JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 2572−2586

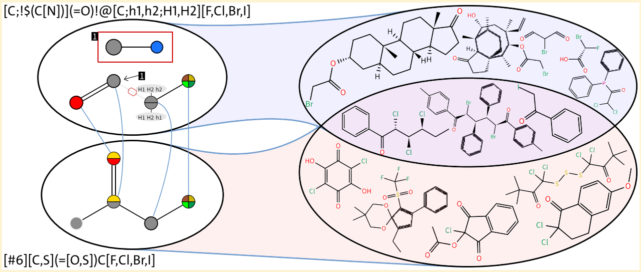
pubs.acs.org/jcim

Comparing Molecular Patterns Using the Example of SMARTS: Applications and Filter Collection Analysis

Emanuel S. R. Ehmki, Robert Schmidt, Farina Ohm,[†] and Matthias Rarey*

ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany

Supporting Information


[C;!$(C[N])](=O)!@[C;h1,h2;H1,H2][F,Cl,Br,I]
1
1
H1 H2 h2
H1 H2 h1
[#6][C,S](=[O,S])C[F,Cl,Br,I]

ABSTRACT: In a recent work, an algorithm to compare chemical patterns, written for example in SMARTS, was presented. This algorithm, called SMARTScompare, is able to assess the identity, subset relation, and similarity of a pair of patterns. Here we used an implementation of SMARTScompare to analyze SMARTS filter sets that were published in the context of, for example, high-throughput screening. We found that the difference in intentions with which the filter sets were designed is mirrored in the similarity values we calculated. The analysis revealed which patterns from one filter set are covered by filters from another set. In one case it became obvious that a filter set is more or less completely covered by another. Furthermore, we analyzed pattern hierarchies for consistency, and we propose a method to remove redundant patterns. SMARTScompare together with SMARTScompareView equips users with powerful methods to visualize, compare, and focus their filter sets.

INTRODUCTION

The size of compound databases has been increasing steadily.[1,2] Equally, the need for sophisticated search and filter methods has been on the rise, too. Most queries are designed to retrieve or exclude classes of molecules or specific types of chemistry. Therefore, many queries are designed as molecular patterns that describe a specific substructure or a group of substructures. In practice, these molecular patterns are often encoded in, for example, SMILES[3] Arbitrary Target Specification (SMARTS)[4] or SYBYL Line Notation (SLN).[5]

Sets of molecular patterns can be used to model unwanted behavior in many fields of application, e.g., high-throughput screening (HTS)[6] or toxicity.[7] Frequently they are used as filters to exclude molecules with unwanted groups before expensive experimental screening starts. Compounds that are matched by any pattern have been shown to express unwanted behavior and can be flagged for further observation or removed right away. Several structural filter sets, mostly in the context of experimental screening, have been published.[2,6−13] With the arrival of large collections of molecular patterns, a whole new set of problems emerged as well. Molecular patterns can be highly complex, and design implications cannot always be foreseen from the start. Therefore, good tools aiding scientists in visualization and browsing,[14,15] designing,[16,17] and comparing molecular patterns are needed.

Considering the importance of SMARTS for the field of cheminformatics, relatively few approaches to compare two SMARTS pattern exist to date. The most straightforward strategy— visually inspecting two patterns—is extremely laborious as pattern sets are growing. A more automatable approach is the comparison of sets of molecules that are matched by two patterns. This makes the analysis of the set relation of two patterns possible. Nonetheless, this approach can only deny subset relation. It cannot prospectively classify one pattern as a sub- or superset of the other. Besides the large computational burden, this makes the approach extremely dependent on the selection of molecules that are used to analyze the patterns. The third and also most algorithmic approach is implemented in part in the Rational Design Kit (RDKit).[18] It enables a node-based comparison of two molecular patterns in the form of SMARTS strings within the boundaries of some restrictions. Roughly sketched, the pattern comparison is possible in cases where the semantics of the SMARTS language are compatible. In cases where the same chemical state (e.g., an sp^3-hybridized carbon) is encoded differently within the SMARTS pattern, the pattern nodes can no longer be mapped correctly. Our newly designed algorithm, called SMARTScompare,[19] is independent of the input form because it maps SMARTS semantics into the general pattern space. This enables us to determine the chemical state of a pattern node, making the comparison independent of the SMARTS string representation. Employing a maximum common subgraph (MCS) algorithm that includes the handling of pattern recursion eventually enables the comparison of SMARTS patterns for nearly all practically relevant scenarios.

For the analysis of structural filters, the possibility to compare chemical patterns with an algorithmic approach is highly desirable. Common structures that are part of several

Received: March 22, 2019
Published: May 24, 2019

DOI: 10.1021/acs.jcim.9b00249

structural filter sets are more likely to be relevant than patterns occurring in exactly one. With hundreds of molecular patterns in a single filter collection, it is challenging to detect patterns related to the same or similar chemical features. When the pattern notations for the same kind of chemistry are different, the analysis of a complete set of patterns becomes an extremely laborious task. Comparing patterns designed by several different scientists for the same purpose can give many insights, sharpen the pattern collection, and help to prevent errors or gaps in the description of substructures of interest. Furthermore, analyses like the one conducted by Capuzzi et al.[20] on PAINS[8] do not have to rely on the molecules that are matched by molecular patterns. Instead, they can be performed directly on the level of molecular patterns, allowing more precise and reliable results.

To be able to conduct comprehensive analyses of molecular patterns, we developed a novel analytic approach.[19] Schomburg et al.[14] presented an intuitive visualization concept for SMARTS patterns. We extended their approach to aid the user in better understanding chemical patterns and structural filter sets. The method is customized to SMARTS and handles nearly all frequently used language elements. With runtimes in the millisecond range, the algorithm can be used interactively even on larger filter collections with a few hundred to thousands of patterns.

THEORY

For a detailed discussion and the proper mathematical formalism, we refer the reader to the companion paper.[19] The following section discusses only briefly the important theoretical aspects that were presented there. Additionally, implications for the implementation and application of the algorithm are discussed. For consistency, the same notation is used.

Theoretical Aspects of SMARTS Pattern Comparison. The general assumption is that a chemical pattern describes a potentially infinite subset of molecules in the chemical space (CS). Therefore, we can approximate the subset relation of two patterns P_1 and P_2 on the basis of how well their subsets of the chemical space, $CS(P_1)$ and $CS(P_2)$, overlap. To get an exhaustive description of the chemical space, we introduce two new description systems. First, the atomtype space represents possible states that atoms can take on in chemical molecules. Second, equivalent to the atomtype space, the bondtype space describes possible states of bonds in molecules. In a graph representation of a molecular pattern, nodes are described via a set of atomtypes and edges via a set of bondtypes. Given that approach, we conclude that if patterns are identical, their sets of matched molecules are identical, too. With additional restrictions applied, the same is valid for the opposite direction. What has been detailed for subset relations can be easily extended to allow similarity analysis.

Since we model chemical information as a set of atomtypes, the similarity of two patterns can be estimated by analyzing the overlap of atomtypes of computed node mappings. Prior to the comparison step, SMARTS string expressions are translated into their graph representations. Each node representation in the SMARTS string is converted to one graph node.

To be able to compare two SMARTS graph structures, each node or edge is described by a fingerprint of atomtypes or bondtypes, respectively. The node and edge fingerprints are of constant size. Each bit represents exactly one atomtype, and the overall size of the fingerprint is determined by the number of atomtypes defined in the chemistry model that is used.

When two SMARTS graphs are compared, first an induced maximum common connected subgraph is computed. The compatibility of nodes depends on the comparison mode, e.g., similarity or set relation. The similarity score of two SMARTS expressions is then calculated on the basis of the fingerprint similarity of compatible nodes. The following sections discuss relevant aspects that have to be considered when designing an implementation of the presented algorithm.[19] The focus lies on handling of aromaticity, wildcards, and ring properties and fine-tuning of similarity calculations.

Implementation-Specific SMARTS Handling. *Wildcards.* Wildcards in SMARTS like "*", "A", and "a" match several elements. In NAOMI,[21] the SMARTS matching supports several options to handle hydrogens during SMARTS matching:

- Hydrogens are not matched at all.
- Wildcards do not match hydrogen atoms.
- The wildcard * matches hydrogen as well.

The second option is enabled by default. Thus, the patterns [#1] and * are considered dissimilar by default.

Ring Size Property. The SMARTS ring count property "R" as defined by Daylight Chemical Information Systems, Inc.[4] describes the number of smallest set of smallest rings (SSSR)[22] rings of which an atom is a part. For the fingerprint approach, however, it is important that the "R<n>" property is unique. Unique ring families (URFs) describe exactly the number of smallest rings of which an atom is a part. Therefore, the NAOMI library describes R′ as the number of URFs[23,24] of which an atom is a part.

In order to describe all possible states, the number of rings and the smallest ring size have to be enumerated up to a constant value. We therefore limit the maximum number of rings to 4 and the maximum smallest ring size to 25. If the values are larger, they can be handled but can no longer be distinguished.

It should be noted that the ring properties increase the size of the fingerprint substantially. Therefore, special care has to be taken if fingerprints are converted to similarity values.

Similarity Calculation. The most simple way to express the similarity of two patterns is to count their matching nodes as determined by the MCS approach. Although this measure might be enough for a rough estimate, it does not do justice to the complexity of the pattern matching problem. A more meaningful value can be achieved if the similarity between two nodes in SMARTS patterns is calculated via the atomtype fingerprints associated with them. Suppose that P_1 and P_2 are the pattern graph structures and M is the matching. Furthermore, suppose that u and v are nodes of graphs, where $u \in P_1$ and $v \in P_2$. The similarity of P_1 and P_2 can then approximately be calculated as shown in eq 1:

$$S(P_1, P_2, M) = \frac{\sum_{(u,v)\in M} S(u, v)}{\text{normalization term}} \qquad (1)$$

Possible normalization modes are listed in Table 1a.

Calculating similarity on fingerprints is a well-established strategy in cheminformatics. In SMARTScompare one can select from several comparison modes in order to achieve an outcome that is most fitting for the problem posed (see Table 1b). Aside from the MCS similarity score, all options work with the fingerprint that is reduced or weighted upon

Table 1. Possible Options for the Normalization and Scoring Level Modes Implemented in SMARTScompare: Each Section Describes a Type of Option That Can Be Selected When the Comparison Mode *Similarity* Is Selected

(a) Normalization Modes
sum of nodes—number of matched nodes
max. number of nodes in both patterns
mean number of nodes in both patterns
min. number of nodes in both patterns
number of nodes in the first pattern
number of nodes in the second pattern
(b) Scoring Level Modes
MCS
elements fingerprint
reduced fingerprint
extended valence state fingerprint
extended valence state weighted fingerprint

comparison. As the name implies, the *elements* fingerprint contains only chemical elements. The *reduced* fingerprint comprises extended valence states as described in the companion paper[19] but has a reduced set of ring membership states that is registered for each node. The *extended valence state* fingerprint is the standard version and contains all possible atomtypes that are part of the NAOMI framework. The *extended valence state weighted* fingerprint is the same as the extended valence state fingerprint but undergoes some weighting before the similarity score is calculated.

