data

MDPI

Data Descriptor

RetroTransformDB: A Dataset of Generic Transforms for Retrosynthetic Analysis

Svetlana Avramova, Nikolay Kochev * and Plamen Angelov

Faculty of Chemistry, University of Plovidv "P. Hilendarski", 24 Tsar Assen Str., 4000 Plovdiv, Bulgaria; avramova@uni-plovdiv.net (S.A.); angelov.plamen@gmail.com (P.A.)

* Correspondence: nick@uni-plovdiv.net; Tel.: +359-32-261-447

Received: 28 March 2018; Accepted: 19 April 2018; Published: 21 April 2018

Abstract: Presently, software tools for retrosynthetic analysis are widely used by organic, medicinal, and computational chemists. Rule-based systems extensively use collections of retro-reactions (transforms). While there are many public datasets with reactions in synthetic direction (usually non-generic reactions), there are no publicly-available databases with generic reactions in computer-readable format which can be used for the purposes of retrosynthetic analysis. Here we present RetroTransformDB—a dataset of transforms, compiled and coded in SMIRKS line notation by us. The collection is comprised of more than 100 records, with each one including the reaction name, SMIRKS linear notation, the functional group to be obtained, and the transform type classification. All SMIRKS transforms were tested syntactically, semantically, and from a chemical point of view in different software platforms. The overall dataset design and the retrosynthetic fitness were analyzed and curated by organic chemistry experts. The RetroTransformDB dataset may be used by open-source and commercial software packages, as well as chemoinformatics tools.

Dataset: https://doi.org/10.5281/zenodo.1209312

Keywords: transforms; retrosynthesis; SMIRKS

1. Summary

Retrosynthetic analysis is one of the main tasks in the planning of organic synthesis and a milestone in the computer-aided synthesis design. Different approaches have been proposed [1] and several software systems have been developed as a solution to this issue, including rule-based expert systems, algorithms that use principles of physical chemistry to predict energy barriers of a reaction and machine learning techniques [2]. Among them, rule-based expert systems have been the most widespread approach used for prediction of retrosynthetic routes [3]. It should be noted that the quality of results from rule-based system strongly depends on the available reactions for the purpose of the retrosynthetic analysis—the so-called transforms [4]. And while many retrosynthesis software systems are based on manually coded rules [5–8], some systems [4,9] attempt to automate the rule (transforms) generation process [10] in order to cover more reactions. Applying such an approach is certainly attractive, but the depth of the predictive models that use it strongly depend on the reaction databases they are working with [11]. Large reaction databases can be used for the extraction of reaction rules, but full access to such databases is often strictly limited or expensive [12–15]. There are some freely-available databases of chemical reactions extracted from patents [16] but they describe non-generic reactions which cannot be employed in a generic fashion right way. Another technique for automatic extraction of reaction rules are so-called matched molecular pairs (MMPs) [17,18]. Typically,

MMPs correspond to chemical transformations that lead to changes of the molecular property values. Predominantly, MMP-based transformations do not make sense in a synthetic context, but they are useful for describing a molecule transformation by means of a software algorithm. There are also other chemoinformatics tools (analogous to some extent to MMPs) for automatic extraction of chemical transformations from reaction datasets based on automatic reaction mapping [19]. Although these approaches are promising, it should be noted that they are strongly dependent on the method used to detect the reaction centers (typically done by identifying common or maximal common substructures of the reactant and product structures). Additionally, in most of the cases there is more than one possible reaction mapping. Furthermore, the automatically-generated transformation does not always describe the correct reaction center, or describes it partially or redundantly. Therefore, the manual curating of the obtained transformations is mandatory.

In this context, collections of transforms manually coded from experts are still a reasonable choice and such approaches are capable of providing very promising results [20,21].

Although several datasets with reactions in the synthetic direction [22–25] have been published, as far as we know there is no publicly-available databases with transforms that can be used for the purposes of retrosynthesis. A few collections of transforms found in the literature are not presented in a computer-readable form, which makes a process of their implementation in retrosynthetic analysis software quite difficult. This conclusion is an indication of the need for a dataset with transforms to be created; even more—(1) it should consist of transforms that correspond to generic chemical reactions (or chemical transformations [26]); and (2) the dataset should be presented in computer-readable form, in which transforms can be easily stored, processed, and applied. In accordance with "Nomenclature for organic chemical transformations" [26] and the terminology used in retrosynthetic analysis [27], in this paper we distinguish the following terms: (i) *chemical reaction*—typically describing a specific chemical reaction with concrete reactants, products, agents and conditions; (ii) *chemical transformation*—describing a generic chemical reaction, i.e., the chemical transformation can be related to a set of many ordinary reactions which share the same reaction center; and (iii) *transform* is a retro-reaction—the reverse transformation of a generic chemical reaction.

