Substituents of life: The most common substituent patterns present in natural products

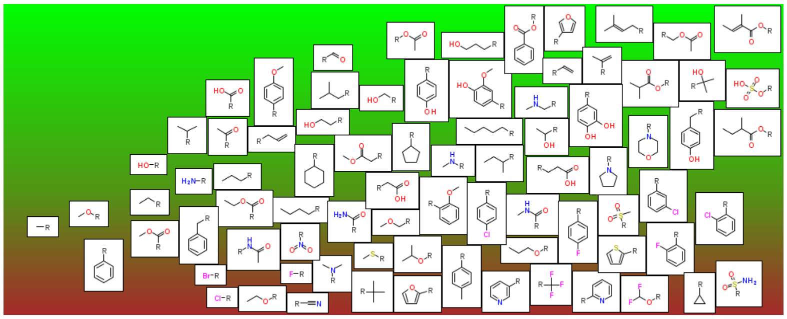
Peter Ertl
https://peter-ertl.com

Novartis Institutes for BioMedical Research, CH-4056, Basel, Switzerland

Abstract

Comparison of substituents present in natural products with the substituents found in average synthetic molecules revealed considerable differences between these 2 groups. The natural products substituents contain mostly oxygen atoms and very little other heteroatoms, are structurally more complex, often containing double bonds and are rich in stereocenters. Substituents found in synthetic molecules contain nitrogen and sulfur atoms, halogenes and more aromatic and particularly heteroaromatic rings. The characteristics of substituents typical for natural products identified here can be useful in the medicinal chemistry context, for example to guide the synthesis of natural product-like libraries and natural product-inspired fragment collections. The results may be used also to support compound derivatization strategies and the design of pseudo-natural natural products.

Graphical abstract



Keywords
natural products, substituents, substructure analysis, chemical space, natural product-likeness, cheminformatics

1. Introduction

It is well known that the structures of natural products (NPs) differ considerably from those of synthetic molecules [1]. This has been documented by several publications focusing on general substructure features [2], scaffolds [3,4] or functional groups [5]. So far, however, no study focused on the detailed analysis of substituents typical for NPs has been performed. The knowledge of substitution patterns that are typical for NPs can be used in several medicinal chemistry activities, including the enrichment of synthetic scaffolds by NP substituents, design of NP-like libraries, or generation of pseudo-natural products - synthetic molecules that contain typical structural features of NPs [6].

Practically all cheminformatics studies of substituents published so far focused on the analysis of drugs or synthetic bioactive molecules. Bemis and Murcko analysed substituents in about 5000 marketed drugs and concluded that their diversity is relatively low with only few common substituents [7]. Similar results were obtained also in a large scale study where 3 million drug-like molecules were processed to identify about 850000 substituents with up to 12 atoms, but only 400 of them were common, i.e. present in more than 0.1% molecules in the database [8]. Effects of substituents on the ligand potency of 30 different protein targets was investigated in the study [9]. The authors found out that certain substituents consistently bias the bioactivity distribution toward higher or lower potency, suggesting the existence of preferred and nonpreferred chemical groups for lead optimization. Influence of substituents on the activity cliffs has been investigated [10] and based on the results a compendium of substitutions with general activity cliff-forming potential was compiled to provide an aid in compound optimization efforts. In the study [11] effect of substituents on the molecule ADME properties, including CYP inhibition, hERG inhibition and permeability through artificial membranes was investigated what helped to formulate "rules of thumb" how to better understand effect of substituent on the ADME properties. In another study this topic was extended to the matched molecular series - groups of molecules having the same core and differing only in one substituent [12]. Recently an analysis of frequent substituents replacement identified from medicinal chemistry literature was performed [13] offering also a list of frequent replacement hierarchies.

The only study focusing on the analysis of substituents in NPs is according to our knowledge a recent publication about sugar substitution patterns in NPs [14] where authors studied the presence of circular, linear, terminal, and non-terminal glycosidic units in NPs, together with their importance in drug discovery.

2. Computational Methodology

In the present study substituents present in natural products were analyzed and compared with the “common” substituents found in bioactive molecules and in various commercial compound collections. The goal of the study was to identify characteristics that are typical for NP substituents and separate them from the common synthetic substituents.

