Sketchy sketches: Hiding chemistry in plain sight

Daniel Lowe, John May and Roger Sayle

NextMove Software

Cambridge, UK

252nd ACS National Meeting, Philadelphia PA, USA 25th August 2016

OVERVIEW

- Motivation for mining sketches
- Tricky cases when interpreting sketches
- Combining text-mining with sketch interpretation

MOTIVATION

- The chemical matter discussed in a document is often critical in determining if it is relevant
- Chemical sketches are not indexed by text-mining
- If chemical sketches can be made “chemistry searchable” this helps with:
  - Identifying relevant documents
  - Prior-art searching

WHAT INPUT SHOULD BE USED?

- Image to structure techniques tools (OSRA/Clide/Imago etc.) work with images
  - Introduces OCR errors on atom labels
  - Crossing bonds present difficulties
  - Often can find chemistry in non-chemical images

- Where the sketch is available as a “computer-readable” format can these issues be avoided?

SOURCES OF CHEMDRAW SKETCHES

- United States patents (2001-present)
  - Over 24 million ChemDraw files!
- Journal articles (albeit in most cases not publicly accessible)
- Thesis (albeit only if the original manuscript is made available)

AMBIGUOUS SYMBOLS

Symbol	Naïve interpretation	Possible meaning
Ac	Actinium	acetyl
Ar	Argon	aryl
B	Boron	Generic label
D	Deuterium	Generic label
P	Phosphorus	Generic label
Ra	Radium	Generic label
Rb	Rubidium	Generic label
V	Vanadium	Generic label
W	Tungsten	Generic label
Y	Yttrium	Generic label

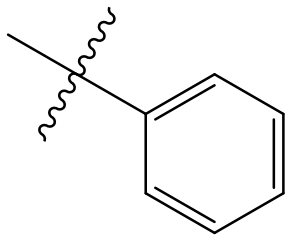
AMBIGUOUS SYMBOLS-CONT.

- Can disambiguate with text-mining:
  - E.g. “B is aryl or heteroaryl”, “B is boron”

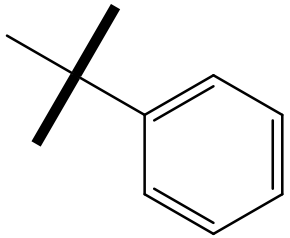
- Can disambiguate by connectivity e.g. is a Yttrium atom with one bond likely?

ATTACHMENT POINT REPRESENTATION

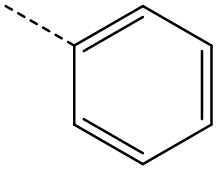
(Below: naïve interpretation)



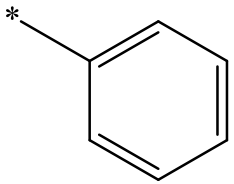
tert-butyl



tert-butyl

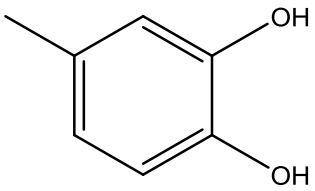


methyl

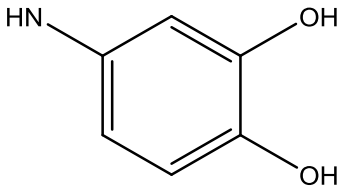


methyl

IMPLICIT ATTACHMENT POINT REPRESENTATION



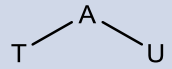
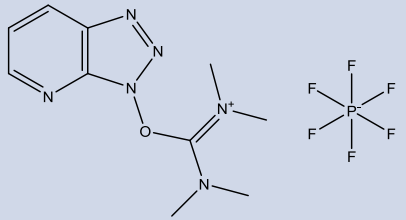
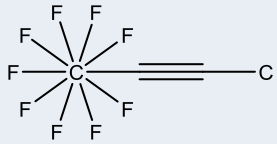
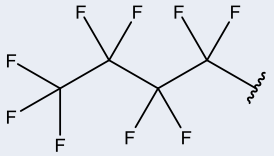
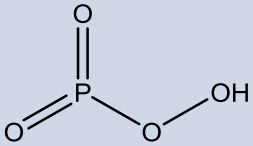
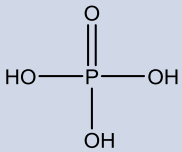
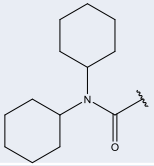
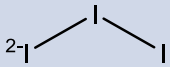
Unlabelled methyl



Under-valent atom

Sketch parser needs to be given a hint that the sketch is a substituent definition!

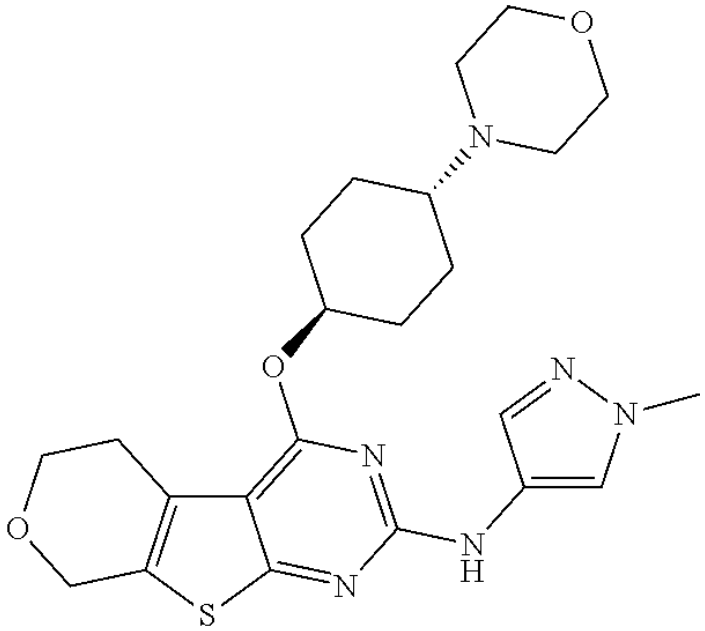
FORMULA INTERPRETATION

Input	ChemDraw 15	This work
HATU		
C_4F_9		
H_3PO_4	 HH	
CON(cHex)2	No result	
III-2		No result