The weighting is based on valence state statistics extracted from a set of molecules that represents the type of chemical space in which one seeks to compare patterns. We used a set of ~370 million molecules comprising all ZINC15[25] 2D compounds that were available on June 8, 2017. In the following we refer to this set as the *validation data set*. The weighting of each atomtype in the fingerprint is defined as the relative frequency of that atomtype in the set of molecules used to calculate the valence state statistic:

$$S_{WT}(X, Y) = \frac{\sum_i w(i)(X_i \wedge Y_i)}{\sum_i w(i)(X_i \vee Y_i)} \tag{2}$$

The standard similarity procedure in SMARTScompare is based on the weighted fingerprint Tanimoto similarity (S_{WT} in eq 2) with a Tanimoto normalization by the nodes of the involved patterns (eq 3):

$$S(P_1, P_2, M) = \frac{\sum_{(X,Y)\in M} S_{WT}(X, Y)}{|P_1| + |P_2| - |M|} \tag{3}$$

The weighting factor $w(i)$ represents the frequency of extended atomtypes as found in the validation data set. This procedure for node similarity is depicted in Figure 1.

This similarity measure can be adjusted by every user of SMARTScompare to a specific kind of chemistry. From every set of molecules, the atomtypes can be extracted with a utility tool and valence state statistics can be generated and employed in SMARTScompare.

At present, fingerprints of pattern edges are not explicitly considered during similarity calculations. Since bondtypes are usually reflected in the fingerprints of the incident nodes, edge properties are implicitly incorporated.

Aromaticity Detection. Some implications of the SMARTScompare aromaticity handling are detailed in the companion paper.[19] Here we discuss why we deemed it necessary to include ancillary aromaticity handling. In SMILES, aromatic systems are detected independently of the specific notation of the input string. With SMARTS, however, bonds are always handled explicitly as aromatic single or double bonds. Aromaticity usually cannot be deduced from SMARTS patterns. Although confusing from the application point of view, there are good reasons for this behavior. While SMILES strings always describe full molecules, SMARTS strings represent substructures, which might not be sufficient for aromaticity detection (also see section 4.7 in the Daylight Theory Manual.[4]). For the SMARTScompare approach, we implemented two *optional* features supporting the handling of aromaticity within SMARTS patterns: *SMILES-like aromaticity detection* and *Detect aromatic bonds*. In SMARTS, C1=CC=CC=C1 is not equivalent to c1ccccc1 or c1:c:c:c:c:c1. However, when parsed as molecules, these three representations are identical. Chemistry toolkits like NAOMI[26] parse a molecule and then determine aromatic ring systems on the basis of the assigned valence states and the Hückel rule. In order to support the chemical intuition of aromaticity (e.g., the π system in a benzene is always aromatic no matter the notation), the SMARTS preprocessing in SMARTScompare includes an aromaticity detection step. This behavior can be enabled with the *SMILES-like aromaticity detection* option. Whenever a ring system is encountered during parsing of a SMARTS pattern, all Kekulé localizations of a SMILES string that are possible with the given ring size and elements are determined. Information on neighboring nodes of the ring system is also included. If any localization is classified as aromatic, the fingerprints of the SMARTS graph are updated to incorporate aromaticity. To all edges a disjunct aromatic bondtype is added, and for all nodes all of the aliphatic node queries are replaced with generic element specifications (e.g., the bond "=" becomes "=,:" and "[C]" becomes "[#6]", whereas "c" is not modified). Additionally, in the case of small rings (eight-membered or smaller) in which each node is exclusively aromatic, the edge fingerprints are updated by resetting all nonaromatic bits to 0. This last step ensures that all aromatic systems are recognized as identical, independent of whether edges are written in implicit or aromatic form. The *Detect aromatic bonds* option allows SMARTScompare to

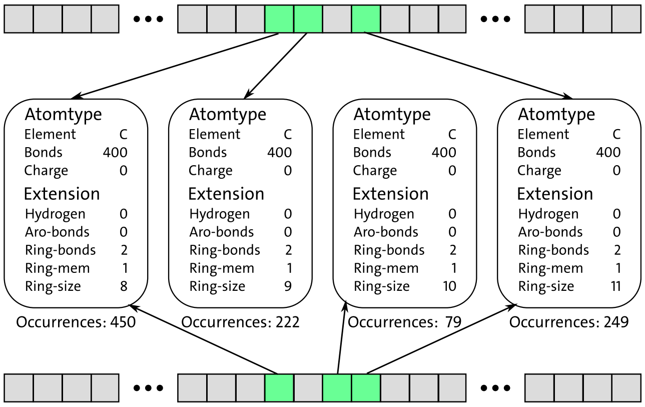


Figure 1. Fingerprints of two nodes. Green squares represent extended atomtypes that exist in the node the fingerprint describes. Beneath each atomtype box, the frequency of the particular atomtype is displayed. It is derived from the validation data set. This value is used as the weight in the calculation of node similarity in case the Tanimoto similarity is selected.

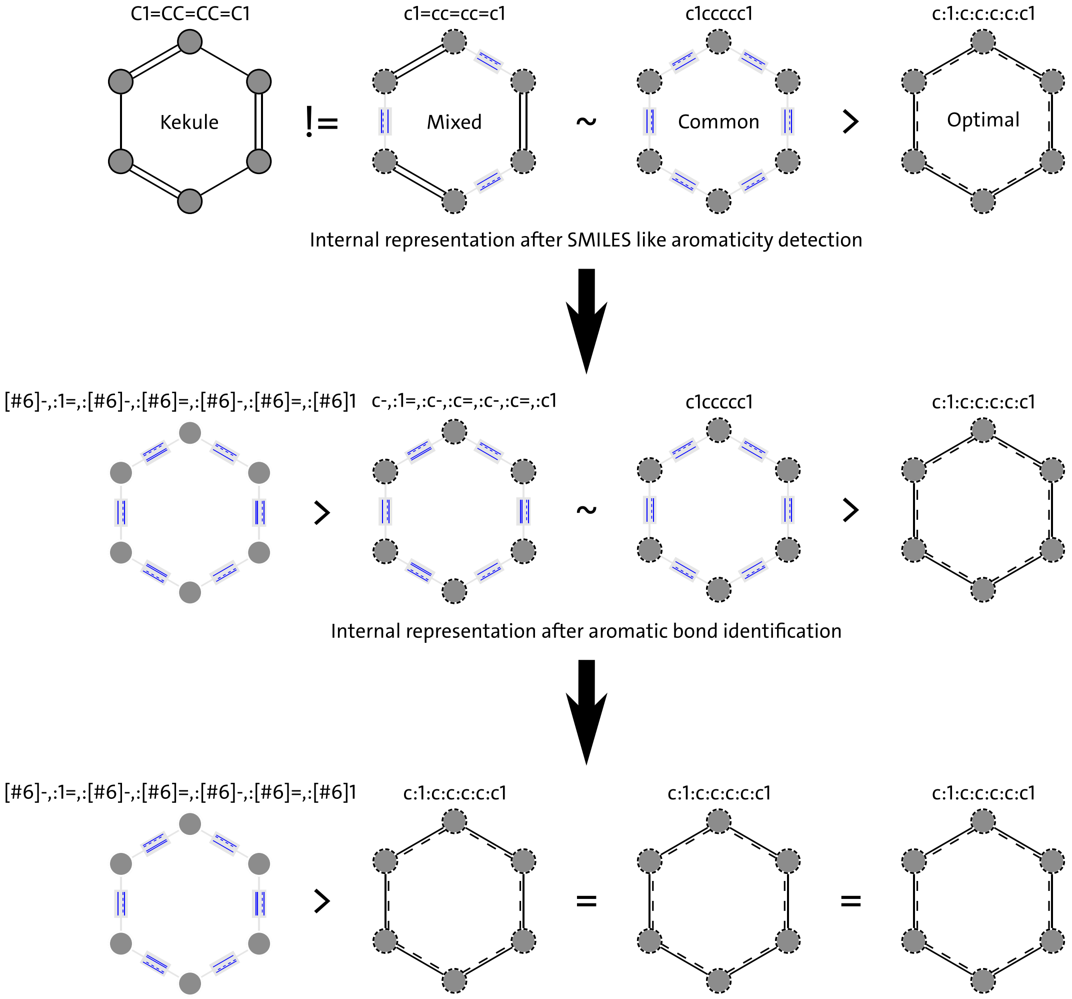


Figure 2. The first row shows possible variations of SMARTS patterns that were designed to match benzene. The second row shows the internal representation of the initial SMARTS pattern in the case where *SMILES-like aromaticity detection* is active. Patterns in the third row represent the internal form when *Detect aromatic bonds* identification is active as well. The logical operators between pattern depictions refer to the compatibility of node and bond fingerprints of the respective pattern pairs: !=, no relation whatsoever; ~, some relation in similarity mode; >, subset relation; =, identity.

transform bond descriptions into one uniform aromatic description. This is achieved by removing all bond information but the aromatic one. Figure 2 showcases this behavior on four notations for matching benzene that are all valid SMARTS strings.

It is important to keep in mind that when this feature of aromaticity handling is used, the underlying model of SMARTS is no longer consistent with the Daylight model.[4] Some internal representations that are needed to detect aromatic systems correctly can no longer be used to match molecules. Additionally, with *SMILES-like aromaticity detection* enabled, the Kekulé form is classified as more generic than its aromatic counterpart. When the *Detect aromatic bonds* option is enabled, nonintuitive behavior is possible in edge cases. An example, that is also described in the companion paper[19] is the handling of bonds that connect two aromatic systems of condensed rings. During pruning of the bond fingerprints they lose the information that they can be single bonds. Only the aromatic state remains, which would lead to wrong subset matching and similarity assessments.

■ METHODS

To validate the similarity concept, we first tested the coverage of enumerated extended atomtypes using the validation data set by assigning extended atomtypes to all atoms. Second, we tested whether we could model pattern similarity in a chemically intuitive manner.

