


# NAME TO MARKUSH: GENERIC CHEMICAL SEARCHES MADE EASY

John Mayfield and Roger Sayle

NextMove Software LTD, Cambridge UK

RSC CICAG Chemical Structure Representations Meeting,
8th April 2026, Burlington House, London, UK



# MOTIVATION

Many tools and techniques exist for retrieving chemical information from databases

Three fundamental operations for a **chemical structure**:

- Exact (***is a***)
- Substructure/Superstructure (***has a***)
- Similarity (***like a***)

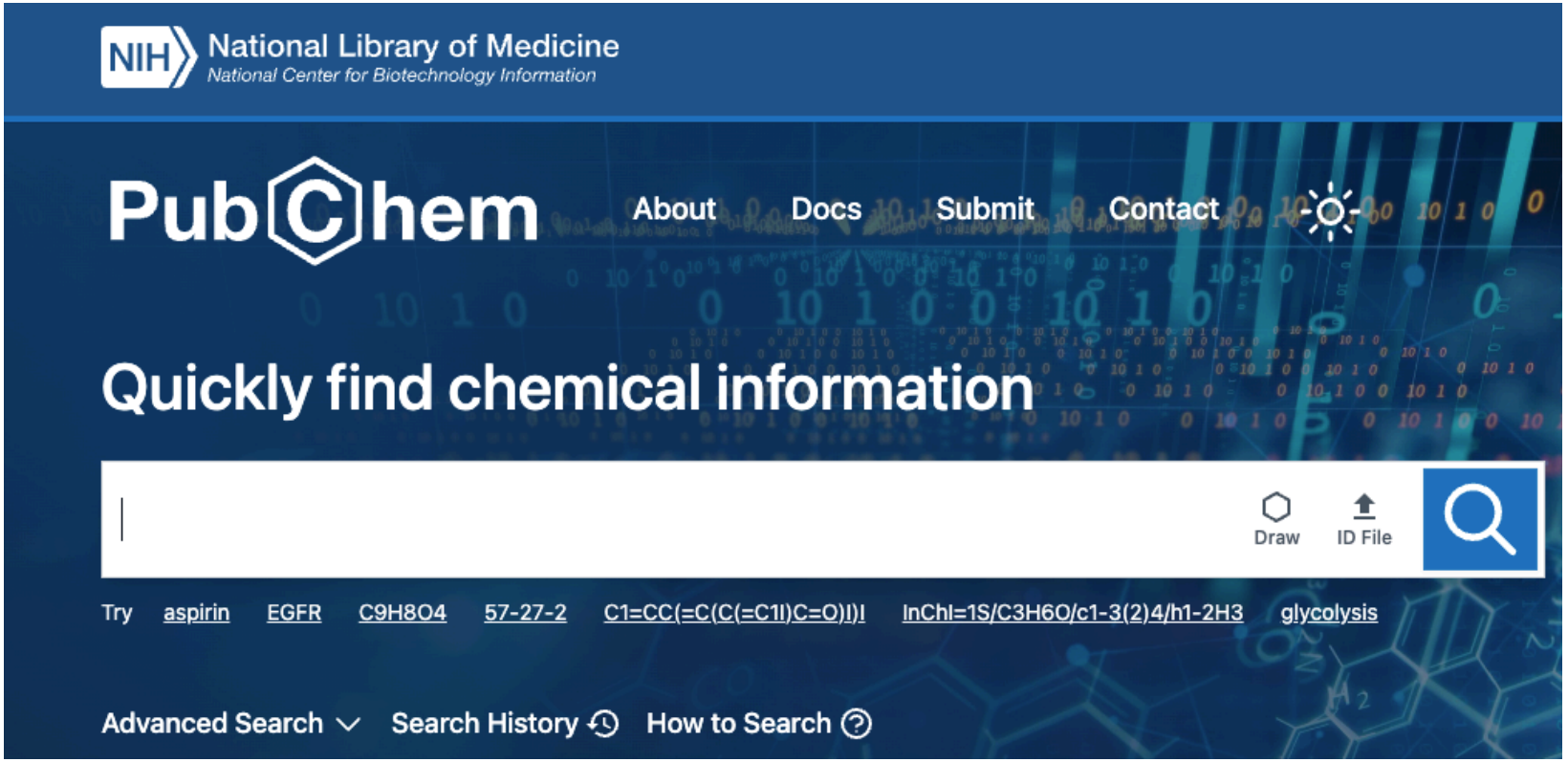
Humans are lazy and **terse text descriptions** are quick to enter:

- Chemical Name (e.g. **abilify**, **7H-purin-8-one**)
- Chemical Identifier (e.g. **129722-12-9**, **CHEMBL1112**)
- SMILES/SMARTS (e.g. **O=C1CCc2ccc(OCCCCN3CCN(c4cccc(Cl)c4Cl)CC3)cc2N1**)



# MOTIVATION

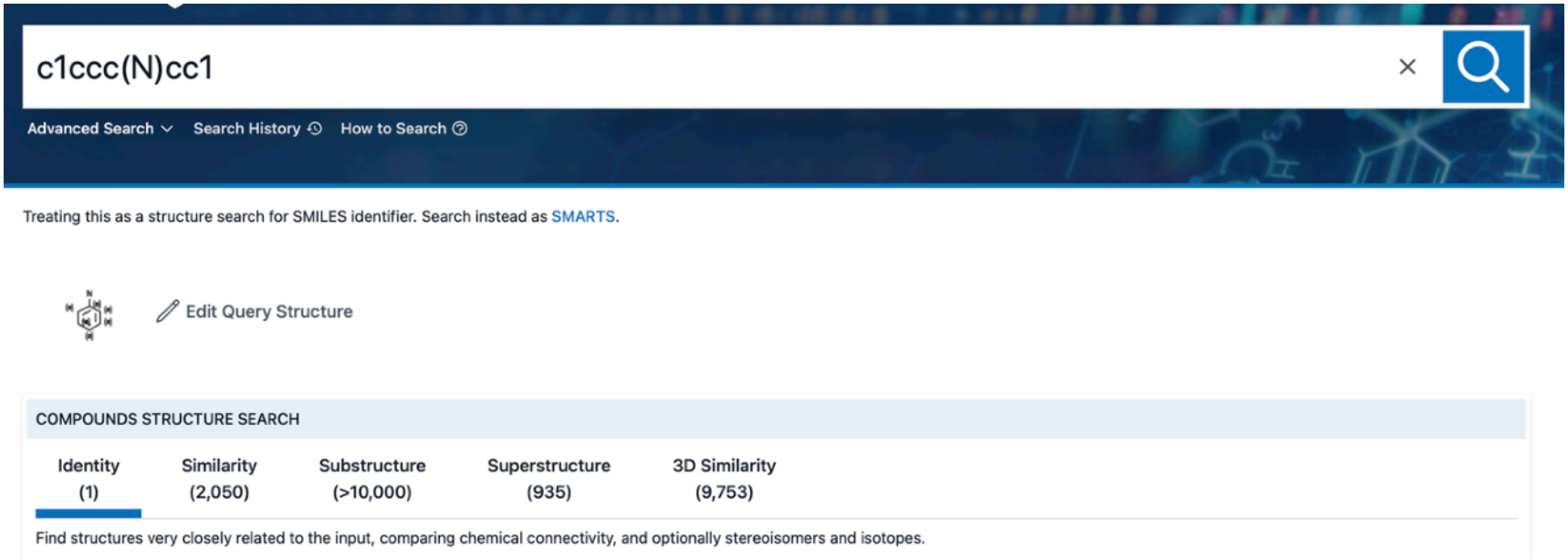
**PubChem's** awesome query box:



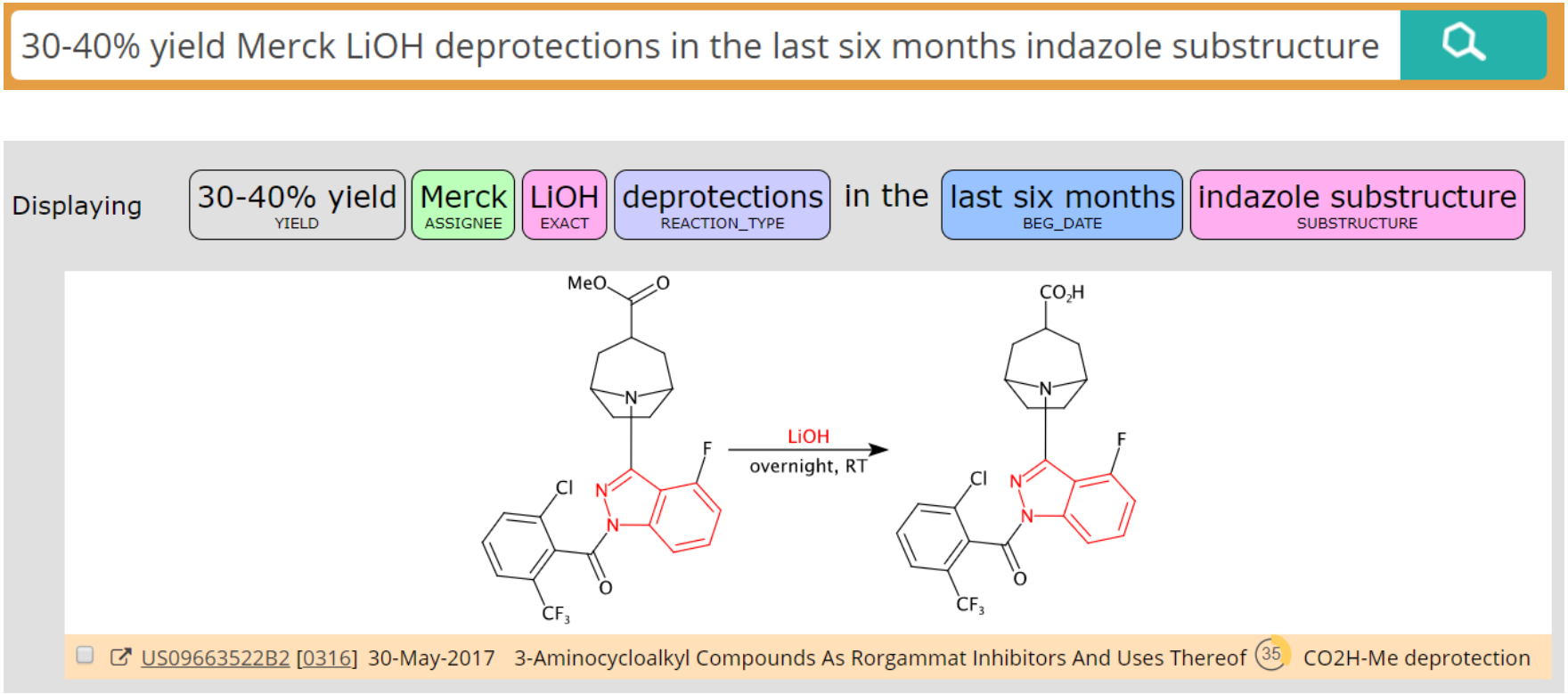

https://pubchem.ncbi.nlm.nih.gov/

# MOTIVATION

**PubChem's** awesome query box:



# MOTIVATION: PISTACHIO OMNIBOX



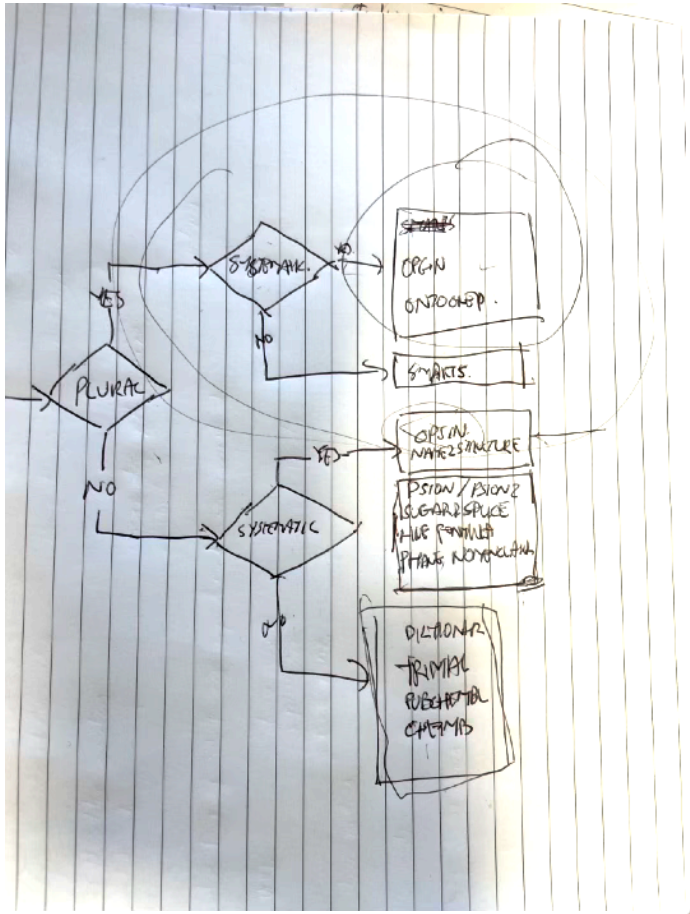

https://www.nextmovesoftware.com/pistachio.html

# MOTIVATION: DISAMBIGUATION

Some customers comment:
*"must be using ChatGPT or similar large language model"*

Pistachio is a rule-based semantic search, **no** free text

Input Text
→ Find Entities / Disambiguate
→ Generate Query



NO ONE GETS PAID FOR A 404 NOT FOUND



# ADVANCED QUERY HANDLING

The trouble with allowing users to type anything is they **will** type anything:



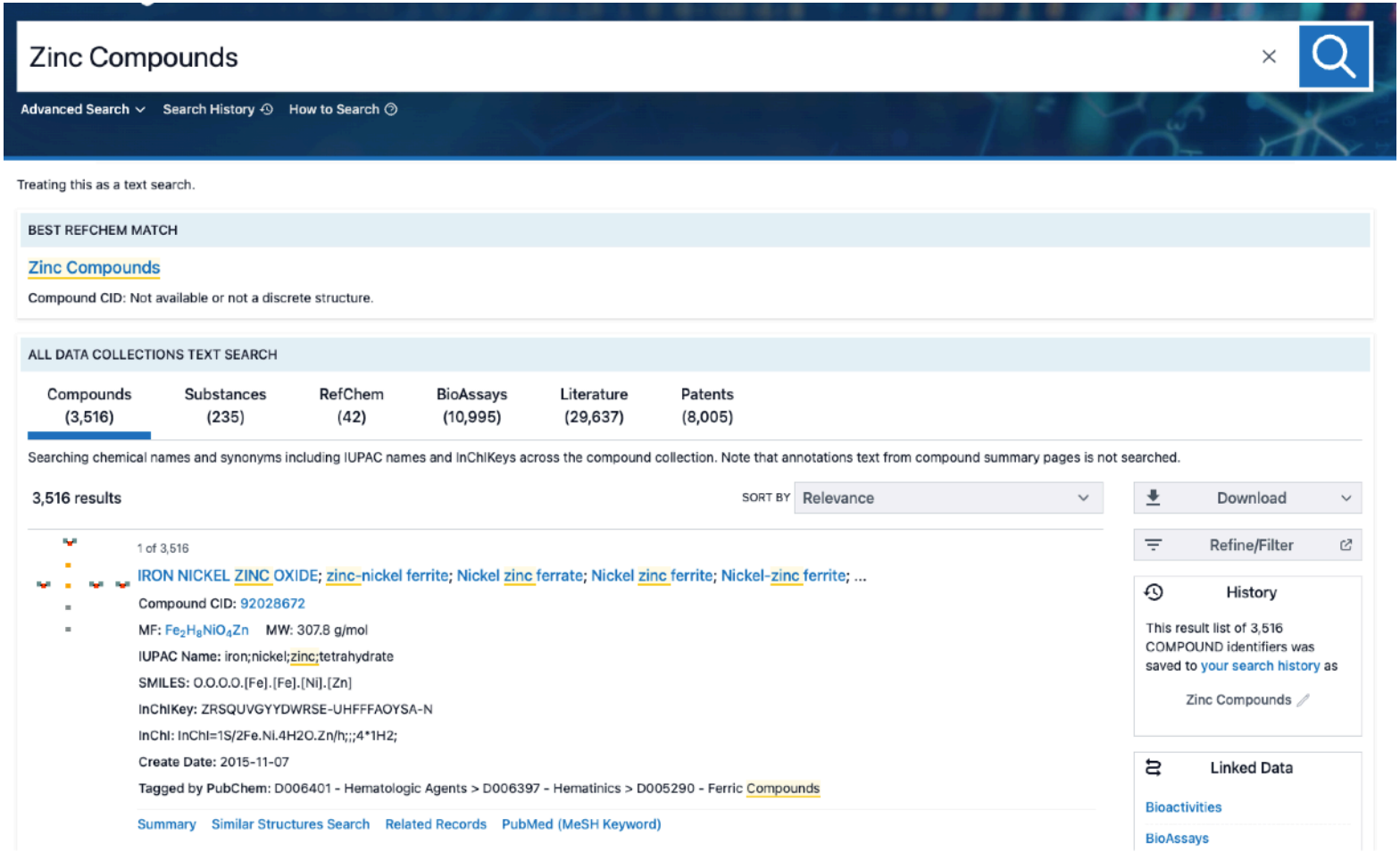
# ADVANCED QUERY HANDLING

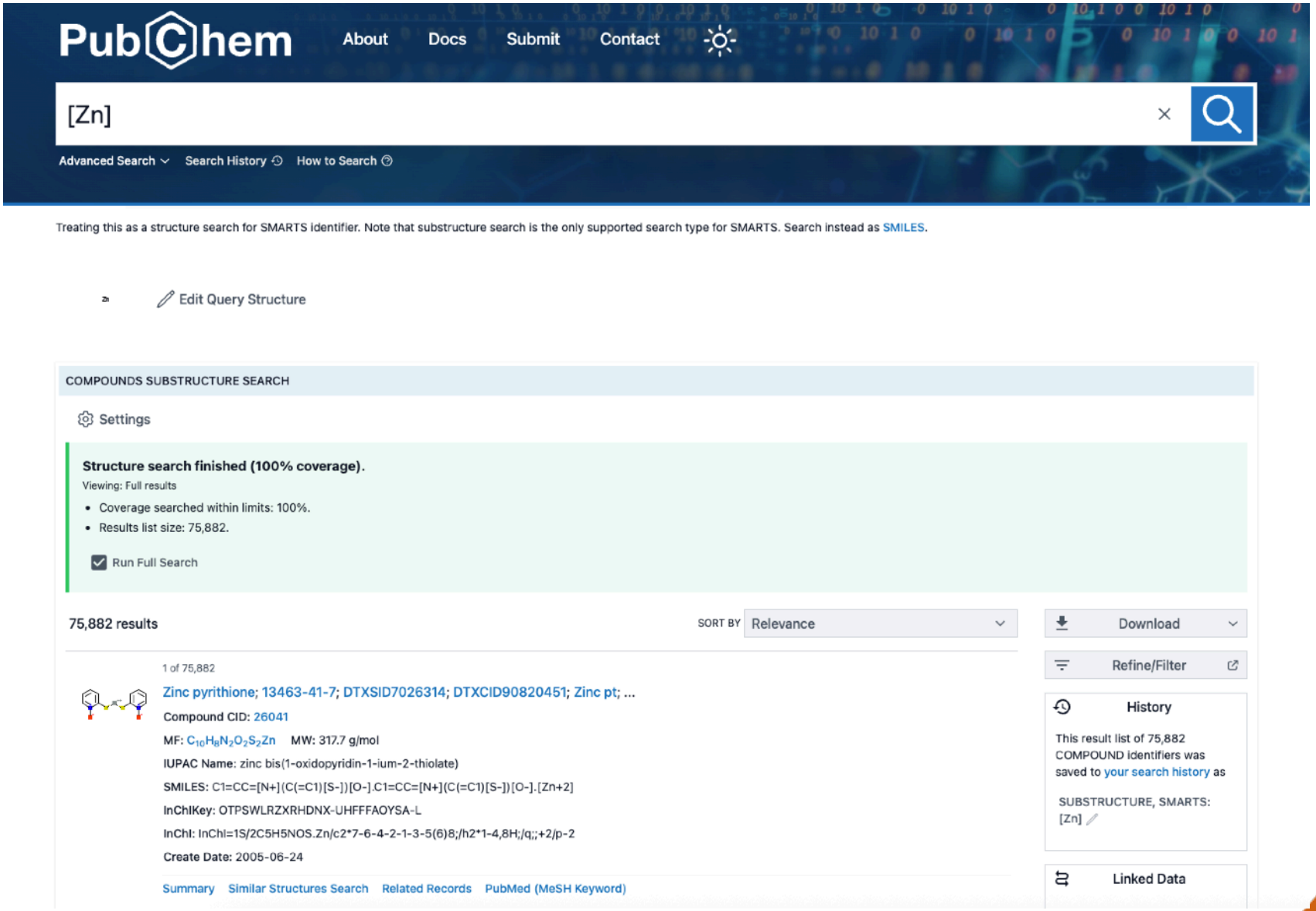
The trouble with allowing users to type anything is they **will** type anything:

- zinc compounds
- boronic acids
- alkanes
- heterocycles
- inorganic acids
- solvents

All these queries refer to the **chemical structure** yet they are *fuzzy* and *non-exact*.