2. Data Description

We present RetroTransformDB (ver.1.0)—a dataset of transforms for retrosynthetic analysis available in tabular data text file as well as in an Excel (Microsoft Corp., Redmond, WA) spreadsheet file. Each row of the *.txt/xlsx file holds a single transform record consisting of five columns describing a generic chemical transformation in a retrosynthetic fashion (Figure 1).

ID	Name	SMIRKS	FunctionalGroup	TransformType
1	Formation of acetals and ketals from aldehy	[C:7][O:6][C:1][O:3][C:4]>>[C:7][O:6][H].[C:	ACETALS and KETALS	FGE
2	Synthesis of acid chlorides From Carboxylic	[C:3][C:1](=[O:2])[Cl]>>[C:3][C:1](=[O:2])[O	ACID CHLORIDES	FGE
3	Synthesis of alcohols from acid chlorides	[#6:7][C:1](=[O:2])[Cl]>>[#6:7][C:1]([H])([H]	ALCOHOLS	FGE
4	Synthesis of alcohols from anhydrides	[#6:7][C:1](=[O:2])[O][C](=[O])[C]>>[#6:7][(	ALCOHOLS	FGE
5	Reduction of esters to alcohols	[#6:7][C:1](=[O:2])[O:3][#6:4]>>[#6:7][C:1](	ALCOHOLS	FGE
6	Hydration of alkenes	[H][C:2][C:1][O][H]>>[C:1]=[C:2]	ALCOHOLS	FGE
7	Ether formation from alcohols	[C:1][O:2][C:3]>>[C:1][O:2][H].[H][O][C:3]	ETHERS	FGE
8	Reduction of aldehydes/ketones	[H,C:4][C:1]([H])([#6:5])[O:2][H]>>[H,C:4][C	ALCOHOLS	FGE
9	Reduction of conjugated aldehydes/ketones	[H,C:4][C:1]([H])([O:2][H])[C:5][C:3]>>[H,C:	ALCOHOLS	FGE

Figure 1. Generic transform records represented as rows in a spreadsheet file.

The first column of the transform record is used as a formal reaction ID and it is intended for a technical use, such as a fast transform reference, the quick storage of reaction sequences, etc. The second column, "NAME", contains the name of the chemical reaction that is associated with the transform. The transform described in a given record corresponds to a transformation of a molecule that is exactly the reverse of the actual generic reaction designated by the record name. Column "NAME" also can be used for transform identification. The third column, designated in the header as "SMIRKS",

contains the SMIRKS linear notation of the transform and it is the most important field of the record. Column "FunctionalGroup" describes which functional groups will be obtained as a result of transform application, e.g., ALCOHOLS, ETHERS, etc. The last column "TransformType" designates what type of transformation is performed in a retrosynthetic direction e.g., FGE (functional group exchange), C-C (disconnecting C-C bond) etc.

The RetroTransformDB dataset consists of more than 100 SMIRKS line notations corresponding to a wide range of well-known and frequently-used retro-reactions. Each transform was manually created and programmatically tested with the Ambit software platform [28–30]. The entire dataset was additionally curated considering all transformations (generic reactions) and their interconnections in a hierarchical fashion. The presented SMIRKS notations can be used by any chemoinformatics system that supports SMIRKS linear notation.

3. Methods

The SMIRKS linear notation [31] is used for describing the transforms in our collection. The SMIRKS notation is intended to present generic reactions [32] that involve one or more changes in atoms and bonds. Using the full SMARTS [33] syntax, SMIRKS notation is flexible enough to define a set of structural constraints that each reactant should fit to in order the encoded transformation to be applied. In addition, SMIRKS linear notation is easy-editable, widely used and implemented in many software packages and toolkits.