Validation of Atomtypes. It is important to keep in mind that the NAOMI ChemBio library was designed to handle chemistry that is common in a medicinal chemistry setting. Any atomtypes that are currently not covered by our internal chemical model impair our ability to calculate the similarity and set relations of patterns. The validation data set has sufficient size and breadth to reliably test all necessary atom states of our chemistry model.

There are two fingerprint-related limitations that we deliberately tolerated. The set of extended atomtypes has limitations related to the maximal ring size and the number of URFs[23,24] to which an atom belongs. In an analysis of the ZINC molecules, we found that limiting URF membership and ring size is a good way to balance the coverage of necessary atomtypes and fingerprint length. Overall, we found 506

DOI: 10.1021/acs.jcim.9b00249
J. Chem. Inf. Model. 2019, 59, 2572−2586

molecules containing atoms with more than four URFs (18 atoms) or in rings larger than 24 atoms (12 896 atoms).

Furthermore, 374 atoms in 353 molecules had more attached hydrogen atoms than expected by the NAOMI chemistry model.[27] Atoms for which no extended atomtype is included in NAOMI were aggregated with the atomtype that has the maximal valid count of hydrogen annotated. These aggregations occurred mainly for sulfur and phosphorus atoms.

Lists of all molecules that gave an error with a short description of the type of error (unmodeled hydrogen counts, exceeding ring size enumeration, exceeding URF count) can be found in the Supporting Information.

Validation of the Pattern Similarity Concept. Most applicants of SMARTS consider patterns to be similar if the substructures matched by them are similar. To evaluate our similarity measure, we used the following experimental setup. For a given pattern P_1, we analyzed the fragments matched by P_1 in all molecules $CS(P_1) \subset CS$. Two patterns P_1 and P_2 are considered similar if all of the matched fragments have corresponding similar fragments in the other set. As a similarity measure of fragments, the diameter 6 extended-connectivity fingerprint (ECFP_6)[28] was applied. We employed the following measurement, the substructure similarity, as the reference similarity for SMARTScompare:

$$\mathrm{FL}(P) = \langle f \subseteq m \mid P \text{ matches } f,\ m \in CS(P) \rangle \tag{4}$$

$$\mathrm{sim}(f_m, f_n) = \frac{|\mathrm{ECFP\_6}(f_m) \cap \mathrm{ECFP\_6}(f_n)|}{|\mathrm{ECFP\_6}(f_m) \cup \mathrm{ECFP\_6}(f_n)|} \tag{5}$$

$$S_{\mathrm{ref}} = \frac{1}{|\mathrm{FL}(P_1)| + |\mathrm{FL}(P_2)|} \left[\sum_{f_m \in \mathrm{FL}(P_1)} \max_{f_n \in \mathrm{FL}(P_2)} \mathrm{sim}(f_m, f_n) + \sum_{f_n \in \mathrm{FL}(P_2)} \max_{f_m \in \mathrm{FL}(P_1)} \mathrm{sim}(f_m, f_n) \right] \tag{6}$$

To validate our similarity approach, we performed two experiments in which we focused on two types of patterns. The first experiment was based on substructure-like patterns with almost no SMARTS-specific features. Patterns that matched at least 10 000 molecules in the validation data set were selected from a set of SMARTS patterns published by Ehrlich and Rarey.[29] Similarity of molecules was calculated on the basis of the Tanimoto similarity of the substructures matched by the SMARTS pattern. The second experiment was based on structural filters. Here we used the SMARTS filters taken from ChEMBL (see Table 2). As set of molecules for this experiment, we took the set of unique molecules from ChEMBL22.[30] Those patterns are usually smaller and use more SMARTS features than those in the first experiment.

Figure 3 shows the correlation plots for the two experiments. The experiments are described and analyzed in more detail in the Supporting Information (see SI1). The plots show that substructure-like patterns (Figure 3a) exhibit a good distribution as well as a good correlation, whereas the similarity for the structural filters (Figure 3b) is dominated by similarity scores below 0.2.

Overall our experiments show that pattern similarities calculated by SMARTScompare correlate with molecule similarities derived from matching of fragments to a large reference data set.

Table 2. Publicly Available Filter Sets As Published in ChEMBL[2] (Excluding SMARTCyp)

name	no. of patterns	publication
Bristol-Myers Squibb	180	Pearce et al.[10]
SMARTCyp	42	Rydberg et al.[11]
Dundee	105	Brenk et al.[13]
Glaxo Wellcome	55	Hann et al.[6]
Inpharmatica	91	no publication available
MLSMR	116	website offline/project discontinued
LINT	57	Blake[12]
PAINS	481	Baell et al.[8]
SureChEMBL OCHEM/ ToxAlerts	166	Sushko et al.[7,39]

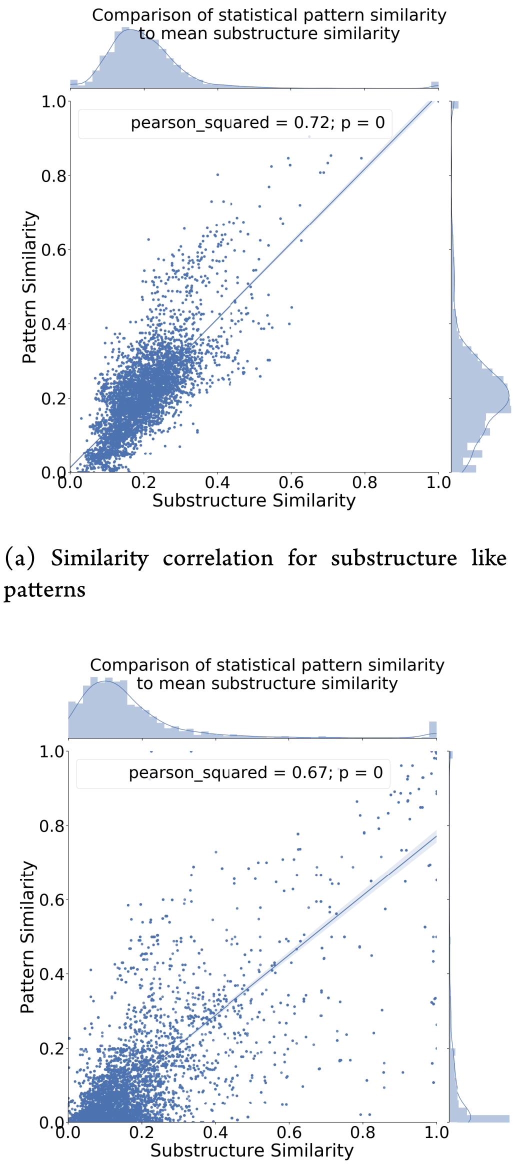
SMARTScompareViewer. Even for an expert user, SMARTS subset relations determined by the SMARTScompare algorithm are sometimes hard to comprehend. To visually support the interpretation, we have developed the SMARTScompareViewer, an intuitive tool that visualizes SMARTS patterns and shows the calculated node mappings that result from the SMARTScompare algorithm. The SMARTScompareViewer is based on the SMARTSView concept and the SMARTSViewer.[14,15] Besides showing the actual node mapping, the SMARTScompareViewer can emphasize the difference between SMARTS nodes. This helps the user to understand unexpected results, meaning that there are more or fewer nodes mapped than expected. SMARTScompareViewer is also available on the web as part of the SMARTSviewServer (https://smarts.plus).

The SMARTS Filter Sets. With an implementation of SMARTScompare, we are able to perform a comprehensive comparison of publicly available filter sets for the first time. Filter sets are used by many pharmaceutical companies and research groups in academia to remove molecules with unwanted properties such as reactivity, toxicity, or assay interference. Since these filter sets are mostly handcrafted, it is interesting to compare different flavors of patterns aimed at the same property. Usually, specific substructural features are responsible for unwanted behavior of compounds. Substructures can be described via SMARTS patterns, allowing computationally easy selection of molecules having such chemical properties.

One of the first pharmaceutical companies to publicly release such a filter set was GlaxoSmithKline (then Glaxo Wellcome Research).[6] The SMARTS patterns published are structured into four groups: reactive functional groups, unsuitable leads, unsuitable natural products, and another group summarizing all SMARTS filters for acids, bases, electrophiles, and nucleophiles. The main purpose of the published filter sets is to avoid pooling of compounds that react with each other within one batch during HTS.

A second set of SMARTS patterns with a similar focus was published by Bristol-Myers Squibb.[10] The authors were looking for a consistent way to remove compounds from HTS screening decks and to flag compounds that do not disturb screenings but are helpful in assay evaluation when flagged as potentially disruptive. Other filter sets were designed with a broader focus in mind.

James Blake of Array BioPharma Inc. investigated how to reduce general attrition during drug development.[12] Apart from the usual suspects drug-likeness, lead-likeness, and



(a) Similarity correlation for substructure like patterns

(b) Similarity correlation for structural filters

Figure 3. Correlation plots for pattern and substructure similarity. Pattern similarity refers to the calculated value of SMARTScompare for two patterns. The substructure similarity is the Tanimoto similarity of the substructures that the nodes of the patterns match. Plot (a) corresponds to the first experiment with substructure-like patterns, and plot (b) corresponds to the second experiment with filters that include more SMARTS features.

implications of computed properties, the author addressed functional groups linked to problematic performance during the drug development process. Blake tied a certain percentage of the overall attrition to "reactive groups and compounds or functionalities that have been shown to be mutagenic or carcinogenic". Those groups also "tend to give false positives in high-throughput screens". Researchers from the University of Dundee in Scotland published a filter set with a similar focus but specializing in neglected diseases.[13]

Baell and Holloway focused solely on pan-assay interference compounds (PAINS).[8] As a result of the increased interest in HTS at the time of their writing, they wanted to develop substructure filters that efficiently encode structural information on problematic compounds. With such a description at hand many interfering compounds and their analogues can be excluded from pending bulk purchases.[8] The substructures were originally published in SLN[5] and later translated into SMARTS patterns.[31] The OCHEM project also translated the SLN notation of the PAINS patterns.[7,32] In this case they were translated by hand. In the Results we discuss the difference in the two translations.