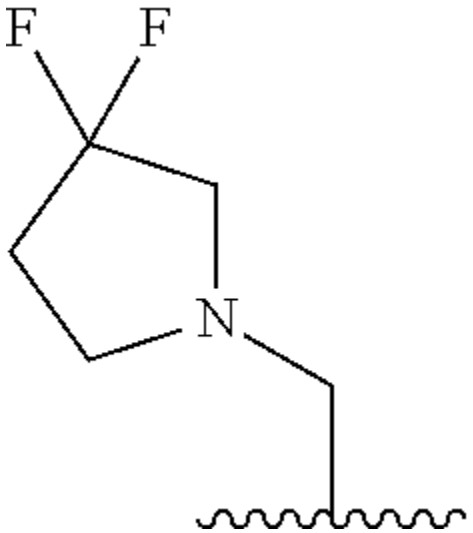
CATEGORISATION

1)	Sketch Type	Molecule Reaction Substituent No connection table
2)	Detail	Specific Generic Unknown
3)	Confidence in interpretation	High Medium Low

EXAMPLES OF CATEGORISATION



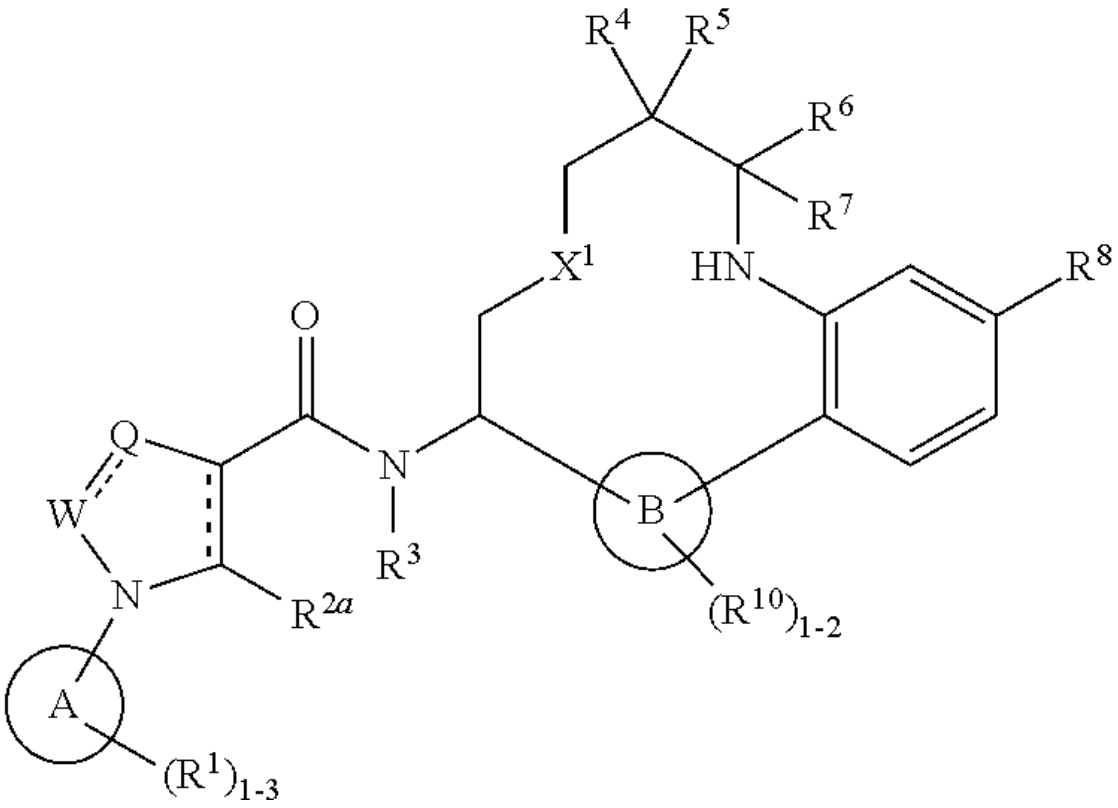
Molecule/Specific/High



Substituent/Specific/High

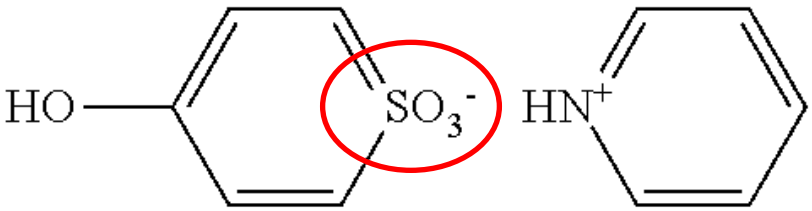
EXAMPLES OF CATEGORISATION

(II)



