Application Note

www.molinf.com

molecular informatics

DOI: 10.1002/minf.202000216

SMARTS.plus – A Toolbox for Chemical Pattern Design

Christiane Ehrt[+],[a] Bennet Krause[+],[a] Robert Schmidt[+],[a] Emanuel S. R. Ehmki,[a] and Matthias Rarey*[a]

Abstract: The number of publications concerning Pan-Assay Interference Compounds and related problematic structural motifs in screening libraries is constantly growing. In consequence, filter collections are merged, extended but also critically discussed. Due to the complexity of the chemical pattern language SMARTS, an easy-to-use toolbox enabling every chemist to understand, design and modify chemical patterns is urgently needed. Over the past decade, we developed a series of software tools for visualizing, editing, creating, and analysing chemical patterns. Herein, we highlight how most of these tools can now be easily used as part of the novel SMARTS.plus web server (https://smarts.plus/). As a showcase, we demonstrate how researchers can apply the web server tools within minutes to derive novel SMARTS patterns for the filtering of frequent hitters from their screening libraries with only a little experience with the SMARTS language.

Keywords: Chemical Patterns · SMARTS Visualization · SMARTS Comparison · Medicinal Chemistry · Filter Collections

Chemical patterns are one of the workhorses of cheminformatics. By describing a generic structural feature of molecules, they are of central importance for classifying and organizing compound collections. In contrast to a classical substructure, a chemical pattern allows logical expressions and atom/bond specifications via properties. Invented in the late 80s by Daylight Information Systems,[1] today, the SMARTS language is the quasi-standard for the description of chemical patterns. Although not complete, SMARTS is very feature-rich allowing chemists to precisely specify a structural pattern they have in mind. Unfortunately, the SMARTS language is quite complex and many researchers struggle in formulating their patterns due to the cryptic nature of SMARTS notations. Furthermore, even for experienced computational chemists, it is sometimes hard to spot errors in SMARTS expressions making their development usually a trial-and-error process.

Over the past decade, we developed a series of software tools supporting researchers in designing and analysing chemical patterns using the SMARTS language. Recently, we developed a web server named SMARTS.plus[2] to circumvent the software installation hurdle making SMARTS analytics available to even occasional users and students. In the following, we will first briefly summarize the functionality, how it can be accessed in SMARTS.plus und will round off with a use case, namely the application of the web server tools to derive novel patterns for the filtering of pan-assay interference compounds.

Although systematic names exist in chemistry, the daily language of chemistry is structure diagrams. Chemical patterns have a lot in common with structure diagrams; roughly spoken, they are just a more generic form. The most important aspect to make chemical patterns comprehensible is therefore an adequate visualization. Following the IUPAC nomenclature for structure diagrams as closely as possible, we carefully designed the graphical depiction of molecules to patterns. Figure 1 shows an example of the resulting SMARTSview image for a complex pattern.[3] Substructural features and structural variances get ascertainable, immediately showing the great value of this approach. Once having had a first look at a SMARTSview image, it becomes evident that a graphical editor is indispensable. Therefore, we developed a powerful graphical editor, SMARTSeditor, as a standalone tool[4] which is available for academic use.[5]

Chemical patterns are mostly generated based on example molecules. There is a class A of molecules having a certain property that a class B of molecules does not have. Often, patterns are designed by continuously monitoring which molecules of class A do not yet match and which of class B do still match. The question arises how this process can be best supported algorithmically. In computer science, several algorithms exist for so-called frequent and contrast pattern mining on graphs.[6] These methods are also applied to molecules (see for example[7]); however, they usually do not end up in SMARTS expressions. Therefore, we developed SMARTSminer[8] as a one-stop solution from sets of molecules to a SMARTS pattern. Although SMARTSminer is not able to make use of all SMARTS features (for example,

[a] *C. Ehrt,[+] B. Krause,[+] R. Schmidt,[+] E. S. R. Ehmki, M. Rarey*
Universität Hamburg, ZBH – Center for Bioinformatics,
Bundesstraße 43, 20146 Hamburg, Germany
phone: +49 40 42838–7351
fax: +49 40 42838–7352
E-mail: rarey@zbh.uni-hamburg.de