2.1 Data Sources

The NP substituents were extracted from molecules in the COCONUT database [15]. This database (COlleCtion of Open Natural prodUcTs) is an aggregation of more than 50 open data sources, containing in its 6th release structures of more than 405,000 molecules. This freely available database is an indispensable resource for NP chemists and cheminformatics scientists alike. A lot of effort has been devoted to curation and validation of the data, but due to the fact that this database contains molecules from many sources of various origin and quality, the data inevitably contains some errors, particularly many molecules that are not NPs but are of synthetic origin. Only 37% of structures in the COCONUT database contain information about the source organism, identifying them clearly as NPs, other entries contain only a chemical structure without provenance information. Therefore before extracting the substituents the effort has been made to remove the structure that are not NPs, by discarding all molecules that have the NP-score [16] less than zero. This procedure, although probably also removing numerous true NPs, provided a cleaner data set with 222000 molecules, where one can be sure that the majority of structures are true NPs. The substituents have been then extracted from this cleaned data set.

The synthetic molecules were represented by one million molecules from the ZINC database [17] that contains commercially available structures. These molecules well represent the chemical space of the “common” synthetic molecules. And finally, the bioactive molecules (i.e. the structures showing bioactivity better than 10 μm on any target) were extracted from the ChEMBL database [18], providing a dataset with 750000 structures.

2.2. Extraction of substituents

Before the actual extraction of substituents the molecules were standardized by neutralizing charges and removing smaller parts and counterions. NP structures were also preprocessed by transforming the glycosylated structures to their aglycons. This has been done by a recursive removal of sugar units, described in detail in ref. [19]. Recently also an open source procedure for in silico deglycosylation has been made available [20]. It is well known that the presence of sugar units, influencing mostly molecule solubility and transport properties, is one of the most typical structural features of NPs [14], in our study, however, we were interested in identifying other, smaller substituents typical for NPs. After molecular preprocessing the substituents were extracted. A group with up to 12 non-hydrogen atoms connected by a single non-ring bond to a ring atom was considered to be a substituent. The extraction worked recursively, for example if a phenol substituent was extracted then also a hydroxy group was extracted from this phenol. The whole procedure is schematically illustrated on Figure 1.

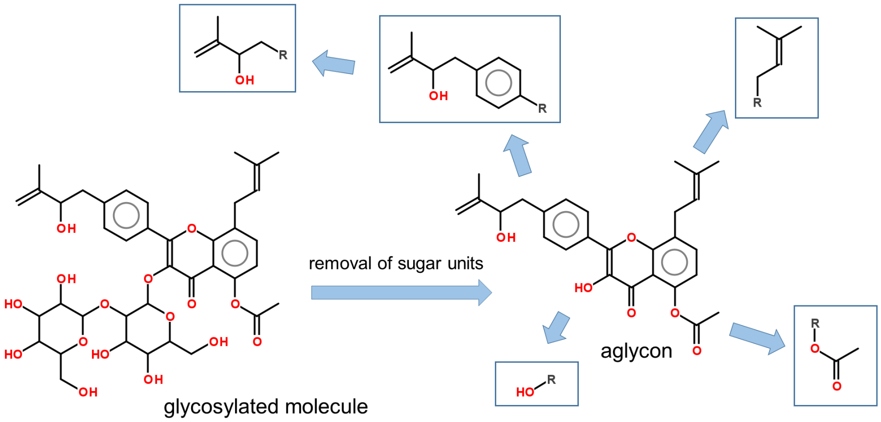


Fig. 1. Scheme illustrating the procedure for substituent extraction.

3. Results and Discussion

The substructure analysis of about 222000 natural products provided 16977 unique substituents. The substituents show a typical power law (or “long tail”) distribution with a few very common substituents and a large number of infrequent substituents. There are only 24 substituents present in more than 1% of the molecules and 157 in more than 0.1% of the molecules, and there are 8963 substituents (52.8%) that are singletons (present in only one molecule). The most frequent substituent present in NPs is the hydroxy group found in 59.3% of all molecules, closely followed by the methyl group (56.7%) and then by the methoxy group (22.1%), acetoxy group (8.7%), hydroxymethyl (8.5%), carboxyl group (5.7%) and isopropyl (3.7%). The 40 most frequent substituents present in NPs are shown in Figure 2. The 2003 substituents typical for NPs that are present in the database 10 or more times may be downloaded as SMILES from https://peter-ertl.com/molecular/data/npsubstituents.txt.