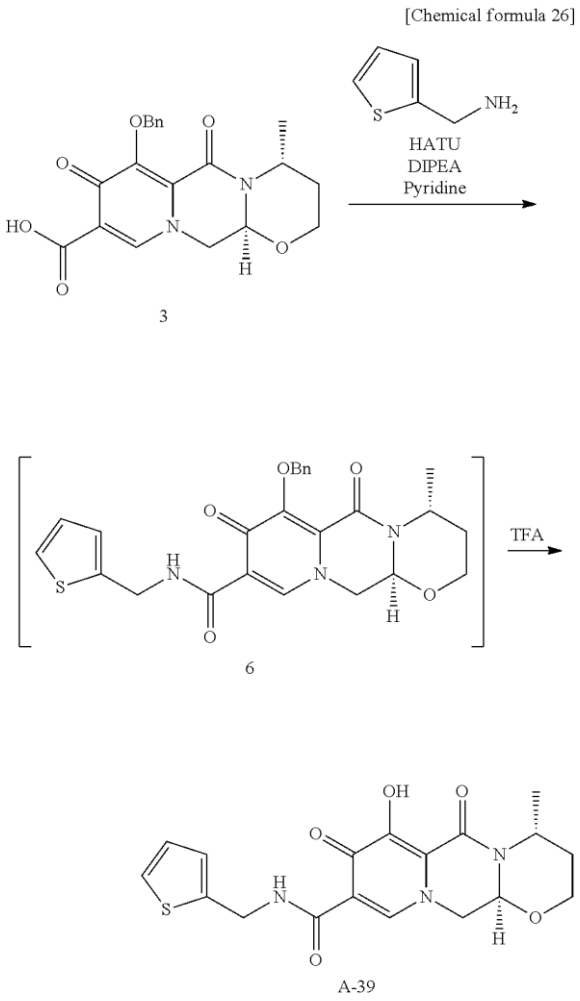
Molecule/Generic/Low

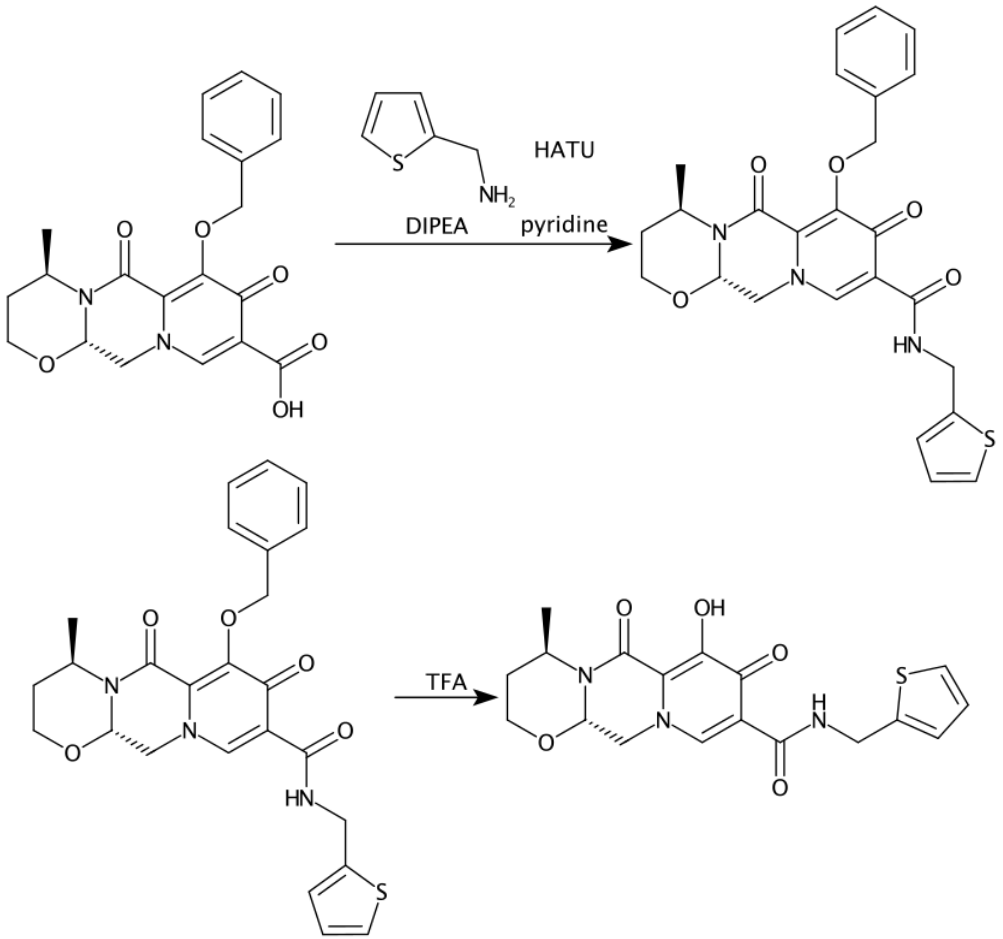
EXAMPLES OF CATEGORISATION



Molecule/Unknown

Formula uninterpretable so can't know for sure whether molecule is specific or generic!





Reaction/Specific/Medium

Two reactions extracted

EXAMPLES OF CATEGORISATION

$Q =$

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	57
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	54	71	194	113	82
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

