SKETCHY SKETCHES II

ADVANCES IN EXTRACTING REACTION INFORMATION

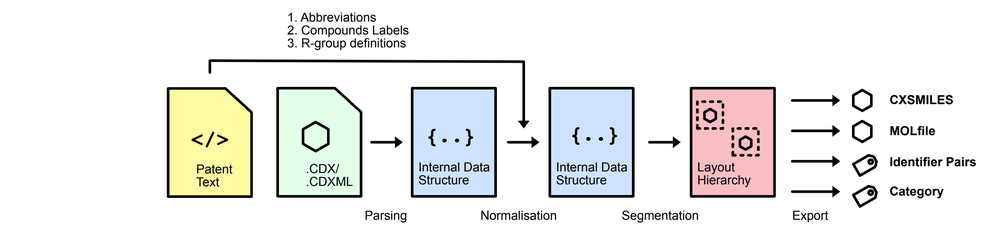
John May, Ingvar Lagerstedt and Roger Sayle
NextMove Software Ltd, Cambridge, UK.

Introduction

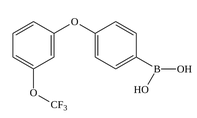
More than **44 million** ChemDraw files are available from the U.S. Patents (2001-2023) as supplementary data. These can be a source high-quality and diverse chemical information that can be extracted. ChemDraw files capture chemical data but their flexibility means they are more akin to a vector graphic format, clean-up and interpretation is needed to extract maximum value[1]. Pistachio[2] is a database of reactions and includes **3.6 million** reactions (2023 Q1) extracted from these sketches. We report on recent enhancements to the processing of the sketches to extract more and better quality reaction data.

Overview

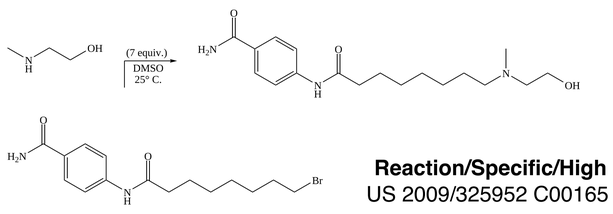
ChemDraw files are processed and normalized using a rule-based system, using additional context information from the patent text. Generic features like R-groups, repeat groups and variable-attachment are captured as CXSMILES.



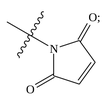
A content/detail/confidence category is assigned to allow quick filtering depending on the use case:



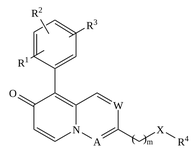
Molecule/Specific/High
US 2015/0344503 C00668



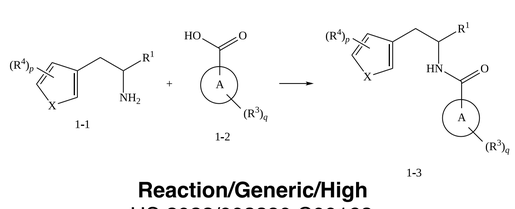
Reaction/Specific/High
US 2009/325952 C00165



Substituent/Specific/High
US 2015/366987 C00008



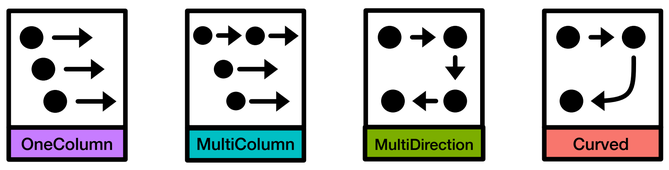
Molecule/Generic/High
US 2007/129372 C00001



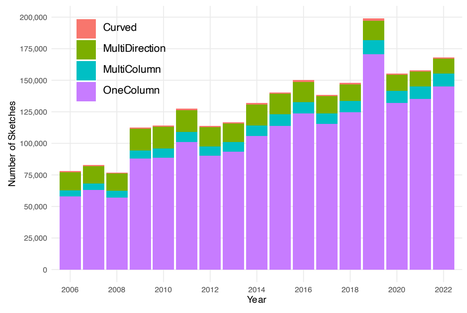
Reaction/Generic/High
US 2022/002280 C00162

