom index encoding for reaction center and remaining group

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
G	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
I	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
J	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
K	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
L	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
M	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
N	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
O	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
P	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
Q	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220
R	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
S	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
T	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280
U	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
V	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320
W	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340
X	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360
Y	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380
Z	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400

Table S2 Atom index encoding for leaving group

Supplementary Table S1 — Atom index encoding for reaction center and remaining group
Supplementary Table S2 — Atom index encoding for leaving group
Supplementary TableS3 — Charge and Isotope encoding

C or i	encoding		C or i	encoding
1	I		-1	H
2	J		-2	G
3	K		-3	F
4	L		-4	E
5	M		-5	D
6	N		-6	C
7	O		-7	B
8	P		-8	A
9	Q		-9	9
10	R		-10	8
11	S		-11	7
12	T		-12	6
13	U		-13	5
14	V		-14	4
15	W		-15	3
16	X		-16	2
17	Y		-17	1

Table S3 Charge and Isotope encoding

Supplementary Table S4 — Atom stereochemistry encoding
Supplementary Table S5 — Bond stereochemistry encoding
Supplementary Table S6 — Bond order encoding
Supplementary Table S7 — Bond change status encoding
Supplementary Table S8 — Atom symbol encoding

Symbol	Type	Shorthand	Numeric shorthand	encoding
@TH1	Tetrahedral	ANTI CLOCKWISE (=LEFT)	1	1
@TH2	Tetrahedral	CLOCKWISE (=RIGHT)	2	2
@AL1	ExtendedTetrahedral	ANTI CLOCKWISE (=LEFT)	1	3
@AL2	ExtendedTetrahedral	CLOCKWISE (=RIGHT)	2	4
@DB1	DoubleBond	OPPOSITE	1	5
@DB2	DoubleBond	TOGETHER	2	6
@CT1	ExtendedCisTrans	OPPOSITE	1	7
@CT2	ExtendedCisTrans	TOGETHER	2	8
@SP1	SquarePlanar		1	9
@SP2	SquarePlanar		2	A
@SP3	SquarePlanar		3	B
@TB1	TrigonalBipyramidal	ANTI CLOCKWISE (=LEFT)	1	C
@TB2	TrigonalBipyramidal	CLOCKWISE (=RIGHT)	2	D
@OH1	Octahedral	ANTI CLOCKWISE (=LEFT)	1	E
@OH2	Octahedral	CLOCKWISE (=RIGHT)	2	F
@AP1	Atropisomeric	ANTI CLOCKWISE (=LEFT)	1	G
@AP2	Atropisomeric	CLOCKWISE (=RIGHT)	2	H

Table S4 Atom stereochemistry encoding

Type	encoding
E	1
Z	2
DOWN INVERTED	3
DOWN	4
UP INVERTED	5
UP	6
E or Z	7
UP or DOWN	8
UP or DOWN INVERTED	9

Table S5 Bond stereochemistry encoding

1	SINGLE
2	DOUBLE
3	TRIPLE
4	QUADRUPLE
5	QUINTUPLE
6	SEXTUPLE
9	AROMATIC
0	NONE

Table S6 Bond order encoding

score	meaning	encoding
0	unmarked	0
-1	not a reaction centre	0
1	a reaction centre	1
4	bond order changes	4
6	bond broken	6
8	bond made	8
5	4 + 1 (is a center and order changes)	5
7	6 + 1 (is a center and bond broken)	7
9	8 + 1 (is a center and bond made)	9
10	6 + 4 (bond is both broken and its order changes)	A
11	10 + 1 (is a center, bond broken and order changes)	B
12	8 + 4 (bond is both made and its order changes)	C
13	12 + 1 (is a center, bond made and order changes)	D

Table S7 Bond change status encoding

-1 is increased to 0
10 is reduced to 4
11 is reduced to 5
12 is reduced to 8
13 is reduced to 9

