

Molecular Networks
Inspiring Chemical Discovery



CSRML – A New Markup Language Definition for Chemical Substructure Representation

Christof H. Schwab

**Molecular Networks GmbH
Henkestraße 91
91052 Erlangen, Germany
www.molecular-networks.com**

Outline

- **Chemical subgraphs**
 - *Representation and use cases*
- ***De facto* standards**
- **Requirements of new definition of subgraph representation**
- **XML-based substructure representation**



Chemical Subgraphs and Substructures



- **Well established concept in chemistry and chemoinformatics**
 - *Ray and Kirsch, Finding Chemical Records by Digital Computers. Science, 1957, 126, 814-819*
 - *Fisanick et al. Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files. J. Chem. Inf. Comput. Sci. 1975, 15 (2), 73-84*
- **Employed by almost all software packages that deal with sets of chemical structures and reactions**





Chemical Substructures – Use Cases

- **Chemical database queries**
 - *Find structure(s) enclosing the query substructure*
 - *Retrieval of analogs or similar structures*
- **MCSS searches**
- **Fingerprinting**
- **Analysis of chemical structures**
 - *Structural alerts*
 - *TTC analysis*
- **Highlighting of functional groups**



Example: Database Lookup

- ChemIDplus
- Query
 - Chlorobenzene
- Search mode
 - Substructure

ChemIDplus Advanced - Windows Internet Explorer

http://chem.sis.nlm.nih.gov/chemidplus/

United States National Library of Medicine ChemIDplus Advanced

News SIS Home | Site | About Us | Contact | Help

Env. Health & Toxicology TOXNET ChemIDplus Lite Advanced

Search Clear History Help

Display 5 results

Substance Identification

Name/Synonym Equals

Data is available for 389,359 records.

Toxicity

Test (any) between (mg/kg or ppm)

Species: (any)

Route: (any)

Effect: (any)

Toxicity data is available for 139,354 records.

Physical Properties

Melting Point between Measurement Type

Physical property data was provided by Syracuse Research Corporation and is available for 25,461 records.

Locator Codes

Structure

View Help

c1ccccc1Cl

Powered by ChemAxon Marvin

Structure Search Options

☒ Substructure Search

☐ Similarity Search 80 %

☐ Exact (parent only)

☐ Flex (parent, salts, mixture) NEW

☐ Hexplus (parent, all variations) NEW

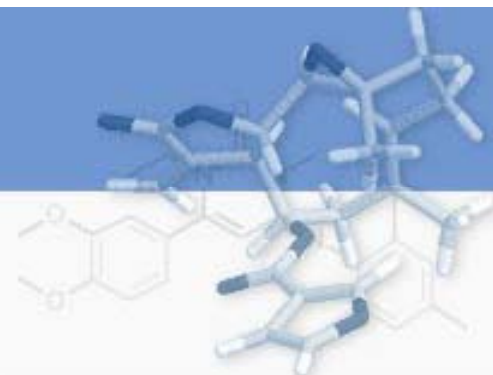
Display structures using

☒ Marvin ☐ Chime Change

Structure data is available for 295,302 records.



Example: Database Lookup



■ 25,512 hits

ChemIDplus Advanced - Windows Internet Explorer

http://chem.sis.nlm.nih.gov/chemidplus/ProxyServlet?chemidheavy

United States National Library of Medicine

ChemIDplus Advanced

News SIS Home | Site | About Us | Contact | Help

Env. Health & Toxicology TOXNET ChemIDplus Lite Advanced

Results: 1 - 5 of 25512

Start New Query
Modify Query
Show Query
Search History
Go To Record Number
TOXNET Home

1 [DDT \[BSI:SO\]](#)
50-29-3

Next Page ►

MW: 354.4901

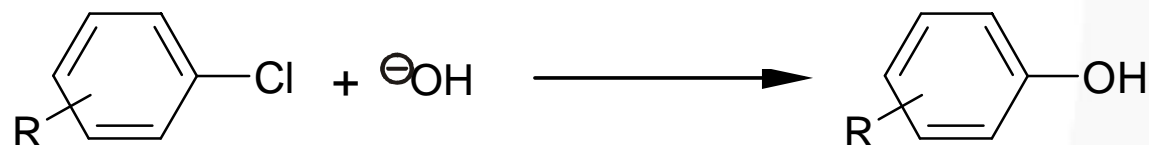
2 [Benzoic acid, 2,6-dichloro-](#)
50-30-6