No Connection Table

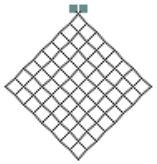
PubChem OPEN CHEMISTRY DATABASE

Compound Summary for CID 21040251

PUBCHEM > COMPOUND > CID 21040251

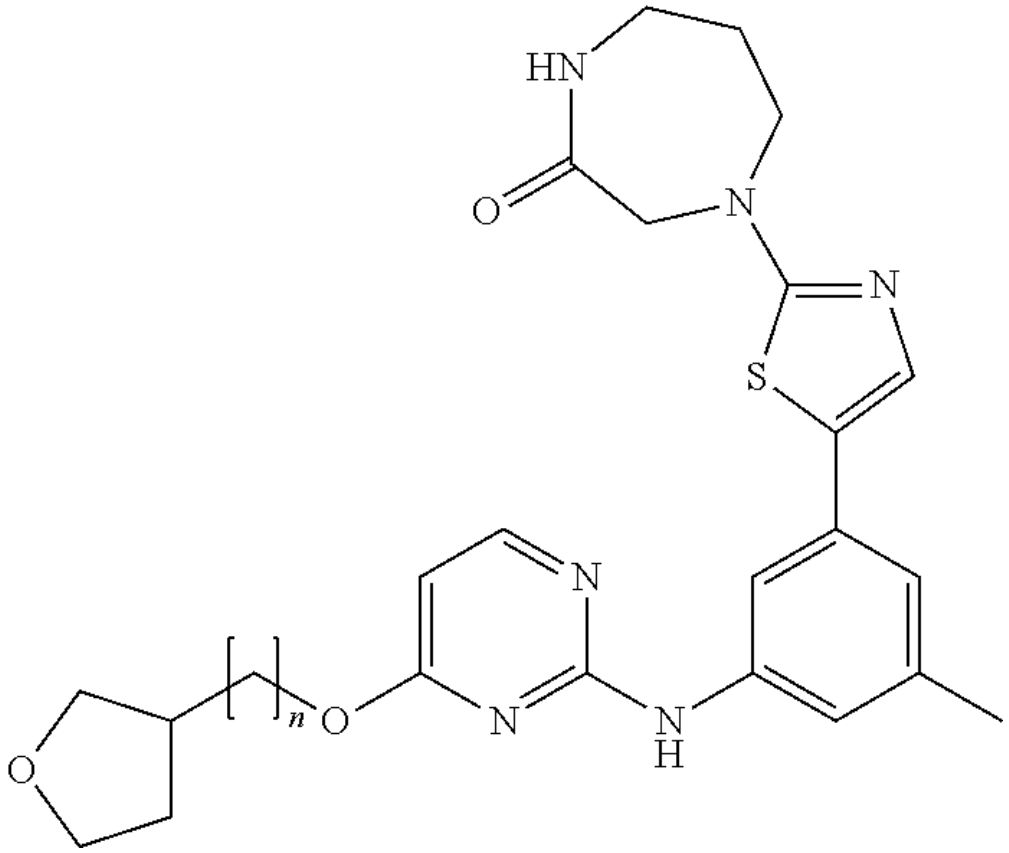
CID 21040251

Patents

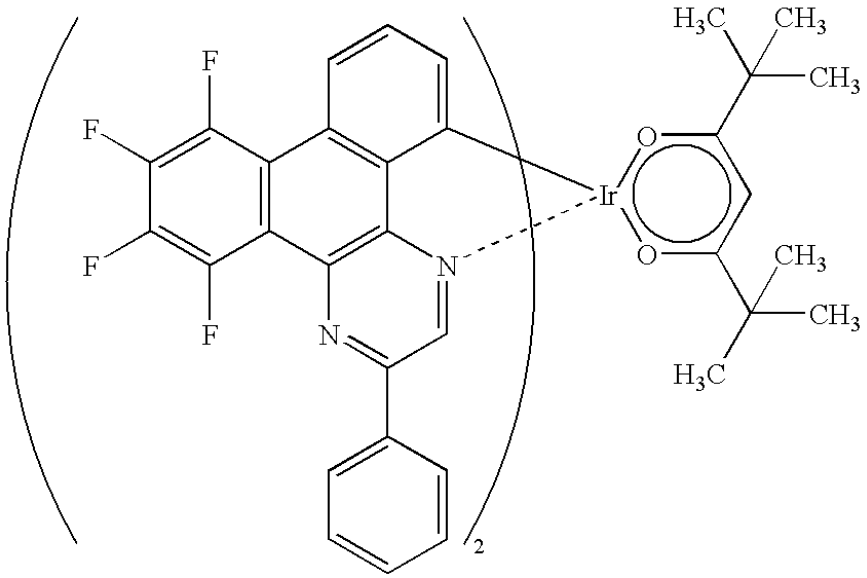


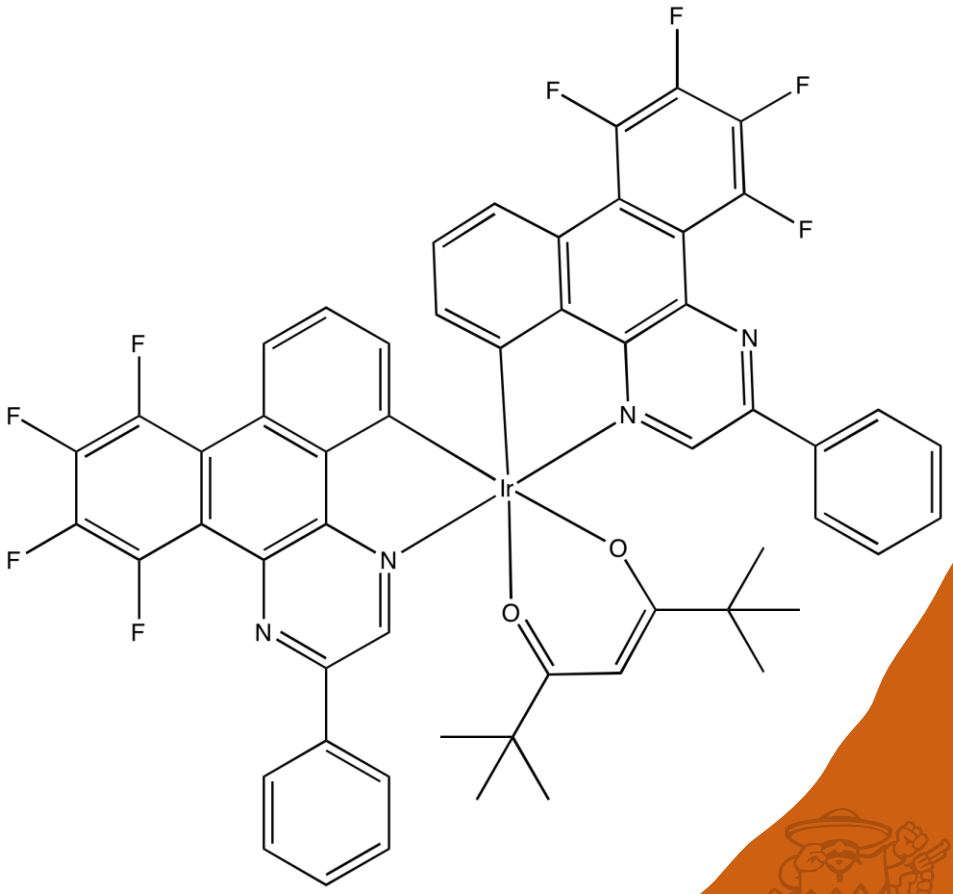
PubChem CID:	21040251
Molecular Formula:	$C_{81}H_{36}$
Molecular Weight:	1009.15254 g/mol
InChI Key:	LQFILANDZPXSQB-UHFFFAOYSA-N
Modify Date:	2016-08-07
Create Date:	2007-12-05