H=1																	He=2
Li=3	Be=4											B=5	C=6	N=07	O=08	F=09	Ne=0A
Na=0B	Mg=0C											Al=0D	Si=0E	P=0F	S=10	Cl=11	Ar=12
K=13	Ca=14	Sc=15	Ti=16	V=17	Cr=18	Mn=19	Fe=1A	Co=1B	Ni=1C	Cu=1D	Zn=1E	Ga=1F	Ge=20	As=21	Se=22	Br=23	Kr=24
Rb=25	Sr=26	Y=27	Zr=28	Nb=29	Mo=2A	Tc=2B	Ru=2C	Rh=2D	Pd=2E	Ag=2F	Cd=30	In=31	Sn=32	Sb=33	Te=34	I=35	Xe=36
Cs=37	Ba=39	La=39	Hf=48	Ta=49	W=4A	Re=4B	Os=4C	Ir=4D	Pt=4E	Au=4F	Hg=50	Tl=51	Pb=52	Bi=53	Po=54	At=55	Rn=56
Fr=57	Ra=58	Ac=59	Rf=68	Db=69	Sg=6A	Bh=6B	Hs=6C	Mt=6D	Ds=6E	Rg=6F	Cn=70	Nh=71	Fl=72	Mc=73	Lv=74	Ts=75	Og=76

La=39	Ce=3A	Pr=3B	Nd=3C	Pm=3D	Sm=3E	Eu=3F	Gd=40	Tb=41	Dy=42	Ho=43	Er=44	Tm=45	Yb=46	Lu=47	
Ac=g9	Th=5A	Pa=5B	U=5C	Np=5D	Pu=5E	Am=5F	Cm=60	Bk=61	Cf=62	Es=63	Fm=64	Md=65	No=66	Lr=67	

R=FF	*=FE	^{2}H=FD	^{3}H=FC

Table S8 Atom symbol encoding

Figures

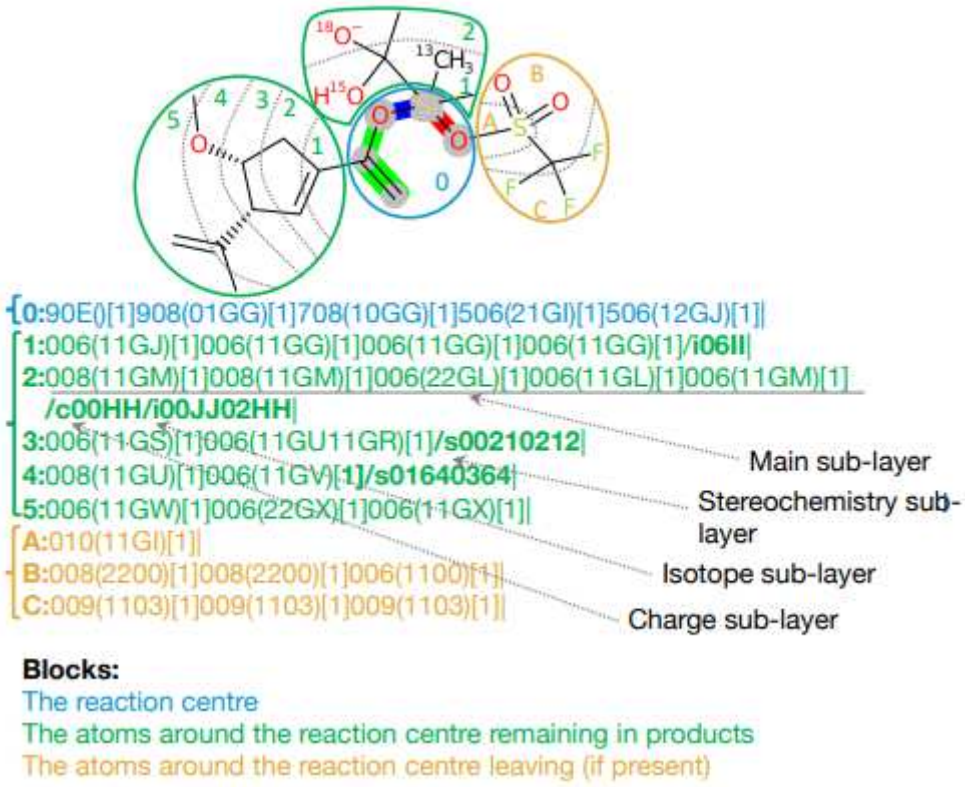


Figure 1

ReactionCode structure The ReactionCode is composed of three blocks. The first one describes the reaction center (shown in blue) and starts with ’0:’, which corresponds to depth 0. The entire reaction center (atoms highlighted in gray) is always stored in this single layer. The block in green illustrates the atoms around the reaction center that are still kept in the product(s). This block can be composed of one or multiple layers and each layer starts with a number followed by ’:’. The figure indicates the depth of the atoms present in this layer in relation to the reaction center. The block shown in yellow encapsulates the atoms around the reaction center which are absent in the product(s). This block can be composed of one or multiple layers and each layer starts with a letter followed by ’:’. The figure indicates the depth of the atoms present in this layer in relation to the reaction center. (The letter A means a depth equal to one.) Each layer is terminated by a ’|’ symbol and is composed of a main sub-layer that starts after the : symbol. The optional layers begin with the ’/’ symbol. /s characterizes the stereochemistry layer. /c describes the charge layers. /i indicates the isotope layer.