[+] *These authors contributed equally to this work.*

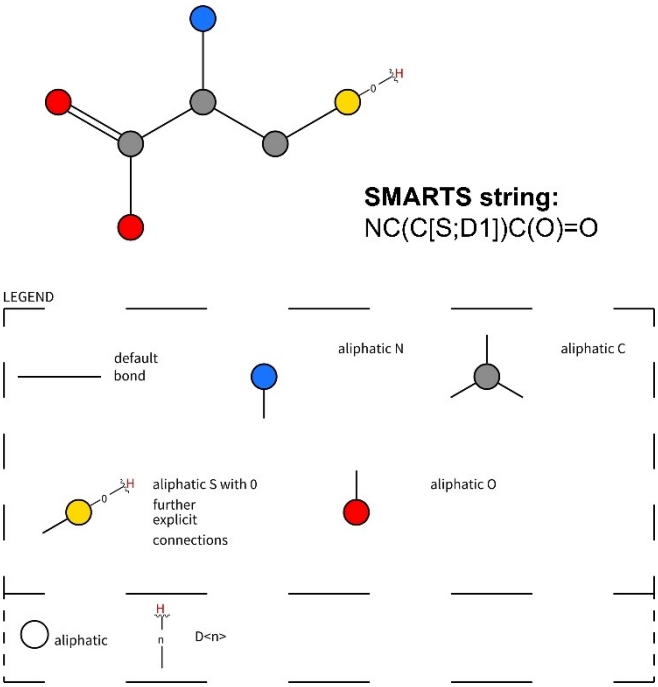

Figure 1. SMARTSview visualization of a typical SMARTS pattern for the exclusion of problematic compounds from molecular datasets. It is extracted from the publication of Pearce and co-workers.[9] The SMARTS pattern describes molecules with a thiol warhead that might covalently modify cysteine residues in a protein in an unselective manner.

recursion is not supported), it simplifies the design of patterns enormously. Within seconds to minutes, it processes sets of hundreds of molecules and creates a SMARTS pattern to enrich or separate them. SMARTSminer is integrated into our standalone editor; to give users a kick-start for pattern design, it is also available on the SMARTS.plus web server.

Comparing two molecules is an almost daily task in cheminformatics. More precisely, we either ask for a substructure relationship (is substructure B contained in molecule A) or a similarity relationship mostly answered with topological fingerprints like the Extended-Connectivity Fingerprints (ECFP).[8] These questions apply to chemical patterns as well and are critical for the analysis of pattern collections. The comparison of SMARTS expressions belongs to the most challenging algorithmic problems in cheminformatics. In 2019, we presented SMARTScompare,[10] an algorithmic approach to address the substructure search and the similarity search on chemical patterns. An atom type fingerprint covering more than 20,000 states is employed in a complex, recursive comparison algorithm.[9] For the first time, a method enables the identification of more specific or generic or just similar chemical patterns in pattern collections - independent of the way they are formulated. SMARTScompare is a command line tool that processes hundreds of pattern comparisons in a few seconds. Recently, we added SMARTScompare to the SMARTS.plus web server for pattern comparison including visualization, as well as for searching public pattern collections.

SMARTS.plus combines SMARTSview, SMARTS-miner and SMARTScompare in an easy-to-use web server based on Rails available at https://smarts.plus. It connects the standalone tools allowing to visually analyse SMARTS expressions. It runs in four modes: In the 'View' mode, the user can enter a SMARTS expression and get a visual representation including a figure legend for less experienced users. In the 'Compare' mode, two expressions can be uploaded and the server calculates substructure relationships and pattern similarity. A graphical depiction of the node mapping helps to comprehend the pattern relationship. In the 'Search' mode, a SMARTS expression is compared to currently nine public pattern collections enabling to browse through the most similar ones. To this end, we used the SMARTS collections as applied for the annotation by the ChEMBL database.[11] These collections include the well-known PAINS filters (PAINS),[12] the SureChEMBL Non-MedChem Friendly SMARTS (SureChEMBL),[13] the Bristol-Myers Squibb HTS Deck filters (BMS),[9] the NIH MLSMR Excluded Functionality filters (MLSMR),[14] the University of Dundee NTD Screening Library Filters (Dundee),[15] filters of unwanted fragments derived by Inpharmatica Ltd. (Inpharmatica),[5] the Pfizer lint filters (lint),[16] and the Glaxo Wellcome Hard filters (Glaxo).[17] Finally, in the 'Create' mode, two compound sets can be uploaded and SMARTS-miner is applied to suggest patterns frequently found in the first set and rarely found in the second. A browser shows the molecules hit for each of the created patterns.

