doi:10.2533/chimia.2022.294 Chimia 76 (2022) 294–297

Automatic Extraction of Reaction Templates for Synthesis Prediction

Amol Thakkar[§]* and Jean-Louis Reymond*

[§]SCS-Metrohm award for best oral presentation in Computational Chemistry

Abstract: Several tools for the computational planning of synthetic routes have been developed over the last 60 years. Traditionally these have been built on manually or automatically extracted reaction rules or templates obtained from a deep knowledge of organic chemistry in the case of the former, and reaction databases for the latter. Herein we give an introductory overview on the process of automatically extracting reaction templates, starting from methods for reaction centre identification, through to their use in computer aided synthesis planning and the *de novo* design of compounds.

Keywords: Computer aided synthesis planning · Computational chemistry · Reaction informatics · Retrosynthesis

Amol Thakkar graduated with a Masters in Chemistry with first class honors from the University of St Andrews; during this time he completed a one-year internship with Pfizer as a synthetic chemist in Process Chemistry. Subsequently, he transitioned into computational chemistry and completed his PhD in the Reymond group at the University of Bern in collaboration with AstraZeneca, graduating summa cum laude. His work has focused on Computer Aided Synthesis Prediction resulting in the open-source retrosynthetic planning tool called AiZynthFinder. He is currently a research scientist at IBM Research Zurich, where he continues to explore models to accelerate materials discovery.

1. Introduction

Reaction templates or rules encode the atom, bond, and bond order changes between a set of substances for a given chemical transformation.[1,2] Thus, a reaction template encodes the reaction centre, the automatic extraction of which was first proposed by Vleduts.[3] It follows that given the ability to extract the reaction centre from a set of reaction examples, a knowledge-base of reaction templates can be constructed codifying organic chemistry.[4] In turn, the knowledge base can be applied to the task of synthesis planning, which starting from a molecule of interest aims to predict the most likely steps for its construction from a set of known building blocks.[5]

Herein, we will give an introductory overview on reaction templates as used in organic chemistry, starting from a generalised method for the identification of reaction centres, exemplified by a Claisen rearrangement, through to their use in computer aided synthesis planning (CASP). For a more exhaustive coverage of methods used for reaction centre identification and extraction, we refer the reader to the references within.[1,6]

2. Automatic Template Extraction

The extraction of reaction templates from a set of examples starts with atom–atom mapping (AAM) to identify correspondence of atoms and bonds in the reaction. The reaction centre (RC) can subsequently be extracted by using AAM to identify the atoms and bonds that have changed. Subsequently, the RC can be encoded into a reaction template, and adapted for downstream modelling tasks as outlined in the following section.

2.1 *Atom–Atom Mapping (AAM)*

The identification of the RC and AAM are two closely related problems. Molecules can be represented as graphs,[7] thus the reaction centre can be represented in the form of sub-graphs describing the atom and bond changes between a set of substances, in this case organic compounds. Consequently, graph matching (isomorphism) techniques can be used to compare two or more sets of molecules. The maximum common subgraph (MCS) is defined as the largest substructure common to the collection of graphs under consideration.[6] In the context of reaction centre detection the MCS algorithm can be used to map atoms in the products to those in the reactants (AAM), and in doing so identify the atoms and bonds that have changed during the course of a reaction.[8] While determination of the correspondence between atoms and bonds may be trivial for a human chemist, the determination of the MCS is too computationally complex to solve exactly. Thus, approximate routines are often used to determine AAM. Several alternatives to MCS-based algorithms exist for AAM and have been comprehensively reviewed elsewhere.[2,9] However, a recent AAM benchmarking study found that RXNMapper, a deep learning-based AAM tool outperformed previous approaches obtaining 83.74% on the benchmark dataset.[9,10]

2.2 *The Reaction Centre*

Given that the RC considers all bond changes occurring during the course of a reaction,[3] several approaches have been developed for its representation.