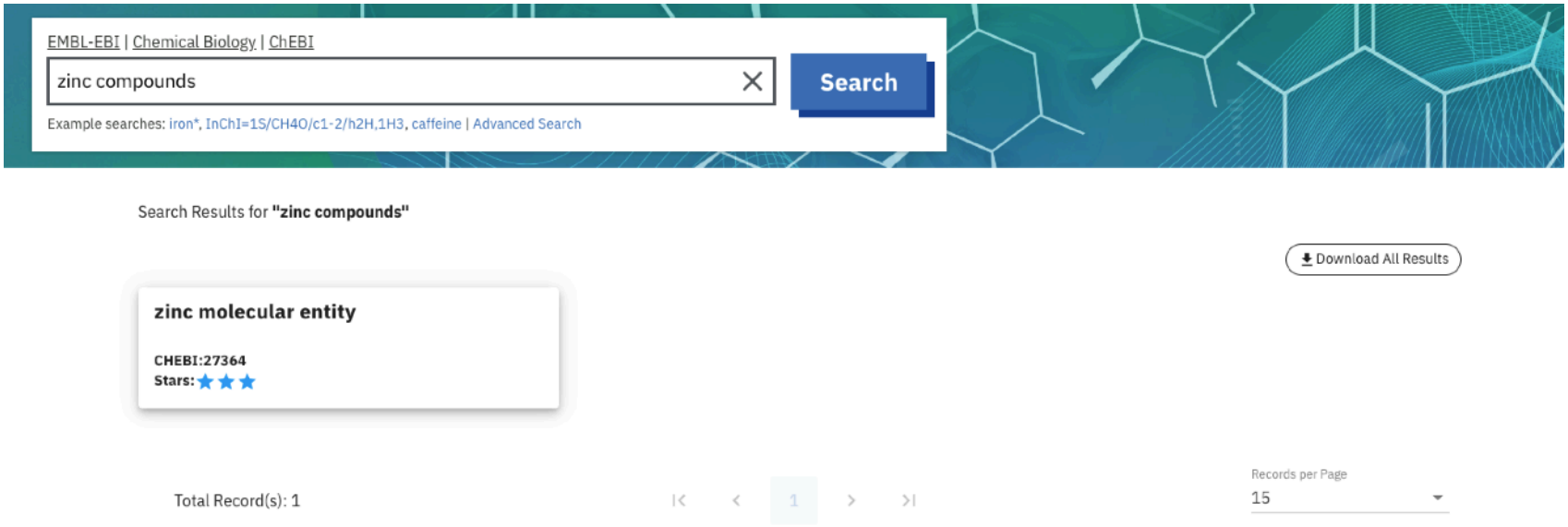



Zinc Compounds
Advanced Search
Search History
How to Search
Treating this as a text search.
BEST REFCHEM MATCH
Zinc Compounds
Compound CID: Not available or not a discrete structure.
ALL DATA COLLECTIONS TEXT SEARCH
Compounds (3,516)
Substances (235)
RefChem (42)
BioAssays (10,995)
Literature (29,637)
Patents (8,005)
Searching chemical names and synonyms including IUPAC names and InChIKeys across the compound collection. Note that annotations text from compound summary pages is not searched.
3,516 results
SORT BY Relevance
Download
Refine/Filter
1 of 3,516
IRON NICKEL ZINC OXIDE; zinc-nickel ferrite; Nickel zinc ferrate; Nickel zinc ferrite; Nickel-zinc ferrite; ...
Compound CID: 92028672
MF: Fe2H8NiO4Zn MW: 307.8 g/mol
IUPAC Name: iron;nickel;zinc;tetrahydrate
SMILES: O.O.O.O.[Fe].[Fe].[Ni].[Zn]
InChIKey: ZRSQUVGYYDWRSE-UHFFFAOYSA-N
InChI: InChI=1S/2Fe.Ni.4H2O.Zn/h;;;4*1H2;
Create Date: 2015-11-07
Tagged by PubChem: D006401 - Hematologic Agents > D006397 - Hematinics > D005290 - Ferric Compounds
Summary
Similar Structures Search
Related Records
PubMed (MeSH Keyword)
History
This result list of 3,516 COMPOUND identifiers was saved to your search history as
Zinc Compounds
Linked Data
Bioactivities
BioAssays

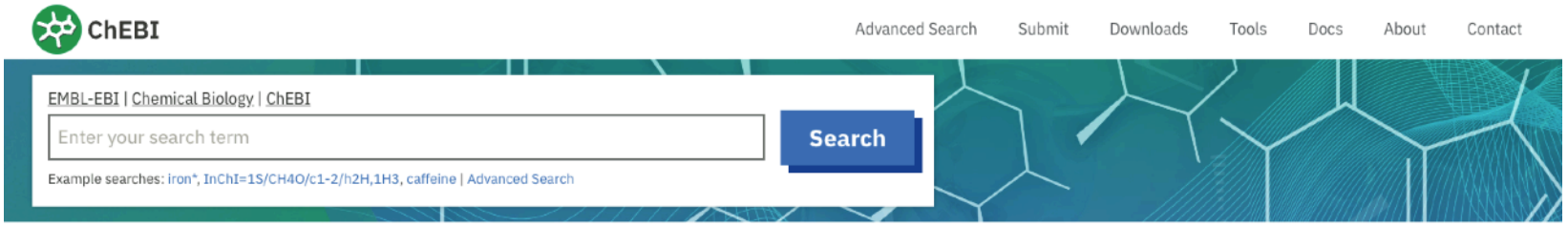

PubChem
About
Docs
Submit
Contact
[Zn]
Advanced Search
Search History
How to Search
Treating this as a structure search for SMARTS identifier. Note that substructure search is the only supported search type for SMARTS. Search instead as SMILES.
Edit Query Structure
COMPOUNDS SUBSTRUCTURE SEARCH
Settings
Structure search finished (100% coverage).
Viewing: Full results
Coverage searched within limits: 100%.
Results list size: 75,882.
Run Full Search
75,882 results
SORT BY
Relevance
Download
Refine/Filter
1 of 75,882
Zinc pyrithione; 13463-41-7; DTXSID7026314; DTXCID90820451; Zinc pt; ...
Compound CID: 26041
MF: C10H8N2O2S2Zn MW: 317.7 g/mol
IUPAC Name: zinc bis(1-oxidopyridin-1-ium-2-thiolate)
SMILES: C1=CC=[N+](C(=C1)[S-])[O-].C1=CC=[N+](C(=C1)[S-])[O-].[Zn+2]
InChIKey: OTPSWLRZXRHDNX-UHFFFAOYSA-L
InChI: InChI=1S/2C5H5NOS.Zn/c2*7-6-4-2-1-3-5(6)8;/h2*1-4,8H;/q;;+2/p-2
Create Date: 2005-06-24
Summary
Similar Structures Search
Related Records
PubMed (MeSH Keyword)
History
This result list of 75,882 COMPOUND identifiers was saved to your search history as
SUBSTRUCTURE, SMARTS: [Zn]
Linked Data

# ChEBI Ontology

**ChEBI** has it's entries classified into an ontology




https://www.ebi.ac.uk/chebi/




# CHEBI:27364 - zinc molecular entity

On this page

- Intro
- Ontology
- Names

| | |
|---|---|
| ChEBI ID | CHEBI:27364 |
| ChEBI Name | zinc molecular entity |
| Stars | ★★★ |
| Last Modified | 12 June 2024 |

## ChEBI Ontology

TEXT | TREE VIEW | GRAPH VIEW

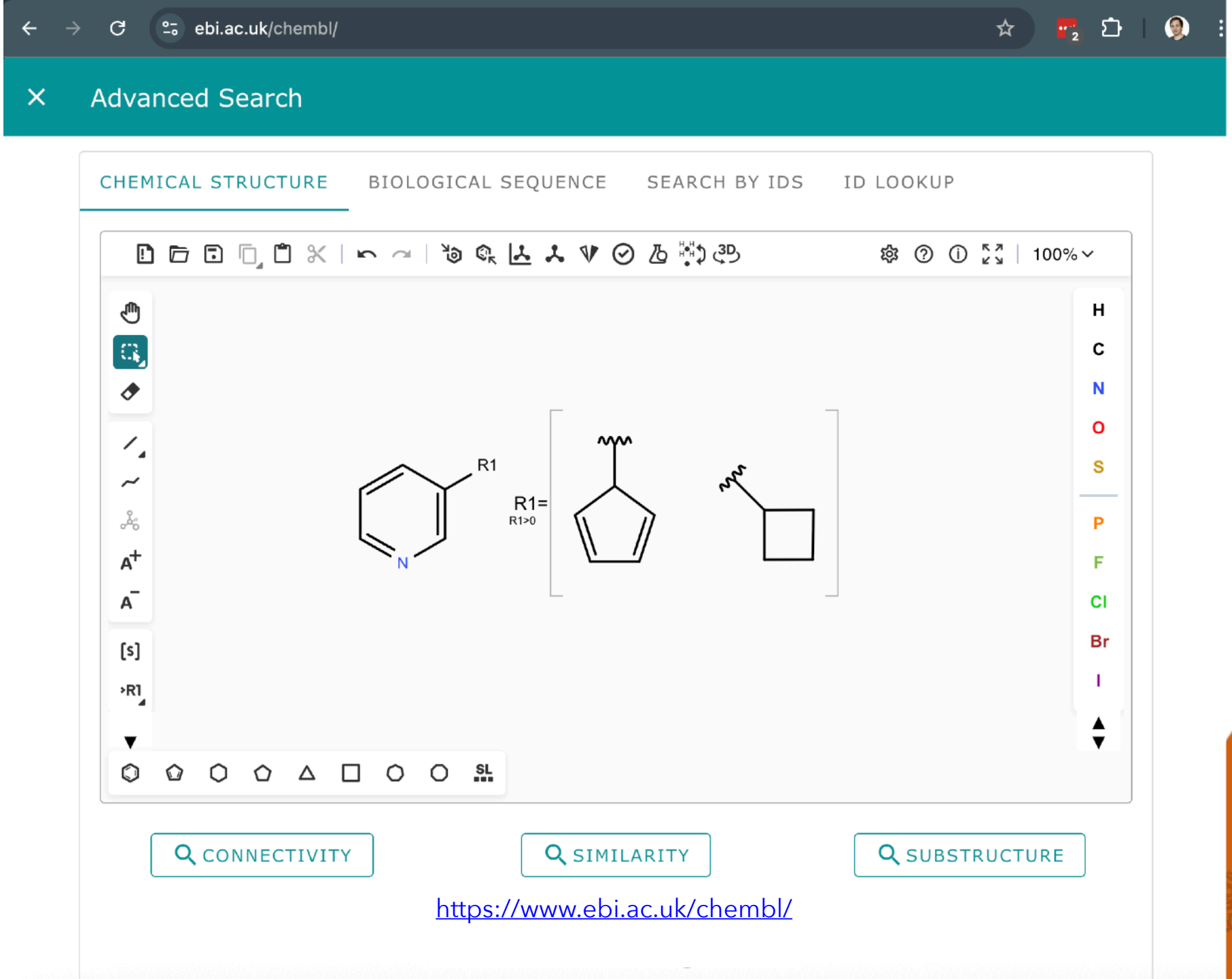
| | |
|---|---|
| Outgoing Relation(s) | zinc molecular entity (CHEBI:27364) **has part** zinc atom (CHEBI:27363)<br>zinc molecular entity (CHEBI:27364) **is a** zinc group molecular entity (CHEBI:33673) |
| Incoming Relation(s) | elemental zinc (CHEBI:37253) **is a** zinc molecular entity (CHEBI:27364)<br>fast blue salt B (CHEBI:87627) **is a** zinc molecular entity (CHEBI:27364)<br>zinc acetate (CHEBI:62984) **is a** zinc molecular entity (CHEBI:27364)<br>zinc coordination entity (CHEBI:36566) **is a** zinc molecular entity (CHEBI:27364)<br>zinc dibutyldithiocarbamate (CHEBI:144323) **is a** zinc molecular entity (CHEBI:27364)<br>zinc dichloride (CHEBI:49976) **is a** zinc molecular entity (CHEBI:27364)<br>zinc diethyldithiocarbamate (CHEBI:144351) **is a** zinc molecular entity (CHEBI:27364)<br>zinc oxide (CHEBI:36560) **is a** zinc molecular entity (CHEBI:27364)<br>zinc salt (CHEBI:231592) **is a** zinc molecular entity (CHEBI:27364)<br>zinc sulfate (CHEBI:35176) **is a** zinc molecular entity (CHEBI:27364)<br>ziram (CHEBI:79736) **is a** zinc molecular entity (CHEBI:27364) |

| Synonyms | Source |
|---|---|
| zinc compounds | ChEBI |
| zinc molecular entities | ChEBI |