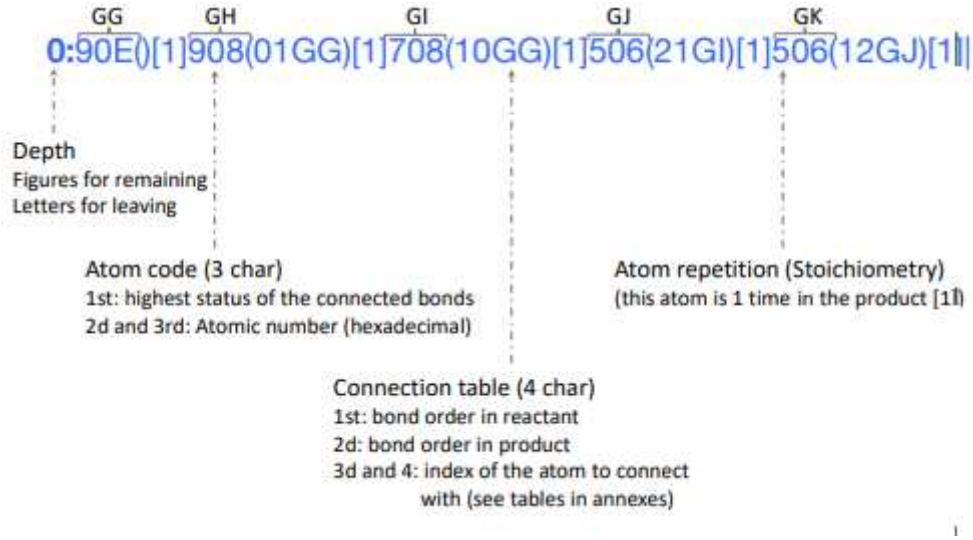


Figure 2

Main sub-layer composition A layer starts with its depth, which is the distance of its atoms from the reaction center (which has depth = 0). The next three characters characterize the atom. The first character indicates the highest bond status among all bonds connected to this atom. (Table S7. For instance, the Si atom (90E) is connected by 4 single bonds: 2 are not part of the reaction centre and are encoded by 0; 1 bond is broken, encoded by 7; and another one is made, encoded by 9, which is the highest bond status. The two other characters encode the atom symbol (Table S8). E.g., 0E stands for Si. The connection table is contained between the brackets. Each bond in the connection table is encoded by 4 characters. E.g., the connection table (11GU11GS) encodes 2 bonds. The first two characters indicate the bond order in reactants for the first one and in products for the second one, respectively (Table S6). 01GG means a bond is made with the atom at index GG. The last two characters represent the index of the other atom to connect to (see Table S2 for the reaction center and the remaining group and Table S1 for the leaving group). The square brackets store the atom stoichiometry, i.e. the number of times a same atom is in the products (Example in SI Figure S1).

