SKETCHY SKETCHES: HIDING CHEMISTRY IN PLAIN SIGHT

John May, Daniel Lowe and Roger Sayle

NextMove Software Ltd, Cambridge, UK.

Introduction

Sketch formats, such as ChemDraw, are often preferred for publication as they provide more freedom to the author. This freedom comes at a cost and what is displayed to the user may not coincide with the stored chemical representation. This can lead to incorrect structures if naïvely exported. More than **24 million** ChemDraw files are available from the U.S. Patents, using these files a high fidelity converter/interpreter has been developed that is able to extract and associate more chemistry at higher precision.

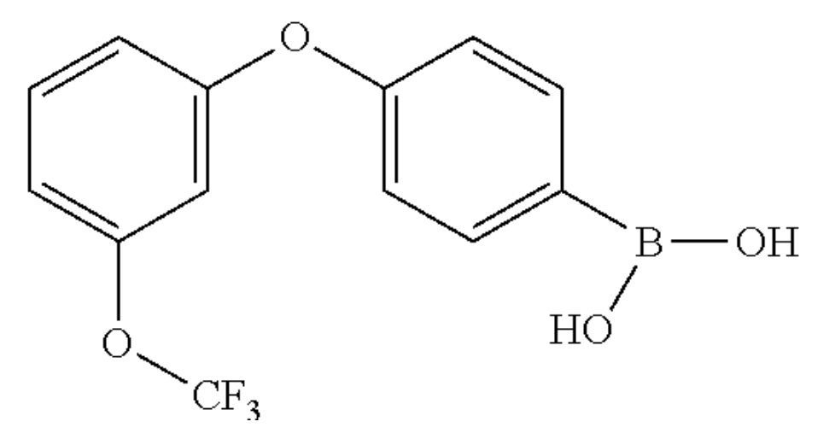
Content Categorisation

A categorisation scheme has been developed to organise and characterise diverse sketch content. Sketches are assigned a content type (1), level of detail (2), and a confidence that the interpretation is correct (3):

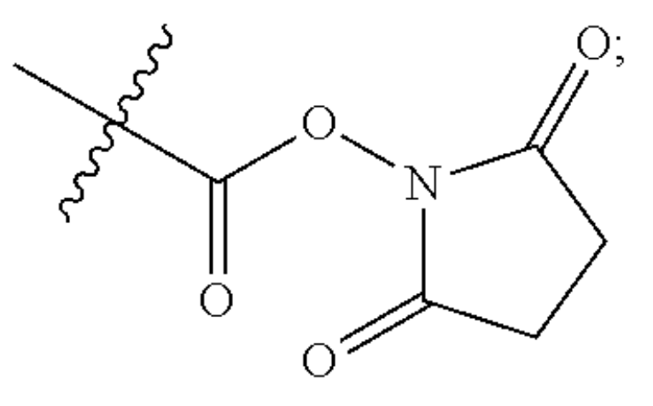
1 Type: Molecule|Reaction|Substituent|NoConnectionTable
2 Detail: Specific|Generic|Unknown
3 Confidence: High|Medium|Low

Substituent is assigned if the entity contains an attachment point and are commonly found in Markush claims. **NoConnectionTable** is assigned to non-chemical sketches and when an extracted connection table is not believed to be real (e.g. chessboardanes[1]).

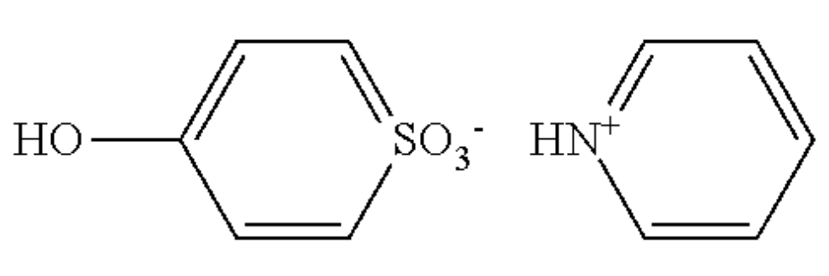
Generic is assigned when substituent, positional, or frequency variation is present. This primarily includes Markush cores but also polymers. **Specific** is assigned when all atom labels have been resolved. An entity with unresolved labels and no variation are **Unknown**, the labels are usually partial, ambiguous, or incorrect condensed formula.