There are two other projects in the context of HTS screening, but no official information regarding their purpose could be found as of the release date of this work. The now-discontinued National Institutes of Health Molecular Libraries Small Molecule Repository (NIH-MLSMR), which was a part of the now-retired Molecular Libraries Probe Production Centers Network (MLPCN), assembled a set of SMARTS filters that is now managed by Evotec.[2,33] The purpose of the assembled filter sets was to remove "chemically reactive functional groups that would interfere with HTS, as well as compounds likely to be promiscuous aggregators"[34,35] with a reference to McGovern et al.[36] ChEMBL has another set of SMARTS filters that were derived by Inpharmatica Ltd.,[37] for which also no official publication exists. The filter set was initially intended to be used on ChEMBL data (when they were still a commerical product of Inpharmatica Ltd.). The purpose was to filter groups that are generally undesirable during drug design.[34]

The following two filter sets have a focus that diverges from those already presented. Sushko et al.[7,32] assembled a web server[38] for structural alerts for toxic chemicals and compounds with potential adverse reactions. The OCHEM/ToxAlerts[7] initiative is an ongoing project, and every user is encouraged to submit new structural alerts. SureChEMBL uses SMARTS patterns for structural highlights on their website that are extracted from the OCHEM/ToxAlerts SMARTS set. These patterns are also part of the ChEMBL releases.[2]

SMARTCyp by Rydberg et al.[11] has a completely different focus from all other SMARTS sets. SMARTCyp is an in silico method that predicts reaction sites of drug-like molecules in cytochrome P450. The idea is to precalculate the reactivity for certain atoms that are part of a chemical group as the activation energy for oxidation reactions. For each atom of the chemical group, a SMARTS pattern is generated and, together with the corresponding energy, stored in a database. When a query molecule is given, a score for each atom is extracted from the precalculated energy values using the fitting SMARTS pattern.

ChEMBL published SMARTS versions of the previously listed structural filters with their 2017 release.[2] A summary of the SMARTS filter sets can be found in Table 2. All of the SMARTS patterns used in this work were taken from one of the SMARTS sets listed in Table 2. Except for SMARTCyp, which was extracted from the publication, all of the filter sets were taken from the MySQL ChEMBL release dump.[2] All of the patterns were used unchanged except for seven patterns from the PAINS filters. These patterns were modified to reduce the number of explicit hydrogens. The replacement of explicit hydrogens by implicit ones substantially reduces the run time of our MCS calculation, which maps all explicit atoms exhaustively. These modified PAINS patterns are provided in the Supporting Information.

In the case of PAINS filters, we had the opportunity to compare two sets of patterns that were both translated from SLN into SMARTS. One set of PAINS SMARTS patterns were taken from the ChEMBL release as stated above. The

DOI: 10.1021/acs.jcim.9b00249

other SMARTS set, used in the publication of Schorpp et al.,[40] was extracted from the OCHEM website.[38]

5. RESULTS AND DISCUSSION

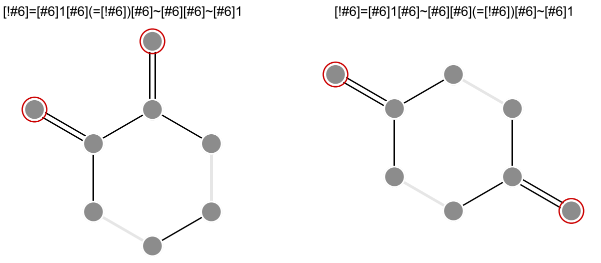
In the following paragraphs we will showcase the capabilities of the algorithm detailed in the companion paper[19] and summarized at the beginning of Theory. The focus of our analysis lies with the set relations of patterns and their similarity. One of our central assumptions is that patterns in filter sets were designed with a specific purpose in mind. We are estimating the intentions of the authors on the basis of the labels that patterns were given and the publications with which they were released, if they exist.

Set Relations and Pattern Redundancy in Filter Sets. One of the simplest application cases of our algorithm is the analysis of existing pattern sets with regard to their content. Often we want to know whether a certain type of chemistry is already contained in the filter set. For this work, four questions are of importance in particular:

- Does our algorithm find all of the related patterns?
- Since SMARTS patterns of the respective filter sets were labeled with the systematic or trivial name, how large is the unwanted overlap between patterns for different purposes?
- Do the patterns actually have the desired form (as specified in the corresponding label)? Sometimes a more precise or general form of a pattern might meet the desired criteria better than the presented one.
- Can we reduce the number of patterns by identifying and removing redundant ones?

To analyze the filter sets regarding their overlap and coverage of chemical functionality, we employed subpattern calculations and string search on the pattern labels.

Consistency of Label and Pattern. Our first experiment focused on patterns that are designed to filter out reactive compounds containing quinone-like substructures. On the one hand, we designed two generic patterns (Figure 4) to find as many quinone patterns as possible. In total, we found 14 patterns, of which 11 match the para variant and three the ortho variant. Eleven of them have annotations referring to quinones. The three patterns without a quinone label are annotated as *keto_keto_gamma(5)* (PAINS), *Dye 1 (1)* (MLSMR), and *Dye 4* (MLSMR), as can be seen in Figure 5. For comparison, we conducted a simple string search on the labels of patterns. Any SMARTS pattern whose label contained the string “quinone” was selected as a result. Seventeen patterns were found. Patterns designed to match derivatives of quinones may still contain the “quinone” substring and are in the solution set of our query. This naming scheme results from the systematic nomenclature that is employed in organic chemistry. One difference in numbers of patterns is attributable to the class of hydroquinones (Figure 6).



(a) The ortho variant of our quinone query patterns (1,2-benzoquinone)

(b) The para variant of our quinone query patterns (1,4-benzoquinone)

Figure 4. Two generic SMARTS patterns to find as many quinone SMARTS strings as possible: (a) ortho version; (b) para version. The focus here lies on finding all patterns that are labeled as quinones and then analyzing the bycatch.

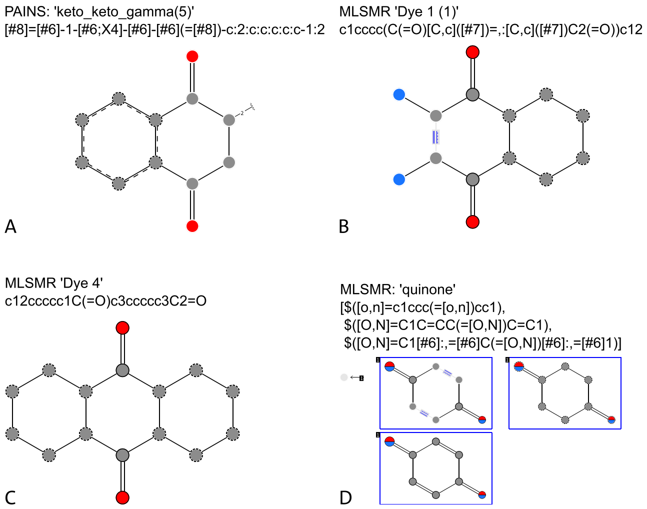
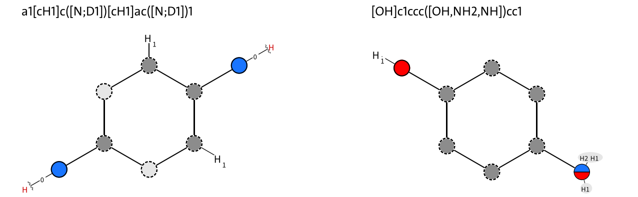

PAINS: 'keto_keto_gamma(5)'
[#8]=[#6]-1-[#6;X4]-[#6]-[#6](=[#8])-c:2:c:c:c:c:c-1:2
2
A
MLSMR 'Dye 1 (1)'
c1cccc(C(=O)[C,c]([#7])=,:[C,c]([#7])C2(=O))c12
B
MLSMR 'Dye 4'
c12ccccc1C(=O)c3ccccc3C2=O
C
MLSMR: 'quinone'
[$([o,n]=c1ccc(=[o,n])cc1),
$([O,N]=C1C=CC(=[O,N])C=C1),
$([O,N]=C1[#6]:,=[#6]C(=[O,N])[#6]:,=[#6]1)]
1
D

Figure 5. Four patterns that are matched by the quinone query patterns shown in Figure 4 without having a quinone label. Pattern A is part of the PAINS filter set. Patterns B and C are part of the MLSMR filter set. Pattern D is a more generic SMARTS pattern that is part of the MLSMR filter set and renders patterns B and C obsolete when used together.



(a) *para_hydroiminiquinone* as given in the BMS filter set.

(b) *hydroquinone* as given in the Dundee filter set.

Figure 6. Two quinone derivatives that are found in the pattern sets of BMS and Dundee.

Quinones and their derivatives, when hydrogenated at the heteroatom, are transformed into their quinol form. These aromatic forms are not matched by our query patterns as we have designed them. Hence, they are not in the list of patterns we find with either of our two query patterns. They could, however, be easily incorporated. Another difference between the string search and the pattern query are the two patterns displayed in Figure 7. The description of the ring and bonds to the heteroatoms diverge drastically from our notation for quinones. It is in general debatable whether the two patterns can still be classified as quinones. Last but not least, simple notation differences have an impact on substring search too. “Quinone” can also be written as “chinone”, which in the case of string matching means we would miss two patterns (77 and 78) from the Dundee filter set.

In general, the pattern labels describe molecules or set of molecules from the class of quinones well. Aside from the two examples shown in Figure 7, the labels are very accurate and descriptive. Depending on the design of the query pattern, all

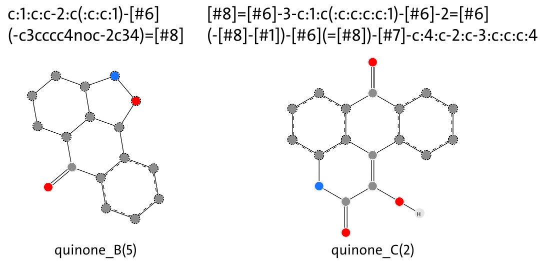

c:1:c:c-2:c(:c:c:1)-[#6]
(-c3cccc4noc-2c34)=[#8]
[#8]=[#6]-3-c:1:c(:c:c:c:c:1)-[#6]-2=[#6]
(-[#8]-[#1])-[#6](=[#8])-[#7]-c:4:c-2:c-3:c:c:c:4
H
quinone_B(5)
quinone_C(2)

Figure 7. Two patterns from the PAINS filter set that we did not find with the generic quinone patterns (Figure 4). Both are labeled as quinone patterns but also include heteroatoms, rendering them unmatchable for our generic query.

of the patterns labeled as quinones or their derivatives can be found. Nonetheless, everything depends on the formulation of the query pattern. However, we did find patterns containing a quinone-like subpattern that we would have been unable to identify without our algorithm. It is important to remember that a subset relation of patterns is manifested as overlapping sets of molecules that are matched by each of the patterns. In our case this would have an impact on the necessity of patterns. In the case where we would include our query patterns into the filter set that also contains the dye patterns (Figure 5), we could remove those and still filter out all molecules that would have been matched by the dye patterns. Furthermore, the MLSMR filter set already contains a very general quinone pattern (Figure 5D). It matches all molecules that are matched by the dye patterns shown in Figure 5B,C. If all three patterns are used together, which they usually are in a filter set, the dye patterns are redundant and may be removed from the filter set without loss.