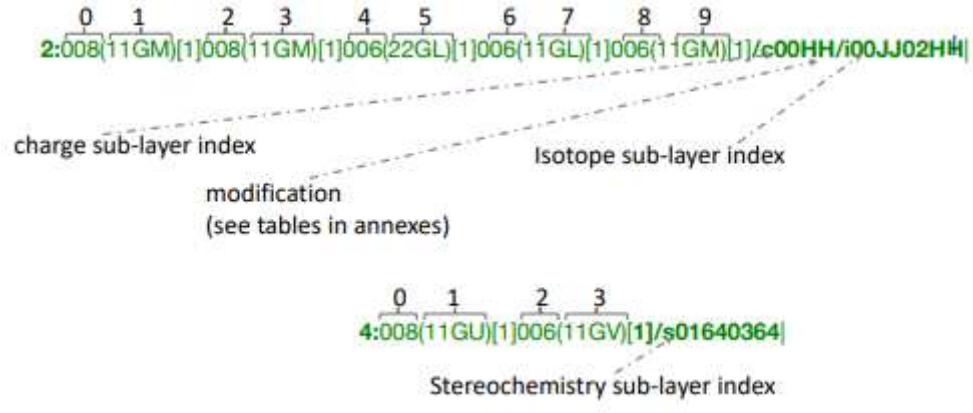


Figure 3

Optional sub-layer composition An optional sub-layer is directly located after the main sub-layer. Such a code is composed by the symbol '/' followed by a letter and indicates the modification (/c for charge, /i for isotope and /s for stereochemistry). Each block qualifying an atom or a bond is composed of four characters. The first two characters are a number (2 characters) indicating the index (decimal system) of the entity (atom or bond) in the current layer which has to be modified. The next 2 characters are encoding the change to apply to reactants (first character) and products (second character) (for isotope and charge see table S3; for atom stereochemistry see table S4; and for bond stereochemistry, see table S5).

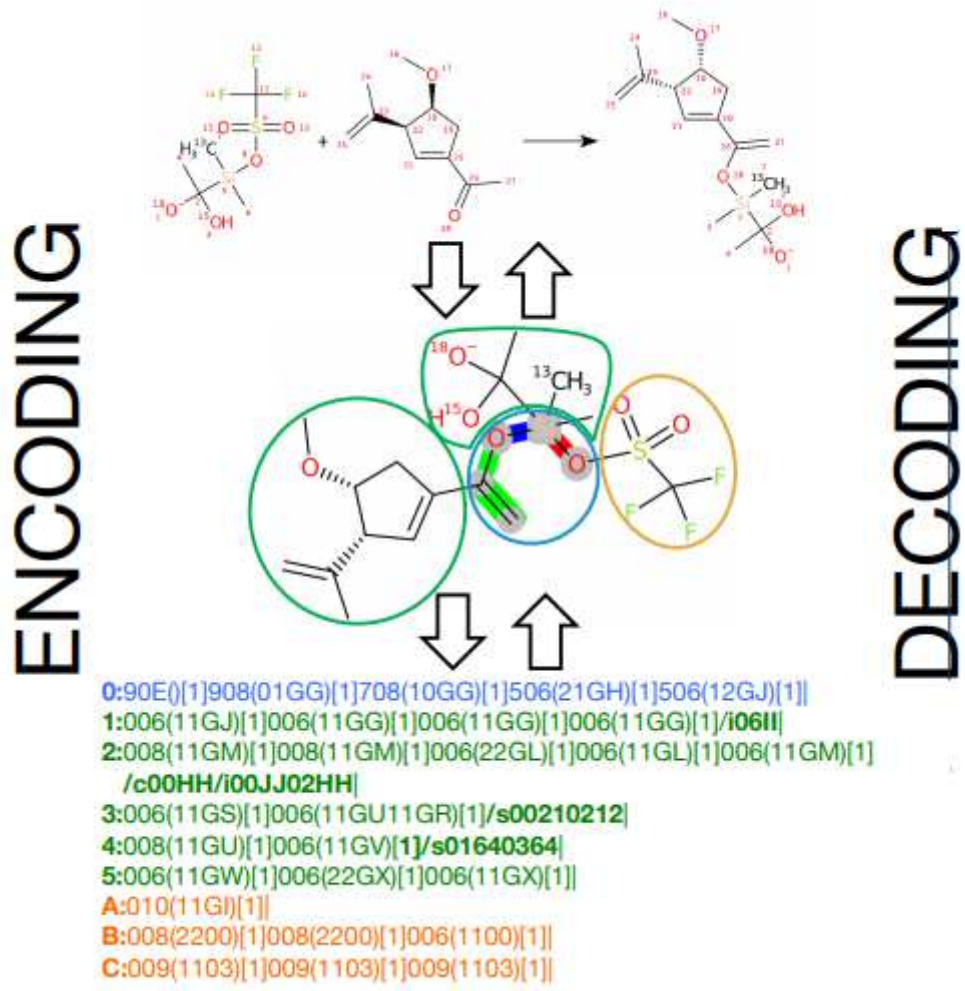


Figure 4

Encoding and decoding process Encoding: The reactants and products of the mapped reaction are aggregated into a pseudo-molecule. The bond changes are annotated: green for a bond order change, blue for a bond made, and red for a bond broken. All atoms and bonds are annotated as part of the reaction center (inside blue circle), remaining in the final product (green circles), or leaving the final product (yellow circle). Finally, all atoms and bonds are encoded into ReactionCode by layers starting from the reaction center to the outermost layer. Decoding: The ReactionCode is transformed into a pseudo-molecule, which allows one to recover the initial reaction