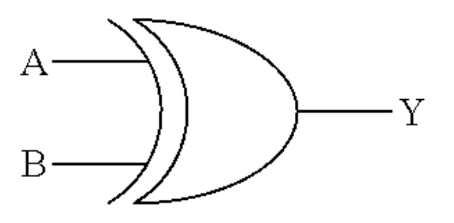
Molecule/Specific/High
US 2015/0344503 C00668



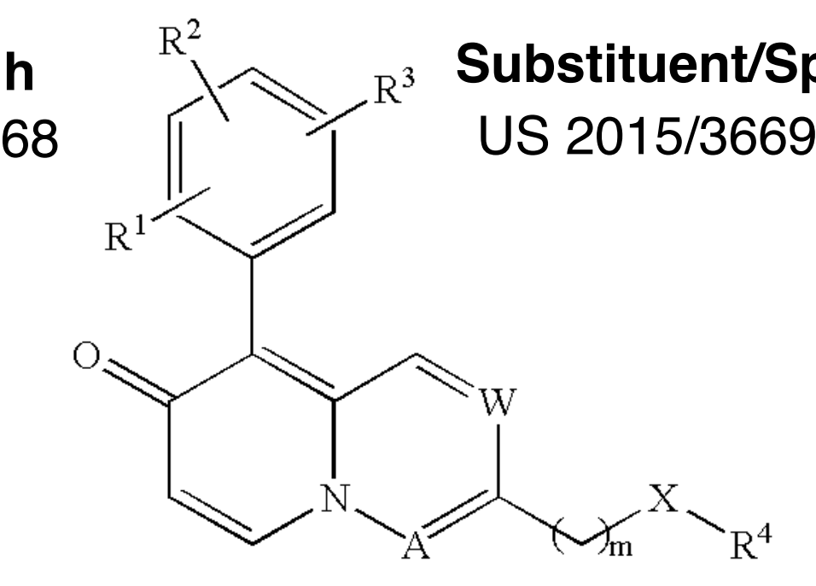
Substituent/Specific/High
US 2015/366987 C00008



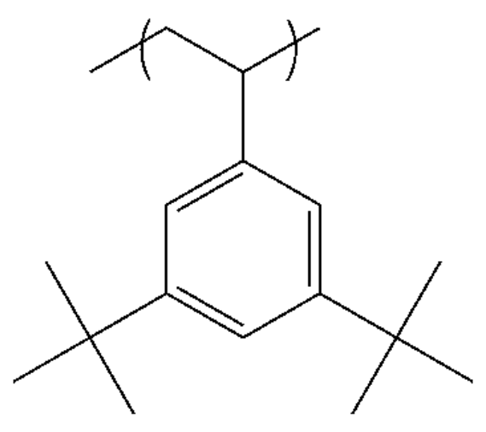
Molecule/Unknown
US 2015/378260 C00003



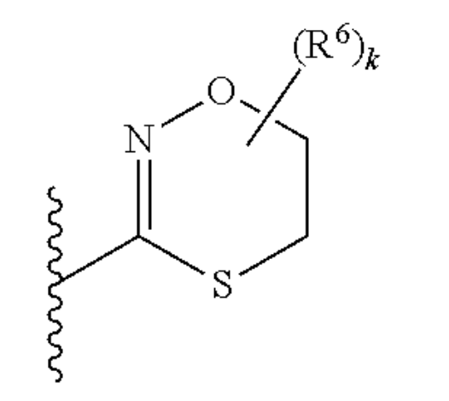
NoConnectionTable
US 2015/370940 C00005



Molecule/Generic/High
US 2007/129372 C00001



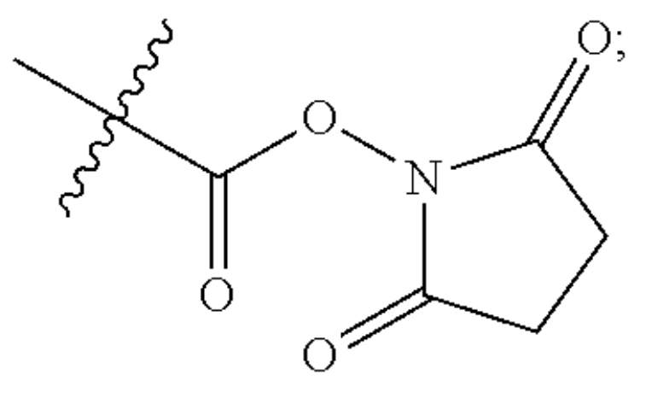
Molecule/Generic/High
US 2015/378257 C00118