REPEATED GROUP DETECTION



ELECTRON LOCALISATION

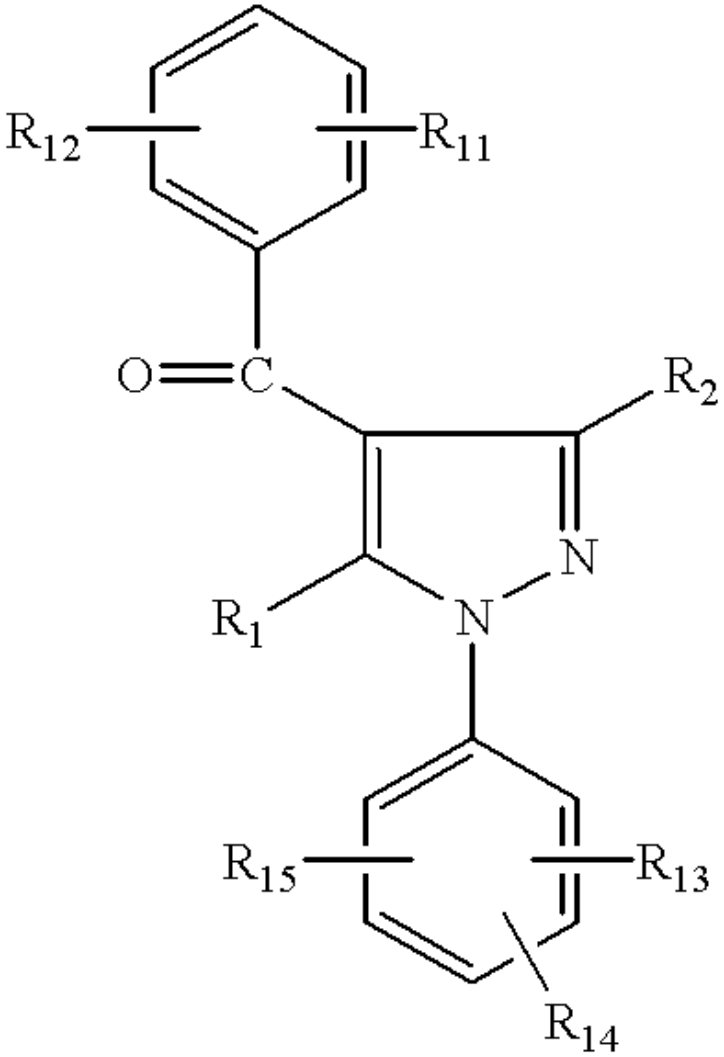




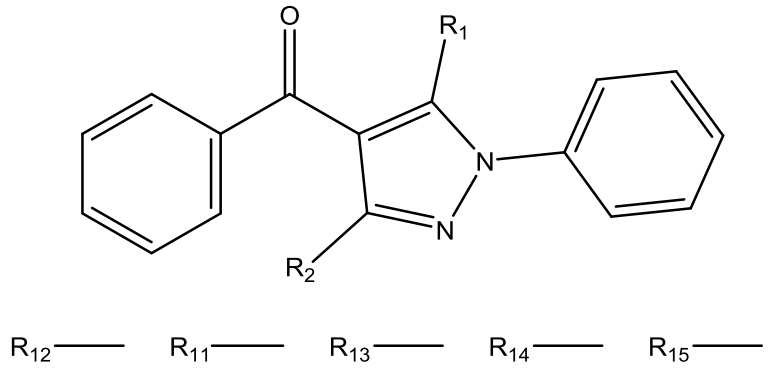
Some delocalised systems don't yield valid SMILES

→convert to localised system

POSITIONAL VARIATION

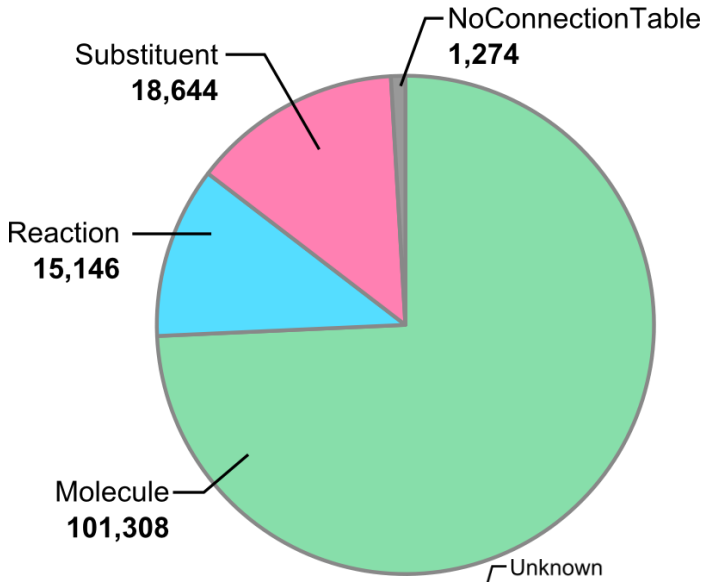


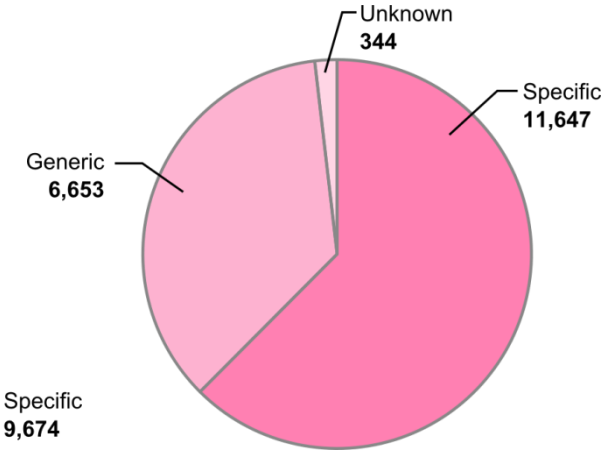
Naïve export:



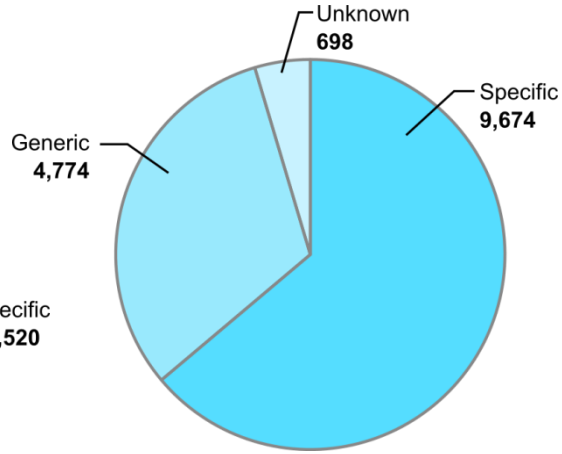
Association of R-groups with ring atoms captured

EVALUATION
(DEC 2015 US PATENT APPLICATIONS)