All of the quinone patterns, including our query patterns, are visualized in the Supporting Information.

Pattern Hierarchies and Redundancy. Our second experiment revolved around the question of which more specific patterns are rendered redundant when we start with a very generic pattern. For this experiment, we selected allenes as example chemical group.

For our search we used the pattern *=C=*, as it is part of the MLSMR and SureChEMBL OCHEM/ToxAlerts filter set. We searched for more specific patterns than the generic allene in all of the filter sets and found 12 matches and eight different patterns in total. Those patterns that occurred in more than one filter set were identical, including the annotation.

When visualizing the hierarchy of the resulting patterns, a specificity tree as shown in Figure 8 can be constructed that shows the most generic pattern at the top and more specific patterns on the subsequent levels below. Each pattern below the top pattern is redundant in the sense that each molecule that would be matched by one of the lower patterns will in every case also be matched by the top pattern. This is especially interesting when we turn our attention toward the origin of patterns that are part of the hierarchy. Five patterns are from the same SureChEMBL OCHEM/ToxAlerts toxicity set, of which four are obsolete and can be removed without compromising the integrity of toxicity filtering mechanism.

In our third experiment, we focused on a bottom-up approach by selecting a relatively specific sulfonyl halide pattern with an anchor wildcard node. Here we were curious whether there were any more generic patterns that might be of interest to our fictive task. The search for more generic patterns found 13 patterns including the query from the Inpharmatica filter set. The 13 matches can be grouped into seven partitions of equal patterns. Those equal patterns differed in the order of the SMARTS nodes. The visualization of the set relation we found is shown in Figure 9.

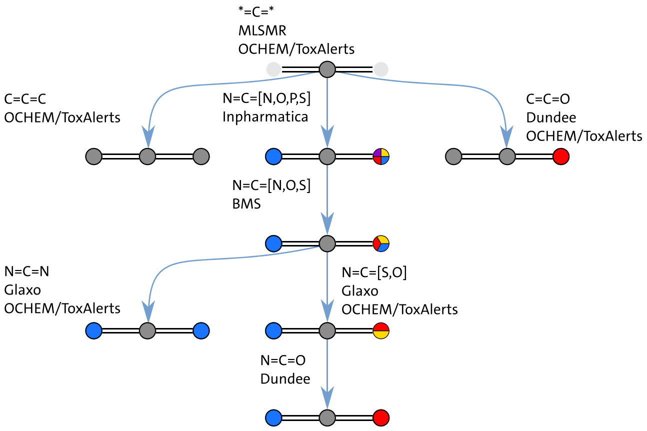


Figure 8. Hierarchical scheme of patterns specifying allene-like chemical groups. The hierarchy starts with the most generic allene pattern at the top and more specific patterns toward the bottom. Equal level does not imply equal specificity, but each pattern on a lower level is more specific than a pattern on the level above.

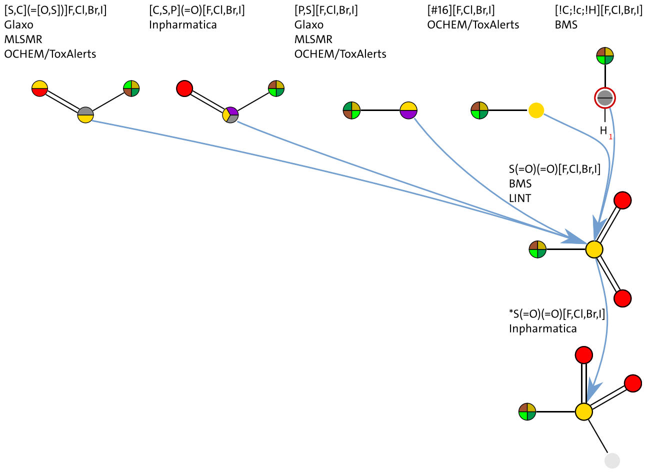


Figure 9. Hierarchical scheme of more generic patterns than the sulfonyl halide pattern. The hierarchy starts again with more generic patterns at the top and gets more specific toward the bottom.

Using SMARTScompare, we are able to analyze the degree to which the intention, derived from the textual label that comes with the pattern, is translated into the final SMARTS pattern. Because of the complexity of the SMARTS language, unforeseen molecules are matched, especially when very generalized patterns are used. The second interesting point is the redundancy of patterns in filter sets. Especially when expressions get more complicated, keeping track of each function that has been cast into a pattern gets increasingly more difficult. By analyzing set relations, we are now able to produce hierarchical schemes that give an easy-to-perceive overview of the subset relations of patterns. This allows us to eliminate redundant patterns and sharpen the focus of the filter set.

Filter Set Similarity Analysis. For generic pattern set comparison, the explicit search for more specific or generic

DOI: 10.1021/acs.jcim.9b00249
J. Chem. Inf. Model. 2019, 59, 2572−2586

patterns is not well-suited. The binary form of the result is too coarse-grained and forbids a subtle analysis of small differences between patterns. This problem is better addressed with a similarity-based approach.

In order to calculate the similarity between two filter sets, we apply the following procedure:

- For each pattern of the query filter set, we compute the similarity to each pattern in the second filter set.
- For each query pattern, we select the most similar pattern from the second set.
- The average of the similarity values of the most similar pattern pairs is calculated between the two sets as an asymmetric similarity measure of those two sets.

The self-similarity of a SMARTS pattern set is calculated as described above, with the one difference that the query pattern was excluded. In this way, we get an idea of the most similar pairs of patterns in a filter set.

The similarity value of two different sets can be considered as a degree of coverage of one filter set by the other. Our analysis of pattern set self-similarity, on the other hand, was conducted with the assumption that higher dissimilarity covers a greater variety of chemical functions and the set contains fewer redundant patterns. Similarity between node fingerprints of a pattern pair was calculated with the weighted Tanimoto coefficient shown in eq 2. The weights were derived from the validation data set as described in Similarity Calculation. In the following paragraphs we discuss the similarity values between filter sets and, if the given data allow for it, tentatively analyze possible reasons.

The Glaxo set constitutes an excellent beginning since it was the first filter set published and also was designed with a similar goal as the majority of the other sets. Figure 10 displays the accumulated values for the Glaxo set's most similar pairs. There is one major peak in the histogram with its maximum at ~0.3 and a local maximum at ~0.5. Since there are no filter pairs with similarity 1.0, the filter set does not contain any duplicates. With a self-similarity mean of 0.34, the Glaxo filter set lies at the lower end of self-similarity values (see Figure 13 for more values).

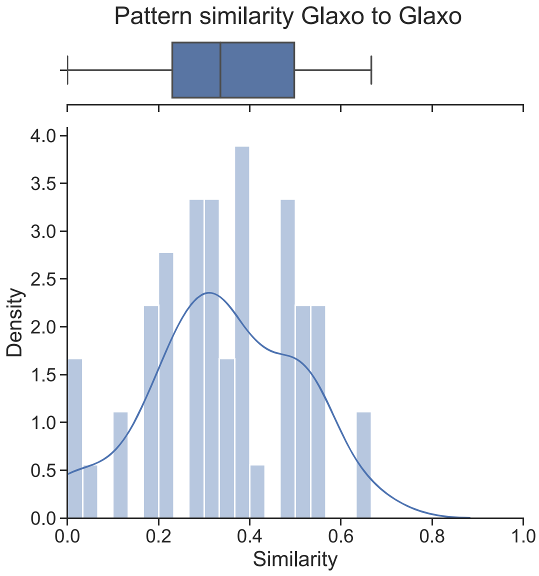


Figure 10. Counts of the self-similarity values for the Glaxo filter set. The mean similarity lies at 0.34. The set has no identical patterns. The density plot has a maximum at ~0.3 and another local maximum at ~0.5.

When we compare the Glaxo set with all of the other filter sets, we get the similarity values displayed in Figure 11. The filter set similarities that draw our attention are the ones with extreme similarity or dissimilarity values. MLSMR and SureChEMBL OCHEM/ToxAlerts are the two with the highest similarity values, while SMARTCyp and PAINS are most dissimilar to the patterns of the Glaxo filter set. The similarity value distributions of these four filter sets are displayed in Figure 12. The similarity between the Glaxo and MLSMR sets is surprisingly high. The extent of the similarity indicates that the Glaxo filter set is almost fully contained in the MLSMR set. However, it does not indicate that most of the MLSMR patterns are from the Glaxo set. In Figure 13 we can see that the similarity value from MLSMR to Glaxo is 31% lower. In contrast, the histograms in Figure 12c,d barely contain similarity values above 0.5, showing that most of the Glaxo patterns are not covered by PAINS and SMARTCyp.