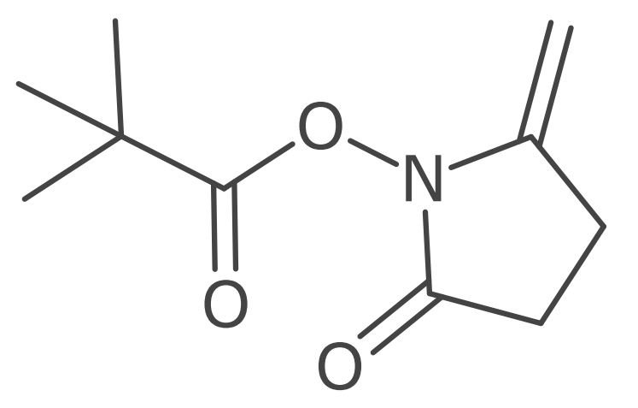
Substituent/Generic/High
US 2015/368236 C00019

Examples of Improved Interpretation

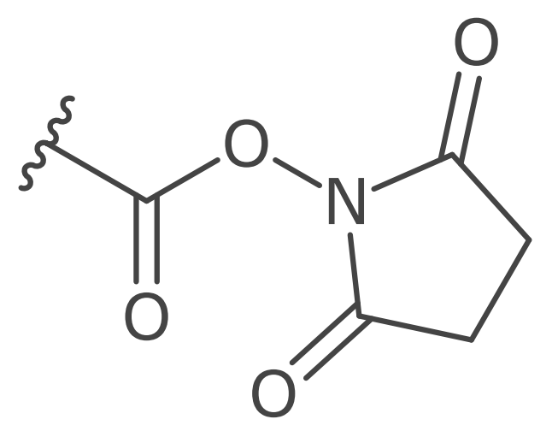
Attachment Points and **Punctuation Striping**. Attachment points may be misrepresented as *tert-butyl*, *iso-propyl,* or *methyl* groups. Punctuation can be used to separate structure lists, if appended to an atom label that label may not be understood.



US 2015/366987 C00008

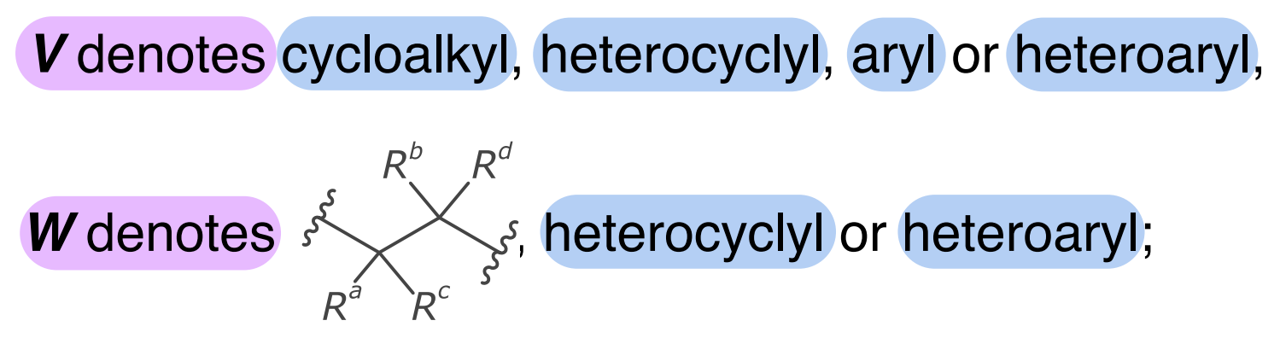


SCHEMBL6726025 / CID 69879541



This Work

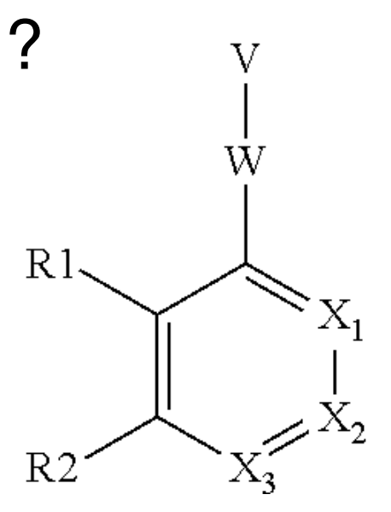
Substituent Labels. Markush definitions are recognised in the patent text and used to resolve ambiguities. W for Tungsten? V for Vanadium?