Most of transforms presented in the form of SMIRKS linear notations are characterized by the following general model:

product >> reactant1.reactant2

where the product of a given reaction in a synthetic direction is the target molecule in retrosynthetic analysis, and reactant1 and reactant2 are the precursors, separated by a period ".". The transforms in our collection are mainly in the format product >> reactant1.reactant2 (two-component) or product >> reactant (single-component). Information on solvents and catalysts, if needed, can be described in SMIRKS itself as: product > solvent > reactant1.reactant2 (there are no such transforms in the current version of the collection). In SMIRKS line notations, one should describe only the substructures that directly participate in the transformation (i.e., the reaction center) or such molecule fragments that are considered essential for its reactivity. This description is the basis for the application of generic reactions, illustrated in Figure 2:

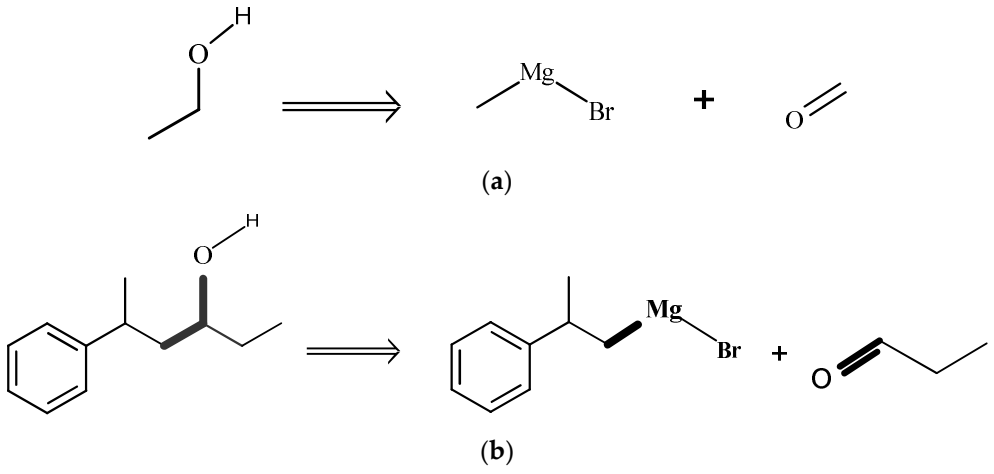

Figure 2. Example of a generic transform (**a**) and its application (**b**).

The rich SMIRKS syntax maintains sufficient functionality for a detailed description of the reaction centers, which is critical to the correct representation of a chemical transformation. The transform for the example illustrated on Figure 1 is as follows:

[C:3][C:1][O:2][H]>>[C:1]=[O:2].[Br][Mg][C:3]

It should be noted that the SMIRKS standard is quite strict, thus, all small details in the linear notation syntax and the encoded logic of chemical expressions must be taken into account. SMIRKS notations in RetroTransformDB are written with explicit H atoms, therefore, it is expected that the used software will apply the SMIRKS transforms against molecules with explicit H atoms. Additionally, the results from the application of a particular SMIRKS may vary in different software systems depending on the level of implementation and the chemoinformatics treatment of the target molecules.

The widespread applicability of some generic reactions was the reason for implementing them in the transforms collection. As a basis for our set, we used the reactions published by D'Angelo and Smith, divided into two main groups: carbon-carbon bond formation and functional group exchange [34]. In addition, transforms are grouped according to the functional group obtained by the application of the transforms. For the sake of completeness of the set of selected reactions, we also used other publications [22–25] to identify additional reactions that would be of interest in retrosynthetic analysis. The choice of reactions for compiling the collection of transforms is based on information from several sources, as each one is relevant to a particular concept. For example, the list of reactions published by Hartenfeller et al. [22], focused on in silico molecule design, does not contain a Diels-Alder cycloaddition, which is, however, one of the most powerful simplifying transforms from the point of view of classical retrosynthesis defined by Corey [27].

The open source application, Ambit-SMIRKS [35], available as a command line interface and GUI, was used for testing the transforms in the RetroTransformDB dataset. An example is given in Figure 3:

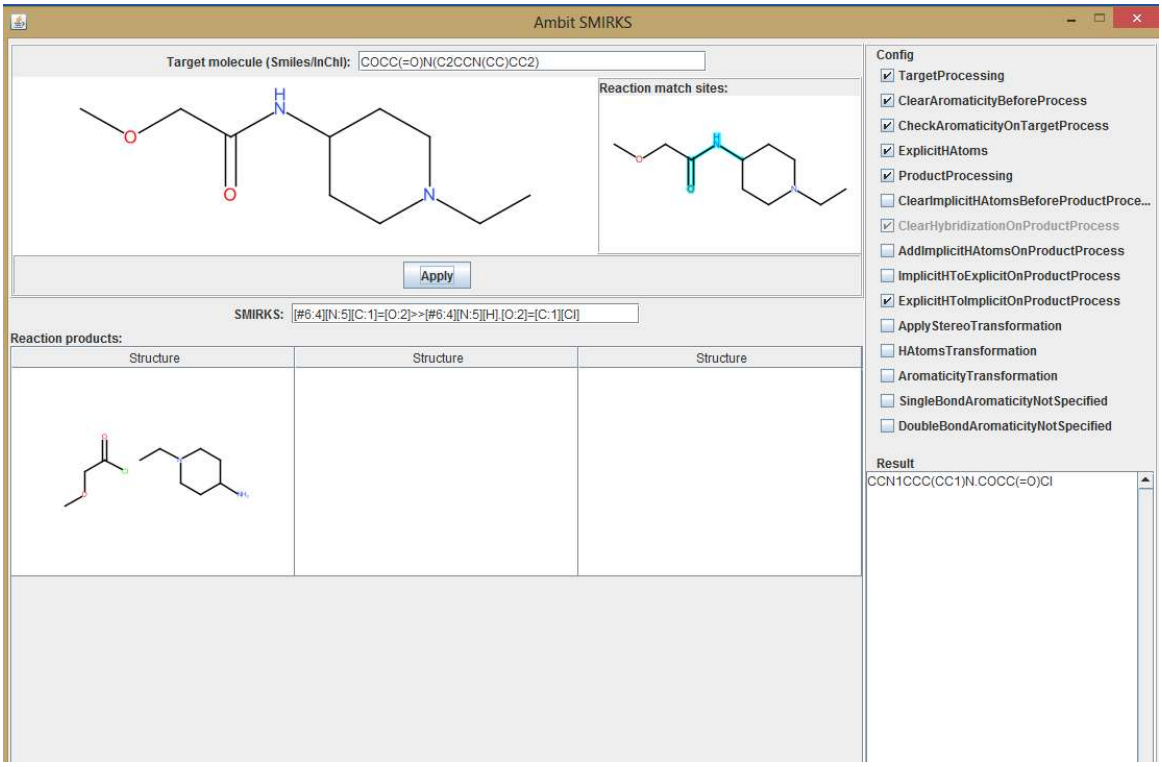

Figure 3. Screenshot of testing transform (corresponding to synthesis of amides from acyl chloride) in the Ambit-SMIRKS GUI.

4. Conclusions

A dataset of appropriate transforms is one of the most crucial elements in every rule-based software for retrosynthetic analysis. Considering the fact that there is not a publicly-available, computer-readable database with retro-reactions, a new collection of transforms (RetroTransformDB) has been presented. The preparation of the transforms selected by the literature using the SMIRKS linear notation is considered to be an appropriate choice as it allows the transforms (retro-reactions) to be clearly defined in a form suitable for computer processing, direct analysis by experts, and edited by users (no special software required). In such a format, the RetroTransformDB collection may be used by multiple software packages and programming tools for molecular modeling.

Further development of the collection (following the Zenoodo future dataset versions) will include a detailed description of more qualitative and quantitative parameters as a type of reaction, yield, experimental conditions, reliability, etc., as well as extending of the dataset with more generic transforms for heterocyclic reactions.

Author Contributions: S.A. prepared all SMIRKS linear notations, performed testing, data analysis, and transforms curation; N.K. performed automatic software tests and SMIRKS syntax and semantic curation; and P.A. performed data analysis, high-level reaction curation, and made the general dataset design from a retrosynthetic point of view. All authors contributed to the writing of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baskin, I.I.; Madzhidov, T.I.; Antipin, I.S.; Varnek, A.A. Artificial intelligence in synthetic chemistry: Achievements and prospects. *Russ. Chem. Rev.* **2017**, *86*, 1127–1156. [CrossRef]
2. Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Nguyen, Q.L.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113. [CrossRef] [PubMed]
3. Segler, M.H.S.; Waller, M.P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. A Eur. J.* **2017**, *23*, 5966–5971. [CrossRef] [PubMed]
4. Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S.Y.; Johnson, A.P.; Major, S.; Wade, R.A.; Ando, H.Y. Route Designer: A retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **2009**, *49*, 593–602. [CrossRef] [PubMed]
5. CMBI—LHASA. Available online: http://cheminf.cmbi.ru.nl/cheminf/olp/history.shtml (accessed on 2 February 2018).
6. Wipke, W.T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. SECS-Simulation and Evaluation of Chemical Synthesis: Strategy and Planning. In *Computer-Assisted Organic Synthesis*; ACS Publications: Washington, DC, USA, 1977; pp. 97–127.
7. Krebsbach, D.; Gelernter, H.; Sieburth, S.M.N. Distributed heuristic synthesis search. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 595–604. [CrossRef]
8. Tanaka, A.; Okamoto, H.; Bersohn, M. Construction of Functional Group Reactivity Database under Various Reaction Conditions Automatically Extracted from Reaction Database in a Synthesis Design System. *J. Chem. Inf. Model.* **2010**, *50*, 327–338. [CrossRef] [PubMed]
9. Huang, Q.; Li, L.L.; Yang, S.Y. RASA: A rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768–2777. [CrossRef] [PubMed]
10. Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. *J. Chem. Inf. Model.* **1999**, *39*, 316–325.
11. Chen, J.H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *J. Chem. Inf. Model.* **2009**, *49*, 2034–2043. [CrossRef] [PubMed]
12. Elsevier, Reaxys Syntheis Planner. Available online: https://www.elsevier.com/solutions/reaxys/how-reaxys-works/synthesis-planner (accessed on 11 February 2018).
13. Reactions—CASREACT. Available online: http://support.cas.org/content/reactions (accessed on 14 January 2018).