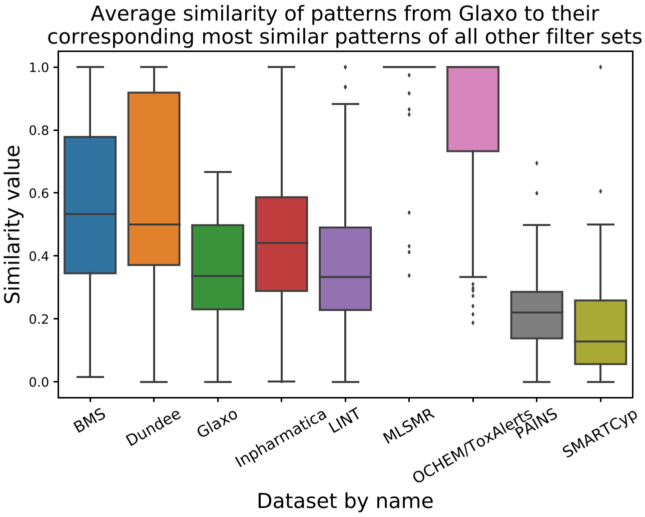
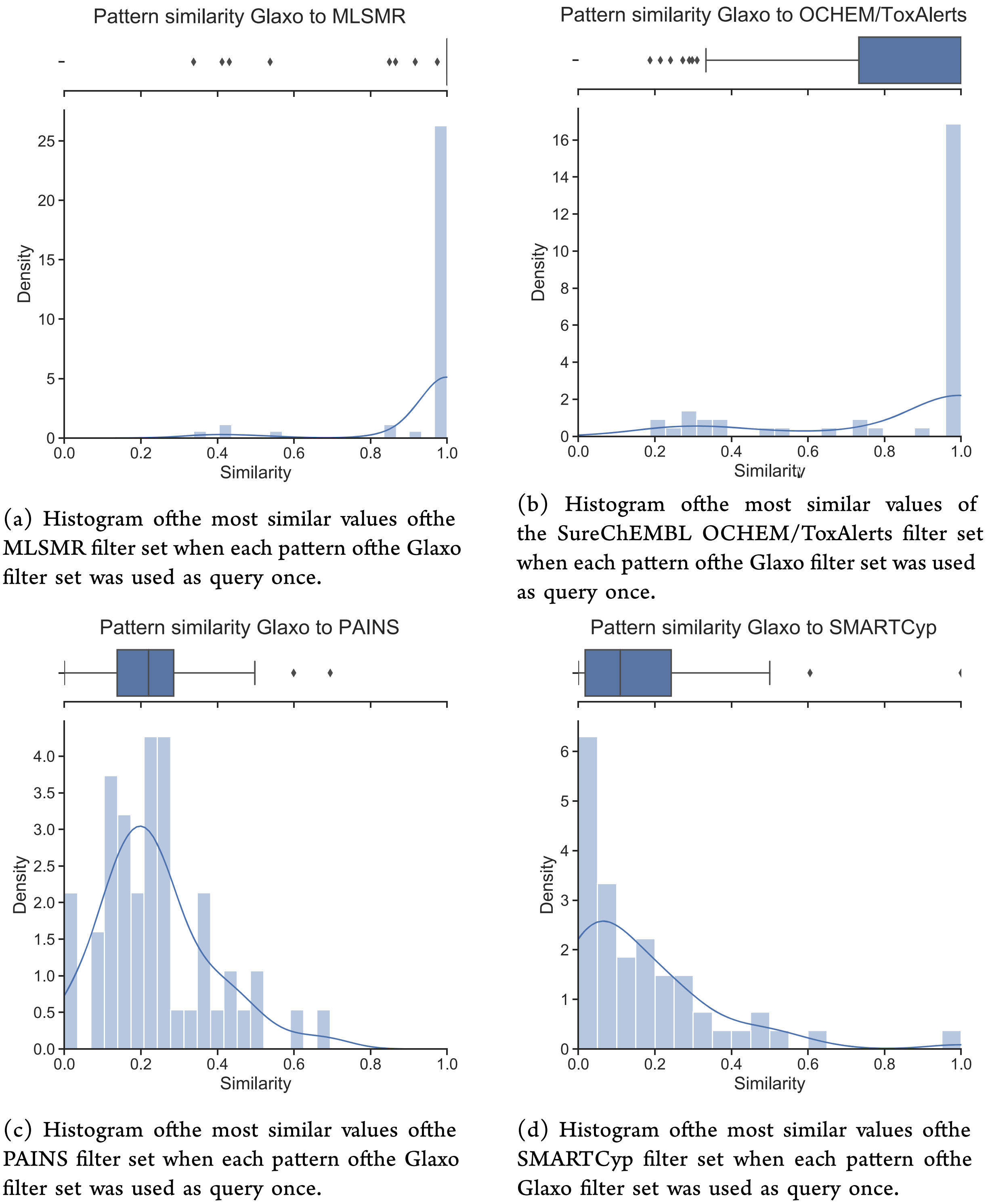


Figure 11. Box plots of the similarity histograms for pattern similarity for all patterns from the Glaxo Wellcome filter set to their corresponding most similar patterns in all of the other filter sets. The Glaxo box plot represents the self-similarity of the Glaxo filter set.

Finally, we extended this experiment to a large-scale all-against-all filter set comparison, as can be seen in Figure 13. Most of the average similarities fall between 0.4 and 0.6. Also, SMARTCyp and PAINS are most dissimilar to all of the other filter sets. The asymmetry of our set similarity measure is clearly visible for the SMARTCyp set. The average similarity for any filter set to SMARTCyp is below 0.25, and in most cases even below 0.2. In the opposite direction, the average similarity from SMARTCyp to any other filer set is above 0.25 except for PAINS (0.16). The other extremes are Glaxo to MLSMR with a similarity of 0.95 and Glaxo to SureChEMBL OCHEM/ToxAlerts with a similarity of 0.83.

There might be many reasons why the Glaxo−MLSMR similarity value is so high. Since the Glaxo set was the first to be published and the MLSMR project was a concerted effort of several institutions only a few years later, it probably was incorporated into the MLSMR filter set almost unchanged. When we were analyzing the two filter sets with the SMARTScompare tool, we found that there are only nine



(a) Histogram ofthe most similar values ofthe MLSMR filter set when each pattern ofthe Glaxo filter set was used as query once.

(b) Histogram ofthe most similar values of the SureChEMBL OCHEM/ToxAlerts filter set when each pattern ofthe Glaxo filter set was used as query once.

(c) Histogram ofthe most similar values ofthe PAINS filter set when each pattern ofthe Glaxo filter set was used as query once.

(d) Histogram ofthe most similar values ofthe SMARTCyp filter set when each pattern ofthe Glaxo filter set was used as query once.

Figure 12. Histograms of similarity values of the Glaxo Wellcome patterns from searches for the most similar ones in four other filter sets. The four filter sets were selected on the basis of their high similarity (a, b) or dissimilarity (c, d).

patterns in the Glaxo set that are not identically included in the MLSMR set. They are shown in Table 3.

Eight of the nine patterns can be observed as differing values in the histogram in Figure 12a and the box plot in Figure 11, where they are classified as outliers by the box plot visualization. The ninth pattern (R22) differs from its most similar MLSMR pattern only in the order of the halogens. The reverse similarity lies at 0.64, which also makes sense when one considers the slightly different focus of the MLSMR filter set. They not only tried to avoid pooling of reactive compounds but also were interested in filtering out promiscuous aggregators as defined by McGovern et al.,[36] leading to a drop of similarity of 31%. The high similarity between the Glaxo and SureChEMBL OCHEM/ToxAlerts filter sets makes sense too if we take into account that many toxic compounds exhibit toxicity because of their reactive nature. When we have

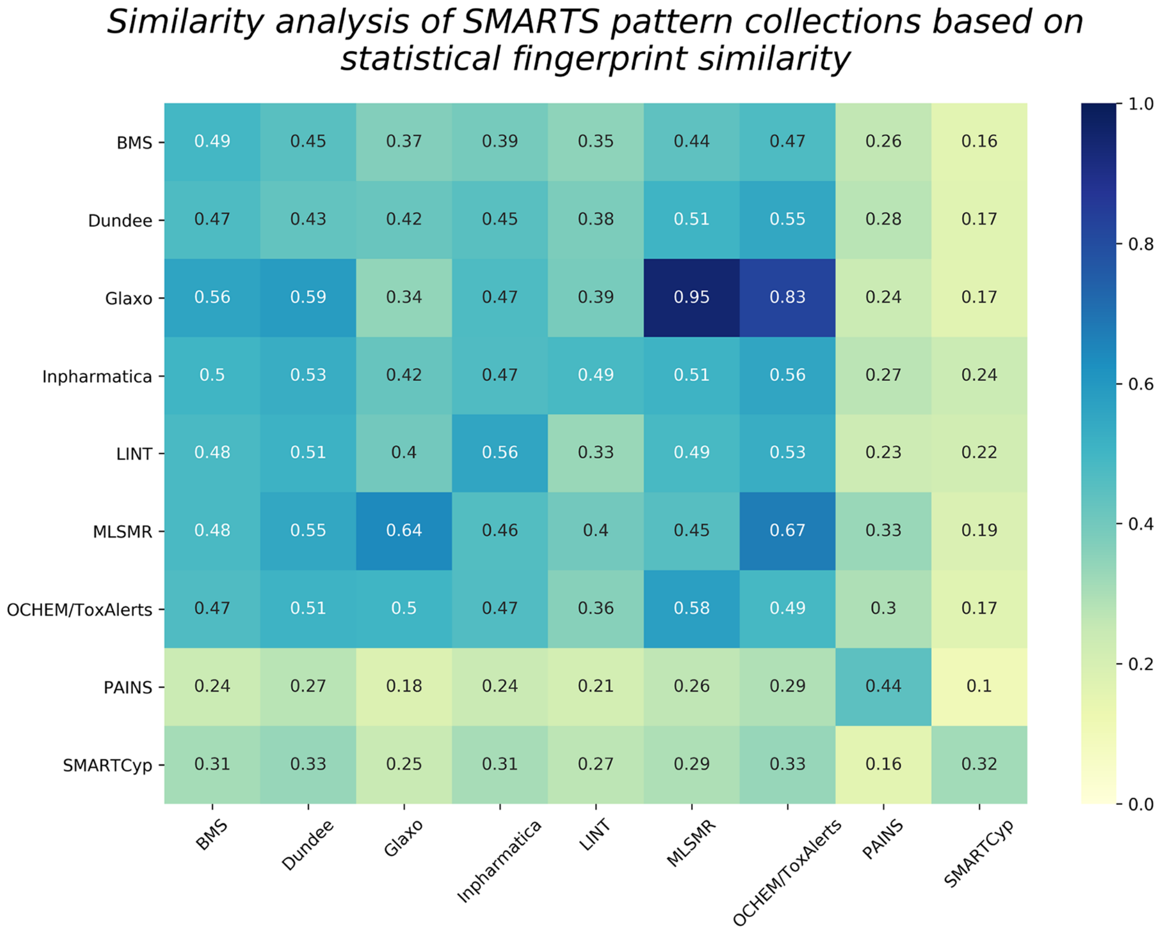


Figure 13. All-against-all filter set similarity matrix plot. Each cell describes the average similarity for patterns from one filter set and their corresponding most similar patterns in the other filter set. On the diagonal, the filter set self-similarity is measured. Here the most similar nonidentical pattern is averaged; diagonal entries of 1 would indicate plenty of pattern duplicates.

Table 3. Patterns from Glaxo That Are Not Identically Part of the MLSMR Filter Set

annotation	pattern
N1 Quinones	O=C1[#6]~[#6]C(=O)[#6]~[#6]1
R1 Reactive alkyl halides	[Br,Cl,I][CX4;CH,CH2]
R3 Carbazides	O=CN=[N+]=[N−]
R23 Carbodiimide	N=C=N
I1 Aliphatic methylene chains 7 or more long	[CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0][CD2;R0]
R11 Isocyanates & Isothiocyanates	N=C=[S,O]
R16 beta carbonyl quaternary Nitrogen	C(=O)C[N+,n+]
R9 Paranitrophenyl esters	C(=O)Oc1ccc(N(=O)~[OX1])cc1
R22 P/S Halides	[P,S][Cl,Br,F,I]

a look at the similarity of SureChEMBL OCHEM/ToxAlerts to Glaxo, on the other hand, we see the difference in reactivity and toxicity ingrained into the similarity value. Toxicity is caused not only by reactive compounds but also compound classes that interfere with signaling processes in the body and compounds that are translated into their toxic form by means of enzymatic activity. This leads to a drop of 33%.

