Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling

Chihae Yang,*[†,‡,∥] Aleksey Tarkhov,[†] Jörg Marusczyk,[†] Bruno Bienfait,[†] Johann Gasteiger,[†] Thomas Kleinoeder,[†] Tomasz Magdziarz,[†] Oliver Sacher,[†] Christof H. Schwab,[†] Johannes Schwoebel,[†] Lothar Terfloth,[†] Kirk Arvidson,[∥] Ann Richard,[⊥] Andrew Worth,[#] and James Rathman[‡,§]

[†]Molecular Networks GmbH, 91052 Erlangen, Germany

[‡]Altamira LLC, Columbus, Ohio 43235, United States

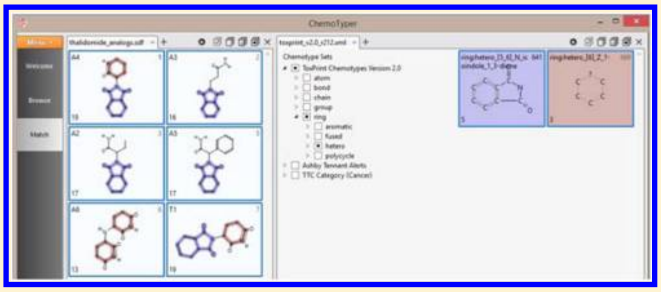
[§]Department of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, Ohio 43210, United States

[∥]US Food and Drug Administration Center for Food Safety and Applied Nutrition, Office of Food Additive Safety (FDA CFSAN OFAS), College Park, Maryland 20740, United States

[⊥]National Center for Computational Toxicology, US Environmental Protection Agency (EPA), Research Triangle Park, North Carolina 27711, United States

[#]EC Joint Research Centre (JRC), I-21027 Ispra, Italy

Ⓢ *Supporting Information*



ABSTRACT: Chemotypes are a new approach for representing molecules, chemical substructures and patterns, reaction rules, and reactions. Chemotypes are capable of integrating types of information beyond what is possible using current representation methods (e.g., SMARTS patterns) or reaction transformations (e.g., SMIRKS, reaction SMILES). Chemotypes are expressed in the XML-based Chemical Subgraphs and Reactions Markup Language (CSRML), and can be encoded not only with connectivity and topology but also with properties of atoms, bonds, electronic systems, or molecules. CSRML has been developed in parallel with a public set of chemotypes, i.e., the ToxPrint chemotypes, which are designed to provide excellent coverage of environmental, regulatory, and commercial-use chemical space, as well as to represent chemical patterns and properties especially relevant to various toxicity concerns. A software application, ChemoTyper has also been developed and made publicly available in order to enable chemotype searching and fingerprinting against a target structure set. The public ChemoTyper houses the ToxPrint chemotype CSRML dictionary, as well as reference implementation so that the query specifications may be adopted by other chemical structure knowledge systems. The full specifications of the XML-based CSRML standard used to express chemotypes are publicly available to facilitate and encourage the exchange of structural knowledge.

■ INTRODUCTION

Precise and meaningful representations of chemical entities are essential to mine chemical and biological information from databases designed to support development of structure–activity inferences and predictive models. Ideally, such chemical representations should enable queries of structure and reactivity moieties that (1) provide broad coverage of the chemical information contained within the database and (2) identify representations that capture salient and influential chemical characteristics, including those associated with biological activity. Existing methods for constructing chemical queries are generally limited in terms of their public accessibility, transparency, lack of standardization, heavy reliance on simple enumeration of chemical features, and inability to carry chemical property information and filters derived from the whole molecule. The ability to construct more complex query representations would enable the incorporation of prior knowledge and structure–activity relationship (SAR) inferences into a flexible data-mining workflow. This, in turn, could empower and more effectively focus user investigations into potentially productive areas of chemical-activity space for model development. The considerable structural and mechanistic (chemical and biological) diversity of public chemical toxicity databases, in particular, pose severe challenges to conventional SAR modeling and data mining approaches, in part due to the limited ability of current chemical

Received: November 4, 2014
Published: February 3, 2015

representations to characterize this diversity in a functionally useful way.

We introduce here the concept of a "chemotype", a new way of representing chemical entities, with the following objectives: publicly accessible; coded in a unique and reproducible manner; and capable of combining both connected and nonconnected chemical patterns as well as atom, bond and molecular-based properties into a single query. To implement chemotypes, we have developed a suite of new publicly available tools and resources: (1) CSRML (Chemical Subgraphs and Reactions Markup Language)—a fully documented, open source, XML-based chemical query language and exchange standard for representing chemotypes; (2) ToxPrint chemotypes—a predefined library of over 700 chemotypes specifically designed to provide excellent generic coverage of public chemical and toxicity databases as well as to capture existing knowledge of chemical frameworks, reactive centers, functional groups, and structural moieties that have informed toxicity prediction models and safety assessment workflows throughout government and industry; (3) ChemoTyper—a software application developed under contract with the U.S. Food and Drug Administration (FDA) to provide an easy-to-use interface for viewing, searching, and filtering structures from an imported chemical structure-data (SD) file using a predefined set of chemotypes (defined in an imported CSRML file of query features, such as the ToxPrint chemotypes) as well as to produce a binary chemical fingerprint output for use in cheminformatics and modeling; and (4) a KNIME node for fingerprinting a structure data set against a set of chemotypes (e.g., ToxPrint) for use in building computational workflows. Together, these public tools and resources enable broad application of chemotypes to data mining and modeling of chemical-biological activity data sets. The ToxPrint chemotypes, in particular, provide a standardized set of chemical query building blocks for use in data mining and filtering that are specifically informed by and designed to facilitate SAR modeling of toxicity, thereby providing a basis for information exchange between projects. Although their development was motivated largely by the need to more effectively address the chemical-toxicity prediction problem, each of these tools and resources are fully extensible and could be tailored and applied to a wide variety of chemical-activity prediction problems.

Motivation for a New Query Language. Current de facto standards for defining chemical queries include specifications such as SMARTS[1] and SMIRKS,[2] which are used to find matching patterns and reaction transformations, respectively. While constructing advanced SMARTS and SMIRKS queries requires considerable expertise, graphic pattern editors can simplify this task.[3] Other available chemical pattern query methods include SYBYL Line Notation (SLN)[4,5] and MOL/SD file extended with query features for patterns[6−8] and reactions,[2] and ChemAxon's Chemical Terms Language, CTL.[9] These are much less widely used than SMARTS since they typically support proprietary chemoinformatics platforms or knowledge-bases. Simple structural fragments representing topological connectivity coded as SMARTS or a modified SD file have been used to represent molecular features associated with biological activities and have been employed as molecular descriptors in various types of structure classification or quantitative SAR (QSAR) models.[10−13] Fragments highly associated with toxicity end points are typically identified as structural alerts and are most often coded in proprietary formats in commercial predictive systems[8,14,15]

Each of the above chemical query approaches has fundamental drawbacks limiting their use in data mining and modeling. One of the most serious deficiencies is the lack of a unique, reproducible standard representation for a single chemical structural moiety or condition. For example, even a simple functional group such as phosphoric acid can be represented in many different ways in SMARTS; Figure 1 enumerates all possible SMARTS representations of the phosphoric acid group.[16]

```
[$(P(=[0X1])([$([0X2H]),$([OX1-]),$([OX2]P)])([$([OX2H]),$([OXl-]),
$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)]),$([P+]([OX1-])([$([OX2H]),$([OX1-]),
$([OX2]P)])([$([OX2H]),$([OX1-]),$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)])]
```

Figure 1. Equivalent SMARTS expressions for the phosphoric acid group.

While the accuracy and efficiency of structure (molecule) searches can be greatly improved by internal rules enforcing a specific "canonical" representation of structures within applications, canonicalization does not provide the same advantage for substructure or pattern searches. Regardless of whether or not molecules are canonicalized, it may be necessary to enumerate SMARTS patterns for a search. Further, if a structure contains several motifs, each requiring an enumeration of possible representations, the complexity of a query expression (e.g., SMARTS patterns) increases significantly.

Another problem with existing query languages is the need to enumerate many different substitution patterns to search for a desired chemical condition. For example, to detect a substituted benzene ring with the potential to undergo electrophilic substitution at ortho- or para-positions to the substituent, most query languages use enumerations of possible substituents based only on structural features. This enumeration process is not only cumbersome, but also assumes that all possible activating conditions are covered. Ideally, the query condition would use computed electronic properties to directly capture the activating effect of a substituent on the ring at the various positions, thus eliminating the need for such enumerations. While conventional queries such as "ortho- or para-substituted ring" are easily implemented in CSRML, the key point is that CSRML enables the definition of chemotypes based on properties, making it possible to create queries not expressible in SMARTS. CSRML therefore allows us to reimagine how chemical patterns are defined and searched.

There have been previous attempts to address the challenges of chemically complex representations. For example, SYBYL Line Notation (SLN)[3,4] employs an annotation mechanism for managing chemical and/or physical properties of structural elements. Although it is compact and concise, this method still requires enumeration of all possible resonance structures. The Molecular Query Language (MQL) provides a feature to annotate user-defined properties;[17] however, it too requires separate enumeration of resonance structures. Attempts to incorporate physicochemical properties to various dialects of SMILES or SMARTS queries have been also reported.[18,19] Another approach to enrich the query specification with physicochemical properties is illustrated by ChemAxon's CTL,[9] where a two-step process is employed to query with features of extensible chemical terms followed by filtering based on calculated descriptor values. Although these various extensions and approaches satisfy their intended purposes, they remain proprietary or at least not publicly available, and also incompatible with each other since there have been no large

Table 1. Initial Set of Query Features in CSRML Specification

domain	feature	description	RL (reference library)	ChemoTyper
atom	atom type	queries the element, or a wildcard	+	+
	atom lists	checks if an atom has the element from the list	+	+
	descriptor value or range	checks the value of a given descriptor on an atom	+	+
	aromaticity/aliphaticity	checks whether an atom is aromatic or aliphatic	+	+
	ring membership/connectivity	checks the ring membership of an atom, counts of rings an atom belongs to, size of rings an atom is member of, or count of adjacent cyclic bonds	+	+
	valency	checks atom valency	+	+
	connectivity	checks atom's connectivity, i.e. count of neighbors	+	+
	formal charge	checks the presence and magnitude of the formal charge on an atom	+	+
	isotope	checks the isotope marks on an atom	+	+
	attached hydrogens	checks the count of attached hydrogen atoms on an atom	+	+
	atom stereo	checks the stereo configuration of an atom in relation to adjacent atoms	+	−
	atom saturation	checks if an atom is adjacent to a double, triple, or aromatic bond	+	+
	attached hetero atoms	checks the count of neighboring hetero atoms attached to an atom, where heteroatom is any atom which is neither carbon nor hydrogen	+	+
	matching query atom	checks if an atom matches another atom already matched in the current substructure query when specifying nonmatching exceptions	+	+
	any combination of above	allows for combining any of the above atom features in an arbitrary logical expression	+	+
bond	bond order	checks the order of a given bond	+	bond types "quadruple" and "coordinate" are parsed but not interpreted
	bond lists	checks whether a bond order is on the list	+	+
	descriptor value or range	checks the value of a given descriptor on a bond	+	+
	aromaticity/aliphaticity	checks whether a bond is aromatic or aliphatic	+	+
	ring membership	checks the ring membership of a bond, the size of rings a bond is member of, etc.	+	+
	any combination of above	allows for combining any of the above bond features in an arbitrary logical expression	+	+
electron system	type of the system	checks the type of an electron system	only π-electron systems	only π-electron systems
	descriptor value or range	checks the value of a given descriptor on an electron system	+	+
	count of π-electrons	checks the count of π-electrons in an electron system	+	+
	any combination of above	allows for combining any of the above electron system features in an arbitrary logical expression	+	+
molecule	descriptor value or range	checks the value of a given descriptor on a molecule or matched fragment	+	+
	element match	checks whether the matched substructure fragment contains a given element or not	+	+
	element count	checks if the matched substructure fragment contains a given element in a given quantity	+	+
	unsaturated bond count	checks if the matched substructure fragment contains a given count of unsaturated bonds	+	+
	connected or disconnected	checks if the fragments of the query are connected to each other or not	+	[a]
	any combination of above	allows for combining any of the above molecule features in an arbitrary logical expression	+	+

[a] The current ChemoTyper v1.0 does not yet support this feature fully. A CSRML query, however, may be specified as a set of disconnected fragments. It can then be matched by ChemoTyper both to a single connected molecule, or to several disconnected molecules provided within a molecular ensemble (in a single CTAB instance, for example).

and coordinated efforts to standardize query languages. The need is clear for an open and freely available query standard, allowing for the description of any arbitrary fragment combined with computed properties.