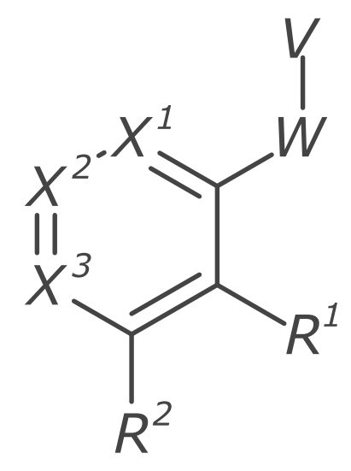
V denotes cycloalkyl, heterocyclyl, aryl or heteroaryl,

W denotes [structure with R^a, R^b, R^c, R^d], heterocyclyl or heteroaryl;

US 2015/376182 [Claim 1] Document

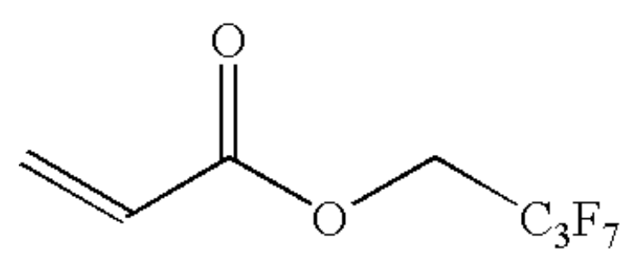


US 2015/376182 C00030

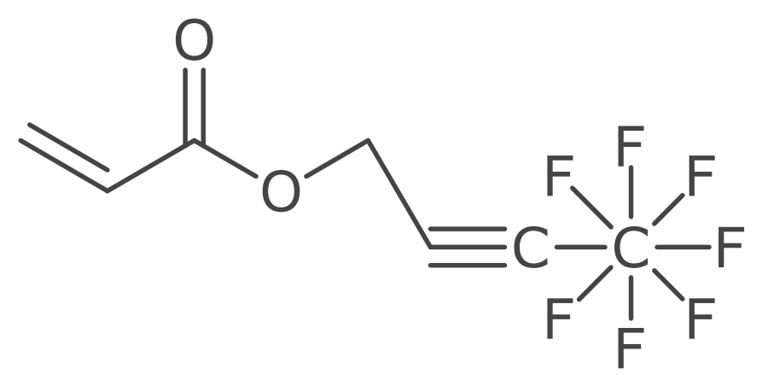


This Work

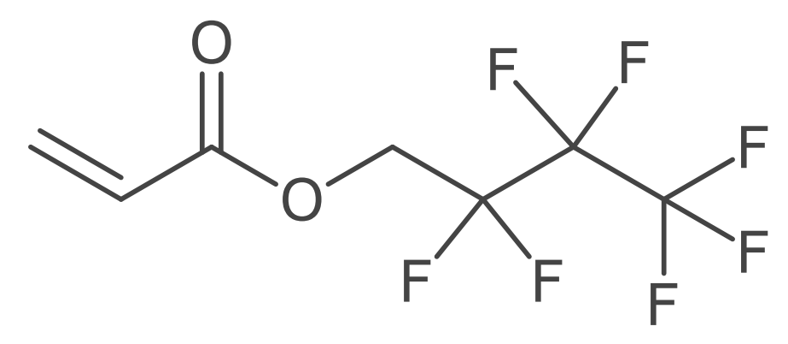
Condensed Formula. Formulas are reinterpreted with a more accurate algorithm. Some labels previously defined will be unrecognised (**Molecule/Unknown**) whilst others are corrected.