The dissimilarity of the Glaxo and SMARTCyp filter sets (0.17) can be explained by the very different focuses of the filter sets. The Glaxo filter set was specifically designed to exclude highly reactive compounds from screening libraries. SMARTCyp was designed to describe very small, often only one atom in size, chemical groups that act as keys in a dictionary of density functional theory calculations, which is orthogonal to the goal of HTS library design.

The dissimilarity of the PAINS and Glaxo filter sets (0.24), on the other hand, is not explained easily, and formulating a sophisticated hypothesis would require a more thorough analysis. One reason about which we are confident is the substantial difference in pattern size. The PAINS pattern set includes many patterns describing whole molecules, while the Glaxo filter set focuses on reactive chemical groups. Figure 14 shows an example of a pattern from the PAINS filter set. The similarity of PAINS to SMARTSCyp (0.1) constitutes the global minimum of our similarity analysis. This supports our hypothesis that large parts of the dissimilarity can be attributed to size difference: large, very specific SMARTS expressions in the case of PAINS filters and one-atom expressions in the case of SMARTCyp.

To deepen our understanding of the filter sets and SMARTS−SMARTS similarity we analyzed highly similar pattern pairs. We searched for the most similar pattern for each of the nine patterns from the Glaxo−MLSMR example that were not found in both sets (Table 3).

We will discuss the patterns *R23 Carbodiimide* and *R11 Isocyanates & Isothiocyanates* exemplarily. Both have the same most similar pattern within the MLSMR filter set, namely, *=C=*, annotated as *allene* (see Figure 15). First, we notice

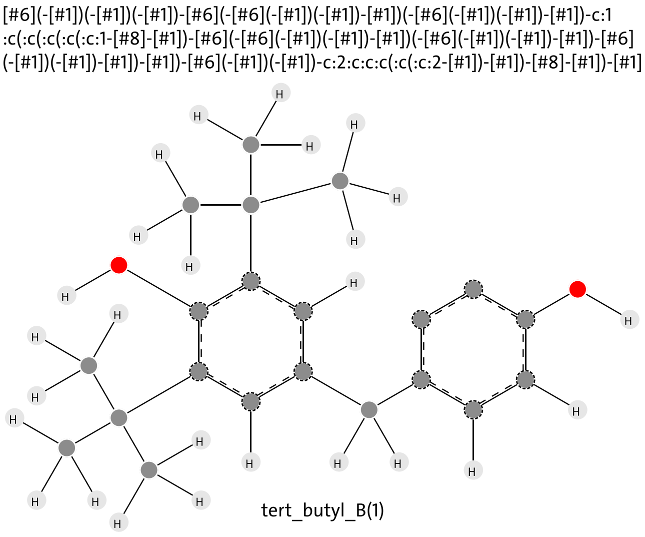

[#6](-[#1])(-[#1])(-[#1])-[#6](-[#6](-[#1])(-[#1])-[#1])(-[#6](-[#1])(-[#1])-[#1])-c:1
:c(:c(:c(:c(:c:1-[#8]-[#1])-[#6](-[#6](-[#1])(-[#1])-[#1])(-[#6](-[#1])(-[#1])-[#1])-[#6]
(-[#1])(-[#1])-[#1])-[#1])-[#6](-[#1])(-[#1])-c:2:c:c:c(:c(:c:2-[#1])-[#1])-[#8]-[#1])-[#1]
H
tert_butyl_B(1)

Figure 14. SMARTS pattern taken from the PAINS filter set (*tert_butyl_B(1)*) that looks like a molecule was translated into a SMARTS pattern.

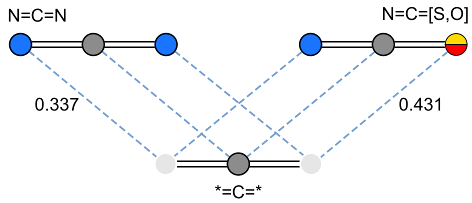


Figure 15. Two patterns from Glaxo with a common most similar pattern from MLSMR. In the upper part of the picture, the two Glaxo patterns are shown with their similarity values to the MLSMR pattern, which is displayed in the lower part of the figure. Although all of the nodes participate in the mapping, both patterns have similarities smaller than 0.5. The wildcards in the MLSMR pattern are the reason for those low similarity scores.

that the MLSMR pattern is more generic than both of the Glaxo patterns. Therefore, it will match all molecules matched by the two Glaxo patterns. When interpreting the SMARTS patterns, we could say that all cumulated double bonds around carbon atoms are considered as reactive by the authors of the MLSMR filter set. The Glaxo patterns, on the other hand, require heteroatoms to be included in order for cumulated double bonds to be deemed too reactive. Moreover, visualization of the Glaxo SMARTS patterns makes it very clear that both patterns could be replaced by a single pattern of N=C=[S,O,N] without losing any filtering power, i.e., we would still match the same molecules.

Another example, this time with high similarity (0.64), was taken from the Inpharmatica to SureChEMBL OCHEM/ToxAlerts filter set comparison (Figure 16). Upon examination of the two patterns displayed, the high similarity value is easily explained chemically. Both patterns describe a carbonyl-like group between two carbons, one of which is connected to one of the most abundant atoms in medicinal chemistry, a halogen (F, Cl, Br, I). Differences are a possible nitrogen at the position α to the carbonyl carbon and a higher specificity for the bond to the atom at the α position in the other direction. Because of the different pattern sizes (one pattern with four nodes and the other with five nodes), the similarity is capped at 0.8. Therefore, the similarity value of 0.64 is relatively high considering the pattern differences. Because of the statistical weighting of individual bits, atomtypes with low occurrence like halogens and sulfur substantially increase the similarity. When the reduced fingerprints without weighting are applied, the similarity drops to 0.29.

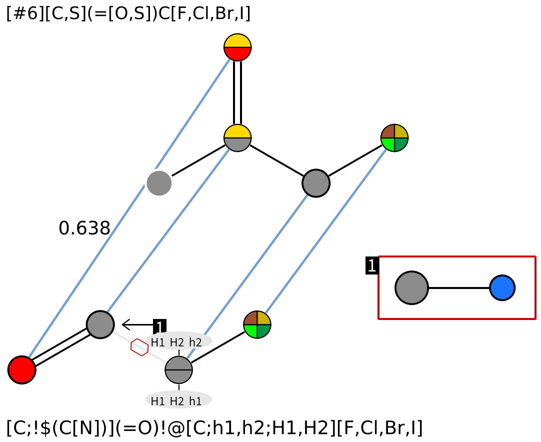

[#6][C,S](=[O,S])C[F,Cl,Br,I]
0.638
1
1
H1 H2 h2
H1 H2 h1
[C;!$(C[N])](=O)!@[C;h1,h2;H1,H2][F,Cl,Br,I]

Figure 16. Two patterns from the Inpharmatica and SureChEMBL OCHEM/ToxAlerts filter sets. They are annotated as *Filter64_-halo_ketone_sulfone* and *Alpha_Halo_Carbonyl*. In contrast to the previous example (Figure 15), there is no subset relation between these patterns. Parts of them are vice versa more specific.

In conclusion, pattern similarity is influenced heavily by the difference in pattern size (the number of nodes) and the weighting of bits of node fingerprints. This behavior is intended and reflects, to the best of our knowledge, the intuitive understanding of pattern similarity. The pure number of common active bits used for molecular similarity is mostly irrelevant for chemical pattern similarity since the number of atomtypes per element varies substantially. The weighting scheme compensates for this effect. We have observed that patterns with high similarity mostly differ in one or a few nodes only. Moreover, most of the time the difference can be attributed to one node allowing exactly one element while the other allows a second or third element alternatively. Reasons for dissimilarity, on the other hand, are so diverse that we do not feel confident in making a generalizing assertion.

Aromaticity Handling. When handling SMARTS comparison strictly as described in theory,[4] patterns that would be perceived as similar or even identical are not classified as such. This is due to different notations for aromatic systems and a lack of techniques to transform them. Earlier in this work we discussed two modes that we included in the implementation of the comparison algorithm. Those two modes introduce aromaticity detection in which patterns with aromaticity are handled more intuitively during comparison.

To demonstrate this effect, we compare two translations of PAINS patterns into SMARTS that were originally published in SLN.[8] The one we took from the MySQL ChEMBL dump was automatically translated with the tool CACTVS.[31] The other set was translated by hand as part of the OCHEM project.[7] Figure 17 shows an example of two translated SMARTS that stem from the same pattern written in SLN. The two patterns on the left show the compatibility of the pattern nodes in similarity mode with both modes for aromaticity correction switched off. Because of the different notations for the imidazo[1,2-*a*]pyridine matching part of the pattern, most of the pattern nodes are classified as dissimilar. Since mismatches of nodes have an influence on the compatibility

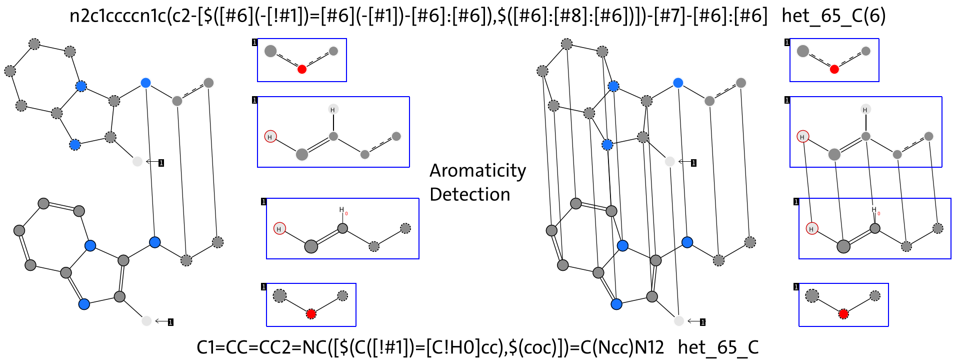

n2c1ccccn1c(c2-[$([#6](-[!#1])=[#6](-[#1])-[#6]:[#6]),$([#6]:[#8]:[#6])])-[#7]-[#6]:[#6] het_65_C(6)
Aromaticity Detection
C1=CC=CC2=NC([$(C([!#1])=[C!H0]cc),$(coc)])=C(Ncc)N12 het_65_C

Figure 17. Comparison of two SMARTS patterns that stem from the same pattern written in SLN and were translated by two different sources.[7,31] The pattern was originally published as SLN N[1]C(=C(NC:C)N[6]C=CC=CC=@1@6)Any[IS=C(Hev)=CHC:C,C:O:C].[8] The two patterns on the left show the SMARTS comparison with both modes for aromaticity correction switched off. The comparison on the right, on the other hand, was conducted with both modes switched on. The patterns at the top of the graphic were taken from the set that was automatically translated by CACTVS[31] and is also stored in ChEMBL. The patterns at the bottom of the figure are the ones taken from OCHEM.[7]

of neighboring nodes, the similarity of other nodes in the patterns decreases as well. When making use of the aromaticity correction modes, this incompatibility is overcome by amending node and bond fingerprints internally.