Definition of Chemotypes. To address the issues described above, a method to support a new chemical query representation, termed "chemotype", has been developed. A chemotype is defined as a structural fragment encoded for connectivity (which may extend beyond a single connected fragment) and also, when desirable, for physicochemical properties of atoms, bonds, fragments, electron systems, and even a whole molecule. This new approach for chemical representation is supported by a new open-source XML-based query language, Chemical Subgraphs and Reactions Markup Language (CSRML), which enables (1) the representation of different resonance structures; (2) the estimation of phys-

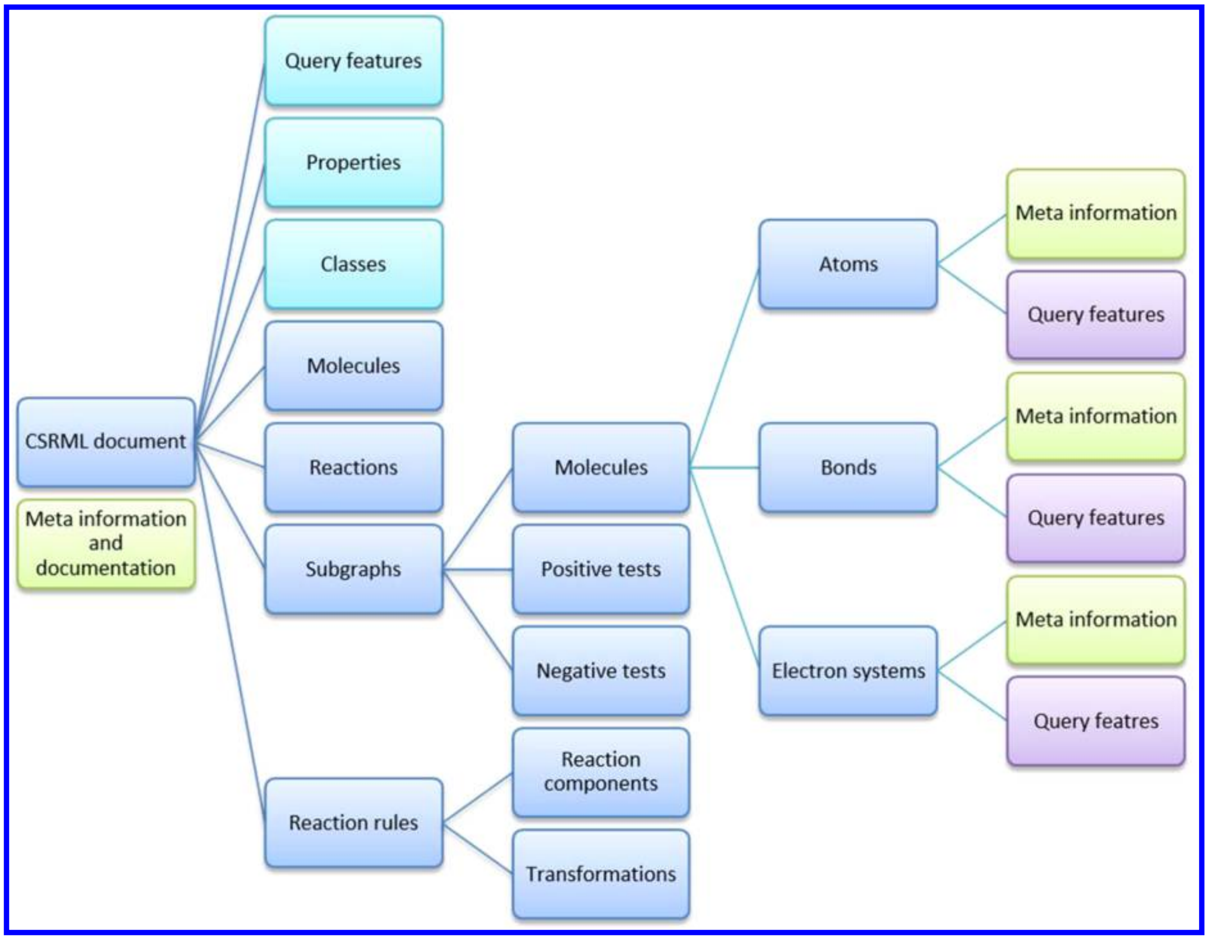


Figure 2. Object model of CSRML.

icochemical properties of atoms, bonds, or structure fragments to avoid enumerations of possible substitution patterns; and (3) standardized exchanges of chemical queries. Chemotypes can be used to define almost all chemical queries possible with SMARTS and SMIRKS, more formally and concisely address the issue of resonance structures, and make possible queries that are either not possible or difficult to define with SMARTS or SMIRKS or other existing methods. Many of the new query features and the language itself were refined in parallel with the development of the public ToxPrint library of chemotypes (precoded queries) intended for use in association with toxicity data mining and prediction modeling efforts. This paper describes the development and specifications of the new query language CSRML and the ToxPrint chemotype library coded in CSRML. Further introduced are the publicly free software applications, ChemoTyper with its reference implementation, and a CORINA Descriptors Community Edition KNIME node, both of which are designed to enable the use of chemotypes for compound profiling, structural knowledge development, and molecular descriptor generation for QSAR applications.[20−22]

■ CSRML AND CHEMOTYPER SOFTWARE

Although the concept of a chemotype is not tied to any particular query language or specification scheme, expressing them in an XML-based language enables a hierarchical, transparent construction, as well as easy translation into a format appropriate for computational work. Additionally, by virtue of their unique and open construction, the logic and accuracy of queries become subject to rigorous validation and testing. Hence, the CSRML language was designed to meet the following criteria:

- capable of encoding molecules, chemical substructures and patterns, reaction queries, structural alerts, and transformation rules in a way that includes not only topology but also, if desired, properties of atoms, bonds, electronic systems, or molecules
- allows for annotation of chemical entities with various textual and/or meta information, enabling extended documentation
- serves as a data exchange format for chemical structures, substructures and patterns, reaction queries, and reactions
- employs XML for modularity and ease of implementation by computer programs and can be used with XML-enabled or native XML databases.

The initial grammar for CSRML was inspired by the Chemical Markup Language (CML).[23] However, the CSRML grammar has been extended to handle subgraph queries, molecules, reactions, and transformation rules to an extent sufficient for performing queries and storing the necessary data. Consequently, the CSRML grammar is more compact and strictly defined than CML, making it easier to control the correctness of CSRML documents even at the stage of a simple, nonchemically aware XML parsing process. To enable the use of chemotypes coded in CSRML, a software application tool, ChemoTyper, was developed for public use under a contract from U.S. FDA's Center for Food Safety and Nutrition

Table 2. Functionality of the Reference Library

feature	description
input	Read and parse a CSRML document validating them against the CSRML schema definition. Invalid documents are not read and an error message indicating the cause of the error is provided.
output	Parsed CSRML object can be output into, e.g, a new file.
validation	The CSRML objects are checked against various rules defined by the CSRML XML schema (XSD), such as (1) syntax checking and XSD conformance, (2) uniqueness of objects' identifiers in their defined scopes, and (3) referential integrity of references to atoms, bonds, molecules, etc.
canonicalization	The CSRML definition of molecules and subgraphs are canonicalized when parsed.
uniqueness check	The canonicalized CSRML objects are used to produce their canonical representation which is employed to check the uniqueness of the queries defined in a document.
test application	A test application is included with the RL that (1) reads a provided set of the query features, (2) reads given example CSRML documents, (3) parses CSRML documents and checks them against the CSRML XSD, (4) canonicalizes the CSRML subgraphs, (5) generates their canonical representations, and (6) prints out the parsed document assembled back anew from the parsed CSRML content.

(CFSAN). ChemoTyper was designed to provide users access to an intuitive and easy-to-use platform for viewing query features (i.e., chemotypes coded in CSRML), performing a wide range of chemotype Boolean searches, and browsing results. Following FDA's specification, a public reference implementation is available to encourage adaptation of the query language in other chemistry development kits. Also publicly available is the CORINA Descriptors Community Edition KNIME node for searching structures and generating fingerprints from a predefined CSRML document. These features enable different user communities to exchange their subgraph queries or alerts in a straightforward manner.

The ToxPrint chemotypes are a set of chemical features and rules derived from various toxicity prediction models and safety assessment guidelines within FDA and other federal agencies and industries. In addition, the need to provide sufficient coverage of public databases pertaining to toxicity interests and concerns to enable robust chemical mining capabilities was a major driving force for the development of the ToxPrint chemotype library. The ToxPrint chemotypes and the need to extend current query technologies, in turn, became a prime motivator for development of CSRML. Hence, both CSRML and the ChemoTyper tools were developed in parallel with the development of the ToxPrint chemotypes, yet each component represents a significant independent scientific contribution.

■ MATERIALS AND METHODS

Object Model of Chemical Subgraphs and Reactions Markup Language (CSRML). The CSRML data model was developed to enable the query requirements of chemotypes. Table 1 summarizes the initial set of query features with CSRML specifications. The object model adopted by CSRML is illustrated in Figure 2. The basic atom and bond blocks similar to the CML architecture were preserved. Matching and testing elements (not shown in Figure 2) are included to assist in the design, development and validation of new chemotypes. In addition to conventional representations based on VSEPR (valence shell electron-pair repulsion) theory,[24] CSRML also provides a novel approach for querying electron systems for cases in which the actual order of the underlying bonds is irrelevant and what matters is the type of the bonding electron system (e.g., σ or π) and the number of electrons. Moreover, at each level of the model there are reserved informational elements that allow for detailed documentation of each object. The documentation of the CSRML grammar by an XML schema definition is available.[25] The schema allows automated class generation and data binding by means of JAXB[26] or similar technologies.

ChemoTyper and Reference Implementation. Chemotypes were developed in three phases. First, a specification of the CSRML language and its query features was created to support the creation of the ToxPrint library. Based on existing substructures and patterns, the set of initial query features was compiled. The CSRML specification, however, foresees a standardized method of adding, interpreting, and transferring query features, allowing for practically unlimited extensibility. The CSRML language was then extended to describe reaction queries. Functionalities of the reference library are listed in Table 2; the canonicalization algorithm used here is that of Weininger et al.[27] A comprehensive set of documentation including the CSRML language specification, initial compilation of supported query features, and an extended user manual has been compiled and made publicly available.[25]

In the second phase, a reference library (RL) supporting input, output, parsing, and XML validation of CSRML documents was designed and implemented in a standard C++ language.[28] The only external dependence employed by the RL is the XML parser Xerces-C[29] which is available on the vast majority of platforms on an open-source basis under Apache Software Foundation license.[30]

Finally, to illustrate the use and encourage end users to adopt the new standard, a reference implementation of CSRML is provided as a module within the MOSES cheminformatics library.[21] ChemoTyper, a freeware application, was developed to enable broader user access to ToxPrint chemotypes.[25] This software application implements most of the chemotype query features listed in Table 1. ChemoTyper can be used to browse chemical structure data sets (i.e., an imported SD file), perform various Boolean operations of subgraph (chemotype) matching, visualize chemotype matches, generate fingerprints, and visualize ToxPrint chemotypes or any other set of CSRML-defined chemotypes against any imported SD file. A detailed list of the functionalities available in ChemoTyper is provided in Table 3.

ToxPrint, a Public Library of Chemotypes. *Chemical Domain.* A library of chemotypes was developed from a chemistry domain defined by large and diverse sources, including public inventories and databases spanning environmental, commercial, and regulated chemicals. The sources related to toxicity data and risk assessment information include US FDA Drugs@FDA,[31] US FDA PAFA,[32] National Toxicology Program,[33] National Library of Medicine ToxNet−CCRIS,[34] ToxNet−IRIS,[35] ToxNet−GeneTox,[36] ToxNet−DART,[37] TERIS,[38] US EPA ECOTOX,[39] US FDA EDKB,[40] Carcinogenicity Potential Database,[41] US EPA's DSS Tox,[42] AcTOR[43] and ToxRefDB,[44] ISS CAN,[45] EU REACH Substance Registration Database,[46] and EU Scientific

Table 3. Functionality of ChemoTyper

feature	description
input	reading CSRML files containing chemotype definitions
	reading various chemical files with molecules to test
substructure search	applying chemotype rules to the molecules
	detecting matches
	counting matches
visualization	visual representation of the matches detected: (1) highlighting the matches over the molecules; (2) indicating the counts of matches for selected chemotypes; (3) indicating the counts of matching molecules for selected chemotypes; (4) zoom
data processing	filtering/searching the chemotypes and molecules by names/identifiers
	combining several chemotypes using logical operators AND, OR, NOT AND, or NOT OR to display only molecules that have all, any, or none of selected chemotypes
	combining several molecules using logical operators AND, OR, NOT AND, or NOT OR to display applicable chemotypes that are found in all, any, or none of selected molecules
	fingerprint generation
output	output of the fingerprints
	export of selected chemotypes into a separate CSRML document
	output of selected molecules into a separate molecule file (in original chemical format such as SDF)

Committee of Consumer Safety.[47] Chemical inventories include US EPA TSCA Chemical Substance Inventory,[48] US EPA Pesticide Inert list,[49] Pesticide PAN,[50] Tox 21 inventory,[51] Canadian Domestic Substance List,[52] and the EU COSING database.[53] Chemical structures were obtained from ChemID Plus,[54] ChemSpider,[55] DSSTox,[42] and US FDA CFSAN CERES.[56] Over 100 000 structures from these inventories were used as an underlying structure domain spanning pharmaceuticals, agrochemicals, food ingredients and additives, cosmetics ingredients, and consumer and industrial chemicals.

Development of the Default Library of Chemotypes. Chemical structures were first roughly grouped by an initial set of simple fragments, publicly available and commonly used for classifications. These fragments include simple organic functional groups,[57,58] chains and ring systems,[59] and metal/organometals.[60] The sources of these initial fragments included the ChEBI ontology classification,[61] PubChem Ontology58, structural categories defined by CFSAN PAFA database[32] reflecting the legacy FDA Redbook guidelines used in safety assessment, EPA mode of action classes,[62] and chemical groups deemed reactive in drug lead structure screening.[63,64] These initial fragments were used as seeds to design more detailed chemical patterns by the following process. They were first coded into SMARTS and used for pattern searching against a target structure set;[65] the structures in each subgroup were then further clustered, for example by employing MACCS keys in the KNIME CDK node.[22,66] The initial SMARTS were refined to differentiate more specific patterns, including alkane (linear and branched) and alkene (linear and branched) chains, aliphatic rings (alkane and alkenes), aromatic rings, heterocyclic rings, polycyclic rings, fused rings, inorganics, metal complexes, organometallics, and metals. These patterns were systematically extended to reflect desired substitution specifications and then were coded into the CSRML-based proto-chemotypes. To precisely retrieve the desired structure groups after the pattern searching, numerous iterations of design, coding, pattern searching, and analysis were performed. During this design stage, electronic systems were defined and atoms and bond properties were calculated by tools provided by Molecular Networks. These proto-chemotypes were validated against the candidate structure sets by assessing their ability to accurately retrieve the intended target structures. Thus, while the process used here to create chemotype definitions began with SMARTS, many of the final CSRML representations provide feature definitions that cannot be coded in SMARTS.