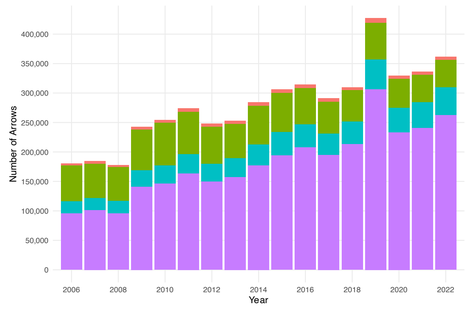
Reaction Layouts

The reaction layout in a sketch can be categorized into the one of four categories depending on: how many, the direction, and type of arrow:



ChemDraw files from USPTO applications (2001-2022) were analysed and categorised based on layout.





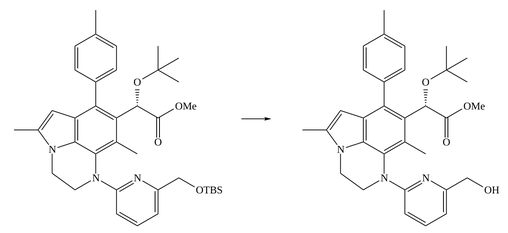
Overall the counts are:

	Sketches		Arrows	
Curved	21,295	0.9%	108,862	2.0%
MultiDirection	300,092	12.3%	1,220,050	22.9%
MultiColumn	157,754	6.5%	672,208	12.6%
OneColumn	1,956,703	80.3%	3,334,708	62.5%
Total	2,435,844		5,335,828	

Previously, only reactions in **OneColumn** layouts were extracted, this is **80%** of the files but more complex layouts are more likely to contain multiple reactions. When the counts are adjusted based on the *number of arrows* in each sketch we see that it is only **62%** that laid out in a single column. We do not currently include **Generic** Reactions (i.e. with R-groups) in the Pistachio data, if we count just those with **Specific** reactions the counts are as follows:

	Sketches		Arrows	
Curved	6,117	0.4%	35,421	1.0%
MultiDirection	105,316	6.2%	470,437	13.4%
MultiColumn	135,351	8.0%	559,204	15.9%
OneColumn	1,447,768	85.4%	2,455,655	69.7%
Total	1,694,552		3,520,717	

Reaction Compound Labels



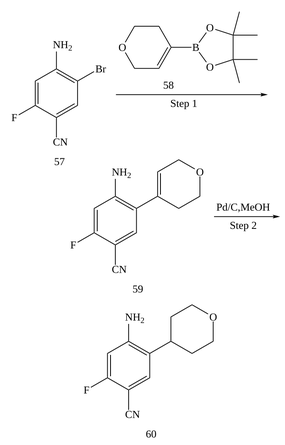
US 2018/0162876 A [0936]

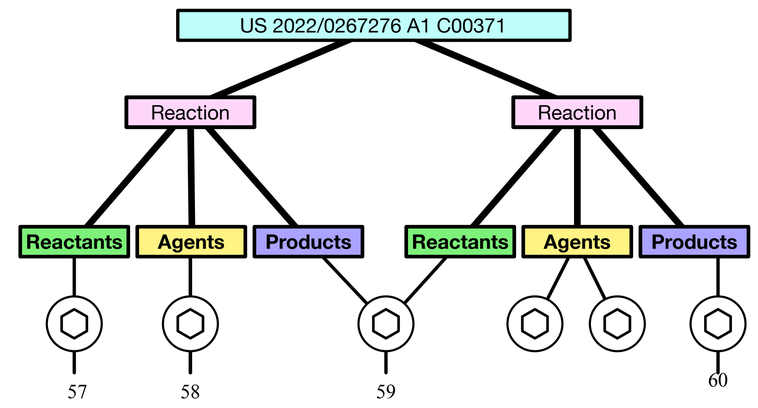
Step 2

To **Compound 152** (134 mg, 0.209 mmol) in THF solution (1.34 mL) was added 0.92 mol/L TBAF in THF solution (0.681 mL, 0.626 mmol), the mixture was stirred at room temperature for 1.5 hours. Water (20 mL) was added thereto, and the mixture was extracted with ethyl acetate (30 mL). The organic layer was washed with water (20 mL) and saturated brine (20 mL), and dried over anhydrous magnesium sulfate. After concentration, the residue was purified by silica gel chromatography (hexane-ethyl acetate) to obtain **Compound 153** (106 mg, 96% yield) as brown foam.