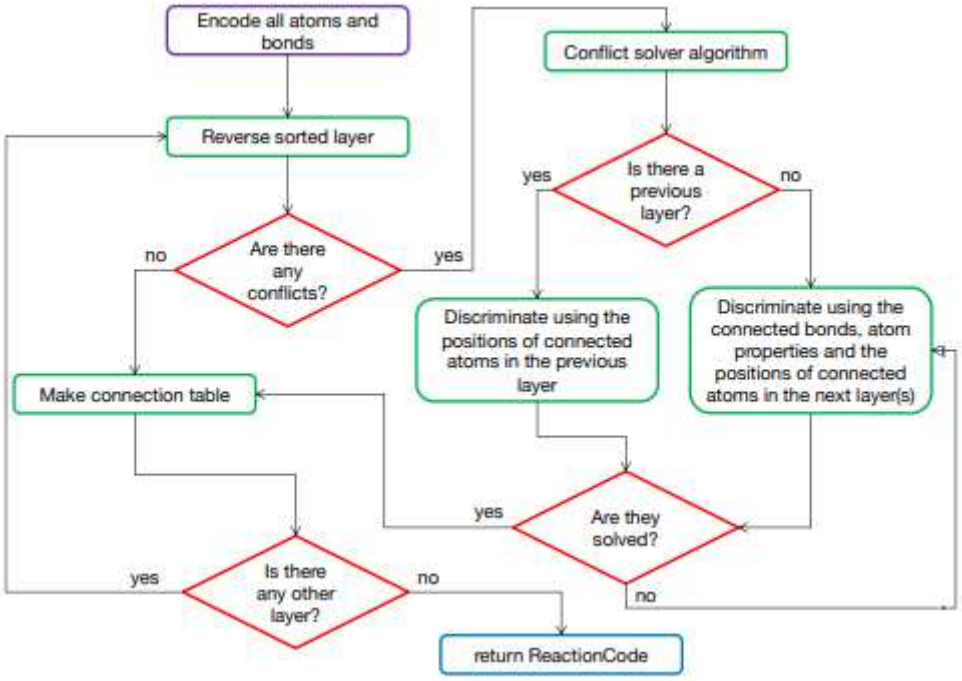


Figure 5

ReactionCode encoding algorithm Once all atoms and bonds are encoded, all encoded atoms present in the reaction center (depth = 0) are reverse sorted. If two or more atoms have the same code, the conflict solver algorithm is started. The conflict algorithm will make a tree of the connected atom (see Figure S2 in Supplementary Information) and compare the atom codes and the bond codes of the atom present in the next layer using a BFS algorithm. The algorithm iterates until the conflicts are solved and a proper order can be set up. Then, the connections between the atoms in the current layer are made. Then, all the atoms present in the depth n+1 (if any), which will be all atoms connected with those in the reaction center in this situation, are reverse sorted. If there are any conflicts, they are first solved by comparing the position of the connected atoms in the previous layer, then using the connected bonds, the atom properties (stereochemistry, charge and isotopy) and finally in the next layer(s) if the conflict cannot be solved. We iterate this process until all layers are processed