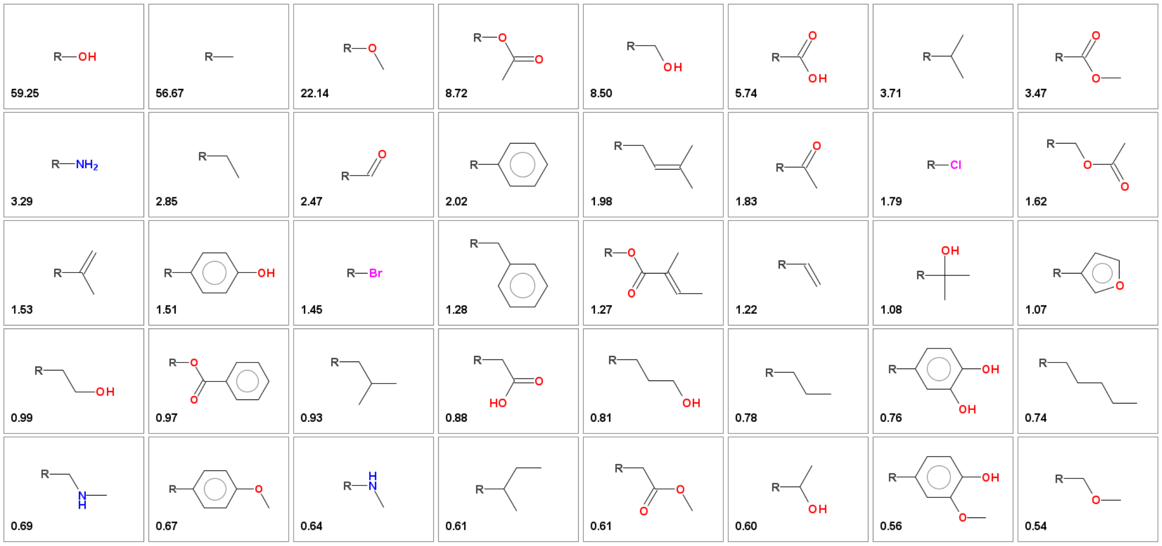


Fig. 2. The most common substituents present in natural products. The number indicates the percentage of molecules having this substituent.

Another question we wanted to answer was how the substituent distribution between natural products and common synthetic molecules differ. The results of this analysis are shown in Figure 3. Here, the distribution of the 50 most frequent substituents from both these sets (80 unique substituents totally) are visualized. The horizontal axis in the diagram represents the substituent frequency (the most common groups are on the left and the less common on the right), whereas the vertical axis indicates the propensity of substituents for natural products (green area at the top) or synthetic molecules (brown area at the bottom). The graph shows clear differences between substituents typical for NPs and those for synthetic molecules. The NP substituents contain mostly oxygen and very little other heteroatoms, are structurally more complex, often containing multiple bonds and are rich on stereocenters. Substituents found in synthetic molecules also contain nitrogen and sulfur and their combinations, halogenes and more aromatic and particularly heteroaromatic rings. More detailed information about the structural differences between the NP substituents and substituents from synthetic molecules is shown in Table 1, where the structural features from the 100 top substituents from each class are compared.