The imaginary transition state (ITS) proposed by Fujita *et. al.* may be intuitive for a chemist to grasp given an understanding

*Correspondence: A. Thakkar and Prof. Dr. J.-L. Reymond, E-mail: amol.thakkar@unibe.ch and jean-louis.reymond@unibe.ch
Dept. Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, CH-3012 Bern

of reaction mechanism.[11] The ITS can be considered a unitary reaction representation, meaning that all molecules involved in the reaction are merged into one molecular graph, and the reaction is considered a pseudomolecule. A more recent example of a unitary representation is the condensed graph of reaction (CGR), by Varnek and co-workers.[12,13] The corresponding Python library can additionally be used for reaction centre extraction and curation.[13] The result is a string representing the reaction in the form of a CGR signature.[13]

So-called 'shell/radius-based' approaches for extracting the reaction centre may also be applied. 'Shell/radius' (*Radius-N*) based approaches capture the reaction centre that examine the atomic environment (AE) up to a pre-specified number of bonds (*N*) away from the reaction centre (Fig. 1c–e), as shown for *Radius-0 to Radius-2* in purple shading. To illustrate this, consider the structural representation of a retro-reaction for the Claisen rearrangement shown in Fig. 1a,b.[14,15] The reaction centre is highlighted in dark purple and constitutes the atoms and bonds that change as a result of the reaction. The atom-mapped reaction SMILES,[16] a string representation commonly used in reaction databases and for modelling tasks is shown alongside the structural representation (Fig. 1b).

The reaction centre can be identified by iterating around the molecule using the atom mapping numbers until a change in the AE can be identified. For each atom, a change in AE is identified by evaluating whether the neighbouring atoms and bonding environment have changed between the reactants and products. If a change has been detected, a component of the RC has been found. The iterations continue until all atoms have been visited, and the fragments identified as having changed are combined to create the reaction centre.

The RC is then extracted in the form of SMARTS patterns (Fig. 1c–e). The simplistic radial approach is followed by the application of heuristics to tune the specificity of the reaction centre by accounting for groups known to influence the reaction.

The open source toolkit RDChiral facilitates reaction centre extraction using the aforementioned approach,[1] in line with approaches such as ARChem (Route Designer),[17] KOSP,[18] and InfoChem's CLASSIFY.[6]