The main two differences we found during analysis of the two sets of SMARTS patterns were different notations for aromaticity and the use of SMARTS language features. While the ChEMBL patterns describe the aromaticity more explicitly with aromatic nodes and edges, the OCHEM patterns use the implicit form of aromaticity. Nodes are written in their aromatic form, and bonds are given implicitly. In terms of use of the features the SMARTS language provides, the OCHEM patterns utilize them way more. Specifications of hydrogens are given as explicit hydrogens in the ChEMBL patterns, while the OCHEM patterns make use of hydrogen counts and declaration of explicit bond counts. A third major difference we found is the use of generic element notation, #<*n*>. CACTVS makes heavy use of this notation, while the handcrafted OCHEM SMARTS contain this kind of description only when no aliphatic/aromatic classification is possible. Other minor differences exist but add no information necessary for understanding the conceptual differences.

Aromaticity handling is a complex problem with many pitfalls. On the side of SMARTS matching, many different implementations of the original Daylight theory exist. Each implementation has it is own aromaticity handling that depends on the chemical model used. To be independent of the notation that is used to describe potentially aromatic systems in SMARTS, the two described methods were introduced. The presented modes cover the theoretically correct way of handling aromaticity as well as the chemically intuitive one. Nonetheless, we recommend being very careful when using these modes, as there might be unexpected side effects.

■ CONCLUSION

We have presented the application of SMARTScompare, which allows analysis of chemical patterns in the form of SMARTS expressions in an unprecedented way. SMARTScompare makes it possible to compare SMARTS patterns independent of their string representation and supports most features of the SMARTS language. It is capable of similarity assessments and can detect subset relations between patterns. Beyond SMARTScompare, we developed the SMARTScompare-Viewer, a tool that provides intuitive depictions of SMARTScompare pattern mappings.

As a first application study, we used our newly developed tools to analyze several published filter sets. For most of the filter sets, publications are available that summarize the purpose for which they were designed. For those filter sets without a publication, we at least had the labels that were given to each pattern contained in the filter sets. This allowed us to analyze the similarities of filter sets against the background of their original purpose. The data suggest that the Glaxo filter set is fully contained within the MLSMR filter set, with only nine patterns that were adjusted. We also can easily see that the SMARTCyp set has a totally different focus compared with the rest of the filter sets. The PAINS filter set consists of many large patterns that seem like they were directly translated from SMILES strings for specific molecules. As a consequence, filters from PAINS are very dissimilar to filters from other sets used in HTS screening. Last but not least, we are able to see differences in design intentions ingrained in the similarity values of the pattern set comparison. Our analysis suggests that the main difference between the Glaxo and SureChEMBL OCHEM/ToxAlerts filter sets is the focus on chemical instability in the first case and the focus on toxicity in the second case.

From an analysis of the factors that have the greatest influence on pattern similarity, pattern size (the number of nodes), the number of atomtypes an element has, and the weighting of bits of node fingerprints emerge as the strongest factors affecting similarity.

Aromaticity poses a problem, as no standard aromaticity detection is included in the SMARTS language. We introduced two optional modes that can be engaged if needed. If in use, aromaticity is handled in a way that makes the comparison of two patterns more independent of their notation.

For cheminformaticians interested in designing filter sets, SMARTScompare offers an easy approach to learn from filter sets and combine rules with own experiences. For chemists interested in understanding why compounds are filtered out, SMARTScompare offers many new opportunities for analysis. With a pattern of interest from one filter set, similar patterns from other sets become searchable, making direct comparisons possible for the first time.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00249.

> Additional chapters providing more detailed information about the similarity validation (AdditionalApplicationChapters.pdf), list of molecules (with ZINC IDs) violating hydrogen properties as specified by our NAOMI chemistry model (UnmodeledHydrogenCounts.smi); list of molecules (with ZINC IDs) exceeding the enumeration limit for the minimum ring size (RingSizeEnumerationExceedance.smi); list of molecules (with ZINC IDs) exceeding the enumeration limit for the number of unique ring families of which an atom is a part (UrfCountEnumerationExceedance.smi); SmartsView depictions of our quinone query patterns and all matched patterns from the filter sets, with ChEMBL pattern ID, filter set, and annotation for each pattern (QuinonePatterns.pdf); seven patterns from the PAINS[2,8] filter set that we modified to speed up the calculations (ModifiedPAINSPatterns.csv) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de. Phone: +49 (40) 428387351.

ORCID

Emanuel S. R. Ehmki: 0000-0003-0457-4856
Robert Schmidt: 0000-0002-7089-4842
Matthias Rarey: 0000-0002-9553-6531

Present Address

[†]F.O.: Evotec AG - Manfred Eigen Campus, Essener Bogen 7, 22419 Hamburg, Germany.

Author Contributions

F.O. and M.R. initially tested the algorithm presented in the companion paper[19] on the pattern sets that are available in ChEMBL. R.S. implemented the SMARTScompare software based on the NAOMI cheminformatics library (current version). E.S.R.E. provided application scenarios for the SMARTScompare approach and analyzed and interpreted the results. E.S.R.E., R.S., and M.R. wrote the manuscript.

Notes

The authors declare the following competing financial interest(s): M.R. declares a potential financial interest in the event that the SMARTScompare software is licensed for a fee to non-academic institutions in the future.

SMARTScompare and SMARTScompareViewer, a tool for visualization of SMARTS relationships, are available for Linux, OS X, and Windows as part of the NAOMI ChemBio Suite (https://uhh.de/naomi) and for free for academic use and evaluation purposes. Furthermore, SMARTScompare and the SMARTScompareViewer are integrated in the SMARTSview serverhttps://smarts.plus. All feedback is greatly appreciated and supports the further development of SMARTScompare.

■ REFERENCES

(1) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(2) Gaulton, A.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(3) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(4) Daylight Chemical Information Systems, Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Aug 20, 2018).

(5) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(6) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Model.* **1999**, *39*, 897−902.

(7) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.

(8) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(9) Bruns, R. F.; Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **2012**, *55*, 9763−9772.

(10) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060−1068.

(11) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(12) Blake, J. F. Identification and evaluation of molecular properties related to preclinical optimization and clinical fate. *Med. Chem. (Sharjah, United Arab Emirates)* **2005**, *1*, 649−655.

(13) Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem* **2008**, *3*, 435−444.

(14) Schomburg, K.; Ehrlich, H. C.; Stierand, K.; Rarey, M. From structure diagrams to visual chemical patterns. *J. Chem. Inf. Model.* **2010**, *50*, 1529−1535.

(15) Center for Bioinformatics Hamburg. SMARTSviewServer. https://smartsview.zbh.uni-hamburg.de (accessed Oct 11, 2018).

(16) Bietz, S.; Schomburg, K. T.; Hilbig, M.; Rarey, M. Discriminative Chemical Patterns: Automatic and Interactive Design. *J. Chem. Inf. Model.* **2015**, *55*, 1535−1546.

(17) Schomburg, K. T.; Wetzer, L.; Rarey, M. Interactive design of generic chemical patterns. *Drug Discovery Today* **2013**, *18*, 651−658.

(18) Landrum, G. RDKit: Open Source Cheminformatics. https://www.rdkit.org/ (accessed Sept 26, 2018)

(19) Schmidt, R.; Ehmki, E. S. R.; Ohm, F.; Ehrlich, H.-C.; Mashychev, A.; Rarey, M. Comparing Molecular Patterns using the Example of SMARTS: Theory and Algorithms. *J. Chem. Inf. Model.* **2019**, DOI: 10.1021/acs.jcim.9b00250.

(20) Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay interference CompoundS. *J. Chem. Inf. Model.* **2017**, *57*, 417−427.

(21) Urbaczek, S.; Kolodzik, A.; Rarey, M. The valence state combination model: A generic framework for handling tautomers and protonation states. *J. Chem. Inf. Model.* **2014**, *54*, 756−766.

(22) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Model.* **1976**, *16*, 40−43.

(23) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique ring families: A chemically meaningful description of molecular ring topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013−2021.

(24) Flachsenberg, F.; Andresen, N.; Rarey, M. RingDecomposerLib: An Open-Source Implementation of Unique Ring Families and Other Cycle Bases. *J. Chem. Inf. Model.* **2017**, *57*, 122−126.

(25) ZINC15. https://zinc15.docking.org (accessed Aug 20, 2018).

(26) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(27) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(29) Ehrlich, H. C.; Rarey, M. Systematic benchmark of substructure search in molecular graphs—From Ullmann to VF2. *J. Cheminf.* **2012**, *4*, 13.

(30) ChEMBL22. http://chembl.blogspot.com/2016/09/chembl-22-released.html (accessed Dec 5, 2018).

(31) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Model.* **1994**, *34*, 109−116.

(32) Sushko, I.; et al. Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.

(33) Evotec AG. https://www.evotec.com/en (accessed Oct 1, 2018).

(34) The ChEMBL Team (ChEMBL Helpdesk). Personal communication, 2018.

(35) National Institutes of Health. NIH Grand Issue Note MLSMR. https://grants.nih.gov/grants/guide/notice-files/NOT-RM-07-005.html (accessed Oct 1, 2018).

(36) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265−4272.

(37) Inpharmatica Inc. http://www.inpharmatica.co.uk/ (accessed Oct 1, 2017).

(38) Sushko, I.; Tetko, I. V.; Salmina, E.; Potemkin, V. A.; Poda, G. OCHEM/ToxAlerts Webserver. https://ochem.eu//alerts/show.do?render-mode=full (accessed April 29, 2019).

(39) SureChEMBL ToxAlerts WebPage. https://www.surechembl.org/knowledgebase/169485-non-medchem-friendly-smarts (accessed Feb 18, 2019).

(40) Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J. K.; Gopalakrishnan, J.; Tetko, I. V.; Gul, S.; Hadian, K. Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens. *J. Biomol. Screening* **2014**, *19*, 715−726.