MW: 191.013

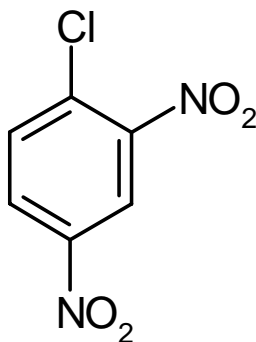
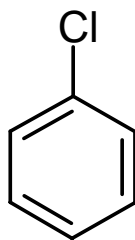


Example: Query with Properties

- Find chlorobenzene derivatives which are easily hydrolyzed at standard conditions

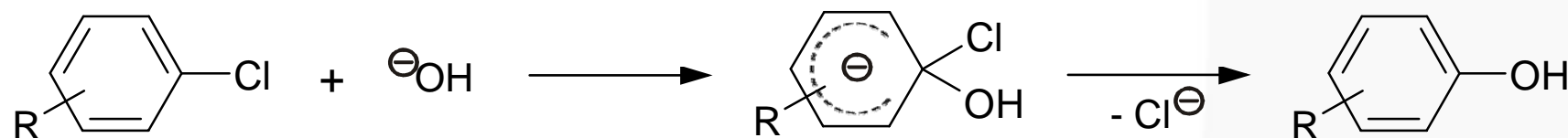


- Substructure based query will return both

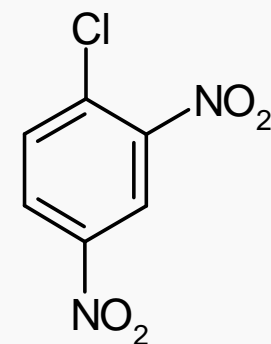
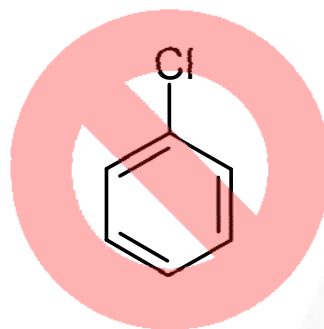


Example: Query with Properties

- **Nucleophilic aromatic substitution**



- **Chlorobenzene does not react at standard conditions**



- **Reaction conditions**
- **Resonance stabilization**

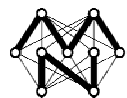
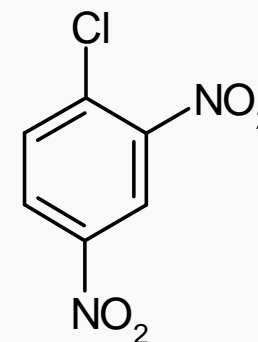
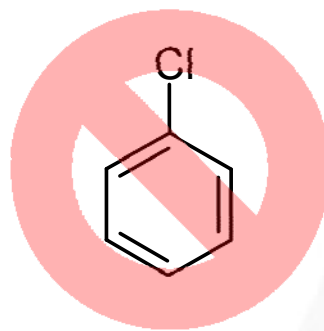
400 °C, 300 bar
0 kJ/mol

room temp.
43.5 kJ/mol



Example: Query with Properties

- It is not sufficient to have queries solely based on substructures



Existing *de facto* Standards

■ SMARTS

- *Substructure specification by text line notation*
- *Definition complex substructure patterns including logical operations, recursion, etc*

■ MDL CTab Query

- *CTab file based query definition*
- *Potentially extendible using SD properties in non-standard way*

■ SYBYL line notation (SLN)

- *Substructure specification by text line notation*
- *Support of property annotations, macros, R-groups, etc*





Limitations of Existing Standards

- **No provision of built-in extension mechanisms**
 - *No support of standardized property annotation (except SLN)*
 - *No support of "inline" test cases*
 - *Limited set of properties for annotation*
- **No built-in support for comments, documentation of queries, etc**





Limitations of Existing Standards

- **No mechanisms to validate queries prior to execution**
 - *Errors – both syntax and semantic ones – first seen when executing the query*
 - **Difficult and error-prone to input**
 - **Proprietary formats**
 - *Very few free/open source libraries and GUI tools*
- ⇒ **Need for a new definition or standard?**



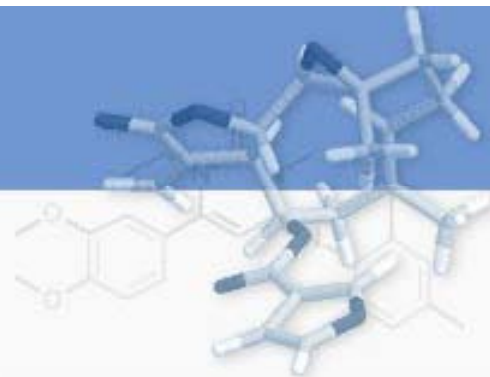
Requirements for New Substructure Representation Definition