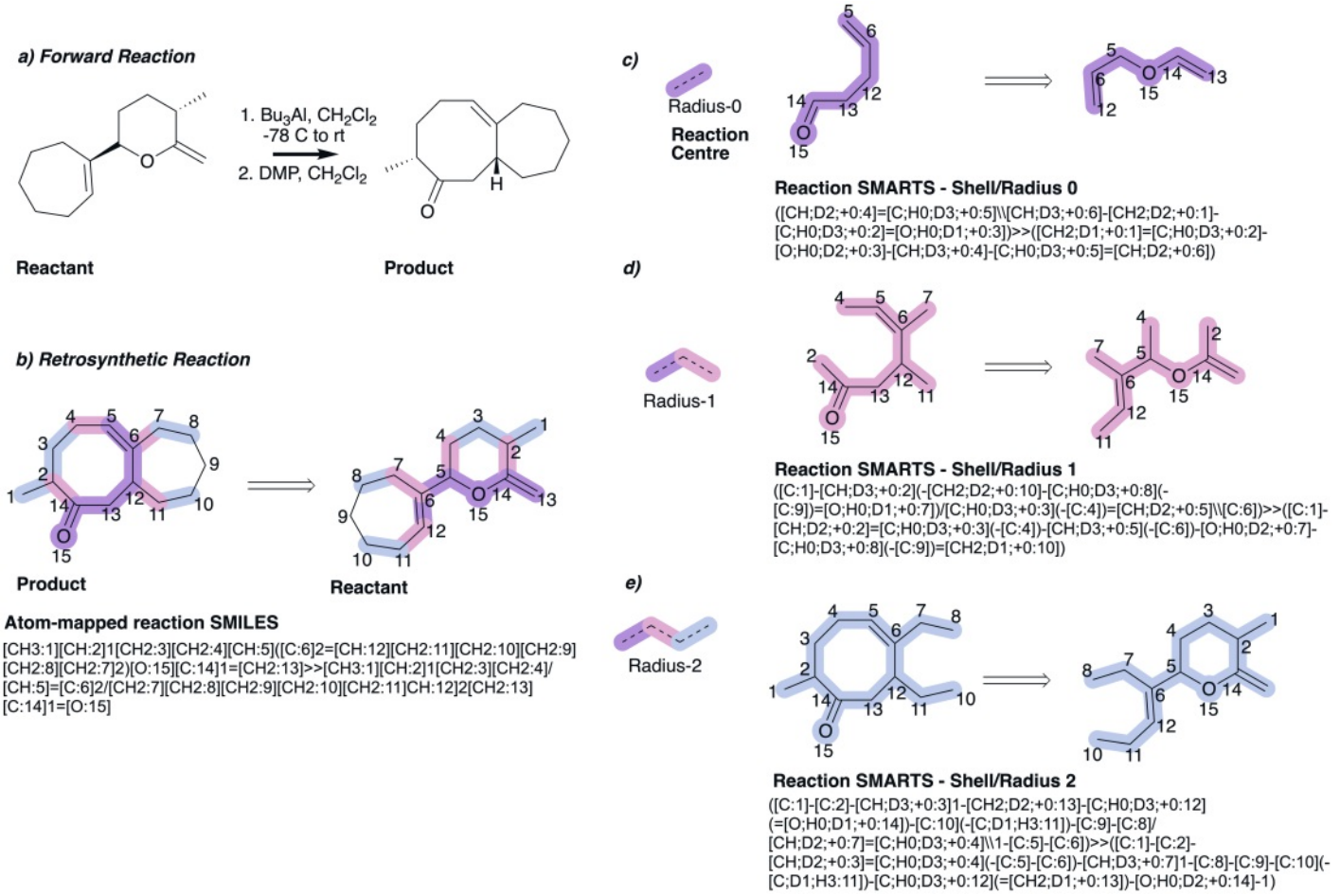

Fig. 1. Automatic template extraction exemplified with a Claisen rearrangement. (a) The Claisen rearrangement as reported in the literature shown in its structural representation (b) The Claisen rearrangement shown as a retrosynthetic reaction alongside the atom-mapped reaction SMILES, commonly used for data-processing tasks. The atom-mapping is annotated on the structure and shows the correspondence between atoms in the reactants and products. The highlighted atoms and bonds correspond to (c) the reaction centre (Radius-0), which constitutes the core atoms and bonds that have changed during the reaction. We see a C–C bond is broken between atoms 12 and 13, and a C–C bond formed between atoms 5 and 15. Additionally all bond orders have been modified. The details of atom, bond, and bond order change are written as reaction SMARTS, shown below the extracted template. The mapping contained in the reaction SMARTS is self-consistent and does not reflect that annotated on the structural representation. (d) depicts the reaction template extracted when the reaction centre is extended one bond away from the reaction centre (Radius-1), alongside the reaction SMARTS. (e) the reaction template extracted when the reaction centre is extended two bonds away from the reaction centre (Radius-2), alongside the reaction SMARTS. The templates correspond to sub-structures extracted from the reaction shown in (b) and have been highlighted accordingly.

2.3 *Size, Specificity, Diversity, and Exclusivity of Reaction Templates*

The size and specificity of reaction templates govern the diversity of reaction centres and their exclusivity. Larger templates, obtained at larger radii, or by extension of the reaction centre are more specific to the substrate from which they were extracted as shown in Fig. 1b–d. This has the advantage of capturing substrate diversity for analytical tasks, however, for synthesis planning the exclusivity of the template means that it cannot broadly be applied to carry out the transformation it encodes. Thus for synthesis planning tasks, non-exclusive templates that describe the same chemical transformation on overlapping sets of molecules are required.[19] In addition, the larger, more specific, and exclusive a template, the greater the number of templates extracted from a reaction database. Thus, the requirements for size, specificity, and exclusivity depend on the downstream task for which the templates are required. In the context of synthesis planning, this can vary depending on the reaction type. For instance, consider enzyme-catalysed reactions for which the bonding motifs required for substrate–enzyme binding are constrained. It is vital to consider both the reaction centre and the groups governing binding to the enzyme active site. Duigou *et. al.* have tackled this issue by specifying a stereochemistry aware set of reaction rules with different levels of specificity.[20]