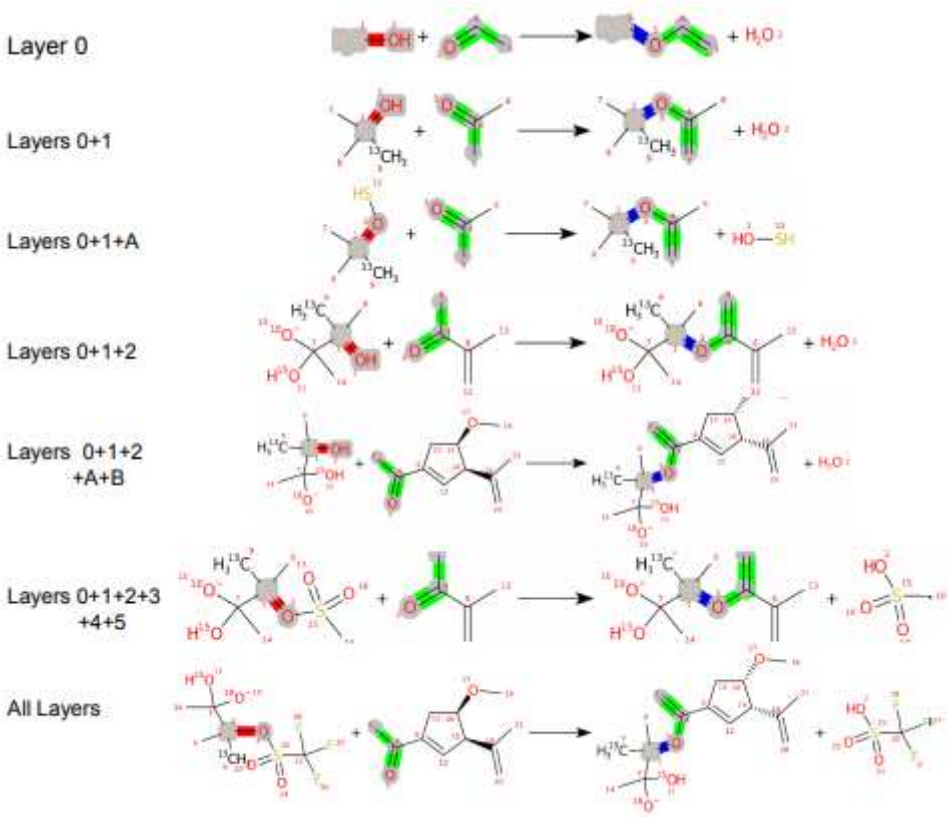
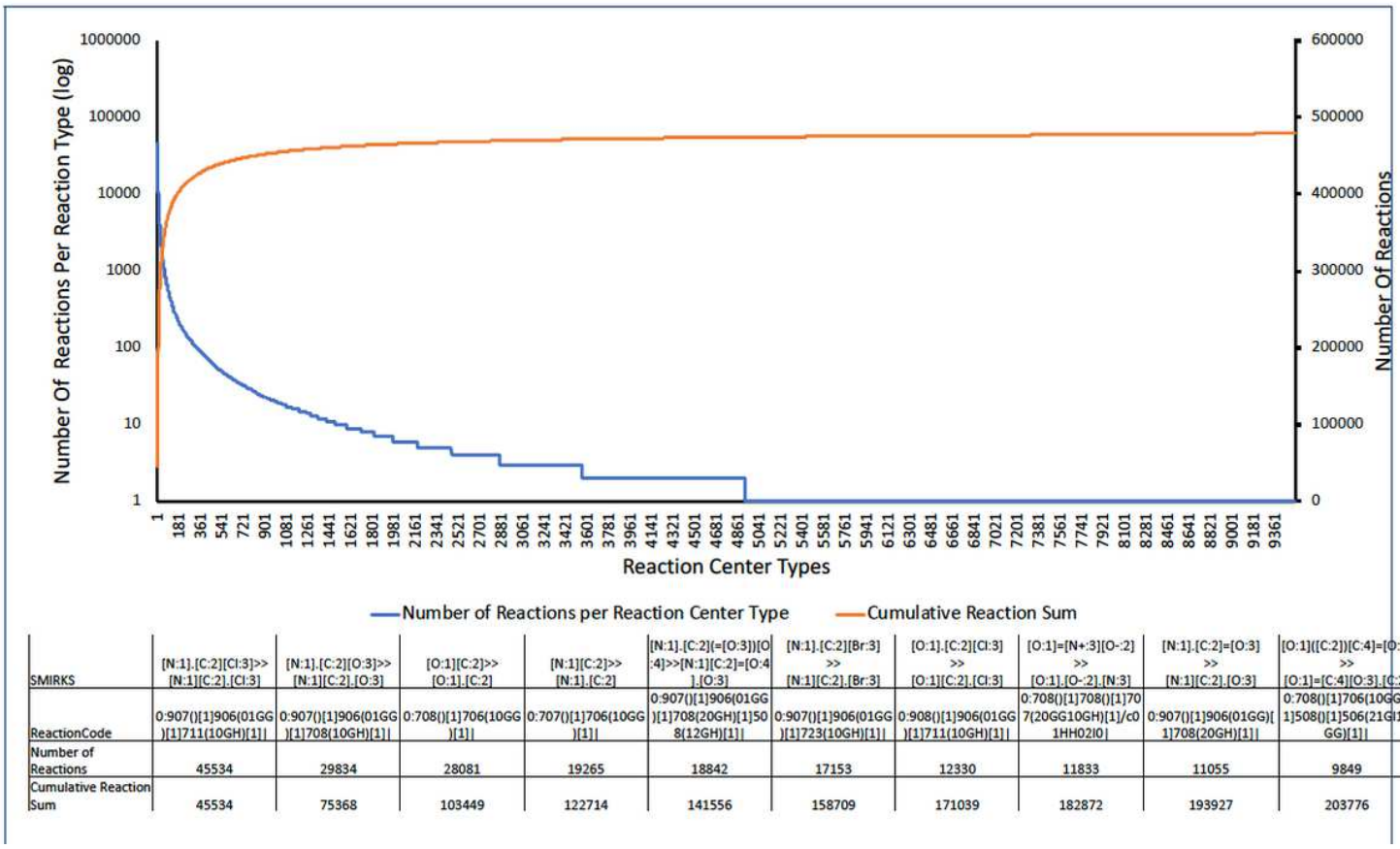