US 2004/101442 C00025



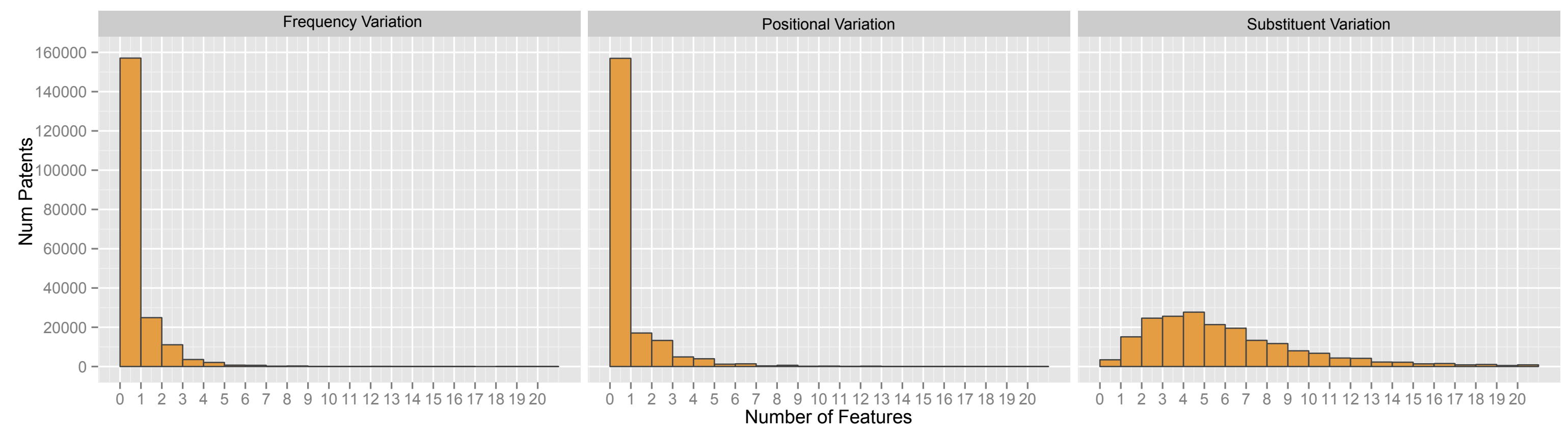
Molfile Contents



This Work

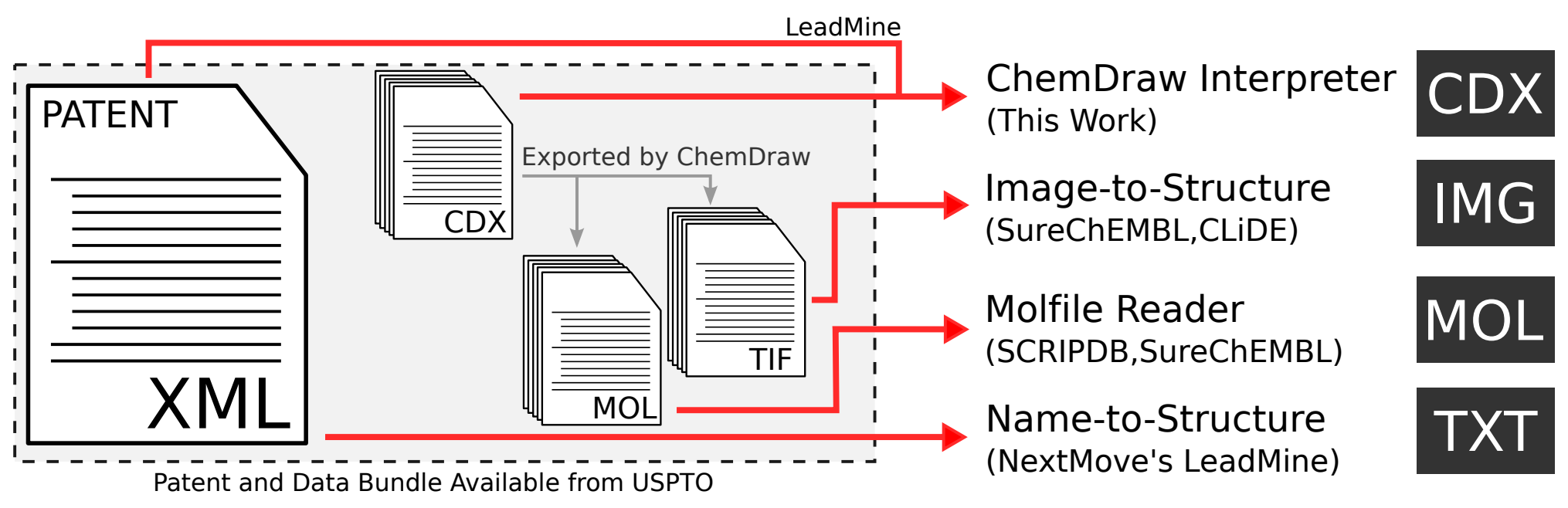
Generic Feature Survey

Generic features are recognised and captured in the output format. This allows a large scale summary of the features over U.S. Patents since 2001. Using the first sketch from the claims of each patent, **200,900** generic molecules with high/medium confidence can be extracted. Their feature distribution is:



Evaluation

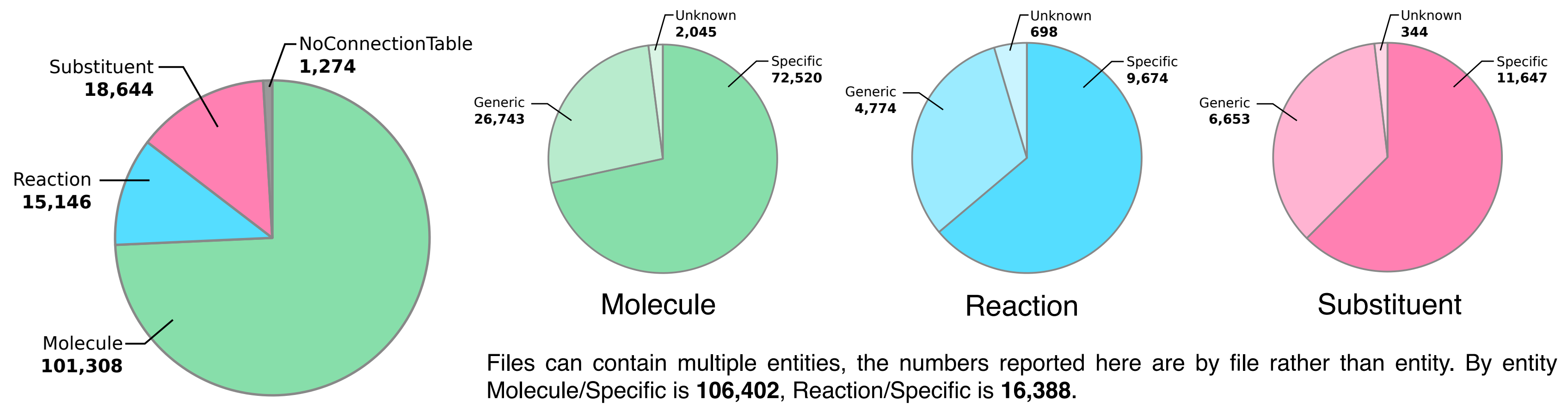
Using **2,528**[3] U.S. Patent Applications published in Dec 2015 the structures extracted by different methods are compared:



ChemDraw Interpreter (This Work)
Image-to-Structure (SureChEMBL,CLiDE)
Molfile Reader (SCRIPDB,SureChEMBL)
Name-to-Structure (NextMove's LeadMine)

SCRIPDB[2] and SureChEMBL[3] include structures from directly converting bundled Molfiles. SureChEMBL also includes structures obtained from raster images using image-to-structure. The raw data is used in this evaluation, SureChEMBL filters structures before indexing in the public interface.

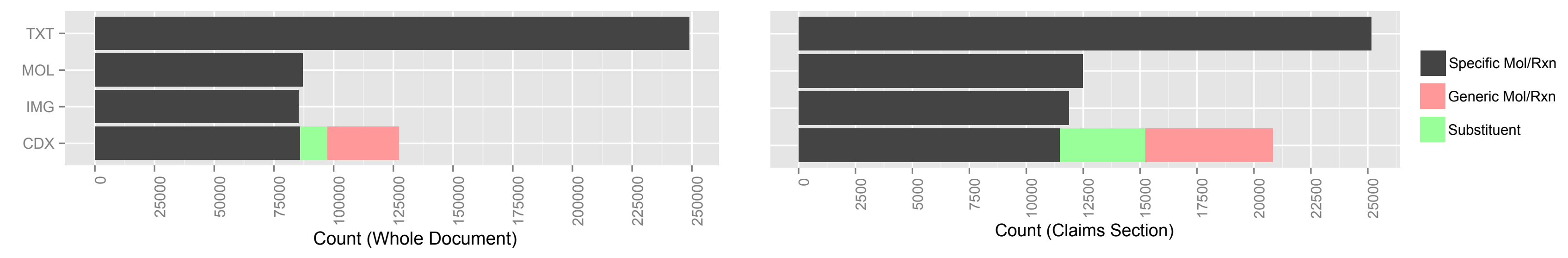
File Content. The **2,528** patents contain **136,372** ChemDraw files with the categories:



Files can contain multiple entities, the numbers reported here are by file rather than entity. By entity Molecule/Specific is **106,402**, Reaction/Specific is **16,388**.