In the following, we describe a workflow (Figure 2) for applying the SMARTS.plus tools to derive novel SMARTS filters for the characterization of frequent hitters or pan-assay interference compounds, i.e., compounds that are frequently found to be active in multiple high throughput screening assays, e.g. due to aggregation or high reactivity. As an example case study, we selected a set of molecules which might unselectively modify cysteine residues in proteins according to the studies of Dahlin et al.[18] The common structural feature of these compounds is the benzodiathiazole scaffold (using the link https://smarts.plus/ you can visualize the corresponding SMARTS pattern: '[*,#1]-[#6]-1=[#6]-[#6](-[*,#1])=[#6](-[*,#1])-[#6]-2=[#7]-[#16]-[#7]=[#6]-1-2'). Whereas six of the analysed compounds were shown to react covalently by monitoring the presence of compound thiol adducts after addition of CoA, 12 further compounds sharing this scaffold did not lead to covalent adducts. SMARTSminer was used to derive a chemical SMARTS pattern that enables differentiation between the positive support structures (thiol-modifying compounds) and the negative support structures (no reaction with free thiols was observed). This can be achieved in the 'Create' mode by uploading both sets to the web server and defining the minimum percentage of

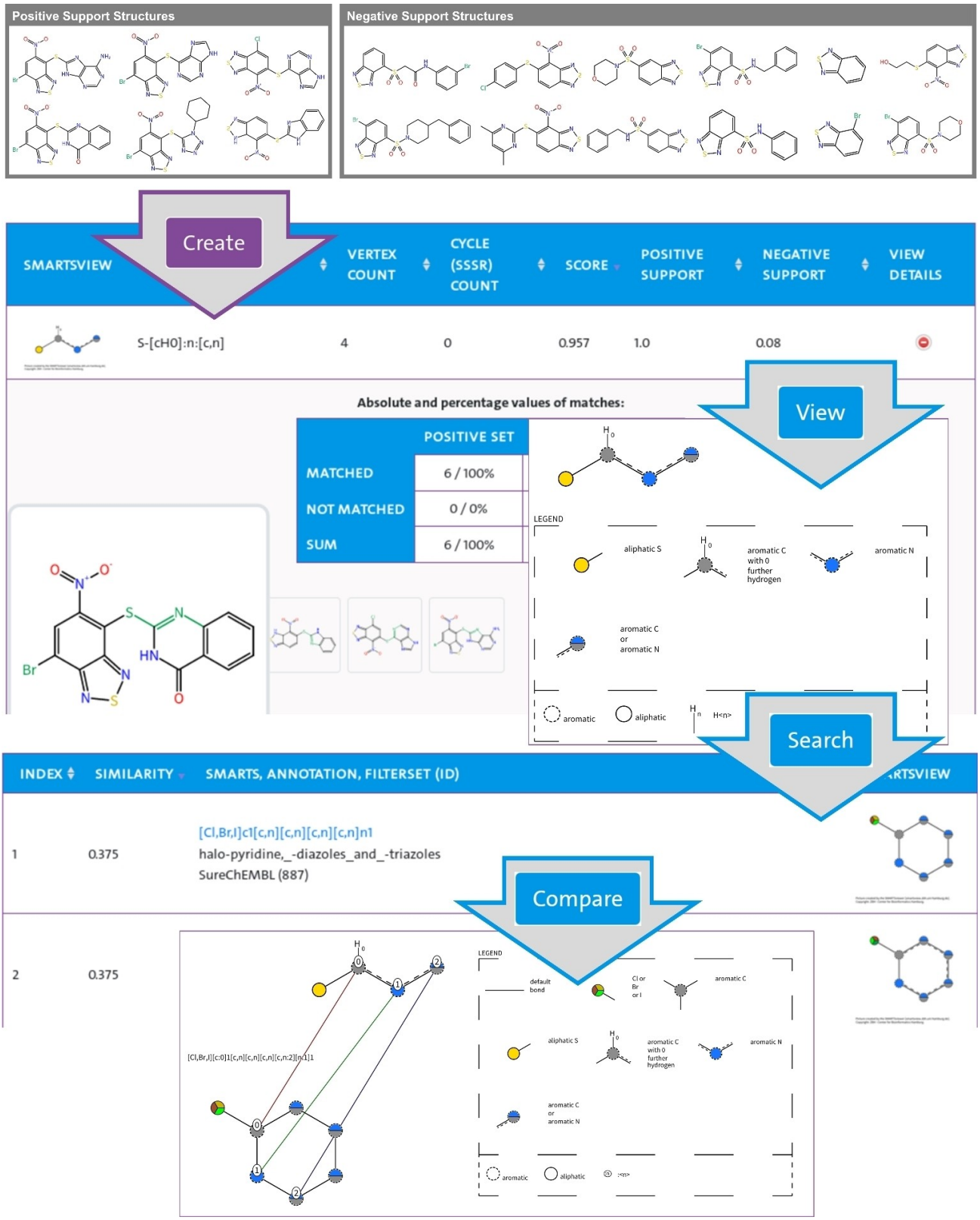

Figure 2. An example workflow utilizing the SMARTS.plus tools in the 'Create', 'View', 'Search', and 'Compare' mode. Here, we show how two molecule sets can be compared creating a SMARTS pattern that enables a good differentiation between both sets. The derived pattern can be immediately visualized in the 'View' mode. Subsequently, multiple filter collections can be searched for similarities to the created pattern. The differences between the pattern of interest and highly similar patterns can finally be visualized in the 'Compare' mode.

compounds from the positive support structures (90% in this example) and the maximum percentage of compounds from the negative support structures (10% in this example) that should be matched by the generated SMARTS pattern. The resulting patterns can be visualized with SMARTSview by one click ('View' mode). However, our search resulted in numerous patterns with identical positive and negative support leading to identical confusion matrices.