The chemotype definition process thus constitutes designing and elaborating the substructure patterns, coding them for proto-chemotypes using CSRML, followed by pattern searching against the target data sets to retrieve the structures containing the intended chemotype. CSRML definitions are iteratively refined until the chemotypes retrieve the structures with intended substituent patterns and properties. Table 4 describes the organization and classes of the chemotypes currently coded in CSRML. Also listed in Table 5 are the various wildcards used to express the queries.

Reaction Rules. Complementary to the regular chemical pattern and property query features of chemotypes, CSRML also offers the possibility for encoding reaction rules. The publicly available ToxPrint chemotype set does not include any reaction rules, and so, we present only a brief introduction to this functionality here. A detailed and thorough description, along with extensive examples, will be provided in a separate publication.

The basic idea is to represent the chemical species participating in a reaction as chemotype queries or standard molecules. Each species is assigned a role of either reactant or product, and the reaction rule also defines a set of transformations that convert reactants into products. Thus, the rules can be used to generate reactions for matching substrates (reactants) to predict, for example, the metabolites of a compound. Figure 3 gives an example template of a reaction rule.

The reaction chemotype data model and an example for ester hydrolysis are illustrated in Figure 4.

Table 4. ToxPrint Chemotype Classes

top class	1st level classes	no. of 1st level classes
atoms	main group element, metals (group I, II, III, transition metals, metalloid, poor metals)	7
bonds	C#N, C(~Z)~C~Q, C(=O)N, C(=O)O, C=N, C=O, C=S, CC(=O)C, CN, CNO, COC, COH, CS, CX, metal, N(=O), N[!C], N=[N+]=[N−], N=C=O, N=N, N=O, NC=O, NN, NN=N, NO, OZ, P(=O)N, P=C, P=O, PC, PO, QQ(Q~O_S), quaternary N, quaternary P, quaternary S, S(=O)N, S(=O)O, S(=O)X, S=O, Se~Q, X(any), X[(any)_!C], X~Z	43 (411)[a]
chains	alkaneBranch (7), alkaneCyclic (5), alkaneLinear (10), alkeneBranch (2), alkeneCyclic (3), alkeneLinear (3), alkyne (1), aromaticAlkane (13), aromaticAlkene (4), oxy-alkaneBranch (1), oxy-alkaneLinear (4)	11 (95)[a]
groups	aminoAcid (21), carbohydrate (9), ligand (4), nucleobase (7)	4 (69)[a]
rings	aromatic (3), fused (6), hetero (36), polycycle (3)	4 (144)[a]

[a] The counts in the parentheses represent the total number of chemotypes within each top class.

Table 5. Wildcards Used in CSRML

wildcards	description
A or *	"any" atom wildcard—will match disregarding atom element information; use it to indicate "connected to something" logic; though it is possible to add extra query features to such an atom, it is better to use the QRY atom type in such cases
R	either hydrogen (H) or carbon (C) atom
Q	heteroatom; this wildcard matches anything except carbon, hydrogen, and hydrogen isotopes—deuterium and tritium
X	halogen atom; this wildcard matches all halogens—F, Cl, Br, I, and At
Z	typical organic heteroatom; currently, the following elements match—N, O, S, and P
E	any of the nonmetals—H, C, N, O, F, P, S, Cl, Se, Br, I, At, and noble gases He, Ne, Ar, Kr, Xe, Rn
M	typical metal atom—elements of groups 1 (except H) to 12 and any poor metals: Al, Ga, In, Tl, Sn, Pb, Bi, and Po
G	any metalloids—B, Si, Ge, As, Sb, and Te
QRY	query atom—indicates that the atom is not required to match any explicit element type but must have at least one query feature annotated that is responsible for the match evaluation; the atom element information on a candidate atom is disregarded

Extensibility by Users. Chemotypes allow diverse and flexible queries. The set of features that can be queried by a CSRML-enabled software must allow users to create and use their own custom chemotype query features. CSRML declares all features that might be employed in queries by means of an annotation dictionary, which describes the features that can be queried, stored in a separate XML file, or embedded into the XML document with the queries. Each query feature must be fully defined in this dictionary, specifying which features, and the possible values associated with each, that may be attached to a queried object. An example definition for such a query annotation in the dictionary is given in Figure 5. More details on declarations of the query features are available in the CSRML documentation.[25]

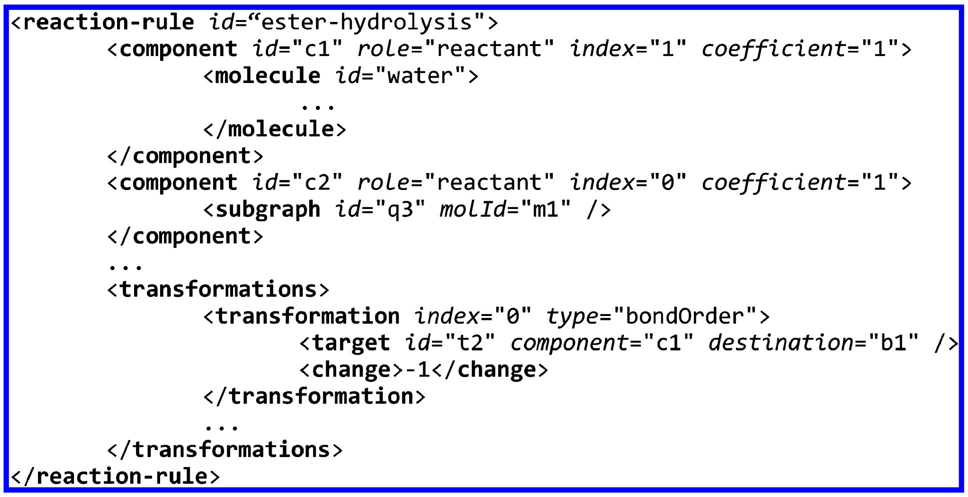
The default set of CSRML query features includes predefined queries common to other substructure/pattern query languages, as presented in Table 1. This set also includes definitions of other specific features, such as substructure exceptions and electron system queries. These "marquee features" of the CSRML language provide means of chemical representation beyond what is attainable in current query specification languages. Each software system that implements CSRML may append its own query features, as needed, to the default dictionary or override the existing ones.

The "severity" attribute of a query feature assigned to a queried object specifies how to handle the situation when the feature cannot be evaluated. Severity in this context refers to the degree to which chemotype searching is restricted when a query is not understood by the implementation because the query feature is not available within the underlying cheminformatics application. CSRML provides a functionality to control what happens in such cases. For example, in a query that specifies criteria for atomic partial charges, the charge specification will be ignored if the severity is set to "skipFeature" and the cheminformatics platform does not support the calculation of partial charges. The query is thus evaluated but the partial charge specification is ignored. If the severity is set to "skipQuery", the query will not be evaluated at all, but all other queries in the document will be processed. This would typically be the default behavior. Finally, if the severity is set to "stop", the processing of entire document is terminated; none of the queries are processed if the underlying chemoinformatics platform does not support all necessary features. A chemotype designer will have the ability to select which of the predefined chemotypes to include in a given implementation and can create new chemotypes or modify existing ones to ensure compatibility with a particular cheminformatics platform. For example, the specific method by which partial charges may be calculated (empirical, semiempirical, ab initio) is not prescribed by the chemotype definition but instead depends entirely on the cheminformatics platform.

■ RESULTS AND DISCUSSION

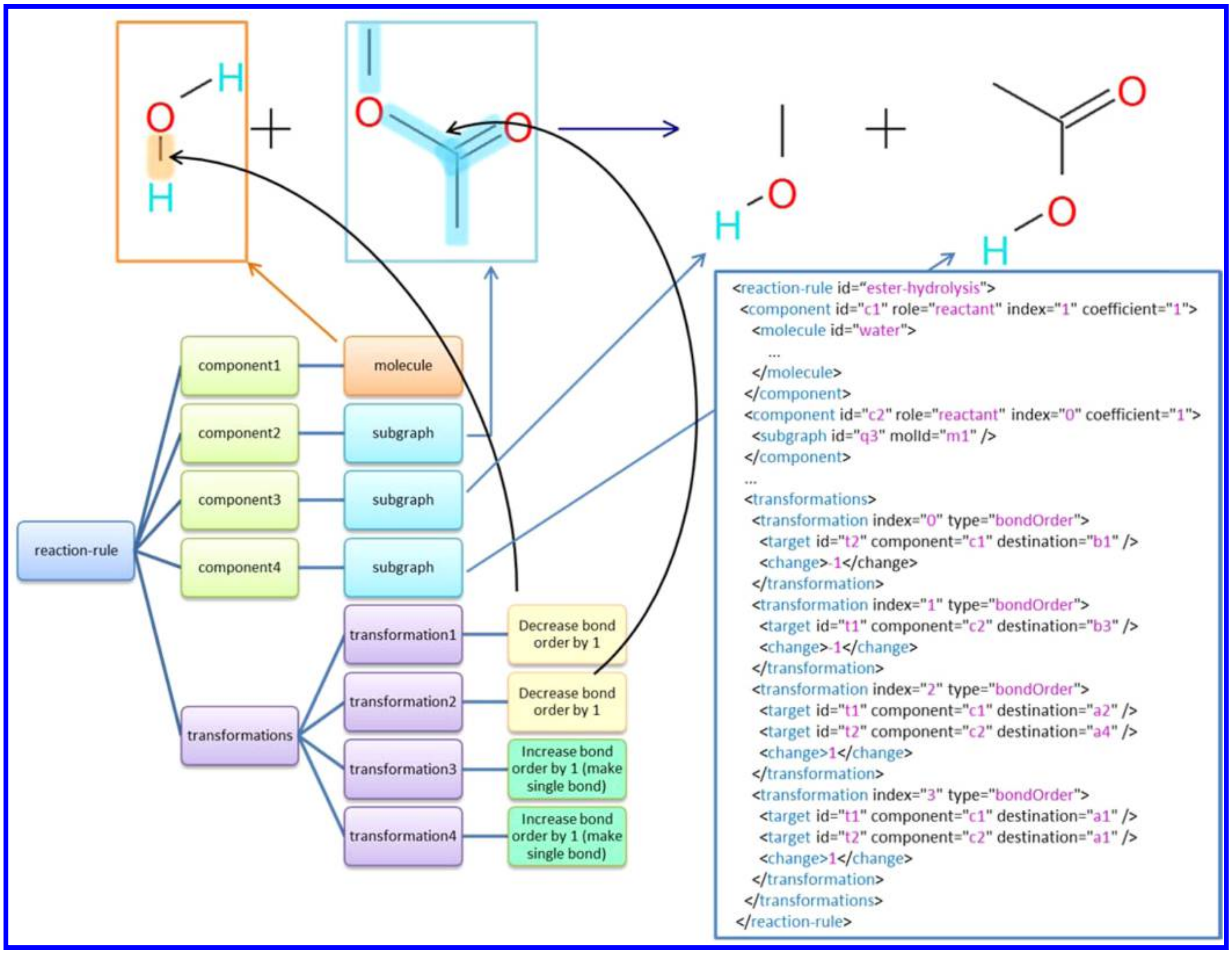
Marquee Query Features of Chemotypes. As described in the Materials and Methods section, chemotypes offer functionalities that extend beyond the existing chemical substructure and pattern query formats, e.g., SMARTS and SMIRKS. The most important marquee features include an electronic system for handling resonance structures and the ability to handle embedded physicochemical properties as described below.