Normalisation. Output from CDX and TXT are split into individual components (e.g. splitting reactions and salts) to make it comparable to IMG and MOL which had already been split. Only components with ≥12 bonds between heavy atoms are considered.

Unique Component-Document Associations



Overlap. Structures are compared using Canonical SMILES (OEChem) of the individual components on a per document basis. For CDX only content categorised as specific molecules and reactions is counted.

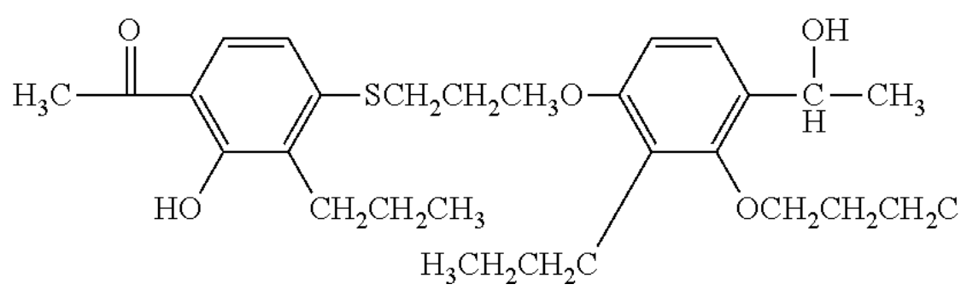
Whole Document

	Not Found in TXT	Also Found in TXT
CDX	49,119	36,829 (**42.8%**)
IMG	49,836	35,545 (**41.6%**)
MOL	58,169	28,926 (**33.2%**)

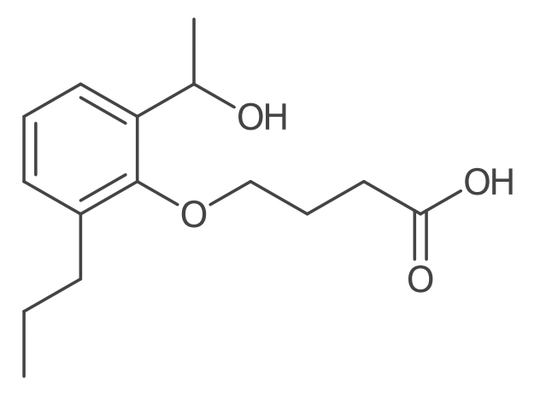
Claims Section

	Not Found in TXT	Also Found in TXT
CDX	10,664	800 (**6.9%**)
MOL	11,078	796 (**6.7%**)
IMG	11,783	689 (**5.5%**)

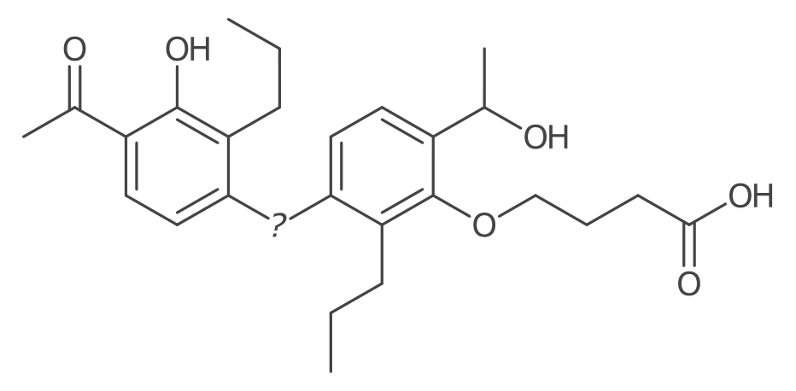
Differences. The biggest contribution to differences between the CDX, IMG and MOL are misinterpreted substituents and condensed formulae (see. left panel). Some differences are a result of human input error.