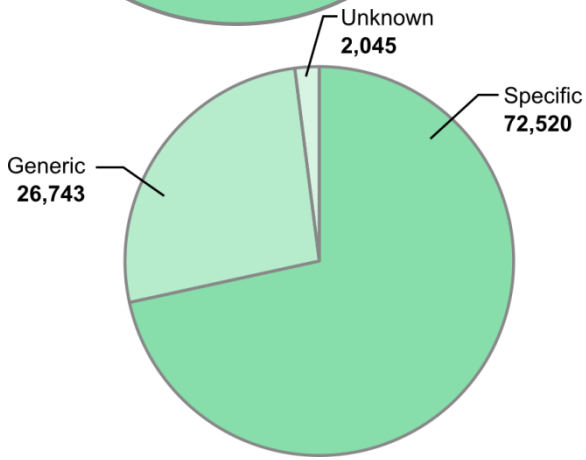




Substituent



Reaction



Molecule

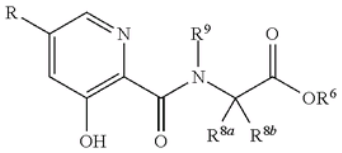
COMPARISON WITH OTHER APPROACHES

	Not found by text-mining	Also found by text-mining
This work (parsing CDX files)	49,119	36,829 (**42.8%**)
Image to structure*	49,836	35,545 (**41.6%**)
ChemDraw exported Mol files*	58,169	28,926 (**33.2%**)

*Results courtesy of the SureChEMBL database

EXEMPLIFIED COMPOUND R-GROUP TABLES

Additional compounds of Formula (I) the formula:



wherein non-limiting examples of R^{8a}, R^{8b}, R^9 and R^6 are further described herein below in Table.

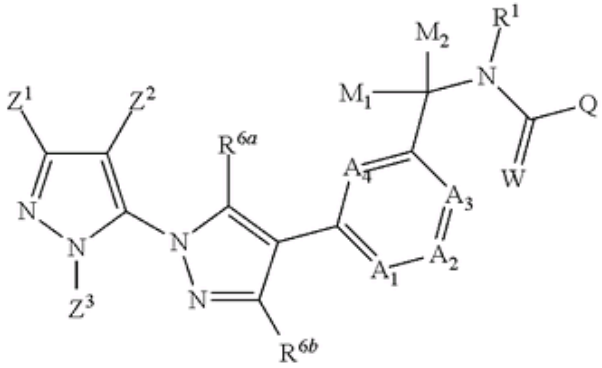
TABLE XIV

No.	R	R^6	R^{8a}	R^{8b}	R^9
559	3-chlorophenyl	—H	$-CH_3$	—H	—H
560	3-chlorophenyl	$-CH_3$	$-CH_3$	—H	—H
561	3-chlorophenyl	—H	$-CH_3$	$-CH_3$	—H
562	3-chlorophenyl	$-CH_3$	$-CH_3$	$-CH_3$	—H
563	3-chlorophenyl	—H	$-CH_3$	$-CH_3$	$-CH_3$
564	3-chlorophenyl	$-CH_3$	$-CH_3$	$-CH_3$	$-CH_3$
565	4-chlorophenyl	—H	$-CH_3$	—H	—H
566	4-chlorophenyl	$-CH_3$	$-CH_3$	—H	—H
567	4-chlorophenyl	—H	$-CH_3$	$-CH_3$	—H
568	4-chlorophenyl	$-CH_3$	$-CH_3$	$-CH_3$	—H
569	4-chlorophenyl	—H	$-CH_3$	$-CH_3$	$-CH_3$
570	4-chlorophenyl	$-CH_3$	$-CH_3$	$-CH_3$	$-CH_3$
571	4-methylphenyl	—H	$-CH_3$	—H	—H
572	4-methylphenyl	$-CH_3$	$-CH_3$	—H	—H
573	4-methylphenyl	—H	$-CH_3$	$-CH_3$	—H
574	4-methylphenyl	$-CH_3$	$-CH_3$	$-CH_3$	—H
575	4-methylphenyl	—H	$-CH_3$	$-CH_3$	$-CH_3$
576	4-methylphenyl	$-CH_3$	$-CH_3$	$-CH_3$	$-CH_3$

APPROACH

- Sketches are extracted to extended SMILES capturing:
  - R-group labels
  - Positional variation
  - Repeat groups
- USPTO tables precisely describe how tables should be displayed but are weak on semantics
  - Heuristics used to determine which lines are the same row
  - Table caption disambiguated from table column headings
  - Column widths used to determine columns
  - Colspans detected
- Name to structure used to interpret chemical names/formulas as R-groups; sketches interpreted as R-groups
- Structure assembled from core and R-groups

CORE VARIATION



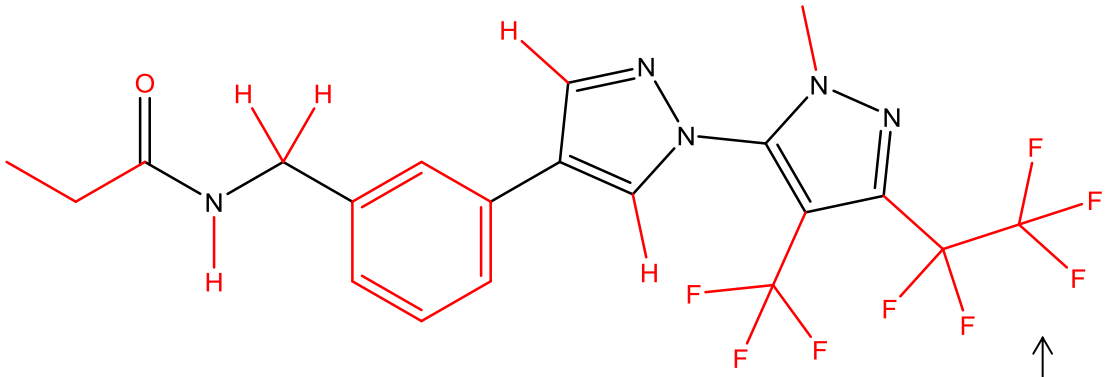
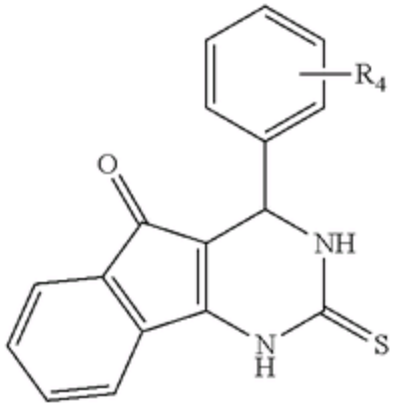