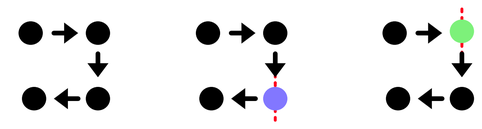
The combination of information from sketches, text, and tables can yield more than would be possible in isolation[3]. An experimental procedure paragraph may reference chemical structures which are **only** described in a sketch (i.e. no systematic name). By extracting the **compound and identifier** pairs the reaction above can be fully extracted.

Results

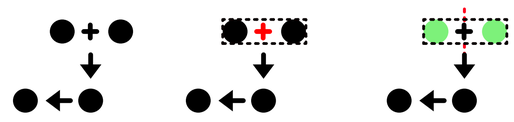




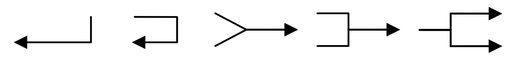
It was previously only possible to export a list of (CX)SMILES from a sketch, the category was also assigned globally per-file. A segmentation step was added which partitions and groups the components of the sketch into a hierarchical *view*. This allows more flexible export options including the extraction of all compound and identifier pairs and simplifies the handling of complex schemes where a component is used multiple times in different contexts.



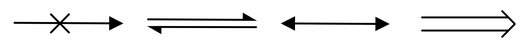
Support was added for processing **MultiColumn** and **MultiDirection** reaction layouts. The reactant/products are determined by casting a ray out from the tail/head of the arrow and identifying which structure(s) are hit. Agents are identified by a zone either side of the arrow. A useful concept when determining proximity is to scale parameters relative to *bond length.*



Component bounds sandwiching a plus (+) are merged to handle multiple reactants and products, often it can be ambiguous.



Relatively rare but problematic are multi-part arrows. There arrows may have multiple heads or tails. We detect and store multi-part arrows as a set of tails, connectors, heads. The heads and tails are then used to cast rays and determine the reactants and products.



Different arrow styles indicating NoGo, Equilibrium, Resonance and Retrosynthetic are now captured exported as a Data Sgroup in CXSMILES.

Conclusion

The extraction of data all sketches from USPTO applications (2001-2022) takes ~**5 hours** (Intel i7-6900K). Previously **2.32 million** (1.23M unique) specific reactions were extracted, with the improvements outlined above **27%** more reactions are now extracted (**2.95 million**, 1.53M unique). **302,646** were novel and not previously seen before, however many of these will be "magic/alchemy" **A → B** reactions where functional groups spontaneously appear. NameRxn[4] was used to classify reactions types and the number of recognized reactions in these set increased from 1.59M (68.5%) to 1.98M (67%). We have mainly focused on patent data but improvements in sketch extraction are also useful for ELN export/analysis.

References

1. John May, Daniel Lowe, Roger Sayle. Sketchy Sketches: Hiding Chemistry in Plain Sight. ***7th Sheffield Conference on Chemoinformatics***. Jul 2016. Poster Available: https://nextmovesoftware.com/talks.html
2. John Mayfield, Ingvar Lagerstedt and Roger Sayle. Pistachio. ***NIH Virtual Workshop on Reaction Informatics***. May 2021. Available: https://nextmovesoftware.com/talks.html
3. Daniel Lowe, John May, Roger Sayle. Sketchy sketches: Hiding chemistry in plain sigh. ***252nd ACS National Meeting & Exposition***. Aug 2016: https://nextmovesoftware.com/talks.html
4. https://www.nextmovesoftware.com/namerxn.html

NEXTMOVE SOFTWARE

www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 320)
Cambridge Science Park
Milton Road, Cambridge
UK CB4 0WG