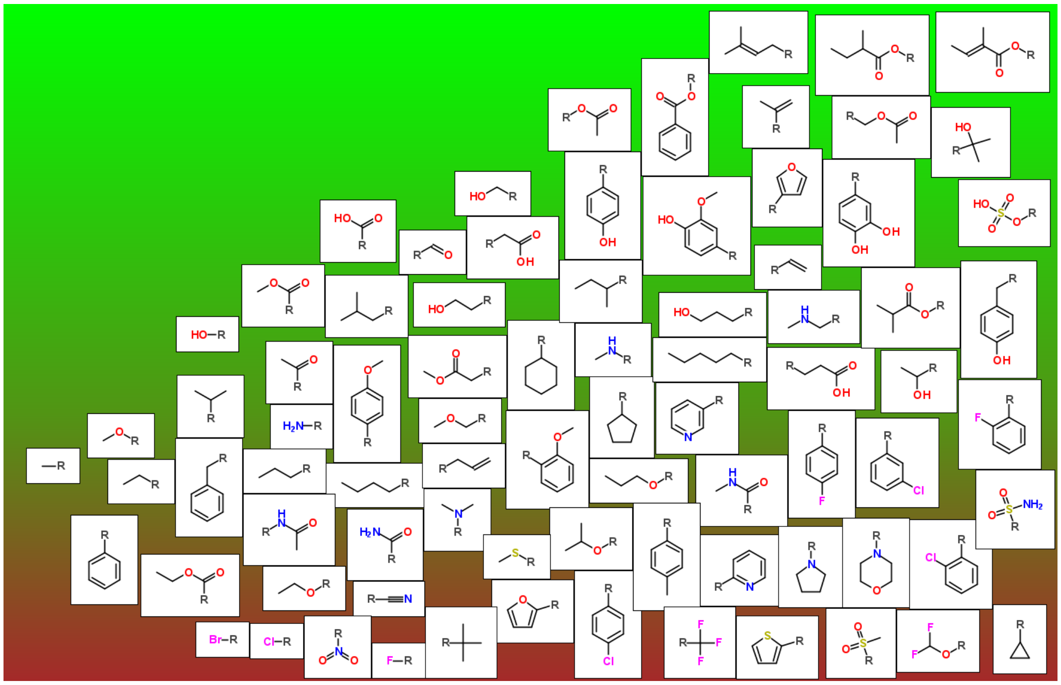


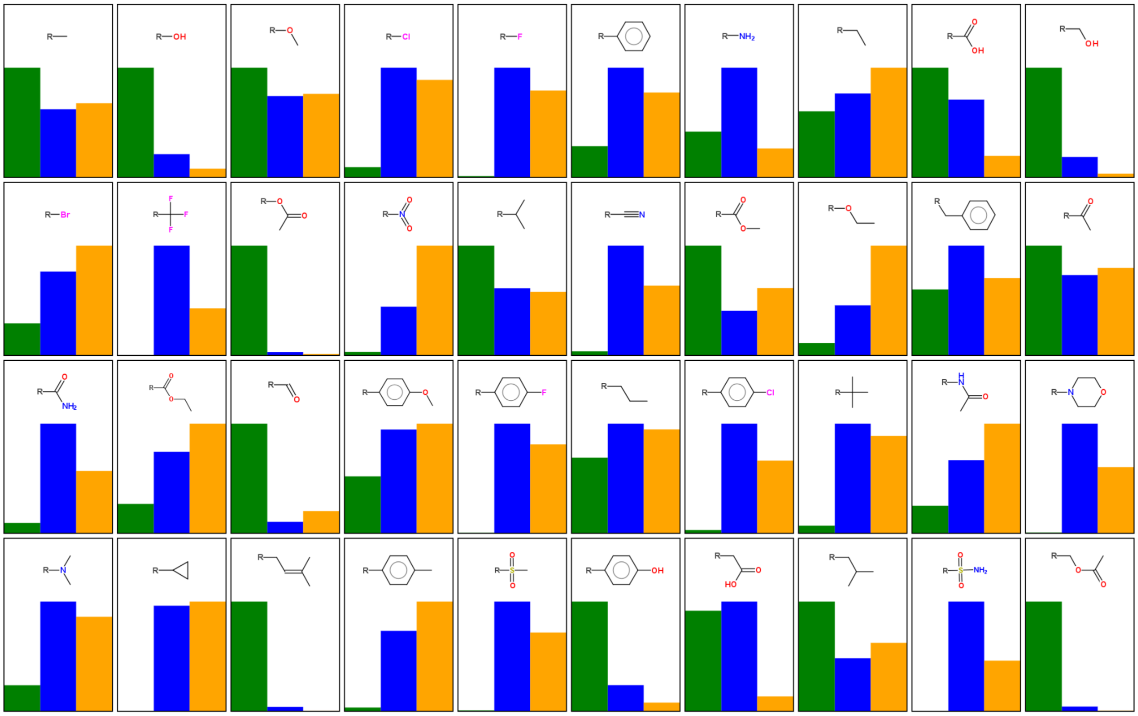
Fig. 3. Plot of common substituents displaying their preference for natural products (green area) and synthetic molecules (brown area). Position on the horizontal axis is proportional to the frequency of substituents - the most common substituents are on the left, less common on the right.

Table 1. Number of substituents from the 100 most common substituents present in natural products (NP) and synthetic molecules (SM), containing specified substructure features.

Substituent contains:	**NP**	**SM**
only carbon	26	18
only carbon and oxygen	56	26
oxygen	60	52
only carbon and nitrogen	9	13

nitrogen	13	37
sulfur	2	10
several heteroatoms	5	24
halogene	3	13
aromatic ring	23	34
heteroaromatic ring	3	10
stereo center	7	0
non-aromatic C=C bond	16	1

Another visualization comparing the distribution of substituents in NPs, bioactive molecules and synthetic molecules is shown in Figure 4, where the frequencies of the 50 most common substituents within these 3 data sets are shown. The heights of the bars in the graph are proportional to the relative frequencies of the particular substituents in these 3 sets. As one could expect, the distribution of substituents in different sets is quite different, with the NP substituents being a clear outlier, as already discussed in the previous analysis..