An issue arising from non-exclusive templates is that multiple templates may encode the same chemical transformation due to variations in the encoding. The variations arise from the existence of multiple solutions to atom-mapping, and the order in which the atoms in the molecule are visited during algorithmic extraction. To address this issue, Heid *et al.* have built upon RDChiral, a template extraction tool, through the development of a canonicalization algorithm to correct automatically extracted templates.[19]

3. Uses in Computer Aided Synthesis Planning

Research into CASP has steadily been making progress since its beginnings in the 1960s.[5] Historically, CASP systems for the prediction of multi-step synthetic routes, and the combinatorial enumeration of virtual libraries has relied upon rule-, or template-based systems, where the terms are used interchangeably.[21–24] Initially, templates were manually encoded based on expert knowledge, however few rule-sets were made publicly available, and those that were, remained limited in chemical diversity and scope.[25,26] The largest rule-base known was developed by the Grzybowski group for the CASP program SYNTHIA (formerly CHEMATICA), consisting of over 70 k manually encoded reaction rules, and has delivered successful synthesis for medicinally relevant and natural product targets.[22,27,28] Given the manpower and time taken to encode such large rule-bases, algorithmic approaches that automatically extract reaction rules have been investigated since the early 1990s.[1,4,6,17,29–34] The debate concerning the quality and scalability of manual versus automatic reaction rule encoding is still ongoing.[35,36] While the process of manual encoding is laborious, the quality of the rules may be higher, and their coverage may be sufficient for the rate at which organic chemistry is growing, argues Grzybowski *et al*.[36] However, purely data-driven approaches to CASP negating the need for templates have now been developed, utilizing the SMILES representation of molecules,[37] combined with developments in natural language processing (NLP) from computer science.[38–40]

As templates encode the reaction centre, they can be applied to a set of reactants to generate the corresponding product (reaction prediction),[34,41] or applied to a product to generate a set of reactants (retrosynthesis).[32,41,42] To determine which reaction template to use for a given set of reactants/products, modern approaches to synthesis planning use neural networks to recommend which reactions, thus templates are suitable for the generation of the desired transformation. The ability of a template to generate an outcome upon application is dependent on there being a substructure match between a template and the substrate to which it is applied. Thus, as discussed above, the requirements for size, specificity, and exclusivity depend on the downstream task for which the templates are required. Approaches for automated template extraction have limitations in that the templates may generalise poorly, as they can be too specific.[19] This has been shown for the task of retrosynthetic prediction, whereby increasing the radius at which the template is extracted, led to a decrease in performance when searching for multi-step synthetic routes.[43]

4. Uses in Synthesis Informed *de novo* Design

Given that templates can be used for reaction prediction and retrosynthesis, it follows that starting from a set of building blocks, templates can be applied strategically to generate *de novo* compounds.[24,25,44–46] These strategies have been used in the past, for instance, for the combinatorial enumeration of virtual libraries. Current approaches combine developments in CASP using neural networks to prioritize and score reaction templates that may be used to generate a set of compounds with a given property profile.[47,48] This overcomes an inherent limitation of enumerative and generative approaches in that synthetic accessibility is considered as a design factor. In doing so, synthetic routes are predicted alongside *de novo* designed compounds.

5. Conclusions

Herein we have given an introductory overview to reaction templates as used in organic chemistry. Starting from methods for the identification of reaction centres, through to their use in computer aided synthesis planning (CASP) and the *de novo* design of compounds. Given the wide use of reaction templates in the field of cheminformatics, there remains a strong interest in developing methods for improved reaction centre identification and extraction. In addition, the development of underlying technologies such as atom-atom mapping, as well as methods addressing the specificity, diversity, and exclusivity of reaction templates are currently under investigation in the community.

Acknowledgements

The authors would like to acknowledge Kleni Mulliri and Aline Carrel for their feedback on the manuscript and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 676434, 'Big Data in Chemistry' ('BIGCHEM,' http://bigchem.eu) for financial support.