Aromaticity, Electronic Systems, and Resonance Structures. CSRML is designed to be as unambiguous as possible. For example, existing substructure query languages explicitly match single or double bonds to both aliphatic and



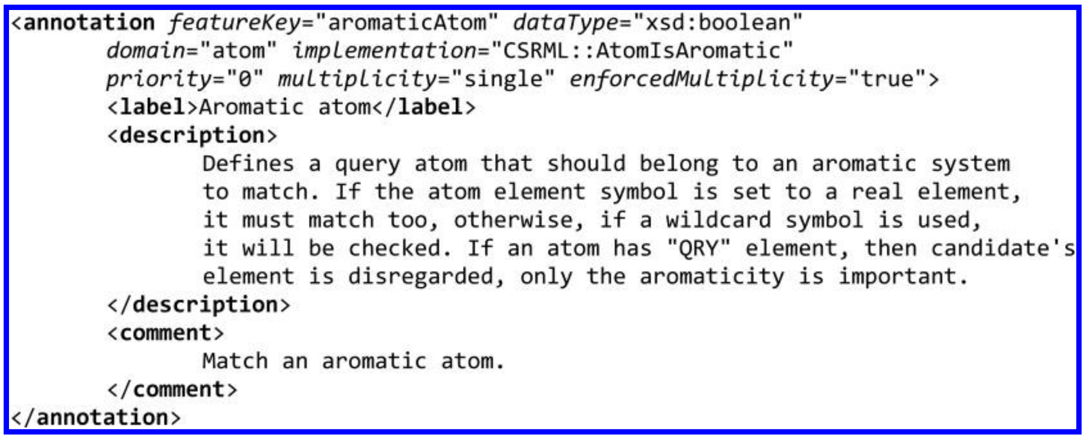
```
<reaction-rule id=“ester-hydrolysis">
       <component id="c1" role="reactant" index="1" coefficient="1">
              <molecule id="water">
                     ...
              </molecule>
       </component>
       <component id="c2" role="reactant" index="0" coefficient="1">
              <subgraph id="q3" molId="m1" />
       </component>
       ...
       <transformations>
              <transformation index="0" type="bondOrder">
                     <target id="t2" component="c1" destination="b1" />
                     <change>-1</change>
              </transformation>
              ...
       </transformations>
</reaction-rule>
```

Figure 3. Example definition of ester hydrolysis reaction rule.



```
<reaction-rule id="ester-hydrolysis">
 <component id="c1" role="reactant" index="1" coefficient="1">
  <molecule id="water">
   ...
  </molecule>
 </component>
 <component id="c2" role="reactant" index="0" coefficient="1">
  <subgraph id="q3" molId="m1" />
 </component>
 ...
 <transformations>
  <transformation index="0" type="bondOrder">
   <target id="t2" component="c1" destination="b1" />
   <change>-1</change>
  </transformation>
  <transformation index="1" type="bondOrder">
   <target id="t1" component="c2" destination="b3" />
   <change>-1</change>
  </transformation>
  <transformation index="2" type="bondOrder">
   <target id="t1" component="c1" destination="a2" />
   <target id="t2" component="c2" destination="a4" />
   <change>1</change>
  </transformation>
  <transformation index="3" type="bondOrder">
   <target id="t1" component="c1" destination="a1" />
   <target id="t2" component="c2" destination="a1" />
   <change>1</change>
  </transformation>
 </transformations>
</reaction-rule>
```

Figure 4. Reaction rules data model and example.



```
<annotation featureKey="aromaticAtom" dataType="xsd:boolean"
        domain="atom" implementation="CSRML::AtomIsAromatic"
        priority="0" multiplicity="single" enforcedMultiplicity="true">
        <label>Aromatic atom</label>
        <description>
                Defines a query atom that should belong to an aromatic system
                to match. If the atom element symbol is set to a real element,
                it must match too, otherwise, if a wildcard symbol is used,
                it will be checked. If an atom has "QRY" element, then candidate's
                element is disregarded, only the aromaticity is important.
        </description>
        <comment>
                Match an aromatic atom.
        </comment>
</annotation>
```

Figure 5. Definition of *aromaticAtom* query feature in CSRML annotations' dictionary.

aromatic bonds due to various possible resonance structures, which may lead to unexpected and undesired substructure query outcomes. In contrast, CSRML, by default, treats single and double bonds as aliphatic, and requires a variety of explicit means to handle aromaticity. In order to find a benzene ring, query bonds, or atoms, or both (atoms and bonds) are clearly and unambiguously specified in CSRML as "aromatic". For more advanced use cases, CSRML also provides a way to combine "aromatic" query features with explicit bond orders by employing the logical disjunction; e.g., a bond can be specified as "double OR aromatic".

For purely aromatic systems, it is sufficient to formulate a CSRML query using the "aromatic atom" or "aromatic bond" features only. Hybrid systems, where some parts of a structure are aromatic and some are not, are typically challenging for any query language. For example, an aromatic ring may reside in a structure that also contains nonaromatic but alternating single−double bond ensembles in a complex, condensed ring system. Consider the structure shown in Figure 6 and an attempt to query the 2H-pyran ring.

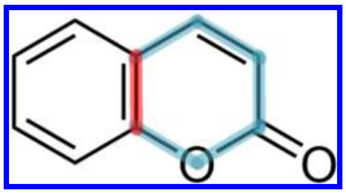
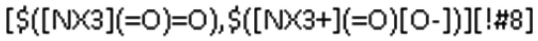


Figure 6. Matching queries in hybrid systems on the example of coumarin and the 2H-pyran ring query.

Double bonds in cyclohexadiene are not aromatic; however, the bond shared by both rings is part of an aromatic system. Thus, the user designing a query needs to decide whether such hybrid species must match the designed query. The use of explicit single and double bonds in CSRML make it possible to limit the query to match only the molecules where none of the bonds in the cyclohexadiene are fused to an aromatic group. Alternatively, the query could be designed to match regardless of aromaticity by the specification "aromatic OR double OR single" flags or by creating an electron system query.

One of the most significant aspects of the CSRML language is its ability to query the electron systems of connected atoms beyond what is possible from the conventional valence bond representation. An electron system is defined as a set of atoms (or even a single atom) and the number of shared electrons. The electron system concept can be applied to encode a radical (one atom, one electron), a carbene (one C atom, two electrons), a π-bond (two atoms, two π-electrons), a benzene ring (six atoms, six π-electrons), etc.[67,68] A chemotype query feature coded in CSRML makes it extremely useful for situations where there are various resonance structures possible. The definition of an electron system query that checks for the presence of a conjugated electron system extended over a given set of atoms with a given count of electrons eliminates the need to enumerate all possibilities. To illustrate the power of electron system queries, consider the simple case of a nitro group, illustrated in Figure 7, where !O denotes any atom other than oxygen. Although there are only two possible resonance structures, encoding these in an SD file or any other approach that depends on numbering the atoms requires separate specification of more than one of the first three structural representations in Figure 7. For example, this query in SMARTS requires two of these structures be specified to ensure that a nitro group is correctly retrieved—see Figure 8.



```
[$([NX3](=O)=O),$([NX3+](=O)[O-])][!#8]
```

Figure 8. SMARTS representation of the nitro group.

In comparison, the chemotype query employs a conjugated electron system "over" the two oxygens and single nitrogen specifying that these atoms must have at least four π-electrons, ignoring the charge distribution. Figure 9 illustrates how the chemotype depicted in Figure 7 is implemented in CSRML.

The electron system defined in Figure 9 ignores the bond orders between specified atoms but instead specifies that they must be part of a conjugated π-electron system with four π-electrons or more. Since the conjugation is only possible if the atoms are in a nitro group, the CSRML query will ensure correct retrieval of varied representations of nitro groups in the target molecules.

CSRML, being XML, is of course much more verbose than SMARTS. However, in contrast to the numerous and only partially compatible versions of SMARTS currently in use, CSRML is a highly structured language that complies with XML standards and that is very easy to read in any XML-aware editor which does simple XML syntax highlighting. A CSRML document can be validated against CSRML XSD scheme to ensure that it is syntactically and grammatically correct; there is no analogous validation method for SMARTS.

Embedded Physicochemical Properties. Currently available substructure query languages are mostly limited to connectivity and topology by predefined atom and bond types based on the VSEPR theory. Situations arise, however, when such predefined features based on "topology and electron-pairs" alone are not sufficient, as in the example given below. The CSRML standard enables the inclusion of any additional atom, bond, electron system, and fragment (including even whole molecule) descriptors that can be calculated by the underlying cheminformatics implementation.

Consider a situation when a substructure search is performed to filter certain reactive compounds matching a given topological pattern. The query performed over a series of compounds will result in both reactive and nonreactive compounds since the difference between them cannot be expressed only by connectivity and topology. Additional estimated factors, such as charge distribution or steric hindrance of putative reactive atoms are needed. In the existing

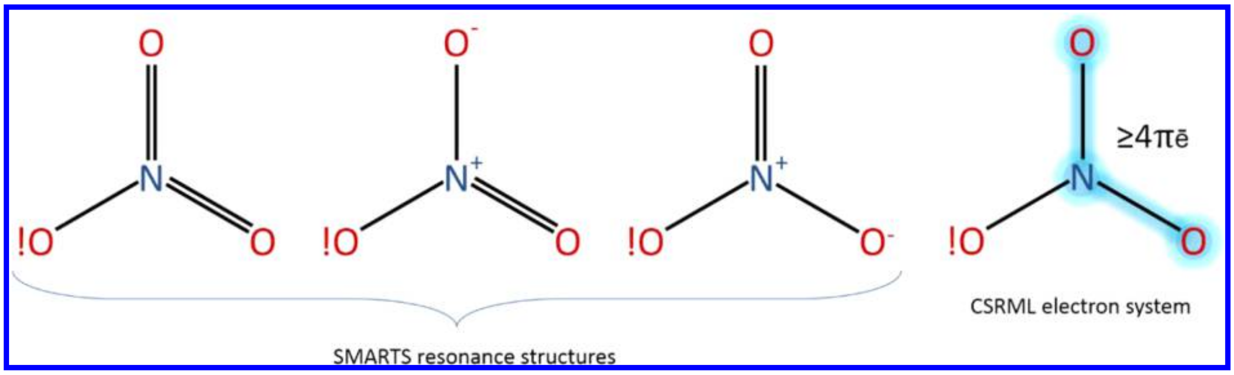
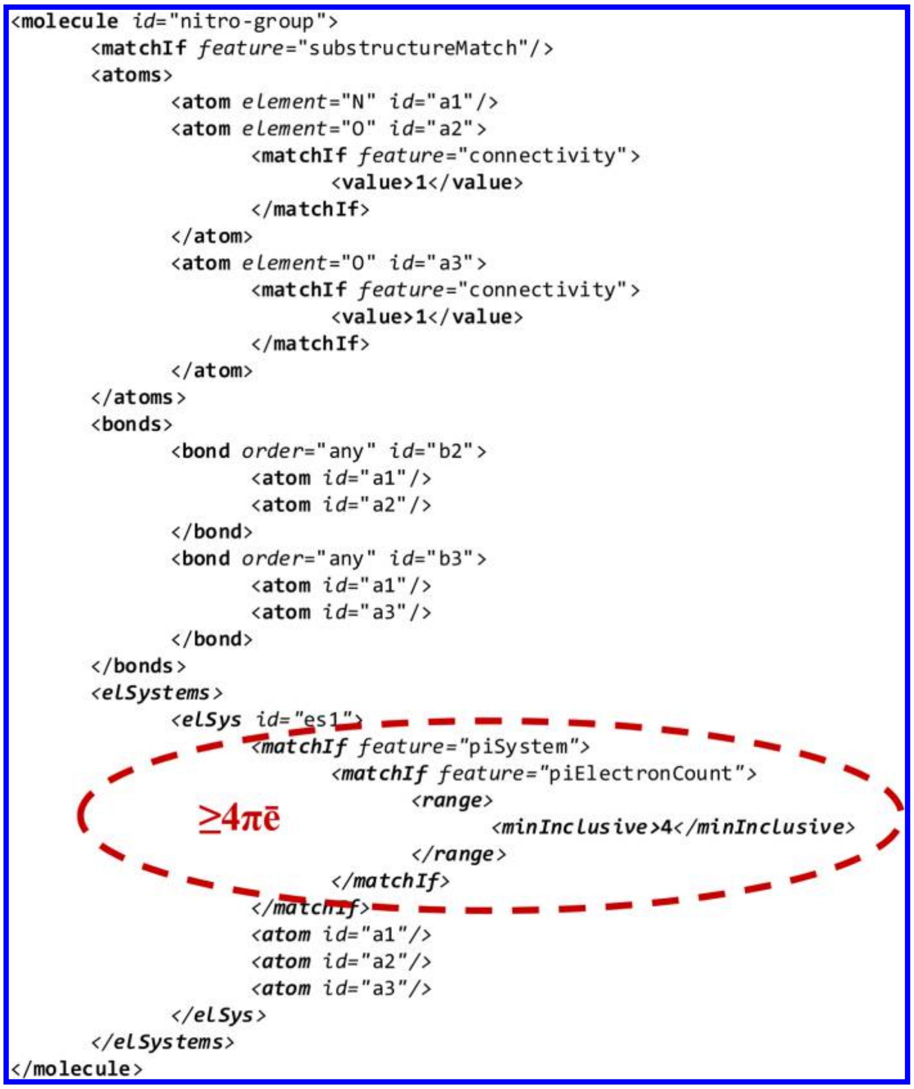
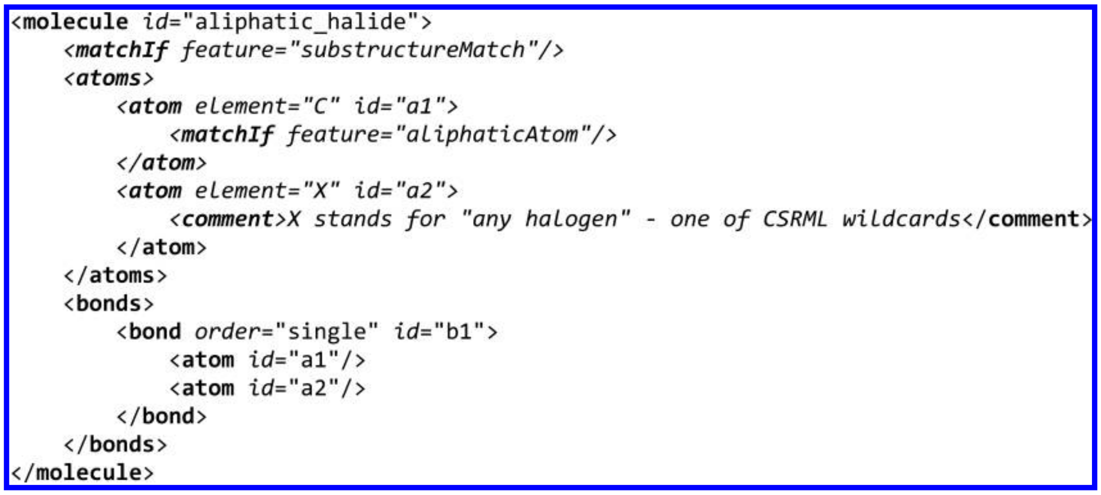


Figure 7. Graphical depiction of the SMARTS and Chemotype to an electron system query.



```
<molecule id="nitro-group">
        <matchIf feature="substructureMatch"/>
        <atoms>
                <atom element="N" id="a1"/>
                <atom element="O" id="a2">
                        <matchIf feature="connectivity">
                                <value>1</value>
                        </matchIf>
                </atom>
                <atom element="O" id="a3">
                        <matchIf feature="connectivity">
                                <value>1</value>
                        </matchIf>
                </atom>
        </atoms>
        <bonds>
                <bond order="any" id="b2">
                        <atom id="a1"/>
                        <atom id="a2"/>
                </bond>
                <bond order="any" id="b3">
                        <atom id="a1"/>
                        <atom id="a3"/>
                </bond>
        </bonds>
        <elSystems>
                <elSys id="es1">
                        <matchIf feature="piSystem">
                                <matchIf feature="piElectronCount">
                                        <range>
                                                <minInclusive>4</minInclusive>
                                        </range>
                                </matchIf>
                        </matchIf>
                        <atom id="a1"/>
                        <atom id="a2"/>
                        <atom id="a3"/>
                </elSys>
        </elSystems>
</molecule>
```

Figure 9. CSRML expression for a nitro group employing an electron system query.