Figure 6

ReactionCode partial and complete decoding The ReactionCode can be decoded by taking all or some layers. The layer 0 corresponds to the decoding of the reaction center only. Layers 0+1 illustrates the decoding of the reaction center and all surrounding atoms in the remaining group having a depth equal to 1, while layers 0+1+A incorporates all surrounding atoms (remaining and leaving groups) with a depth of 1. Layers 0+1+2 and 0+1+2+A+B are considering all surrounding atoms present in the depth lower or equal to 2. Layers 0+1+2+3+4+5 is an example where the reaction is decoded in its entirety but without the leaving group. Finally, ”All Layers” represents the decoding of the complete reaction.



SMIRKS	[N:1].[C:2][Cl:3]>>[N:1][C:2].[Cl:3]	[N:1].[C:2][O:3]>>[N:1][C:2].[O:3]	[O:1][C:2]>>[O:1].[C:2]	[N:1][C:2]>>[N:1].[C:2]	[N:1].[C:2](=[O:3])[O:4]>>[N:1][C:2]=[O:4].[O:3]	[N:1].[C:2][Br:3]>>[N:1][C:2].[Br:3]	[O:1].[C:2][Cl:3]>>[O:1][C:2].[Cl:3]	[O:1]=[N+:3][O-:2]>>[O:1].[O-:2].[N:3]	[N:1].[C:2]=[O:3]>>[N:1][C:2].[O:3]	[O:1]([C:2])[C:4]=[O:3]>>[O:1]=[C:4][O:3].[C:2]
ReactionCode	0:907()[1]906(01GG)[1]711(10GH)[1]|	0:907()[1]906(01GG)[1]708(10GH)[1]|	0:708()[1]706(10GG)[1]|	0:707()[1]706(10GG)[1]|	0:907()[1]906(01GG)[1]708(20GH)[1]508(12GH)[1]|	0:907()[1]906(01GG)[1]723(10GH)[1]|	0:908()[1]906(01GG)[1]711(10GH)[1]|	0:708()[1]708()[1]707(20GG10GH)[1]/c01HH02I0|	0:907()[1]906(01GG)[1]708(20GH)[1]|	0:708()[1]706(10GG)[1]508()[1]506(21GH12GG)[1]|
Number of Reactions	45534	29834	28081	19265	18842	17153	12330	11833	11055	9849
Cumulative Reaction Sum	45534	75368	103449	122714	141556	158709	171039	182872	193927	203776

Figure 7

Reaction center diversity analysis The UPSTO contains 9532 different reaction centers (by extension 9532 reaction types). The graph illustrates the diversity of the reaction types of UPSTO dataset. The x-axis corresponds to the reaction types, where 1 indicates the reaction center, which is the most frequent in the dataset and 9432 the one, which is the less frequent. For instance, the reaction center is common to 45534 reactions. The blue line shows the number of reaction center types in the logarithmic scale (the reaction center 1 is present in 45534 reactions, and the reaction center 2 in 29834 reactions). The orange line depicts the cumulative sum of reactions. For example, the cumulative sum for the reaction center 2 is 75368 as the reactions center 1 is identified in 45534 reactions, and the reaction center 2 in 29834 (45534 + 29834 = 75368). The table details the 10 most frequent reaction centers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ReactionCode1.0.0jarwithdependencies20200828.jar.zip
- encodedreactionsCode.csv.zip
- figConflictSolver.pdf
- figStoichiometry.pdf