Received: February 1, 2022

[1] C. W. Coley, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 2529, https://doi.org/10.1021/acs.jcim.9b00286.
[2] W. L. Chen, D. Z. Chen, K. T. Taylor, *Wiley Interdiscip. Rev. Comput. Mol. Sci*. **2013**, *3*, 560, https://doi.org/10.1002/wcms.1140.
[3] G. É. Vléduts, *Inf. Storage Retr.* **1963**, *1*, 117, https://doi.org/10.1016/0020-0271(63)90013-5.
[4] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 34, https://doi.org/10.1021/ci00023a005.
[5] E. J. Corey, W. T. Wipke, *Science* **1969**, *166*, 178, https://doi.org/10.1126/science.166.3902.178.
[6] H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk, H. Saller, *J. Chem. Inf. Model*. **2013**, *53*, 2884, https://doi.org/10.1021/ci400442f.
[7] W. A. Warr, *Wiley Interdiscip. Rev. Comput. Mol. Sci*. **2011**, *1*, 557, https://doi.org/10.1002/wcms.36.
[8] M. F. Lynch, P. Willett, *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154, https://doi.org/10.1021/ci60015a009.
[9] A. Lin, N. Dyubankova, T. I. Madzhidov, R. I. Nugmanov, J. Verhoeven, T. R. Gimadiev, V. A. Afonina, Z. Ibragimova, A. Rakhimbekova, P. Sidorov, A. Gedich, R. Suleymanov, R. Mukhametgaleev, J. Wegner, H. Ceulemans, A. Varnek, *Mol. Inform*. **2021**, 2100138, https://doi.org/10.1002/minf.202100138.
[10] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, *Sci. Adv.* **2021**, *7*, eabe4166, https://doi.org/10.1126/sciadv.abe4166.