14. InfoChem—SPRESI—Storage and Retrieval of Chemical Structure and Reaction Information. Available online: http://www.infochem.de/products/databases/spresi.shtml (accessed on 14 January 2018).
15. Chen, L.; Nourse, J.G.; Christie, B.D.; Leland, B.A.; Grier, D.L. Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1296–1310. [CrossRef] [PubMed]
16. Daniel Lowe, Chemical Reactions from US Patents (1976–Sep 2016). Available online: https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed on 12 April 2018).
17. Hu, Y.; Bajorath, J. Chemical Transformations That Yield Compounds with Distinct Activity Profiles. *ACS Med. Chem. Lett.* **2011**, *2*, 523–527. [CrossRef] [PubMed]
18. Hu, Y.; Bajorath, J. Hierarchical Analysis of Bioactive Matched Molecular Pairs, Encoded Chemical Transformations, and Associated Substructures. *Mol. Inform.* **2016**, *35*, 483–488. [CrossRef] [PubMed]
19. Chen, W.L.; Chen, D.Z.; Taylor, K.T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 560–593. [CrossRef]
20. Szymkuć, S.; Gajewska, E.P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B.A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937. [CrossRef] [PubMed]
21. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M.P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E.P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522–532. [CrossRef]
22. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51*, 3093–3098. [CrossRef] [PubMed]
23. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. Dogs: Reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380. [CrossRef] [PubMed]
24. Masek, B.B.; Baker, D.S.; Dorfman, R.J.; Dubrucq, K.; Francis, V.C.; Nagy, S.; Richey, B.L.; Soltanshahi, F. Multistep Reaction Based de Novo Drug Design: Generating Synthetically Feasible Design Ideas. *J. Chem. Inf. Model.* **2016**, *56*, 605–620. [CrossRef] [PubMed]
25. Schürer, S.C.; Tyagi, P.; Muskal, S.M. Prospective exploration of synthetically feasible, medicinally relevant chemical space. *J. Chem. Inf. Model.* **2005**, *45*, 239–248. [CrossRef] [PubMed]
26. Jones, R.A.Y.; Bunnett, J.F. Nomenclature for organic chemical transformations (Recommendations 1988). *Pure Appl. Chem.* **1989**, *61*, 725–768. [CrossRef]
27. Corey, E.J. *The Logic of Chemical Synthesis*; John Wiley & Sons: Toronto, ON, Canada, 1989.
28. Jeliazkova, N.; Kochev, N.; Jeliazkov, V. ambitcli-3.0.2. 14 April 2016. Available online: https://zenodo.org/record/173560#.WjlcRyvfHVq (accessed on 19 December 2017).
29. Ideaconsult Ltd., AMBIT. Available online: http://ambit.sourceforge.net/ (accessed on 19 December 2017).
30. Jeliazkova, N.; Jeliazkov, V. AMBIT RESTful web services: An implementation of the OpenTox application programming interface. *J. Cheminform.* **2011**, *3*, 18. [CrossRef] [PubMed]
31. Daylight, SMIRKS: A Reaction Transform Language. Available online: http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed on 20 December 2017).
32. Daylight, Reaction Toolkit. Available online: http://www.daylight.com/products/reaction_kit.html (accessed on 20 September 2017).
33. Daylight, SMARTS: A Language for Describing Molecular Patterns. Available online: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed on 19 September 2017).
34. Angelo, J.D.; Smith, M.B. *Hybrid Retrosynthesis*; Elsevier: New York, NY, USA, 2015.
35. Ideaconsult Ltd. Ambit-SMIRKS. Available online: http://ambit.sourceforge.net/smirks.html (accessed on 20 April 2018).