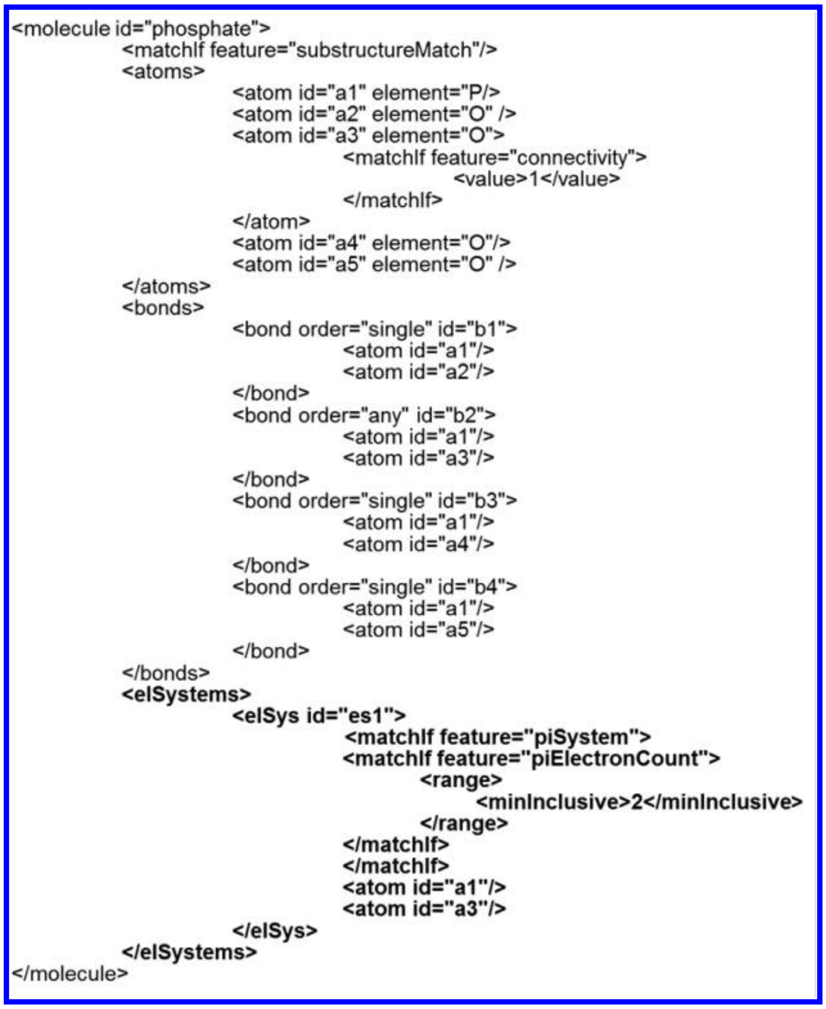


```
<molecule id="aliphatic_halide">
    <matchIf feature="substructureMatch"/>
    <atoms>
        <atom element="C" id="a1">
            <matchIf feature="aliphaticAtom"/>
        </atom>
        <atom element="X" id="a2">
            <comment>X stands for "any halogen" - one of CSRML wildcards</comment>
        </atom>
    </atoms>
    <bonds>
        <bond order="single" id="b1">
            <atom id="a1"/>
            <atom id="a2"/>
        </bond>
    </bonds>
</molecule>
```

Figure 10. CSRML expression for an aliphatic halide chemotype query.

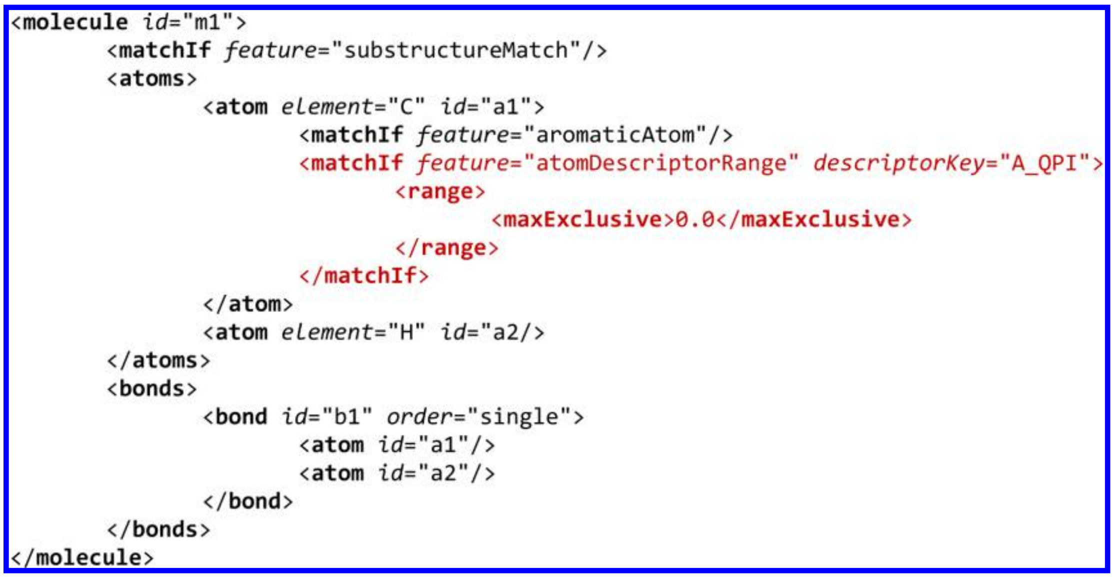
languages, the screening procedure requires at least two separate steps: first identification of all topologically matching candidates, followed by computation of the reactivity-related properties for selected atoms to further filter the correct candidates.

CSRML offers a way to combine these two steps into a one-step process. The query definitions of CSRML may employ both topology and predefined features with calculable descriptors. The implementing software is instructed by the CSRML query specification to compute and evaluate the



```
<molecule id="phosphate">
    <matchIf feature="substructureMatch"/>
    <atoms>
        <atom id="a1" element="P/>
        <atom id="a2" element="O" />
        <atom id="a3" element="O">
            <matchIf feature="connectivity">
                <value>1</value>
            </matchIf>
        </atom>
        <atom id="a4" element="O"/>
        <atom id="a5" element="O" />
    </atoms>
    <bonds>
        <bond order="single" id="b1">
            <atom id="a1"/>
            <atom id="a2"/>
        </bond>
        <bond order="any" id="b2">
            <atom id="a1"/>
            <atom id="a3"/>
        </bond>
        <bond order="single" id="b3">
            <atom id="a1"/>
            <atom id="a4"/>
        </bond>
        <bond order="single" id="b4">
            <atom id="a1"/>
            <atom id="a5"/>
        </bond>
    </bonds>
    <elSystems>
        <elSys id="es1">
            <matchIf feature="piSystem">
            <matchIf feature="piElectronCount">
               <range>
                  <minInclusive>2</minInclusive>
               </range>
            </matchIf>
            </matchIf>
            <atom id="a1"/>
            <atom id="a3"/>
        </elSys>
    </elSystems>
</molecule>
```

Figure 11. CSRML expression for phosphate, phosphoric acid and its esters, etc.



```
<molecule id="m1">
        <matchIf feature="substructureMatch"/>
        <atoms>
                <atom element="C" id="a1">
                        <matchIf feature="aromaticAtom"/>
                        <matchIf feature="atomDescriptorRange" descriptorKey="A_QPI">
                                <range>
                                        <maxExclusive>0.0</maxExclusive>
                                </range>
                        </matchIf>
                </atom>
                <atom element="H" id="a2/>
        </atoms>
        <bonds>
                <bond id="b1" order="single">
                        <atom id="a1"/>
                        <atom id="a2"/>
                </bond>
        </bonds>
</molecule>
```

Figure 12. Simplified CSRML expression to detect an activated ortho−para position.

defined values or value ranges of specific descriptors in order to produce a substructure match. Thus, any atom, bond, electron system, or fragment descriptors supported by the underlying cheminformatics platform may be used on demand when necessary for substructure match evaluation. Such "descriptor-augmented structural motifs", i.e., property embedded structural fragments, are quite important for a variety of uses including reactivity prediction, binding affinity modeling, or exploring reactive-metabolite mediated toxicity.

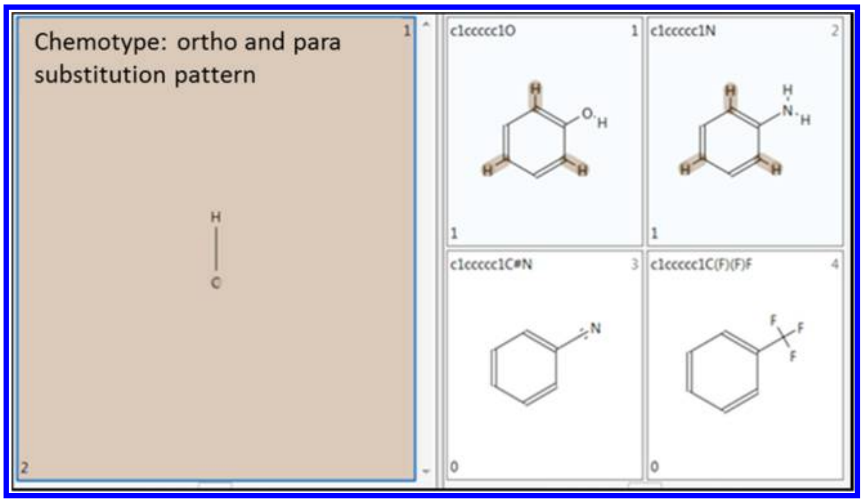


Figure 13. Matching of carbons at both ortho- and para-positions of the substitution.

Examples of Substructure Chemotypes. The following section illustrates some basic examples of CSRML usage in defining chemotypes.

Aliphatic Halides. Figure 10 illustrates a relatively simple case in which a carbon atom is marked with an "aliphaticAtom" flag. The second atom is expressed by a wildcard X for any halogen. A complete list of wildcards is given in Table 5.

Phosphoric Acid. As shown in Figure 1, a phosphoric acid group can be represented by multiple equivalent representations using rather complex SMARTS. In comparison, the phosphoric acid chemotype is uniquely represented using the CSRML electronic system, as shown in Figure 11. This expression will match varied representations of phosphate, including free acidic, anions, or any phosphoric esters or polyphosphates.

Ortho–Para-Electrophilic Substitution Patterns. Chemotypes additionally allow the detection of reactivity patterns in compounds. For example, consider an ortho- or para-substitution reaction in an aromatic ring. Whereas most substructure query languages require a series of structural fragments enumerating all possible ortho–para-activating substituents CSRML chemotypes achieve pattern recognition by specifying partial charges on the atoms in the aromatic ring as shown in Figure 12.

A distinctive feature of chemotypes is the ability to include atomic partial π-charges in the query if the calculation of such charges is supported by the underlying cheminformatics platform. The carbon atoms should have at least one attached hydrogen atom to be able to participate in an electrophilic substitution, and also need to be activated. The carbon activation can be queried, for example, by atomic partial charge due to π-electrons in the conjugated system (denoted as A_QPI in Figure 12). The specification is tested by the **atomDescriptorRange** feature annotated on carbon atom a1 with a **maxExclusive** value of 0 set to require that the partial π-charge on this atom is negative, i.e. an excessive electron density is present on these atoms. Although this is a simplification of reactivity in aromatic systems, this example illustrates how chemotypes encoding reactivity can be created using CSRML.

Applying the above rule to substituted benzenes in ChemoTyper, a specific matching of aromatic carbons at ortho- and para-positions to a substitution can be visualized as shown in Figure 13. The panel in the right side is the result of applying such a chemotype (shown in the left panel, where calculation of partial charges has been enabled and is implicit in the chemotype shown) to the four target structures with matching C–H bonds ortho or para to –OH or –NH substitutions in benzene ring. The top two structures on the right panel satisfy the condition, whereas the bottom two do not. This simple example illustrates the true power of chemotypes and their implementation in CSRML, and the advantage over SMARTS or other substructure query languages in which this type of precise representation is simply not possible.

Examples of Reaction Chemotypes. *pK_a of Substituted Benzoic Acid.* As illustrated in Figures 3 and 4, CSRML enables a user to specify reactions and reaction rules. The dependence of pK_a on para-substitutions of benzoic acid provides another simple and elegant example of the usefulness of CSRML.