[11] S. Fujita, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205, https://doi.org/10.1021/ci00052a009.
[12] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput. Aided Mol. Des*. **2005**, *19*, 693, https://doi.org/10.1007/s10822-005-9008-0.
[13] R. I. Nugmanov, R. N. Mukhametgaleev, T. Akhmetshin, T. R. Gimadiev, V. A. Afonina, T. I. Madzhidov, A. Varnek, *J. Chem. Inf. Model*. **2019**, *59*, 2516, https://doi.org/10.1021/acs.jcim.9b00102.
[14] T.-W. Sun, W.-W. Ren, Q. Xiao, Y.-F. Tang, Y.-D. Zhang, Y. Li, F.-K. Meng, Y.-F. Liu, M.-Z. Zhao, L.-M. Xu, J.-H. Chen, Z. Yang, *Chem. – Asian J*. **2012**, *7*, 2321, https://doi.org/10.1002/asia.201200363.
[15] Y.-D. Zhang, W.-W. Ren, Y. Lan, Q. Xiao, K. Wang, J. Xu, J.-H. Chen, Z. Yang, *Org. Lett*. **2008**, *10*, 665, https://doi.org/10.1021/ol703126q.
[16] Daylight Theory: SMILES, https://daylight.com/dayhtml/doc/theory/theory.smiles.html#RTFrxn5.
[17] J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model*. **2009**, *49*, 593, https://doi.org/10.1021/ci800228y.
[18] K. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 316, https://doi.org/10.1021/ci980147y.
[19] E. Heid, J. Liu, A. Aude, W. H. Green, *J. Chem. Inf. Model*. **2022**, *62*, 16, https://doi.org/10.1021/acs.jcim.1c01192.
[20] T. Duigou, M. du Lac, P. Carbonell, J.-L. Faulon, *Nucleic Acids Res*. **2019**, *47*, D1229, https://doi.org/10.1093/nar/gky940.
[21] D. A. Pensak, E. J. Corey, in 'Computer-Assisted Organic Synthesis', Vol. 61, American Chemical Society, **1977**, pp. 1, https://doi.org/10.1021/bk-1977-0061.ch001.
[22] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed*. **2016**, *55*, 5904, https://doi.org/10.1002/anie.201506101.
[23] W. T. Wipke, G. I. Ouchi, S. Krishnan, *Artif. Intell*. **1978**, *11*, 173, https://doi.org/10.1016/0004-3702(78)90016-4.
[24] H. M. Vinkers, M. R. De Jonge, F. F. D. Daeyaert, J. Heeres, L. M. H. Koymans, J. H. Van Lenthe, P. J. Lewi, H. Timmerman, K. V. Aken, P. A. J. Janssen, *J. Med. Chem*. **2003**, *46*, 2765, https://doi.org/10.1021/jm030809x.
[25] M. Hartenfeller, M. Eberle, P. Meier, C. Nieto-Oberhuber, K.-H. Altmann, G. Schneider, E. Jacoby, S. Renner, *J. Chem. Inf. Model*. **2011**, *51*, 3093, https://doi.org/10.1021/ci200379p.
[26] S. Avramova, N. Kochev, P. Angelov, *Data* **2018**, *3*, 14, https://doi.org/10.3390/data3020014.
[27] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem* **2018**, *4*, 522, https://doi.org/10.1016/j.chempr.2018.02.002.
[28] B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83, https://doi.org/10.1038/s41586-020-2855-y.
[29] E. S. Blurock, *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 505, https://doi.org/10.1021/ci00068a024.
[30] A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev*. **2015**, *19*, 357, https://doi.org/10.1021/op500373e.
[31] C. D. Christ, M. Zentgraf, J. M. Kriegl, *J. Chem. Inf. Model.* **2012**, *52*, 1745, https://doi.org/10.1021/ci300116p.
[32] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604, https://doi.org/10.1038/nature25978.
[33] H. Gelernter, J. R. Rose, C. Chen, *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 492, https://doi.org/10.1021/ci00068a023.
[34] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci*. **2017**, *3*, 434, https://doi.org/10.1021/acscentsci.7b00064.
[35] K. Molga, E. P. Gajewska, S. Szymkuć, B. A. Grzybowski, *React. Chem. Eng*. **2019**, *4*, 1506, https://doi.org/10.1039/C9RE00076C.
[36] S. Szymkuć, T. Badowski, B. A. Grzybowski, *Angew. Chem. Int. Ed*. **2021**, *60*, 26226, https://doi.org/10.1002/anie.202111540.
[37] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31, https://doi.org/10.1021/ci00057a005.
[38] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci*. **2017**, *3*, 1103, https://doi.org/10.1021/acscentsci.7b00303.
[39] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci*. **2019**, *5*, 1572, https://doi.org/10.1021/acscentsci.9b00576.
[40] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, E. Ozkirimli, *Drug Discov. Today* **2020**, *25*, 689, https://doi.org/10.1016/j.drudis.2020.01.020.
[41] M. H. S. Segler, M. P. Waller, *Chem. – Eur. J.* **2017**, *23*, 5966, https://doi.org/10.1002/chem.201605499.
[42] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, E. Bjerrum, *J. Cheminformatics* **2020**, *12*, 70, https://doi.org/10.1186/s13321-020-00472-1.
[43] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Chem. Sci.* **2020**, *11*, 154, https://doi.org/10.1039/C9SC04944D.
[44] H. Patel, W.-D. Ihlenfeldt, P. N. Judson, Y. S. Moroz, Y. Pevzner, M. L. Peach, V. Delannée, N. I. Tarasova, M. C. Nicklaus, *Sci. Data* **2020**, *7*, 384, https://doi.org/10.1038/s41597-020-00727-4.
[45] Y. Zabolotna, D. M. Volochnyuk, S. V. Ryabukhin, K. Gavrylenko, D. Horvath, O. Klimchuk, O. Oksiuta, G. Marcou, A. Varnek, *J. Chem. Inf. Model*. **2021**, https://doi.org/10.1021/acs.jcim.1c00754.
[46] G. M. Ghiandoni, M. J. Bodkin, B. Chen, D. Hristozov, J. E. A. Wallace, J. Webster, V. J. Gillet, *Mol. Inform.* **2021**, 2100207, https://doi.org/10.1002/minf.202100207.
[47] W. Gao, R. Mercado, C. W. Coley, *ArXiv Prepr.* ArXiv211006389 2021.
[48] S. Krishna Gottipati, B. Sattarov, S. Niu, Y. Pathak, H. Wei, S. Liu, K. M. J. Thomas, S. Blackburn, C. W. Coley, J. Tang, S. Chandar, Y. Bengio, *arXiv* **2020**, arXiv:2004.12485.