Back to top

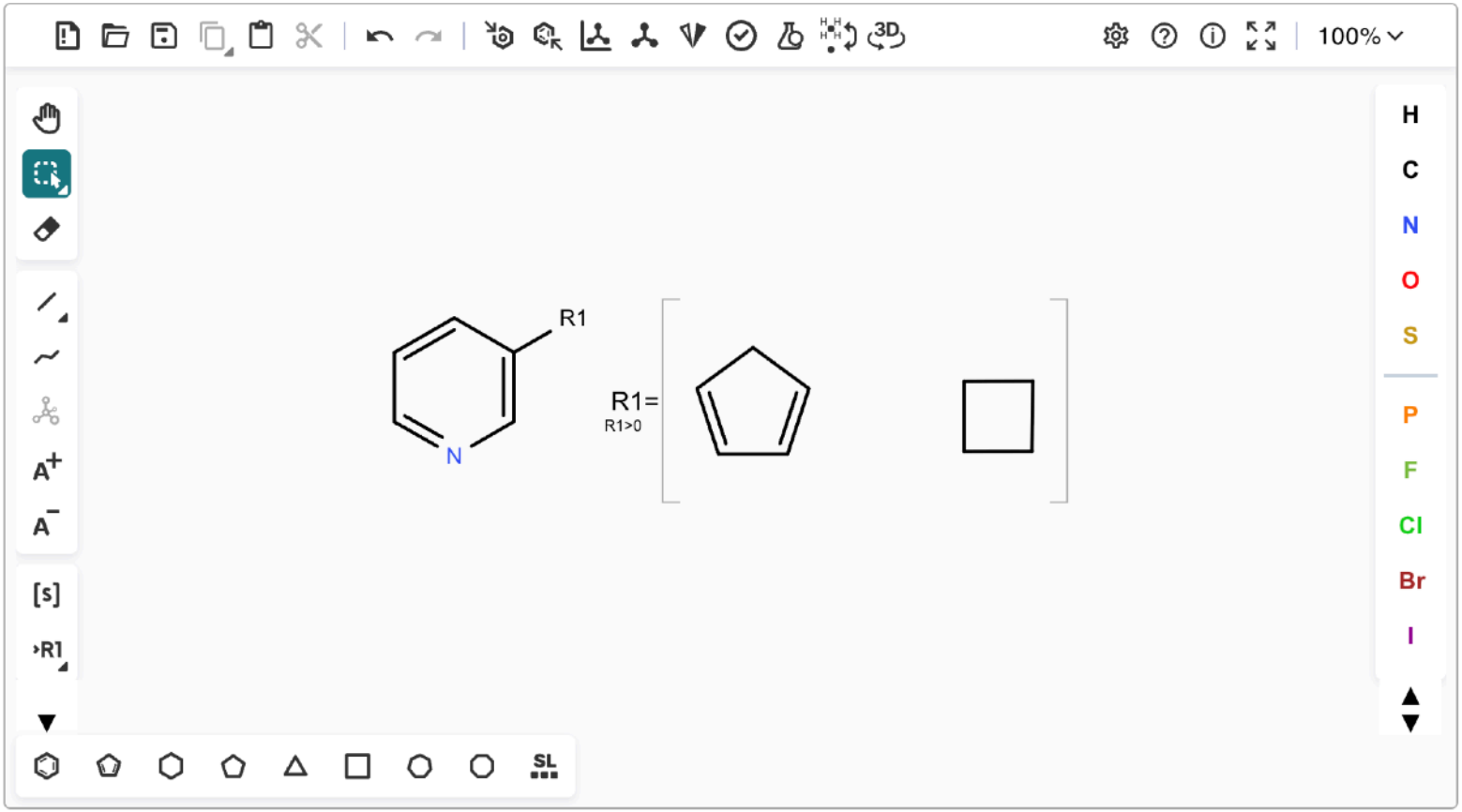



ebi.ac.uk/chembl/
Advanced Search
CHEMICAL STRUCTURE
BIOLOGICAL SEQUENCE
SEARCH BY IDS
ID LOOKUP
100%
R1
N
R1=
R1>0
H
C
N
O
S
P
F
Cl
Br
I
CONNECTIVITY
SIMILARITY
SUBSTRUCTURE
https://www.ebi.ac.uk/chembl/

ebi.ac.uk/chembl/

# Advanced Search

CHEMICAL STRUCTURE | BIOLOGICAL SEQUENCE | SEARCH BY IDS | ID LOOKUP




There was an error while translating the molecule into SMILES, please try again. (Request failed with status code 422)

# MARKUSH QUERY COMPLICATIONS

ChEMBL / ChEBI / SureChEMBL did used to support generic (Markush) queries

Either no one told them…

… or no one knew?

… or no one used it?

SciFinder (STNExpress) / Reaxys offer training to make sure you do it properly

Should the human adapt to the computer or the computer to the human?



# ONTOGREP

NextMove Software's **OntoGrep** technology demonstrator allows handling of selected generic queries

- nitrogen containing heterocycles
- cationic ring systems
- cyclic alkanes
- branched acyclic alkanes
- transuranic elements
- zinc compounds
- binary compounds
- Lewis acids
- atropisomers
- polyspiro ring systems
- polyatomic elements
- inorganic salts
- radioactive compounds
- radioactive elements
- monocyclic ring systems
- neutral compounds
- uncharged compounds
- zwitterionic compounds
- carbon containing inorganics
- zinc oxides
- metal chlorides
- iron halides
- inorganic salts

*… and many more*

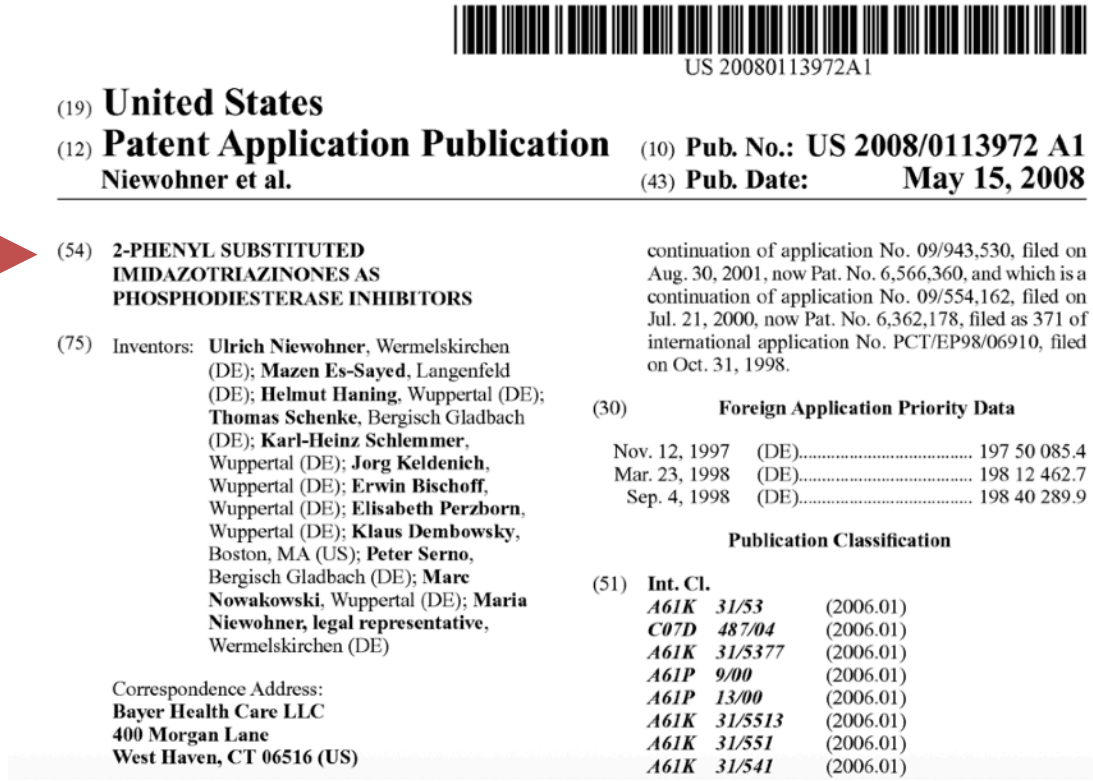


# ONTOGREP LIMITATIONS

Bayer patent for Vardenafil (Levitra)

*"2-phenyl substituted imidazotriazinones"*

*"phosphodiesterase inhibitors"*



US 20080113972A1

(19) **United States**

(12) **Patent Application Publication**

**Niewohner et al.**

(10) **Pub. No.: US 2008/0113972 A1**

(43) **Pub. Date: May 15, 2008**

(54) **2-PHENYL SUBSTITUTED IMIDAZOTRIAZINONES AS PHOSPHODIESTERASE INHIBITORS**

(75) Inventors: **Ulrich Niewohner**, Wermelskirchen (DE); **Mazen Es-Sayed**, Langenfeld (DE); **Helmut Haning**, Wuppertal (DE); **Thomas Schenke**, Bergisch Gladbach (DE); **Karl-Heinz Schlemmer**, Wuppertal (DE); **Jorg Keldenich**, Wuppertal (DE); **Erwin Bischoff**, Wuppertal (DE); **Elisabeth Perzborn**, Wuppertal (DE); **Klaus Dembowsky**, Boston, MA (US); **Peter Serno**, Bergisch Gladbach (DE); **Marc Nowakowski**, Wuppertal (DE); **Maria Niewohner, legal representative**, Wermelskirchen (DE)

Correspondence Address:
**Bayer Health Care LLC**
**400 Morgan Lane**
**West Haven, CT 06516 (US)**

continuation of application No. 09/943,530, filed on Aug. 30, 2001, now Pat. No. 6,566,360, and which is a continuation of application No. 09/554,162, filed on Jul. 21, 2000, now Pat. No. 6,362,178, filed as 371 of international application No. PCT/EP98/06910, filed on Oct. 31, 1998.