Any substituent (A) stabilizing the ion will increase the acidity of benzoic acid—see Figure 14. As shown by the Hammett constants, the electron withdrawing power of the substituent increases in order from OH < OCH_3 < CH_3 < H < Cl < Br < I < C(=O)H < C#N < NO_2. A similar trend is observed with the summed total σ partial charge plus π partial charge at the carbon atom labeled with an asterisk (*) in Figure 15.

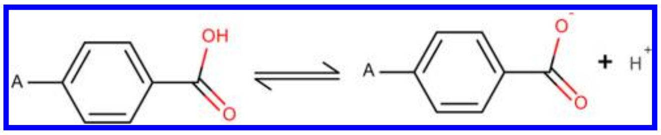


Figure 14. Simplified scheme of hydrogen dissociation in para-substituted benzoic acid.

The effect of any substituent on the acid dissociation constant can thus be captured with remarkable simplicity without enumerating the electron withdrawing and donating

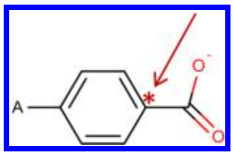


Figure 15. para-Substituted benzoic acid/benzoate.

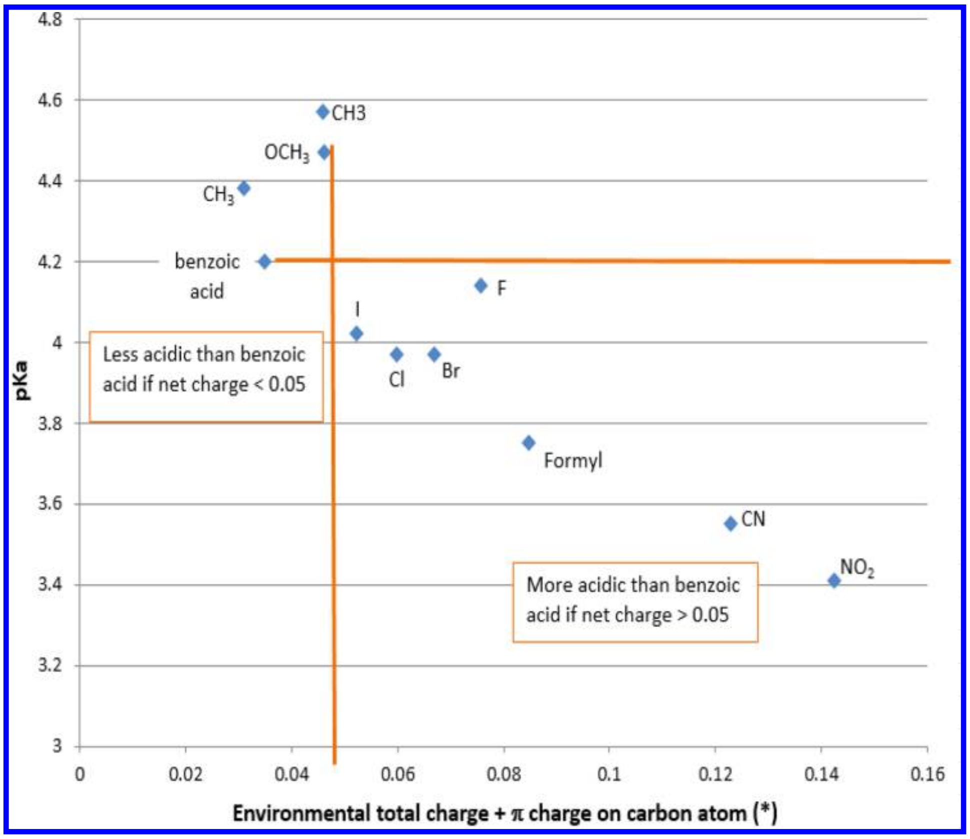
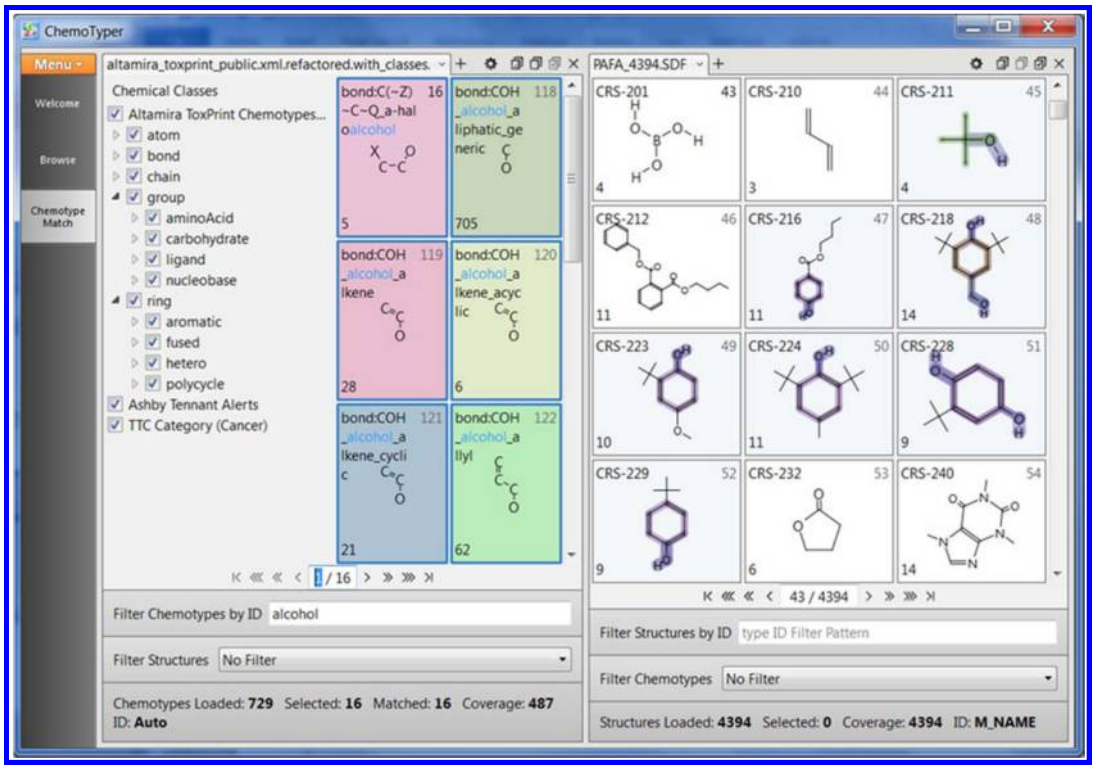


Figure 16. Effect of substituents on pK_a of benzoic acid.



Filter Chemotypes by ID alcohol

Filter Structures No Filter

Chemotypes Loaded: **729** Selected: **16** Matched: **16** Coverage: **487**
ID: **Auto**

Filter Structures by ID type ID Filter Pattern

Filter Chemotypes No Filter

Structures Loaded: **4394** Selected: **0** Coverage: **4394** ID: **M_NAME**

Figure 17. PAFA structures matching with ToxPrint in ChemoTyper.

groups by embedding atomic charges into a chemotype defined using CSRML. The chemotype definition relies on calculations of the summed total σ charge and atomic π partial charge (A_QPI) for the (*) carbon atom in each target molecule that matches the para-substituted benzoic acid substructure. The quantity equals 0.05 for benzoic acid; thus, as shown in Figure 16, compounds for which this sum is >0.05 have a lower pK_a

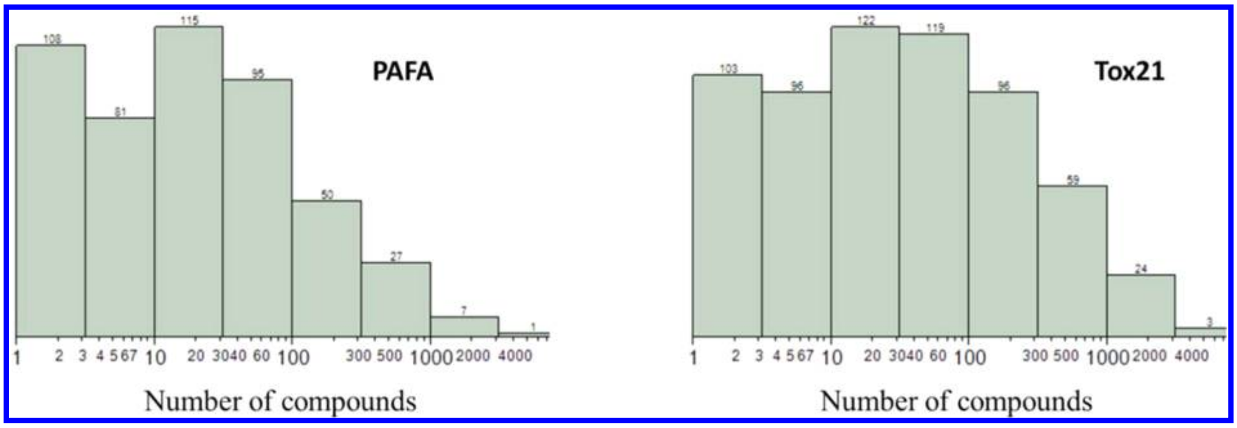


Figure 18. Profiling different inventories with ToxPrint chemotypes. Histogram bar heights are the number of ToxPrint chemotypes that match the corresponding number of compounds. For example, in the PAFA inventory, 77 chemotypes hit between 100 and 1000 compounds, while in the Tox21 inventory 155 chemotypes do the same.

than benzoic acid (increased acidity), whereas compounds for which this sum is <0.05 have a higher pK_a (decreased acidity).

Hence, this classical SAR example to account for the effect of a series of substituents is efficiently handled by a single chemotype definition coded in CSRML that embeds the appropriate charge descriptors for the (*) carbon atom.

ToxPrint Public Chemotype Library. *Coverage Validation.* ToxPrint is a publicly available, default set of chemotypes, which are coded for connectivity, valence bond principles, and electronic systems. It is provided in ChemoTyper, CORINA Descriptors Community Edition KNIME node, and as a separate CSRML file. The chemotype classes are organized by atoms (elements, including metals and metalloids), bonds (bonds involved in functional groups), chains (aliphatic, alicyclic, aromatic−aliphatic, oxy-aliphatic), ring systems (aromatic, polycyclic, heterocyclic, fused ring), and groups (carbohydrate, nucleobase, ligands). These chemotype classes were further refined by electronic system functionalities in CSRML. Table 4 lists the five top-level classes and a total set of 729 chemotypes defined in the core public library. Also included in the ToxPrint set are the Ashby−Tennant structural alerts[69] and structural category definitions identified for the cancer TTC (threshold of toxicological concerns) approach.[70] These additional chemotypes represent alerting features for genotoxic carcinogens. The left panel of Figure 17 shows a sample of the ToxPrint chemotype classes and chemotypes alerts.

Although the ToxPrint set was developed from a chemical space in which toxicity studies have been reported, most of ToxPrint chemotypes are not structural alerts; the majority (about 90%) represent generic query features intended for broader use in substructure searching, clustering, and to generically mine a data set for enriched areas of chemical and biological activity. The invariant set of ToxPrint chemotypes is additionally useful for comparative profiling of diverse data sets, as well as across studies and projects. The generic ToxPrint chemotypes are designed to provide broad coverage of structural features for a broad variety of substance types constituting public chemical inventories, including complex drug molecules, agrochemicals, cosmetics ingredients, food ingredients and additives, and industrial chemicals. In Figure 17, a few structures from the FDA PAFA[32] inventory are displayed in the right panel. A set of 16 alcohol chemotypes (with 6 shown in the left panel) was selected to locate the structures matched with such queries; the right panel displays the structures where matched bonds are highlighted with colors of the selected chemotypes with 6 selected alcohol chemotypes (in the left panel, only 6 are shown).

The excellent coverage of diverse data sets of interest provided by the ToxPrint library was confirmed by projecting these chemotypes onto various chemical substance inventories. Figure 18 depicts the histogram of matches against two large public inventories. The US FDA PAFA database[32] consists of a highly diverse set of indirect food additives containing more than 7200 test substances, 4394 of which can be assigned chemical structures. The structurable subset of PAFA is completely covered by 484 (out of 729 total) ToxPrint chemotypes as shown in Figure 18, i.e., each of the 4394 structures contains 1 or more of 484 ToxPrint unique chemotypes. The Tox21 inventory[51] consists of agrochemicals, drugs and drug candidates, cosmetics, food additives, fragrances, consumer and industrial chemicals, and known or presumed toxicants. Of the 8193 chemicals in Tox21 (TOX21S_v2a-8193[42]), 8165 had connection tables. The full set of structurable chemicals in the Tox21 inventory are completely covered by 622 (out of 729 total) unique ToxPrint chemotypes. The relative coverage of unique ToxPrint chemotypes, 66% in the case of PAFA, and 85% in the case of Tox21, also provides a quantitative assessment of absolute and relative diversity within each of the two data sets, with Tox21 correctly seen as being the larger and more structurally diverse of the two according to this measure. Figure 18 shows the histograms of the chemotype hits in each of the two separate data sets (numbers converted to log scale). On average, a single structure in PAFA hits 67 chemotypes whereas a single structure in Tox21 hits 161 (note in this example that a chemotype may occur and be counted multiple times within the same structure). These high numbers are a reflection of the variety and large number of overlapping chemotypes, enabling many types of possible query results, as well as the high incidence of relatively large, complex structures in the two data sets, with significantly more of these apparently present in Tox21 than in PAFA. In addition, note that there are 245 chemotypes not present in any of the 4394 PAFA structures, whereas only 107 chemotypes are unused in the 8165 Tox21 structures. These “absent” chemotype subsets not only provide a direct measure of the extent of structural diversity of the two data sets but could be used to explore missing chemical space from a biological activity standpoint.