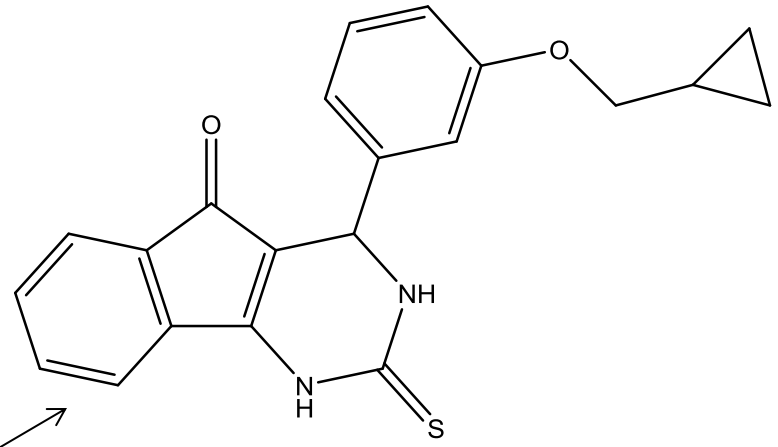
TABLE 2

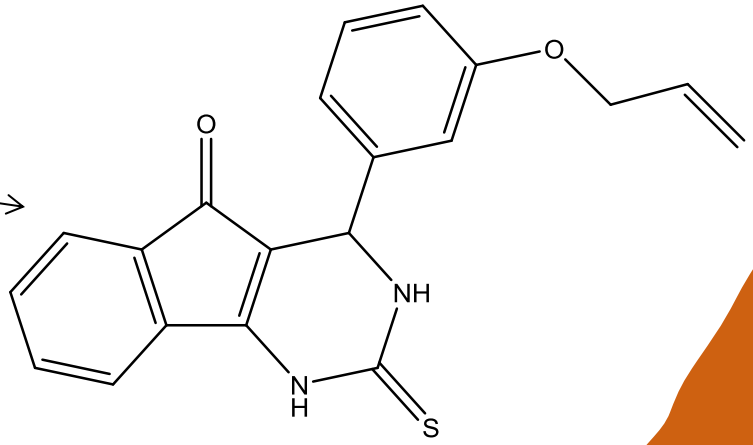
Ex. No.	Z^1	Z^2	Z^3	R^1	R^{6a}	R^{6b}	A_4	A_3	A_2	A_1	M_1	M_2	W	Q	logP a)	Mass a)
Ic-1	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	C—H	C—H	H	H	O	ethyl	3.59	496.2
Ic-2	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	ethyl	3.72	514.1
Ic-3	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	1,3-thiazol-yl	3.93	569.0
Ic-4	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	CH_3	3.40	500.1
Ic-5	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	2,2,2-trifluoroethyl	3.93	568.0
Ic-6	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	trifluoromethyl	4.24	554.0
Ic-7	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—F	C—H	C—H	H	H	O	thiophen-2-ylmethyl	4.12	582.0
Ic-8	C_2F_5	CF_3	CH_3	H	H	H	C—H	C—H	N	C—H	H	H	O	CH_3	2.35	483.0

POSITIONAL VARIATION



Compound ID	R4	Activity
1	3-OCH2cPr	A
2	3-OCH2CH=CH2	A
3	3-OEt	A
4	3-OCF3H	A
5	3-OMe	A
6	3-OCH2CH2OCH3	A





Incorrect formula

SUBSTITUENTS DEFINED AS SKETCHES

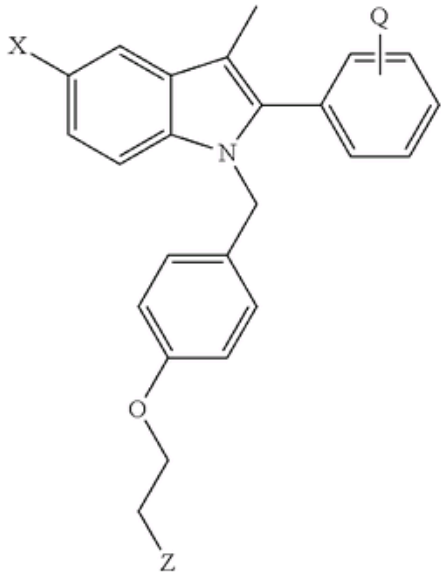
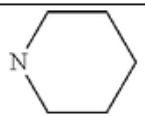
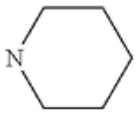