To further filter the results, the option 'Small patterns' can be selected on the bottom of the results page. In case of multiple patterns with the same support values, small patterns are preferred, large ones eliminated. Here, this leads to a reduction of the results to three SMARTS patterns. The chosen pattern can subsequently be compared to already available SMARTS filter collections to find potential patterns already covering the compounds of interest.

This can be achieved in the 'Search' mode where we can insert the SMARTS pattern of interest and compare it to all patterns within the previously mentioned SMARTS filter collections: PAINS, SureChEMBL, BMS, MLSMR, Dundee, Inpharmatica, lint, and Glaxo. In our example, the maximum similarity is 0.375; so we might conclude that we found a novel SMARTS pattern for the characterization of typical frequent hitters. However, it is often difficult to compare the most similar patterns by eye. Therefore, we can finally use the 'Compare' mode to get a better understanding of the basic differences between the newly created and the most similar pattern. We can click on the most similar SMARTS string and a graphical representation of the differences is issued. Now, we can immediately see that the previously defined pattern of the MLSMR collection is missing the sulfur atom which is crucial in our newly designed SMARTS pattern. In consequence, we were able to derive a novel structural pattern for the filtering of unselectively reacting compounds in screening datasets based on experimental data.

The described workflow can be pursued on our SMARTS.plus web server within minutes. It enables interested researchers to derive appropriate conclusions based on their experimental results in a highly intuitive way. We hope that current efforts to derive new SMARTS patterns for the description of frequent hitters will benefit from our freely available web server.

Acknowledgement

Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

None declared.

References

[1] Daylight Chemical Information Systems, Inc., https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Sep 30, 2020).
[2] Center for Bioinformatics Hamburg. SMARTS.plus, https://smarts.plus/ (accessed Sep 30, 2020).
[3] K. T. Schomburg, H. C. Ehrlich, K. Stierand, M. Rarey, *J. Chem. Inf. Model.* **2010**, *50*, 1529–1535.
[4] K. T. Schomburg, L. Wetzer, M. Rarey, *Drug Discovery Today* **2013**, *18*, 651–658.
[5] Center for Bioinformatics Hamburg, NAOMI ChemBio Suite, uhh.de/naomi (accessed Sep 30, 2020).
[6] C. Jiang, F. Coenen, M. Zito, *The Knowledge Engineering Review* **2012**, *28*, 75–105.
[7] C. Borgelt, M. R. Berthold, in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, **2002**, pp. 51–58.
[8] S. Bietz, K. T. Schomburg, M. Hilbig, M. Rarey, *J. Chem. Inf. Model.* **2015**, *55*, 1535–1546.
[9] B. C. Pearce, M. J. Sofia, A. C. Good, D. M. Drexler, D. A. Stock, *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.
[10] R. Schmidt, E. S. R. Ehmki, F. Ohm, H. C. Ehrlich, A. Mashychev, M. Rarey, *J. Chem. Inf. Model.* **2019**, *59*, 2560–2571.
[11] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magarinos, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res.* **2017**, *45*, D945–D954.
[12] J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719–2740.
[13] SureChEMBL, https://www.surechembl.org/knowledgebase/169485-non-medchem-friendly-smarts (accessed Sep 30, 2020).
[14] MLSMR, https://www.yumpu.com/en/document/view/12367541/mlsmr-excluded-functionality-filters-nih-molecular-libraries- (accessed Sep 30, 2020).
[15] R. Brenk, A. Schipani, D. James, A. Krasowski, I. H. Gilbert, J. Frearson, P. G. Wyatt, *ChemMedChem* **2008**, *3*, 435–444.
[16] J. F. Blake, *Med. Chem.* **2005**, *1*, 649–655.
[17] M. Hann, B. Hudson, X. Lewell, R. Lifely, L. Miller, N. Ramsden, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.
[18] J. L. Dahlin, J. W. Nissink, J. M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M. A. Walters, *J. Med. Chem.* **2015**, *58*, 2091–2113.

Received: August 18, 2020
Accepted: September 28, 2020
Published online on October 8, 2020