Along these lines, the chemical space of inventories can be further profiled and compared using chemotypes. To illustrate this process, a subset of ToxPrint chemotypes well-suited for profiling the PAFA and Tox21 inventories were selected using three criteria: (1) representing common organic groups including aldehyde, alcohol, amine, carboxylic acid, ester, amide, organohalides, heterocyclic (5 and 6 membered rings); (2) present in approximately 10−40% of the compounds (a percentage that may vary depending on the data sets being analyzed; note that chemotypes present in nearly all or in very few compounds in the data sets are in general not considered useful or discriminating); (3) enable differentiating queries across the two inventories (i.e., that hit many more compounds in one inventory than the other).

Figure 19 shows that chemicals in Tox21 contain more aromatic rings, halides (both aliphatic and aromatic), carboxamide, aromatic ether, and heterocylic rings. These structure class observations are consistent with the Tox21 inventory being more highly enriched with drugs, drug candidates, and agrochemicals than PAFA, which contains mostly food additives and cosmetics colorants. In comparison, PAFA chemicals are more enriched with classes including varying alkyl chains (e.g., C4, C10, C12), alkene, carboxylic ester, and aldehydes. This example illustrates how using an invariant set of ToxPrint chemotypes to profile inventories and data sets provides a powerful tool for the comparison of chemical space.

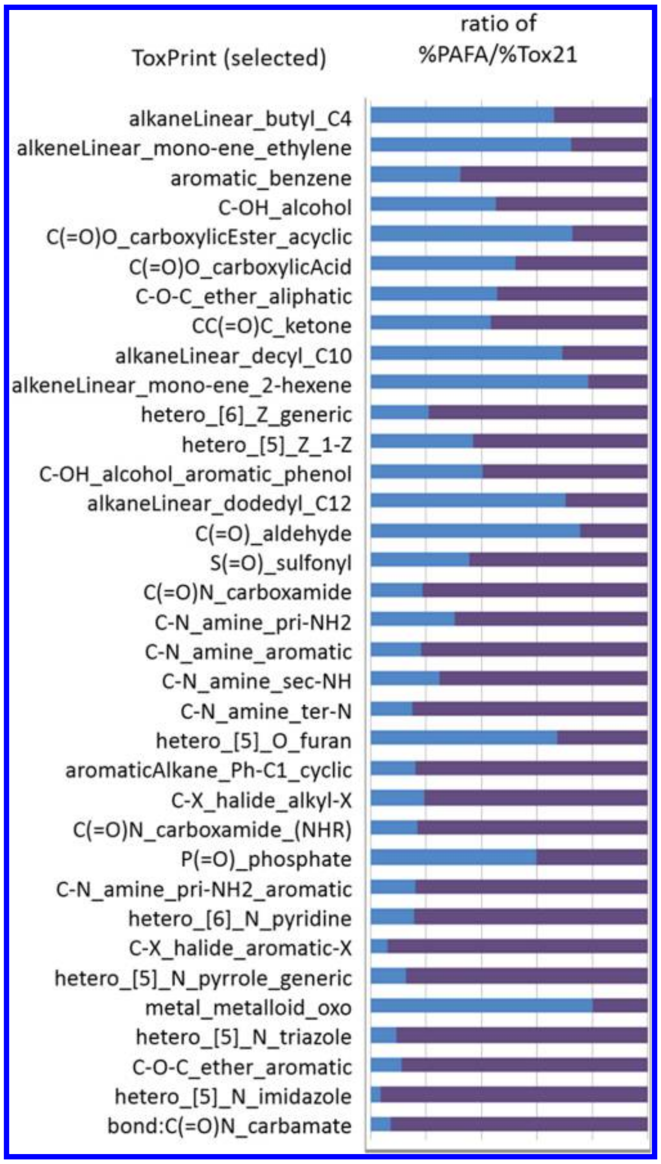
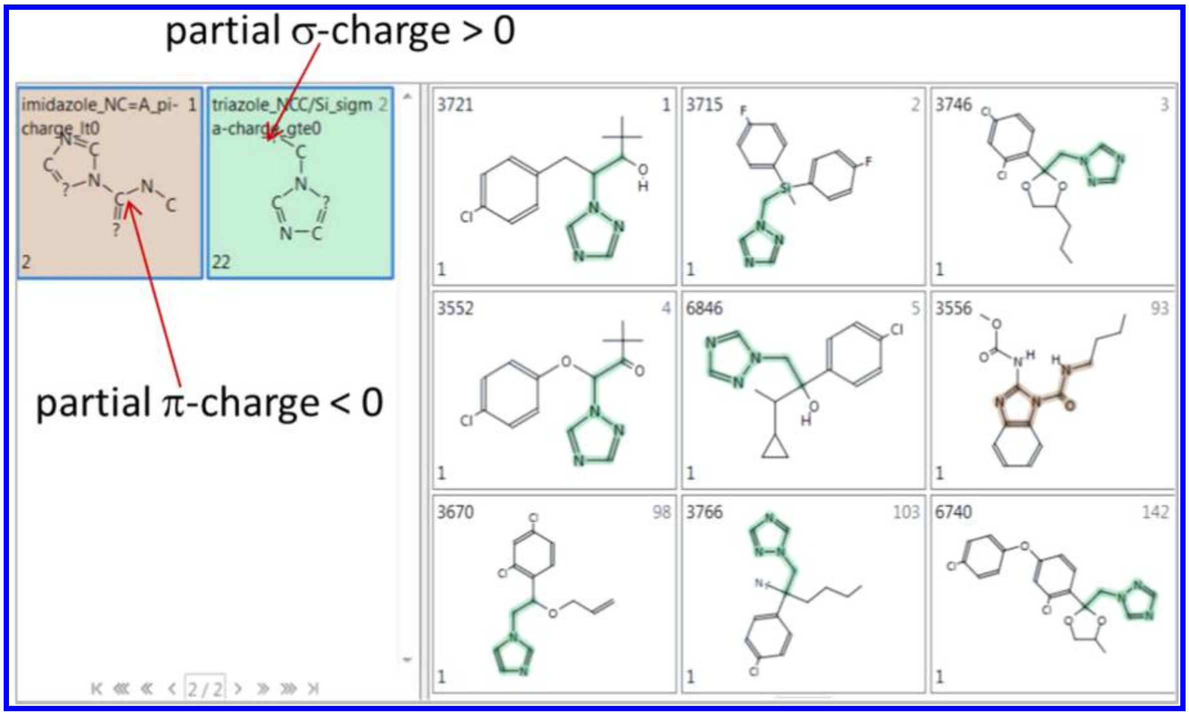


Figure 19. Representative ToxPrint chemotypes used for profiling the FDA CFSAN PAFA and Tox21 inventories.

Structural Alerts and Alerting Chemotypes. As discussed earlier, chemotypes can be used to represent results of SAR studies, such as in the pK_a illustration. On the other hand, whereas structural alerts are not individually considered to constitute a SAR model, they embody elements of SAR and have proven useful for screening and prioritizing. As part of ToxPrint, a well-known set of genotoxic carcinogenicity features based on Ashby−Tennant alerts[69] are included. Also included are the alerting category features from the cancer Threshold of Toxicological Concerns (TTC).[70] Both are considered genotoxic carcinogen alerting chemotypes in the present treatment. Of the 4394 PAFA structures, 325 were matched with Ashby−Tennant alerts and 495 with cancer TTC category chemotypes. When applied to the Tox21 inventory of 8165 structures, 1572 were found to contain Ashby−Tennant alerts and 2478 contained cancer TTC category chemotypes. Since PAFA is a food additive inventory, whereas Tox21 would be expected to contain more overt toxicants, these results are consistent with the observation that much smaller fraction of structures in PAFA are alerted with genotoxic carcinogen potential. Using potentially alerting chemotypes to compare different data sets is also a powerful use-case of ToxPrint and can either be used to focus investigations into areas of higher probability for carcinogenic activity or to focus development efforts away from this chemical space.

Although chemotypes can be used to represent conventional structural alerts, it is important to emphasize that most of the ToxPrint chemotypes should not be considered structural alerts even though they were developed from a toxicity data rich chemical space. For example, the FDA Redbook considers imidazoles and triazoles as potentially "toxicologically active" classes, but these need to be further defined for specific end points before associating them with structural alerts. As a specific case, *N*-alkyl substituted triazoles shown in Figure 20 are conazole inhibitors in sterol biosynthesis, which can lead to cleft palate abnormalities.[71,72] The triazole class can be considered as a rough structural alert in this case and could lead to more refined SAR investigation.

When exploring alerts, an important advantage of chemotypes over other query representations is the ability to embed nontopological criteria directly into the alert. An alerting chemotype therefore can extend beyond conventional "structural alert" approaches to incorporate more sophisticated SAR considerations. For example, the triazole class mentioned above can be further refined for greater improvement in detecting potential rodent cleft palate positives by considering influences on likely sites of reactivity, i.e. the partial charges of the carbon atoms due to electron withdrawing or donating groups. Two possible conazole substitution patterns are captured by the two chemotypes in Figure 20, where partial charges on the carbon adjacent to the triazole N most likely provide a discriminating feature with respect to probable reactivity within the subsets. There are 81 triazoles detected in the Tox21 inventory; 19



partial σ-charge > 0

partial π-charge < 0

Figure 20. *N*-Alkyl-substituted triazole.

structures had rodent cleft palate data with 6 of these reporting positive and 13 reporting negative for cleft palate.[73] The two alerting chemotypes in Figure 20 match 28 compounds within the larger set of 81 triazoles. Of these 28, 5 are among the 6 confirmed cleft palate actives, whereas 9 were among the confirmed negatives (the remaining 14 had no test data). Hence the cleft palate alert augmented with the partial atomic charge condition significantly enriched the proportion of cleft palate actives within the triazole alerting subset, from 6/81 (7%) confirmed positives to 5/28 (18%).[73] Further enrichment or discrimination of these triazoles could be based on biological inputs or additional reactivity considerations and used to build a SAR predictive model for cleft palate. In this scenario, chemotypes are used to generate and refine hypotheses and can undergo end point-specific customization by refining the structural motifs to include discriminating and predictive physicochemical properties, resulting in more predictive alerting chemotypes. Hence, a chemotype can range in function from a generic structure query feature (e.g., phosphoric acid group), to an alerting feature informed by biological activity associations (such as an Ashby–Tennant alert, primarily used for hazard screening and activity enrichment), to a component of a predictive SAR model. The last application would, in turn, require more stringent validation and delineation of domain of applicability, consistent with accepted standards of SAR practice.

FUTURE WORK AND CHALLENGES

By design, the initial release of CSRML did not support recursive queries since, as noted in several of the examples, CSRML allows queries to be constructed in a way that eliminates the need to enumerate multiple possible substitution patterns. However, since SMARTS users commonly use recursion to, for example, define a list of substituents attached to an atom, the ability to construct recursive queries has been added to CSRML. Due to how CSRML has been designed, adding this functionality simply involved defining an additional query feature and did not require any change to the CSRML syntax. This functionality is somewhat limited in comparison to SMARTS because the CSRML implementation does not support multilevel (nested) recursion; i.e., one cannot define fragments attached to fragments attached to the main substructure.

ChemoTyper does not currently support stereospecific queries, although a framework for these query features in CSRML has been designed. Implementation will be straightforward since it requires definition of new query features but no modification of the CSRML syntax.

Incorporating CSRML into other platforms presents the same challenges expected whenever new technology is integrated with existing systems. The reference implementation and XSD schema provide a solid foundation for incorporating the CSRML support into toolkits. The substructure search machinery will have to be added or adjusted in these toolkits, as these depend on toolkit-specific query representations. For databases, cartridges and/or plug-ins to support native CSRML queries on the database server side will need to be developed.

CONCLUSION

To overcome the limitations of the current representation methods for chemical patterns (SMARTS) or reactions (SMIRKS), chemotypes, a new representation method for chemical molecules, substructures, patterns, reaction rules, and reactions, have been developed. Chemotype query features are expressed in an XML-based language and can be encoded not only with connectivity and topology, but also with properties of atoms, bonds, electronic systems, or molecules. The language has been developed in parallel with a public set of chemotypes, i.e., ToxPrint, such that the query features are comprehensive and realistic. The ToxPrint chemotypes represent the chemical space of toxicity data-rich resources and can be further designed to produce end point specific chemotype alerts, based both on connectivity/topology and mechanistic properties. A software application, ChemoTyper has also been developed and made publicly available to enable chemotype searching and fingerprinting against a target structure set. The public ChemoTyper houses the ToxPrint chemotype CSRML dictionary, and a reference implementation has been made available so that the query specifications may be adopted by other chemical structure knowledge systems. A ToxPrint fingerprinter is also available in the CORINA Descriptors Community Edition KNIME node to be used in computational workflows. The full specifications of the XML-based CSRML standard used to define chemotypes are also publicly available to facilitate and encourage the exchange of structural knowledge between various knowledge systems.

The above capabilities are generic and applicable to a wide range of potential uses. However, never before has such a complete suite of capabilities and resources been publicly available specifically for use in toxicity modeling. We believe that these new chemistry capabilities will enable a greater engagement and participation of toxicologists and modelers in mining an expanding body of public toxicological data resources, such as high-throughput screening data being generated within public efforts such as Tox21 and EPA's ToxCast program.[74] It is also our hope that these capabilities will both facilitate and guide development of chemotypes for a wide variety of toxicologically relevant end points, and that the public sharing and distribution of such chemotypes will foster greater collaboration and fuel discovery in this field.

■ ADDITIONAL INFORMATION

The full specifications of the XML standard used in the CSRML language and a software application, ChemoTyper, to enable chemotype searching and fingerprinting against a target structure set are publicly available at https://chemotyper.org. Also available for download at this site are the ToxPrint chemotypes and a CSRML reference implementation to facilitate adoption of CSRML query specifications by other chemical structure knowledge systems.