(30) **Foreign Application Priority Data**

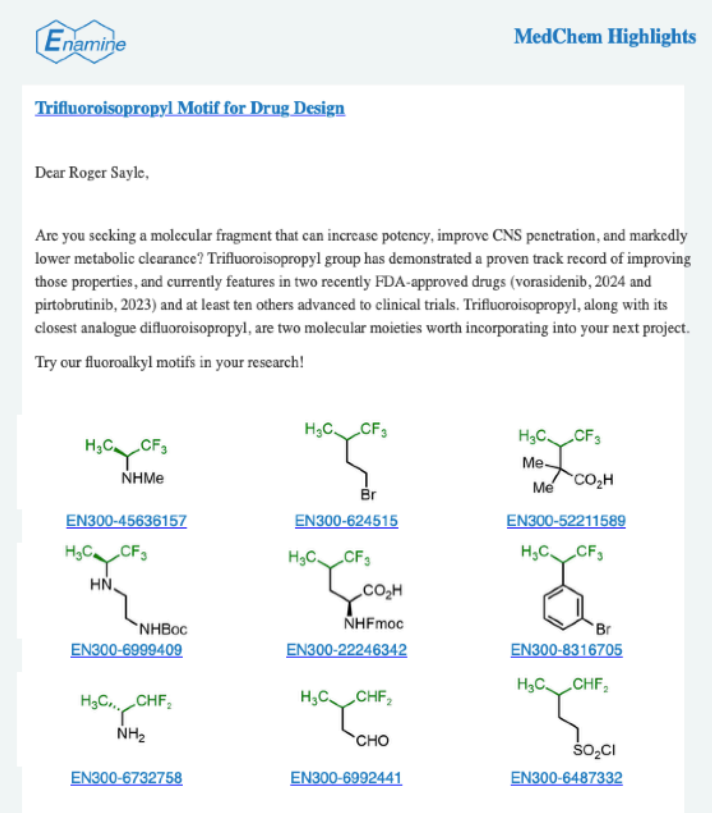
| | | |
|---|---|---|
| Nov. 12, 1997 | (DE) | 197 50 085.4 |
| Mar. 23, 1998 | (DE) | 198 12 462.7 |
| Sep. 4, 1998 | (DE) | 198 40 289.9 |

**Publication Classification**

(51) **Int. Cl.**

| | | |
|---|---|---|
| *A61K* | *31/53* | (2006.01) |
| *C07D* | *487/04* | (2006.01) |
| *A61K* | *31/5377* | (2006.01) |
| *A61P* | *9/00* | (2006.01) |
| *A61P* | *13/00* | (2006.01) |
| *A61K* | *31/5513* | (2006.01) |
| *A61K* | *31/551* | (2006.01) |
| *A61K* | *31/541* | (2006.01) |

# "I SEE GENERIC CHEMICAL NAMES"... ..."THEY DON'T KNOW THEY'RE GENERIC"



Enamine

MedChem Highlights

**Trifluoroisopropyl Motif for Drug Design**

Dear Roger Sayle,

Are you seeking a molecular fragment that can increase potency, improve CNS penetration, and markedly lower metabolic clearance? Trifluoroisopropyl group has demonstrated a proven track record of improving those properties, and currently features in two recently FDA-approved drugs (vorasidenib, 2024 and pirtobrutinib, 2023) and at least ten others advanced to clinical trials. Trifluoroisopropyl, along with its closest analogue difluoroisopropyl, are two molecular moieties worth incorporating into your next project.

Try our fluoroalkyl motifs in your research!

EN300-45636157 EN300-624515 EN300-52211589

EN300-6999409 EN300-22246342 EN300-8316705

EN300-6732758 EN300-6992441 EN300-6487332

trinitrotoluene (OPSIN default example)

2-phenyl imidazotriazinones (US6362178B1 - Bayer)

triazolopyridine (WO2017221092A1 - Novartis)

quinolinyltriazole (CCNM - 12 November 2025)

halogen-substituted pyrazolone (Eugene Markush 1923)

diazotized unsulphonated material (Eugene Markush 1923)

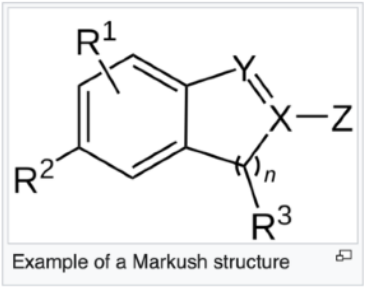
1,2-dihydro-2-(para-substituted aryl/heteroaryl)-(5 or 6)-(aryl/heteroaryl)-3H-indazol-3-ones (US2007129374)

trifluoroisopropyl (Enamine MedChem Highlights March 2026)

bisazolepyridines (https://doi.org/10.1107/S2056989025010977)

# "THE" MARKUSH

Markush structures today – typically think of a core **structure diagram** with different types variation



Example of a Markush structure

https://en.wikipedia.org/wiki/Markush_structure

The original Patent from which the name is coined doesn't contain a single **structure diagram!**

## Eugene Markush (US1506316A)

Claims:

1. The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

2. The process for the manufacture of

*1. The process for the manufacture of dyes which comprises coupling with a **halogen-substituted pyrazolone**, a **diazotized unsulphonated material** selected from the group consisting of **aniline**, **homologues of aniline** and **halogen substitution products of aniline**.*

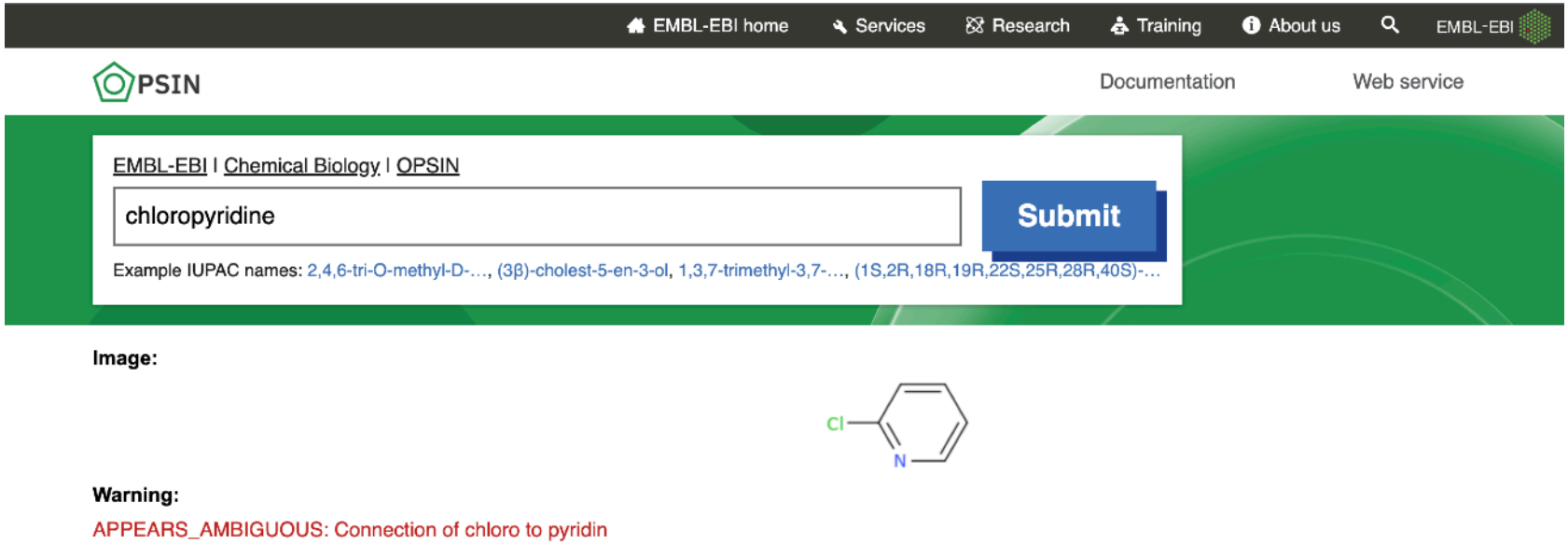
OPGIN



# OPSIN AMBIGUITY HANDLING

OPSIN: Open Parser for Systematic IUPAC Nomenclature

- ambiguity detection

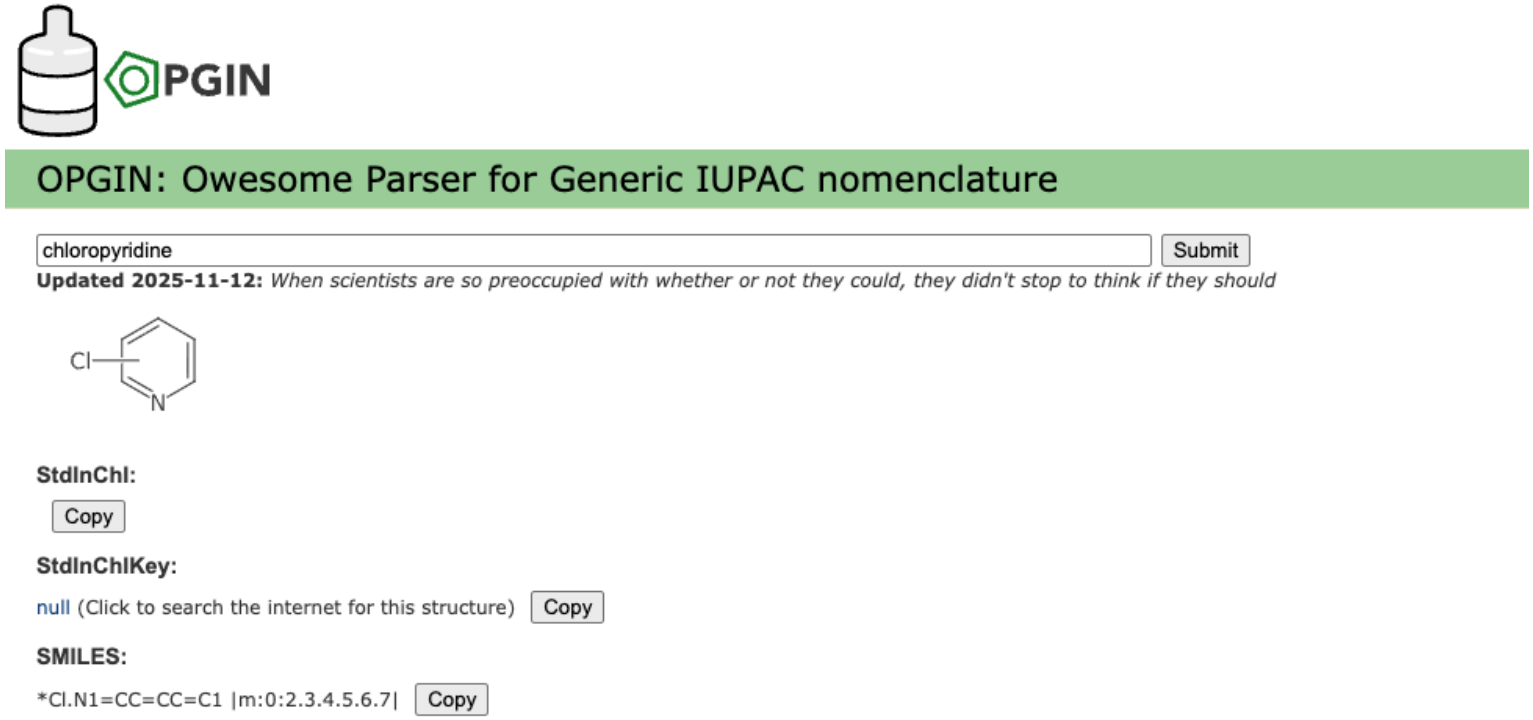



[ebi.ac.uk/opsin](ebi.ac.uk/opsin)

Daniel Lowe. Extraction of chemical structures and reactions from the literature.
https://doi.org/10.17863/CAM.16293

# OPGIN

**OPGIN: Owesome Parser for Generic IUPAC Nomenclature** adds the option to interpret ambiguous names as *"Markush"* structures



# CAPTURING GENERIC FEATURES

Capturing generic features is possible in multiple file formats:

- CTfile (RGfile/MOLfile)
- ChemAxon SMILES/SMARTS Extensions (CXSMILES/CXSMARTS)
- ChemDraw CDX/CDXML
- *…VarInChI?*

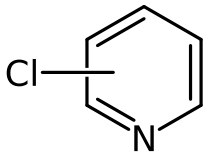
We do **not** have/need coordinates - CXSMILES/CXSMARTS is the natural choice:

- Toolkit support: ChemAxon, RDKit, CDK, EPAM Indigo

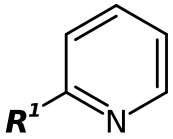


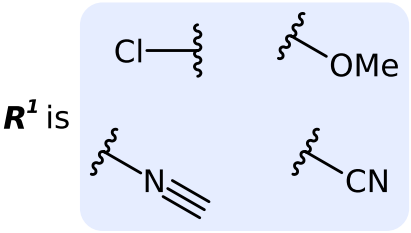
# STRUCTURAL VARIATION

The **4** types of generic structure variation



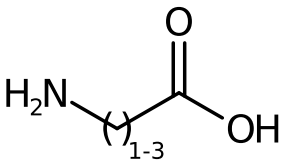

**P**ositional




**S**ubstituent

aka Alternative Group



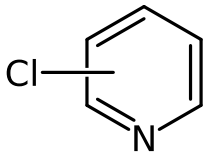

**F**requency

***heteroaryl***—Cl

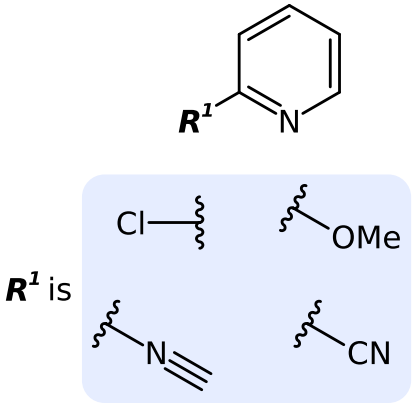
**H**omology

# STRUCTURAL VARIATION

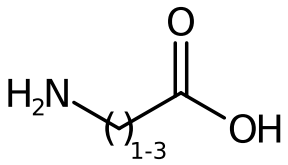
Example names




chloropyridine




2-(chloro/methoxy/cyano/isocyano)-pyridine




???

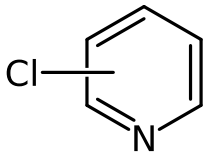
amino(poly(methyl))carboxylic acid?

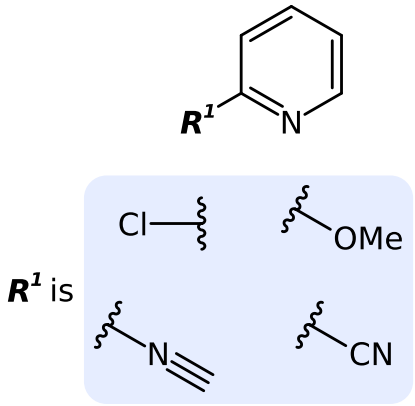
***heteroaryl***—Cl

heteroaryl chloride

# INTRODUCTION TO CXSMILES

## CXSMILES/CXSMART



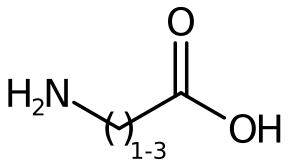




**P**ositional

```
n1ccccc1.*Cl |m:6:0.1.2.3.4.5|
```

**S**ubstituent

```
*c1ncccc1 |$R1_p$,RG:_R1={Cl},{OC},{N#C},{C#N}|
```




**F**requency

```
NCC(=O)O |LN:1:1.3|
```

***heteroaryl***—Cl

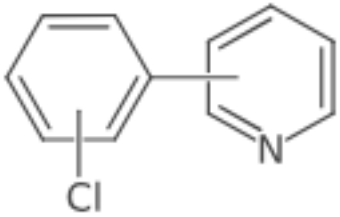
**H**omology

```
*Cl |$heteroaryl_p$|
```

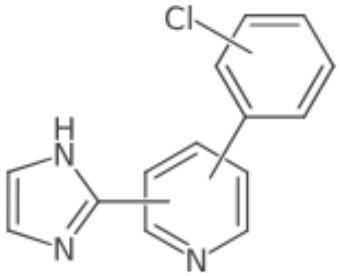
*CXSMILES encodings and depictions using CDK Depict*

# OPGIN BASIC EXAMPLES

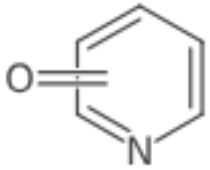
Unlocanted prefixes/suffixes handled by positional variation



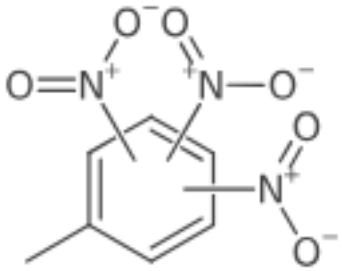

(chlorophenyl)pyridine



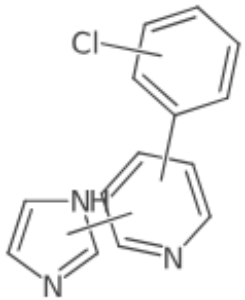

imidazol-2-yl(chlorophenyl)pyridine




pyridone




trinitrotoluene

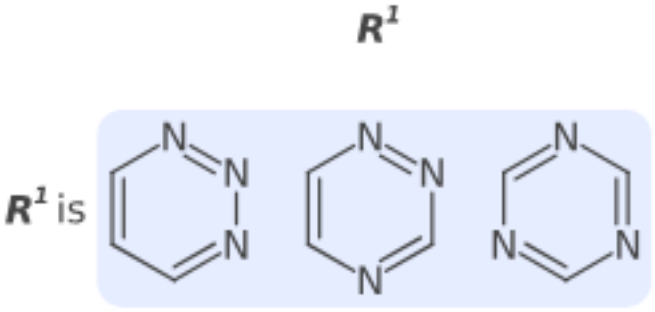



imidazolyl(chlorophenyl)pyridine

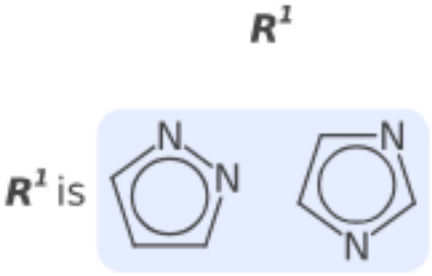
**Not possible in RGfile!**

# OPGIN: HANTZSCH-WIDMAN

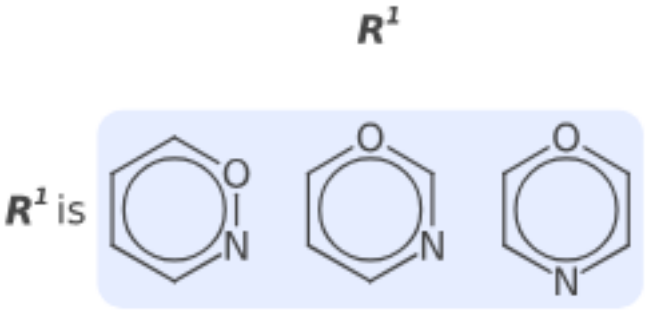
**Hantzsch-Widman** is a nomenclature for naming heterocyclic parents, these are currently handled by enumeration



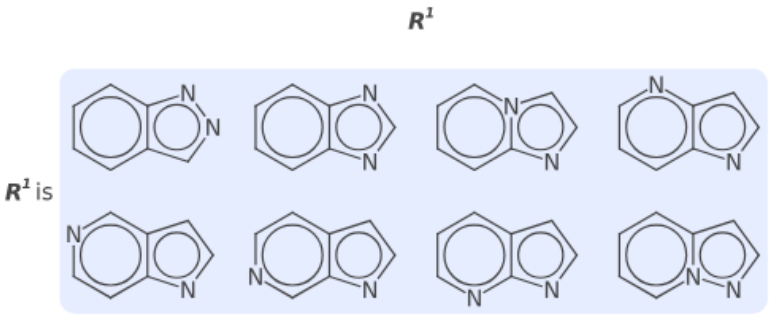

triazine




diazole



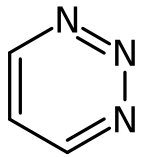

azoxaine



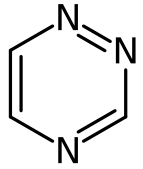

azaindole

# OPGIN: HANTZSCH-WIDMAN

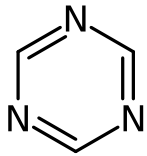
**Hantzsch-Widman** is a nomenclature for naming heterocyclic parents, these are currently handled by enumeration



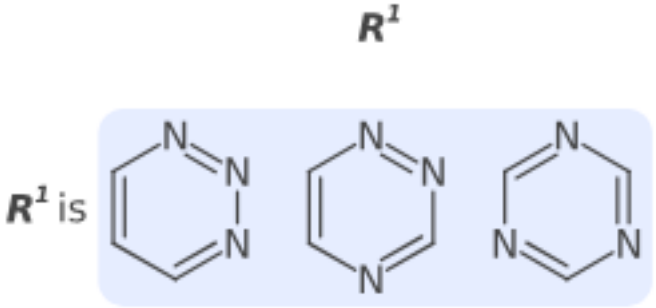
1,2,3-triazine



1,2,4-triazine



1,2,5-triazine

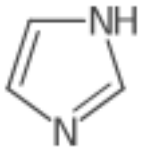



triazines

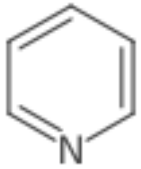
```
[*] |$_R1$,RG:_R1={n1nnccc1},{n1ncncc1},{n1cncnc1}|
```

# OPGIN: RING FUSIONS

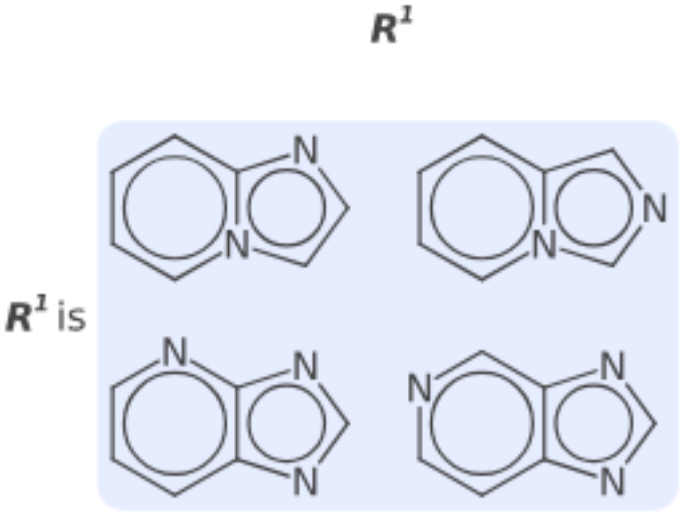
Ring fusion nomenclature allows you to build complex parent rings using a *fusion descriptor*