- **Well defined representation of (sub)structures**
 - *Unambiguous interpretation*
 - *Clear document structure*
- **Support of (any) property annotation, query logic, etc**
 - *E.g., physicochemical properties, toxicity alerts, etc*
- **Support of comments, documentation, etc**



Requirements for New Substructure Representation Definition

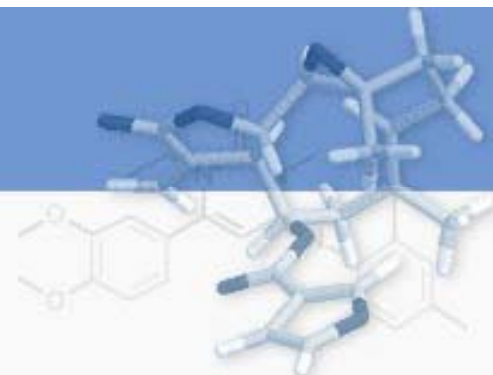


- **Built-in validation**
 - *Mechanisms to validate the syntax of queries*
 - *Test cases to validate the semantics of queries*
- **Conversion of queries into existing standards**
- **Built-in support for future extensions**
- **Non-proprietary, open format**

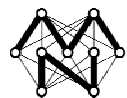
⇒ **XML (?)**



Advantages of XML-based (Sub)Structure Representation



- **Structured representation of structures, test cases, *etc***
- **Native support of**
 - *Syntax validation*
 - *Comments and documentation (extensible)*
- **Easy to**
 - *Transfer/exchange*
 - *Integrate into other XML-based languages*
 - *Extend and modify*
- **XML open standard as well**
 - *Large number of software available to work with*