TABLE 7

Example No.	X	Q	Z
No. 85	H	H	
No. 86	H	4'-OH	

CURRENT RESULTS

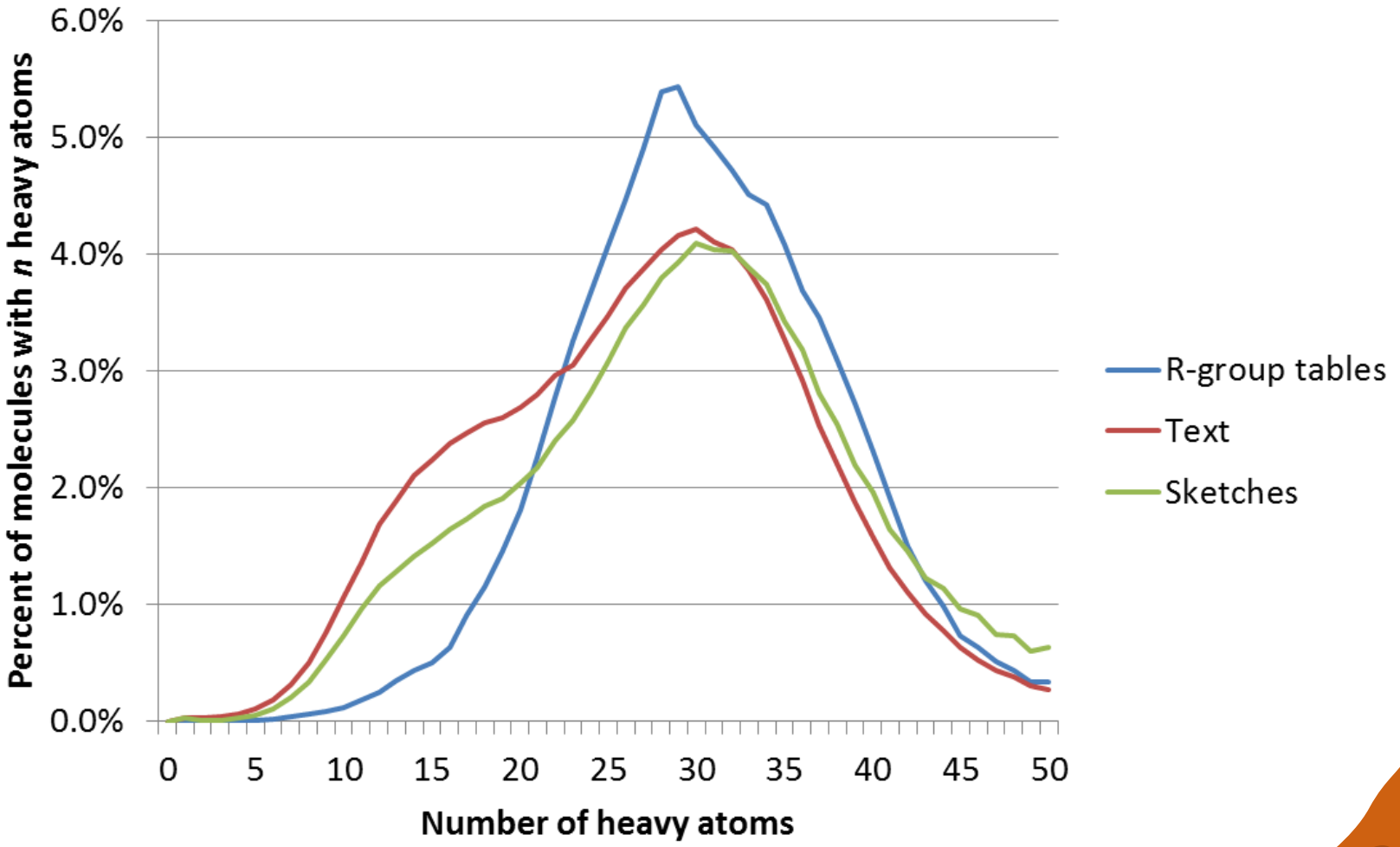
- 2001- June 2016 USPTO patent applications:
  - 1.96 million potential table entries detected
  - 1.13 million (57.9%) converted to specific chemical structures
  - 621 thousand unique chemical structures

NOVELTY OF RESULTS (VERSUS OTHER PIPELINES)

Data type	Unique Compounds	Not found in text /sketches	Not found in text	Not found in sketches
Exemplified compound R-group tables	**621,140**	**529,417 (85.2%)**	**541,974 (87.3%)**	**590,889 (95.1%)**
Text	4,759,009	0%	0%	2,960,937 (62.2%)
Sketches	4,479,113	0%	2,681,041 (59.9%)	0%

Structural identity checks performed using StdInChI

HEAVY ATOM COUNT DISTRIBUTION



NOVELTY OF RESULTS (VERSUS PUBCHEM)

Data type	Unique Compounds	Not in PubChem	Not in PubChem (SureChEMBL)
Exemplified compound R-group tables	621,140	**496,831 (80.0%)**	532,166 (85.7%)
Text	4,759,009	**564,886 (11.9%)**	911,976 (19.2%)
Sketches	4,479,113	**886,991 (19.8%)**	1,179,229 (26.3%)

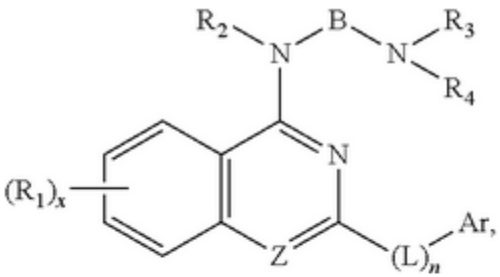
Structural identity checks performed using StdInChI

CURRENT LIMITATIONS

- Application of variable repeat groups
- Obtuse ways of depicting attachment points
- R-groups defined in terms of other R-groups
- R-groups defined elsewhere in the document
- Positional variation R-group representing multiple groups e.g. “3,4-diCl”
- Formulas involving substituted rings e.g.“4-ClPh”
- “Formulas” that mix systematic names with formula e.g. “4-OMe-phenyl”
- Algorithmic number of simple ring-systems (for positional variation)
- Ditto mark

CURRENT LIMITATIONS

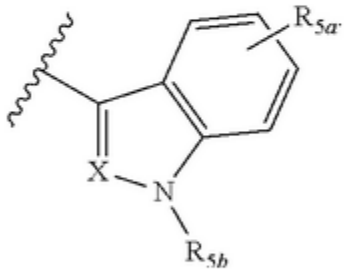
[0971] Table 14 depicts compounds of Formula Ia:



x implicitly 1?

Partially defined by this text and the table

wherein Z is N, n is 0, and Ar is



Nested R-group definition

TABLE 14

No.	R_1	R_2	R_3	R_4	B	R_{5a}	R_{5b}	X	MS*
560	8-CH_3	H	ethyl	ethyl	$CH(CH_3)(CH_2)_3$	7-Cl	CH_3	C	464
561	8-CH_3	H	CH_3	CH_3	$(CH_2)_3$	7-Cl	CH_3	C	408

Which is position 8?

CONCLUSIONS

- Direct interpretation of ChemDraw files can provide precision benefits over using ChemDraw exported Mol files or optical structure recognition approaches

- Structures from R-group tables are not handled by existing text-mining approaches (e.g. SureChEMBL)

- Extracting structures from R-group tables is complementary to existing approaches

ACKNOWLEDGEMENTS

- George Papadatos
- Funding provided by:

Thank you for your time!

http://nextmovesoftware.com
http://nextmovesoftware.com/blog
daniel@nextmovesoftware.com