Updates for the ToxPrint chemotypes can be downloaded separately from ChemoTyper at https://toxprint.org.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Version 2.0.r307 of the CSRML XML schema. This is the version used during the preparation of this manuscript. This material is available free of charge via the Internet at http://pubs.acs.org. As noted in the Additional Information section, the most current version of the schema is always available free of charge at https://chemotyper.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chihae.yang@molecular-networks.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The CSRML language, reference library, and reference implementation software ChemoTyper were developed by Molecular Networks GmbH under a contract from US FDA CFSAN in collaboration with Altamira LLC. The ToxPrint chemotypes were developed by Altamira LLC and US FDA. The authors thank the CERES team members at US FDA CFSAN for the support, especially, Dr. Annette McCarthy and Dr. Dennis Keefe. The authors are also indebted to Dr. Mitchell A. Cheeseman for his inspiration. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Food and Drug Administration, the U.S. Environmental Protection Agency, or other research institutions. Mentioning of trade names or commercial products does not constitute endorsement or recommendation for use.

■ REFERENCES

(1) *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*. Daylight Chemical Information Systems, Inc. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Nov 2, 2014).

(2) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Model.* **1999**, *39*, 1161−1172.

(3) Schomburg, K. T.; Wetzer, L.; Rarey, M. Interaction Design of Generic Chemical Patterns. *Drug Discovery Today* **2013**, *18*, 651−658.

(4) 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. Model.* **1997**, *37*, 71−79.

(5) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 2294−2307.

(6) *CTfile Formats*. MDL Information Systems, Inc. http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php (accessed Nov 2, 2014).

(7) *ChemDraw*. PerkinElmer Inc. http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ (accessed Nov 2, 2014).

(8) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Model.* **2000**, *40*, 1302−1314.

(9) *MarvinSketch*. ChemAxon−cheminformatics platforms and desktop applications. http://www.chemaxon.com/products/marvin/marvinsketch/ (accessed Nov 2, 2014).

(10) Maguna, F. P.; Nunez, M. B.; Okulik, N. B.; Castro, E. A. Methodologies QSAR/QSPR/QSTR: current state and perspectives. *Advances in Chemical Modeling* **2011**, *1*, 411−433.

(11) Sukumar, N.; Das, S.; Krein, M.; Godawat, R.; Vitol, I.; Garde, S.; Bennett, K. P.; Breneman, C. M. Molecular Descriptors for Biological Systems. In *Computational Approaches in Cheminformatics and Bioinformatics*; Guha, R., Bender, A., Eds.; Wiley: 2012; pp 107−143.

(12) Madan, A. K.; Bajaj, S.; Dureja, H. Classification Models for Safe Drug Molecules. *Computational Toxicology Volume II*; Methods in Molecular Biology; Springer: New York, 2013; Vol. *930*, pp 99−124.

(13) Yang, C.; Cross, K.; Myatt, G. J.; Blower, P. E.; Rathman, J. F. Building Predictive Models for Protein Tyrosine Phosphatase 1B Inhibitors Based on Discriminating Structural Features by Reassembling Medicinal Chemistry Building Blocks. *J. Med. Chem.* **2004**, *47*, 5984−5994.

(14) Sanderson, D. M.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System. *Human & Experimental Toxicology* **1991**, *10*, 261−273.

(15) Chakravarti, S. K.; Saiakhov, R. D.; Klopman, G. Optimizing Predictive Performance of CASE Ultra Expert System Models Using the Applicability Domains of Individual Toxicity Alerts. *J. Chem. Inf. Model.* **2012**, *52*, 2609−2618.

(16) *Daylight > SMARTS Examples*. http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html#P (accessed Nov 2, 2014).

(17) Proschak, E.; Wegner, J. K.; Schüller, A.; Schneider, G.; Fechner, U. Molecular Query Language (MQL)A Context-Free Grammar for Substructure Matching. *J. Chem. Inf. Model.* **2007**, *47*, 295−301.

(18) Karabunarliev, S.; Nikolova, N.; Nikolov, N.; Mekenyan, O. Rule Interpreter: A Chemical Language for Structure-Based Screening. *Journal of Molecular Structure: THEOCHEM* **2003**, *622*, 53−62.

(19) 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.

(20) Leist, M.; Lidbury, B. A.; Yang, C.; Hayden, P. J.; Kelm, J. M.; Ringeissen, S.; Detroyer, A.; Meunier, J. R.; Rathman, J. F.; Jackson, G. R.; Stolper, G.; Hasiwa, N. Novel Technologies and an Overall Strategy to Allow Hazard Assessment and Risk Prediction of Chemicals, Cosmetics, and Drugs with Animal-Free Method. *ALTEX* **2012**, *29* (4), 373−388.

(21) *MOSES—Extensive cheminformatics platform|Inspiring Chemical Discovery*. https://www.molecular-networks.com/moses (accessed Nov 2, 2014).

(22) *KNIME*. http://www.knime.org/ (accessed Nov 2, 2014).

(23) Murray-Rust, P.; Rzepa, H. S.; Wright, M.; Zara, S. A Universal Approach to Web-Based Chemistry Using XML and CML. *Chem. Commun.* **2000**, 1471−1472.

(24) Gillepsie, R. J. The valence-shell electron-pair repulsion (VSEPR) theory of directed valency. *J. Chem. Educ.* **1963**, *40*, 295−301.

(25) ToxPrint ChemoTypes. https://toxprint.org/; ChemoTyper Community Website. https://chemotyper.org/ (accessed Dec 19, 2014).

(26) Java Architecture for XML Binding (JAXB). http://www.oracle.com/technetwork/articles/javase/index-140168.html (accessed Nov 2, 2014).

(27) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(28) Stroustrup, B. *The C++ Programming Language*, 3rd ed.; Addison-Wesley: Reading, Mass, 1997; ISBN 0-201-88954-4 OCLC 59193992.

(29) Xerces-C++ XML Parser. https://xerces.apache.org/xerces-c/ (accessed Nov 2, 2014).

(30) Apache License, Version 2.0. https://www.apache.org/licenses/LICENSE-2.0.html (accessed Nov 2, 2014).

(31) US FDA/Center for Drug Evaluation and Research. http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm (accessed Nov 2, 2014).

(32) Benz, R. D.; Irausquin, H. Priority-Based Assessment of Food Additives Database of the U.S. Food and Drug Administration Center for Food Safety and Applied Nutrition. *Environ. Health Perspect.* **1991**, *96*, 85−89.

(33) National Toxicology Program—NTP. http://ntp.niehs.nih.gov/ (accessed Nov 2, 2014).

(34) Chemical Carcinogenesis Research Information System Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/ccrisfs.html (accessed Nov 2, 2014).

(35) Integrated Risk Information System Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/irisfs.html (accessed Nov 2, 2014).

(36) GENETOX. http://toxnet.nlm.nih.gov/newtoxnet/genetox.htm (accessed Nov 2, 2014).

(37) Developmental and Reproductive Toxicology Database Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/dartfs.html (accessed Nov 2, 2014).

(38) TERIS, Teratogen Information System. https://depts.washington.edu/terisweb/teris/ (accessed Nov 2, 2014).

(39) US EPA ECOTOX Database. http://cfpub.epa.gov/ecotox/ (accessed Nov 2, 2014).

(40) US FDA Endocrine Disruptor Knowledge Base. http://www.fda.gov/ScienceResearch/BioinformaticsTools/EndocrineDisruptorKnowledgebase/default.htm (accessed Nov 2, 2014).

(41) The Carcinogenic Potency Project (CPDB). http://toxnet.nlm.nih.gov/cpdb/ (accessed Nov 2, 2014).

(42) US EPA DSSTox, Computational Toxicology Research Program (CompTox). http://www.epa.gov/ncct/dsstox/index.html (accessed Nov 2, 2014).

(43) US EPA ACToR. http://actor.epa.gov/actor/faces/ACToRHome.jsp (accessed Nov 2, 2014).

(44) US EPA ToxRefDB, Computational Toxicology Research Program (CompTox). http://www.epa.gov/ncct/toxrefdb/ (accessed Nov 2, 2014).

(45) Benigni, R.; Bossa, C.; Richard, A.; Yang, C. A novel approach: chemical relational databases, and the role of the ISS CAN database on assessing chemical carcinogenicity. *Ann. Ist Super Sanità* **2008**, *44* (1), 48−56.

(46) ECHA REACH Substance Registration Database. http://echa.europa.eu/information-on-chemicals/registered-substances (accessed Nov 2, 2014).

(47) Scientific Committee on Consumer Safety (SCCS). http://ec.europa.eu/health/scientific_committees/consumer_safety/index_en.htm (accessed Nov 2, 2014).

(48) US EPA TSCA Chemical Substance Inventory. http://www.epa.gov/oppt/existingchemicals/pubs/tscainventory/index.html (accessed Nov 2, 2014).

(49) US EPA Pesticide Inert Ingredients. http://www.epa.gov/opprd001/inerts/ and http://www2.epa.gov/pesticide-registration/inert-ingredients-overview-and-guidance (accessed Nov 2, 2014).

(50) PAN Pesticide Database. http://pesticideinfo.org/ (accessed Nov 2, 2014).

(51) Tice, R.; Austin, C.; Kavlock, R. J.; Bucher, J. R. Improving the Human Hazard Characterization of Chemicals: A Tox21 Update. *Environ. Health Perspect.* **2013**, *121*, 756−765.

(52) Domestic Substances List—Acts & Regulations—Environment Canada. http://www.ec.gc.ca/lcpe-cepa/default.asp (accessed Nov 2, 2014).

(53) CosIng Database. http://ec.europa.eu/consumers/cosmetics/cosing/ (accessed Nov 2, 2014).

(54) ChemIDplus. http://chem.sis.nlm.nih.gov/chemidplus/ (accessed Nov 2, 2014).

(55) ChemSpider. http://www.chemspider.com/ (accessed Nov 2, 2014).

(56) Chemical Evaluation and Risk Estimation System (CERES). http://www.accessdata.fda.gov/FDATrack/track-proj?program=cfsan&id=CFSAN-OFAS-Chemical-Evaluation-and-Risk-Estimation-System (accessed Nov 2, 2014).

(57) Feldman, H. J.; Dumontier, M.; Ling, S.; Haider, N.; Hoque, C. W. CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules. *FEBS Lett.* **2005**, *579* (21), 4685−4691 Aug 29.

(58) PubChem structure search (includes SMARTS search capability), National Center for Biotechnology Information. https://pubchem.ncbi.nlm.nih.gov/search/search.cgi#.

(59) Heterocylic ring classificitaon (Genetontology). https://database.riken.jp/sw/en/heterocycle_biosynthetic_process/cria250u18130i/ (accessed Nov 2, 2014).

(60) Elschenbroich, C. *Organometallics*, 2nd ed.; Wiley-VCH, 2006.

(61) Database and ontology of Chemical Entities of Biological Interest (ChEBI). http://www.ebi.ac.uk/chebi/.

(62) US EPA Office of Pesticide Programs chemical classes. http://www.epa.gov/oppsrrd1/registration_review/explanation.htm.

(63) 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.

(64) Rishton, G. M. Nonleadlikeness and Leadlikeness in Biochemical Screening. *Drug Discovery Today* **2003**, *8*, 86−96.

(65) RDKit. http://rdkit.org/ (accessed Nov 2, 2014).

(66) The Chemistry Development Kit. http://sourceforge.net/projects/cdk/ (accessed Nov 2, 2014).

(67) Randic, M. Aromaticity and Conjugation. *J. Am. Chem. Soc.* **1977**, *99* (2), 444−450.

(68) Bauerschmidt, S.; Gasteiger, J. Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species. *J. Chem. Inf. Model.* **1997**, *37*, 705−714.

(69) Ashby, J.; Tennant, R. W. Chemical Structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis among 222 Chemicals Tested in Rodents by the U.S. NCI/NTP. *Mutation Research/Genetic Toxicology* **1988**, *204*, 17−115.

(70) Kroes, R.; Renwick, A.; Cheeseman, M.; Kleiner, J.; Mangelsdorf, I.; Piersma, A.; Schilter, B.; Schlatter, J.; van Schothorst, F.; Vos, J. .; Würtzen, G. Structure-Based Thresholds of Toxicological Concern (TTC): Guidance for Application to Substances Present at Low Levels in the Diet. *Food Chem. Toxicol.* **2004**, *42*, 65−83.

(71) Volarath, P.; Little, S.; Yang, C.; Martin, M.; Reif, D.; Richard, A. Features Analysis of ToxCast Compounds. *Fall National ACS Meeting*, Boston, MA, Aug 22−26, 2010.

(72) Sipes, N. S.; Martin, M. T.; Reif, D. M.; Kleinstreuer, N. C.; Judson, R. S.; Singh, A. V.; Chandler, K. J.; Dix, D. J.; Kavlock, R. J.; Knudsen, T. B. Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. *Toxicol. Sci.* **2011**, *1*, 109−27.

(73) Yang, C.; Arvidson, K.; Richard, A.; Worth, A.; Tarkhov, A.; Ringeissen, S.; Marusczyk, J.; Gasteiger, J.; Rathman, J.; Schwab, C. Chemotypes and Chemotyper: a new structure representation standard to include atomic/bond properties into structural alerts for toxicity effects and mechanisms. Poster presentation. *52nd Annual Meeting of the Society of Toxicology*, San Antonio, TX, Mar 10−14, 2013.

(74) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstrauer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; Richard, A.; Rotroff, D.; Sipes, N.; Dix, D. Update on EPA's ToxCast Program: Providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* **2012**, *25*, 1287−1302.