imidazole



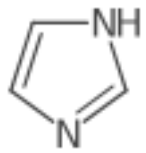

pyridine



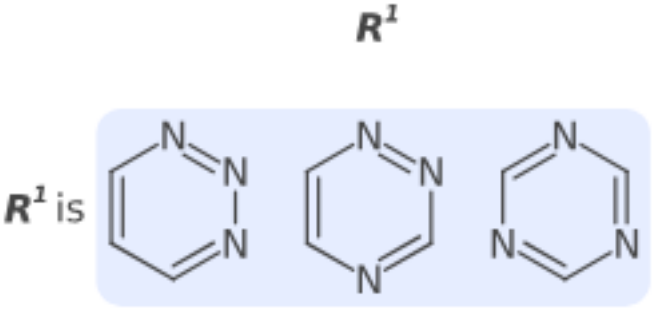

imidazopyridine

# OPGIN: RING FUSIONS

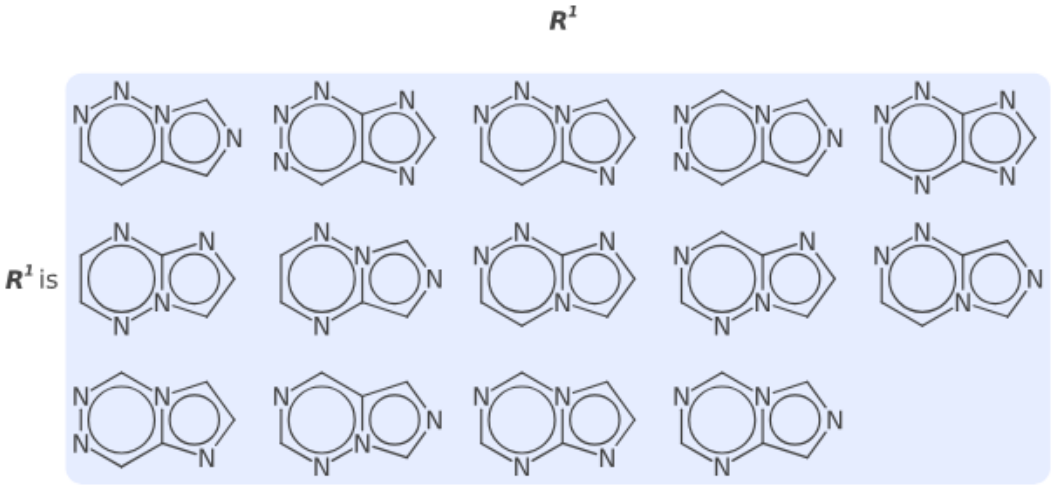
Need to handle fusion of already variable rings



imidazole



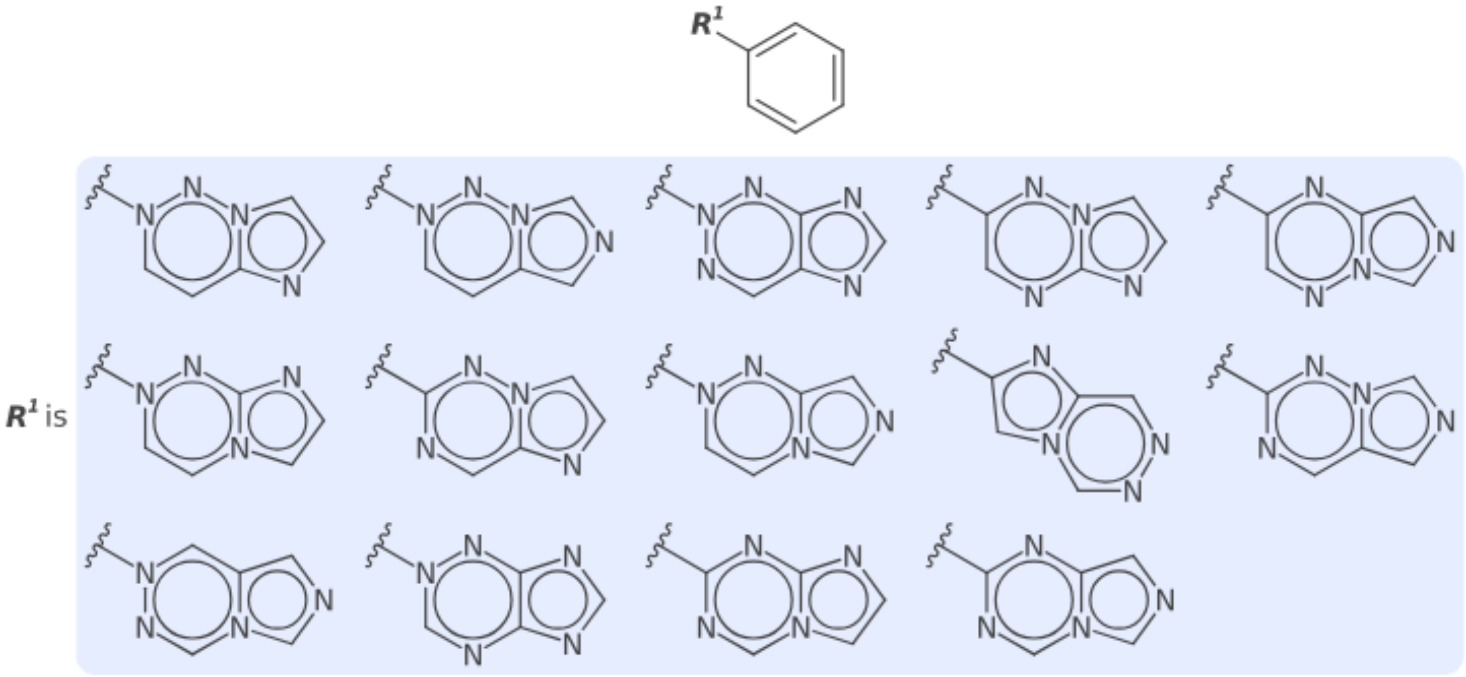

triazine




imidazotriazine

# OPGIN: SUBSTITUTING R-GROUPS

Locanted prefix onto alternative groups.

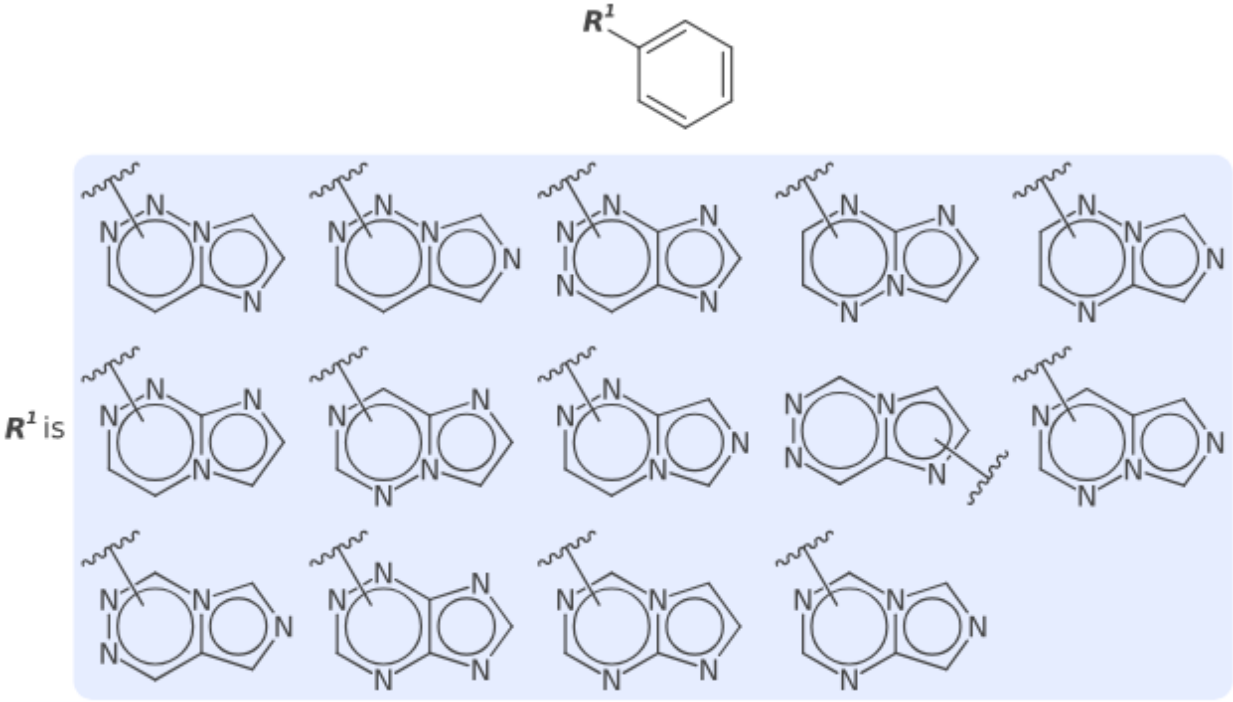



2-phenyl imidazotriazine

```
*c1ccccc1 |$_R1$,RG:_R1={*n1nn2c(cc1)ncc2 |$_AP1$|},{*n1nn2c(cc1)cnc2 |$_AP1$|},{*n1ncc2c(n1)ncn2 |$_AP1$|},
{*c1cnc2n(n1)ccn2 |$_AP1$|},{*c1nc2n(nc1)cnc2 |$_AP1$|},{*n1nc2n(cc1)ccn2 |$_AP1$|},{*c1nn2c(cn1)ncc2 |$_AP1$|},
{*n1nc2n(cc1)cnc2 |$_AP1$|},{*c1nc2n(cnnc2)c1 |$_AP1$|},{*c1nn2c(cn1)cnc2 |$_AP1$|},{*n1ncn2c(c1)cnc2 |$_AP1$|},
{*n1nc2c(nc1)ncn2 |$_AP1$|},{*c1nc2n(cn1)ccn2 |$_AP1$|},{*c1nc2n(cn1)cnc2 |$_AP1$|}|
```

# OPGIN: SUBSTITUTING R-GROUPS

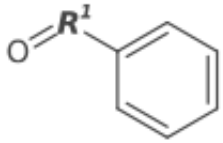
Unlocanted prefix onto alternative groups.