Fig, 4. Relative frequencies of common substituents present in natural products (green), bioactive molecules (blue) and common synthetic molecules (orange).

The effect of substituent patterns on the NP-like character of molecules is illustrated in Figure 5. In this figure 6 different sets of molecules are shown, with the molecules in each set having the same central core and differing only in the substitution patterns. The molecules in the top row of each set originate from the ZINC database and contain common synthetic substituents, while the molecule in the bottom row come from the COCONUT database and have the NP-like substituents. One can nicely see how the different substituent patterns influence the NP-like character of the whole molecule. This example also illustrates that the introduction of NP substituents to a standard synthetic scaffold can be used to create libraries with good NP-likeness.

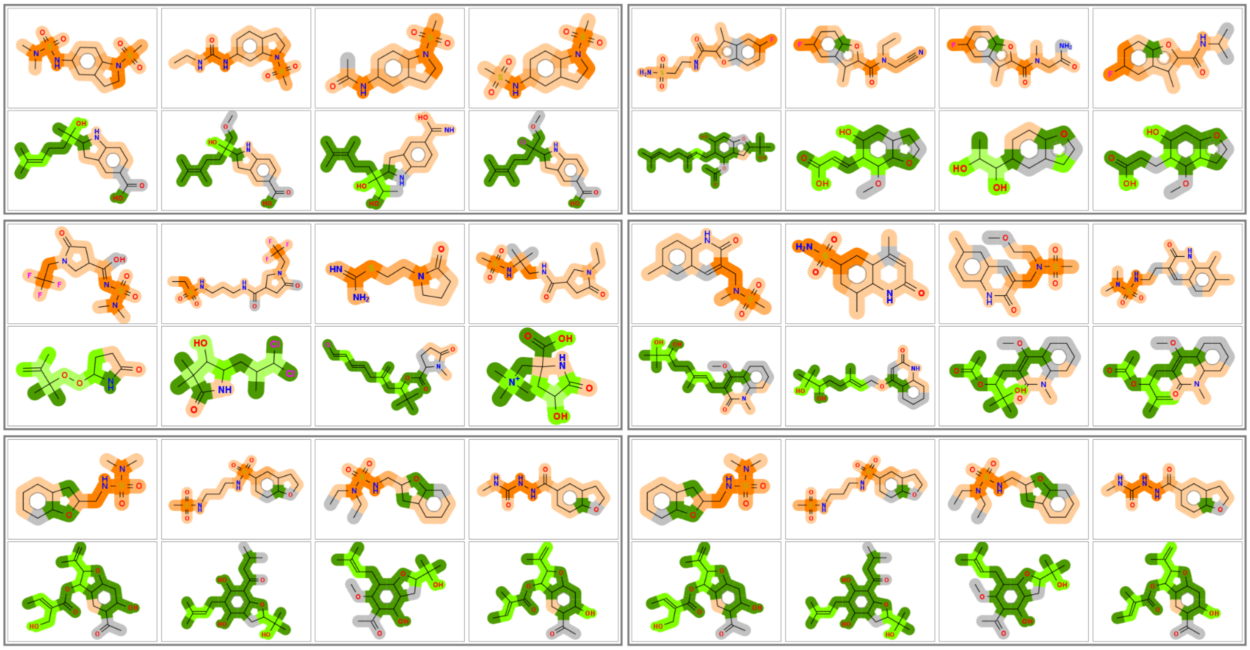


Fig.5. Example of how the substituents influence the overall NP character of the molecules. The image shows 6 sets of molecules, each set has the same central scaffold and the molecules differ only in their substituents; in the top row of each set molecules have synthetic substituents, in the bottom row the NP substituents. The color indicates the NP-like character: green - NP-like, gray - neutral, orange - synthetic.

4. Conclusions

A cheminformatics analysis of substituents extracted from a large collection of natural products was performed. Properties of nearly 17000 NP substituents were compared with the properties of substituents in bioactive molecules and common synthetic molecules and the features that are typical for the NP substituents were identified. This information may be used for the design of NP-like libraries, NP-inspired fragments or pseudo-natural products. The list of typical NP substituents may also be used to drive creation of novel NP-like molecules in generative chemistry applications. The most common NP substituents identified in this work are available for download at https://peter-ertl.com/molecular/data/npsubstituents.txt.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Chen, J. Kirchmair, Cheminformatics in Natural Product-based Drug Discovery, Mol. Inform. 39 (2020) 2000171. https://doi.org/10.1002/minf.202000171.