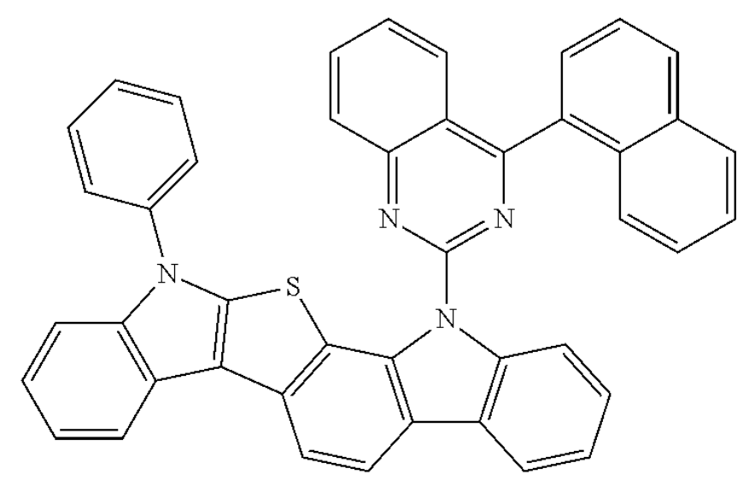
US 2015/342913 C00015



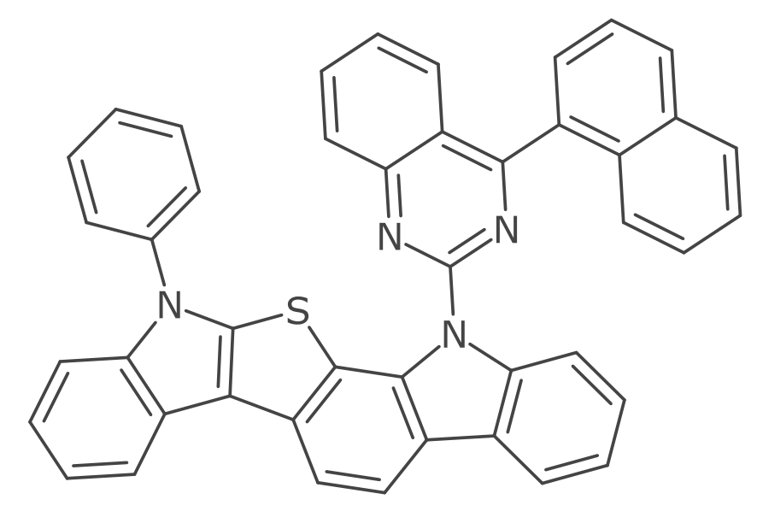
SCHEMBL17244875 / CID 118501484



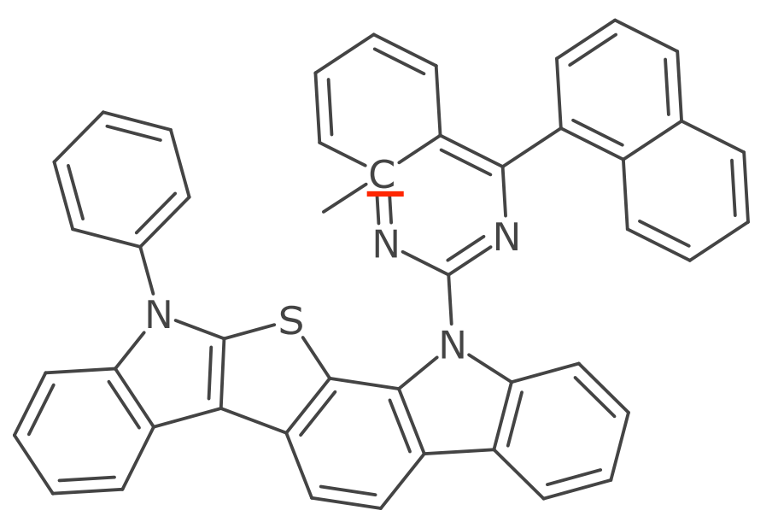
This Work



US 2015/380663 C00381



SCHEMBL17366912 / CID 118602343



This Work

Conclusion

Interpreting ChemDraw files shows closer agreement with chemistry mined from the patent text than image-to-structure and much closer agreement than Molfile conversion. Reading the ChemDraw files, rather than the derived Molfiles, allows extraction of complex concepts such as reaction schemes and generic structures. The categorisation helps identify problematic structures and highlights areas to be improved. A major advantage of the ChemDraw interpretation is speed, taking **5h** on a single-core to process the back catalogue of **24+ million**.

Acknowledgements

We would like to thank MineSoft Ltd for funding and George Papadatos for providing bulk downloads of structures from SureChEMBL for the Dec 2015 Patent Applications.

Bibliography

1. https://cdsouthan.blogspot.co.uk/2015/07/chessboardane-and-other-strange-patent.html
2. Heifets, A and Jurisica, I. SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents. ***Nucl. Acids. Res***. 2012
3. Papadatos, G *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. ***Nucl. Acids. Res***. 2015
4. All U.S. patents applications in Dec 2015 with chemical complex work units.

NEXTMOVE SOFTWARE

www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
UK CB4 0EY