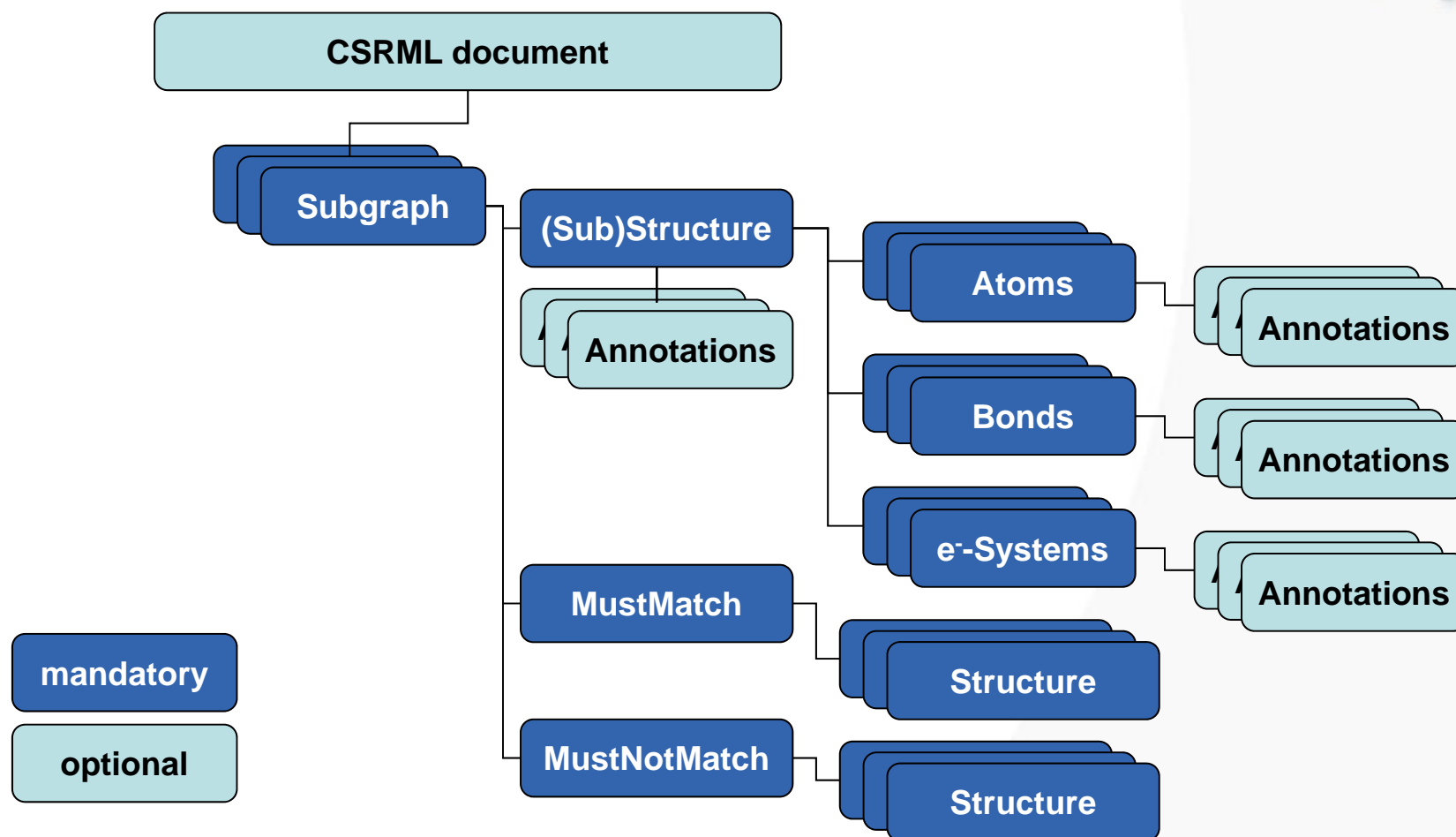
XML-based (Sub)Structure Representation



- Chemical Subgraph Representation Markup Language
- CSRML



CSRML Object Model



Model of Single Subgraph Representation



- Target (sub)structure, molecule, or a disconnected graph which represents the query
 - *Connectivity (atoms, bonds, e⁻-systems)*
 - *Annotated query features and other properties*
 - *Logical constructs*
- Test structure(s) that **MUST** match the target
- Test structure(s) that **MUST NOT** match the target



XML Grammar Definition for CSRML



- **Enables easy validation of query definitions**
 - *XML documents have to be well-formed*
 - *XML documents can be validated against data model (DTD or XSD)*
- **Additional checks to validate the query prior to processing**
 - *Referential integrity checks*
 - *Unique or distinct constraints*





Query with Properties – Example

```
<mol id="M1">  
<atomArray>  
  <atom id="A1" element="N/A" x="0" y="0">  
    <query feature="atomList">  
      <value>N</value>  
      <value>O</value>  
    </query>  
    <query feature="piCharge" logic="AND">  
      <range>  
        <min>-0.6</min>  
        <max>-0.1</max>  
      </range>
```

...





Query with Properties

- **Easy accessible annotated query features**
 - **Easy nesting and logical combination of query features**
 - **Automatic validation of query syntax by XML parser**
 - *Based on XML schema (grammar)*
 - *Partial validation of query semantics*
 - **No chemical validation at this step!**
- ⇒ **The more checking is done by XML parser, the less checking has to be done by implementing library!**



Annotation Model



■ Example of definition for a CSRML annotation

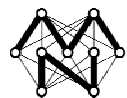
```
<annotation domain="bond" featureKey="ringBond"
  dataType="xsd:boolean" implementation="M_RING_BOND_IMPL"
  priority="2" severity="skip">
  <label>
    Ring bond
  </label>
  <description>
    Bond in ring system; bond order disregarded.
  </description>
  ...
</annotation>
```





Storage of CSRML Queries

- **Storage as XML documents**
 - *Single or multiple queries in a single document*
- **Storage in (XML) databases**
 - *Substructure searches in chemistry-aware XML databases*
- **Integration into other XML-based formats, e.g., ToxML**



Exchange of CSRML Queries



- **Transfer between different applications as XML documents**
 - *Regular files*
 - *Internet (SOAP, HTTP, ...)*
- **Conversion into existing formats**
 - *Omission or separate export of not-supported features*
 - *Transformation into query depicts (SVG)*
- **Conversion from existing formats**
 - *E.g., from SMARTS to CSRML*



Current Status

- **First draft of CSRML definition**
 - *Data model & schema design*
 - *Design of annotation model*
 - *Default set of query features*
- **Beta version of reference implementation**
 - *LGPL or similarly licensed library to support I/O and class structures for query documents, query objects and features*



Next Steps

- **Development of graphical input tool**
 - *Chemical Subgraph Editor, CSE*
- **Publishing everything on the Web**
 - *Announced via Newsletter / RSS feed*



CSRML – Summary



- **Universal and extensible platform for specifying advanced substructure queries**
 - *Connectivity (atoms, bonds, e⁻-systems)*
 - *Annotated query features and other properties*
 - *Logical constructs*
- **Open standard for easy exchange of substructure queries between different applications and databases**
- **Encourage developers to use and distribute CSRML**



Acknowledgements



- **Molecular Networks (co-authors)**

- *Bruno Bienfait, Johann Gasteiger, Thomas Kleinöder, Joerg Marucszyk, Oliver Sacher, Aleksey Tarkhov, Lothar Terfloth*

- **Chihae Yang (co-author)**

- *Discussions about the chemical subgraph definition*

- **US FDA CFSAN**

- *Kirk Arvidson
(Contract for development of the Chemical Subgraph Editor)*



Thank You!

- www.molecular-networks.com