[2] S. Wetzel, A. Schuffenhauer, S. Roggo, P. Ertl, H. Waldmann, Cheminformatic Analysis of Natural Products and their Chemical Space, Chimia. 61 (2007) 355–360.

[3] M. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, Charting biologically relevant chemical space: a structural classification of natural products (SCONP)., Proc Natl Acad Sci U A. 102 (2005) 17272–17277. https://doi.org/10.1073/pnas.0503647102.

[4] P. Ertl, T. Schuhmann, Cheminformatics Analysis of Natural Product Scaffolds: Comparison of Scaffolds Produced by Animals, Plants, Fungi and Bacteria, Mol. Inform. 39 (2020) 2000017.

https://doi.org/10.1002/minf.202000017.

[5] P. Ertl, T. Schuhmann, A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products, J. Nat. Prod. 82 (2019) 1258–1263. https://doi.org/10.1021/acs.jnatprod.8b01022.

[6] G. Karageorgis, D.J. Foley, L. Laraia, S. Brakmann, H. Waldmann, Pseudo Natural Products—Chemical Evolution of Natural Product Structure, Angew. Chem. Int. Ed. 60 (2021) 15705–15723. https://doi.org/10.1002/anie.202016575.

[7] G.W. Bemis, M.A. Murcko, Properties of Known Drugs. 2. Side Chains, J Med Chem. 42 (1999) 5095–5099. https://doi.org/10.1021/jm9903996 S0022-2623(99)00399-4.

[8] P. Ertl, Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups, J. Chem. Inf. Comput. Sci. 43 (2003) 374–380. https://doi.org/10.1021/ci0255782.

[9] P.J. Hajduk, D.R. Sauer, Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency, J. Med. Chem. 51 (2008) 553–564. https://doi.org/10.1021/jm070838y.

[10] A.M. Wassermann, J. Bajorath, Chemical Substitutions That Introduce Activity Cliffs Across Different Compound Classes and Biological Targets, J Chem Inf Model. 50 (2010) 1248–1256. https://doi.org/10.1021/ci1001845.

[11] P. Gleeson, G. Bravi, S. Modi, D. Lowe, ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters, Bioorg. Med. Chem. 17 (2009) 5906–5919. https://doi.org/10.1016/j.bmc.2009.07.002.

[12] M. Awale, S. Riniker, C. Kramer, Matched Molecular Series Analysis for ADME Property Prediction, J. Chem. Inf. Model. 60 (2020) 2903–2914. https://doi.org/10.1021/acs.jcim.0c00269.

[13] K. Takeuchi, R. Kunimoto, J. Bajorath, Systematic mapping of R-group space enables the generation of an R-group replacement system for medicinal chemistry, Eur. J. Med. Chem. 225 (2021) 113771. https://doi.org/10.1016/j.ejmech.2021.113771.

[14] J. Schaub, A. Zielesny, C. Steinbeck, M. Sorokina, Description and Analysis of Glycosidic Residues in the Largest Open Natural Products Database, Biomolecules. 11 (2021) 486. https://doi.org/10.3390/biom11040486.

[15] M. Sorokina, P. Merseburger, K. Rajan, M.A. Yirik, C. Steinbeck, COCONUT online: Collection of Open Natural Products database, J. Cheminformatics. 13 (2021) 2. https://doi.org/10.1186/s13321-020-00478-9.

[16] P. Ertl, S. Roggo, and A. Schuffenhauer, Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries, J Chem Inf Model. 48 (2008) 68–74. https://doi.org/10.1021/ci700286x.

[17] J.J. Irwin, K.G. Tang, J. Young, C. Dandarchuluun, B.R. Wong, M. Khurelbaatar, Y.S. Moroz, J. Mayfield, R.A. Sayle, ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery, J. Chem. Inf. Model. 60 (2020) 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.

[18] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C.J. Radoux, A. Segura-Cabrera, A. Hersey, A.R. Leach, ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Res. 47 (2019) D930–D940. https://doi.org/10.1093/nar/gky1075.

[19] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M.A. Koch, H. Waldmann, The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification, J Chem Inf Model. 47 (2007) 47–58. https://doi.org/10.1021/ci600338x.

[20] J. Schaub, A. Zielesny, C. Steinbeck, M. Sorokina, Too sweet: cheminformatics for deglycosylation in natural products, J. Cheminformatics. 12 (2020) 67. https://doi.org/10.1186/s13321-020-00467-y.