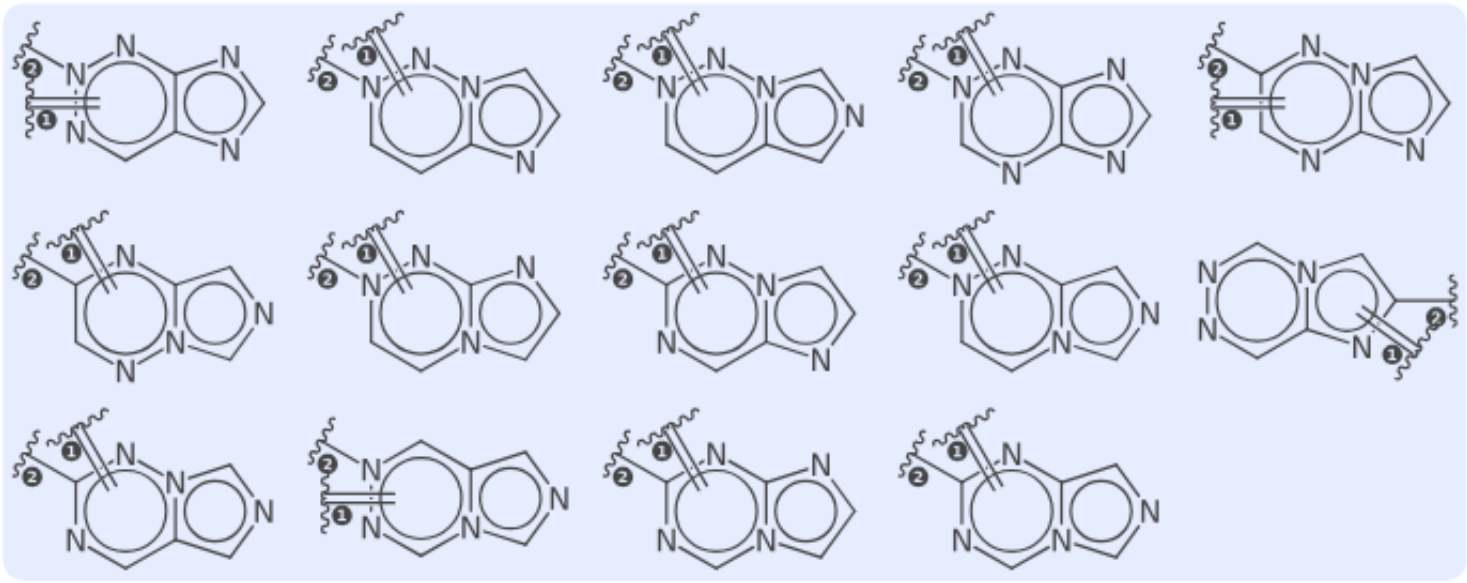

phenyl imidazotriazine

**Very tricky to handle variable attachment point, many tools reject!**

# EXAMPLES



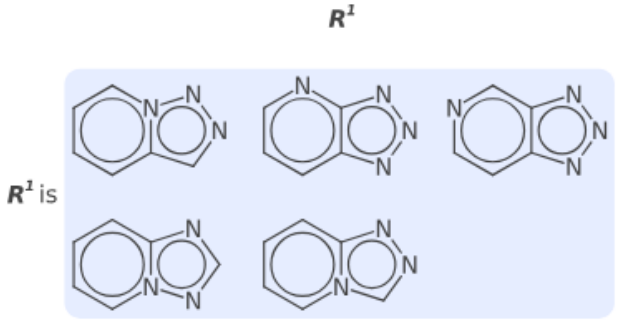

$R^1$ is



2-phenyl imidazotriazinones

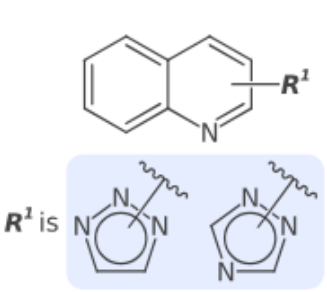
(US6362178B1 - Bayer)

# EXAMPLES

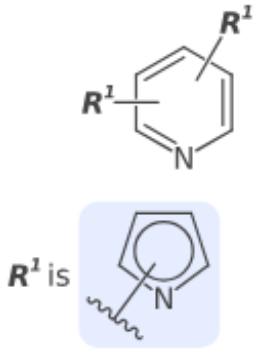



triazolopyridine

(WO2017221092A1 - Novartis)

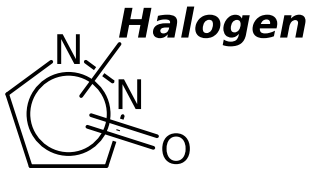



quinolinyltriazole

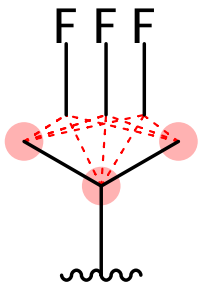



bisazolepyridine

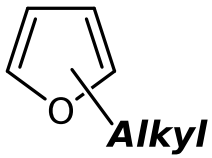
bisazolylpyridine



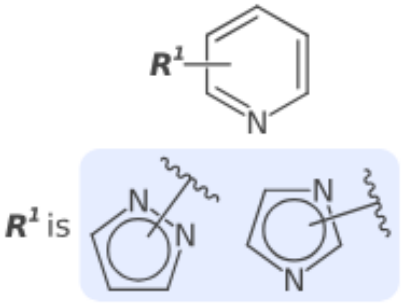

halogen-substituted pyrazolone




trifluoroisopropyl




alkylfurane



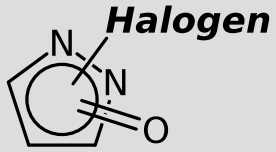

diazolylpyridine

# SEARCHING

*"You should never have to see a SMILES"* - David Weininger

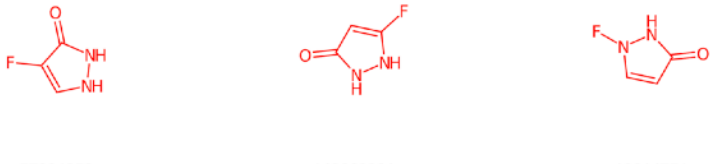
**User Input:** "halogen-substituted pyrazolone"

**Internal:**




**.*=O.n1nccc1 |$Halogen_p$,m:1:4.5.6.7.8,2:4.5.6.7.8|

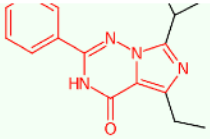
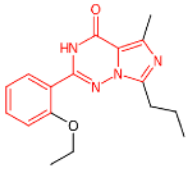
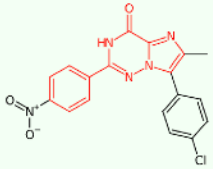
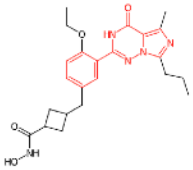
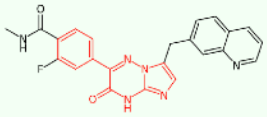
**Results:**



# SEARCHING CHEMBL

*Query: "2-phenyl imidazotriazinones"*

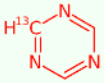
Matched **16** in **0** ms

| | | |
|---|---|---|
| ☐ 2 |  | $C_{17}H_{18}N_4O$<br>294.351 |
| ☐ 3 |  | **CHEMBL352722**<br>$C_{17}H_{20}N_4O_2$<br>312.366 |
| ☐ 4 |  | **CHEMBL5864568**<br>$C_{18}H_{12}ClN_5O_3$<br>381.773 |
| ☐ 5 |  | **CHEMBL4208487**<br>$C_{23}H_{29}N_5O_4$<br>439.507 |
| ☐ 6 |  | **CHEMBL4088155**<br>$C_{23}H_{17}FN_6O_2$<br>428.418 |

# SEARCHING PUBCHEM

*Query: "triazines"*

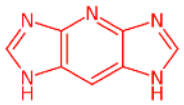
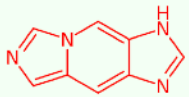
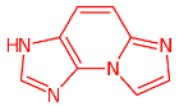
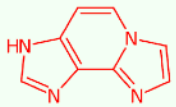
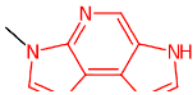
Matched **1,243,024** in **1136** ms

| | | | |
|---|---|---|---|
| ☐ 1 |  | **67520** | $C_3H_3N_3$ 81.076 |
| ☐ 2 |  | **123047** | $C_3H_3N_3$ 81.076 |
| ☐ 3 |  | **9262** | $C_3H_3N_3$ 81.076 |
| ☐ 4 |  | **67057101** | $C_3H_3N_3$ 82.069 |
| ☐ 5 |  | **22718955** | $C_3H_4N_3$ |

# SEARCHING PUBCHEM

*Query: "bisimidazopyridines"*

Matched **406** in **1651** ms

| | | |
|---|---|---|
| ☐ 1 |  | **85774410**<br>$C_7H_5N_5$<br>159.148 |
| ☐ 2 |  | **70391557**<br>$C_8H_6N_4$<br>158.160 |
| ☐ 3 |  | **45120535**<br>$C_8H_6N_4$<br>158.160 |
| ☐ 4 |  | **141532098**<br>$C_8H_6N_4$<br>158.160 |
| ☐ 5 |  | **144220593**<br>$C_8H_7N_5$ |

# SUMMARY

Chemical search systems are advancing and improving

A search box may not do what you think/expect

Avoid drawing complex structure diagrams

Can exploit the ambiguity in chemical nomenclature to provide generic structure representation and better search results

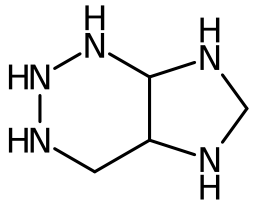


# FUTURE WORK

Struggle against OPSIN trying to do the sensible thing

Handling of and/or lists in IUPAC names and merging with existing OntoGrep semantics *"unsubstituted triazine"*

Hydrogen complications - ideally emit aromatic/tautomeric invariant representation




*"1H,2H,3H,4H,4aH,5H,6H,7H,7aH-imidazotriazine"*

Performance improvements and corner case handling in Arthor

# ACKNOWLEDGEMENTS

Visit the **booth** to see some demos!


- Daniel Lowe
- Egon Willighagen
- NextMove Software Team
  - Delia Sayle
  - Ingvar Lagerstedt
  - Rachael Pirie
  - Lina Heinzke




NextMove Software (Aug 2023)

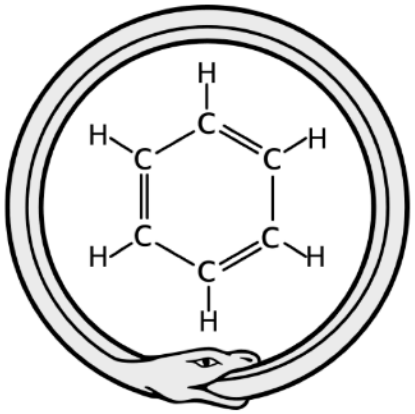
Slides

# KEKULÉ'S DREAM

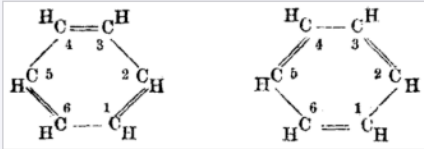
## Kekulé's dream [ edit ]

The German organic chemist August Kekulé described the eureka moment when he realised the structure of benzene, after he saw a vision of Ouroboros:[25]

> I was sitting, writing at my text-book; but the work did not progress; my thoughts were elsewhere. I turned my chair to the fire and dozed. Again the atoms were gamboling before my eyes. This time the smaller groups kept modestly in the background. My mental eye, rendered more acute by the repeated visions of the kind, could now distinguish larger structures of manifold conformation: long rows, sometimes more closely fitted together; all twining and twisting in snake-like motion. But look! What was that? One of the snakes had seized hold of its own tail, and the form whirled mockingly before my eyes. As if by a flash of lightning I awoke; and this time also I spent the rest of the night in working out the consequences of the hypothesis.



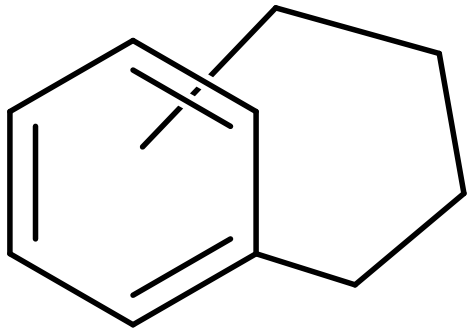
The ouroboros, Kekulé's inspiration for the structure of benzene



Kekulé's proposal for the structure of benzene (1872)

# MAYFIELD'S DREAM

With positional variation, there is nothing in the file formats which prohibits the creation of a ring:



*"ouroboros"*

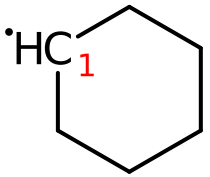
`c1ccccc1CCCC* |m:10:0.1.2.3.4|`

# BONUS

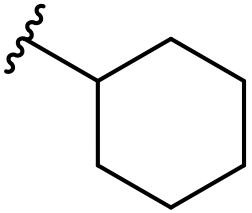


# CTFILE VS CXSMILES

There are philosophical differences in how attachment/phantom atoms are stored:




**CTfile** store the attachment point as an annotation on an existing atom



**CXSMILES** store the attachment as an explicit node in the connection table

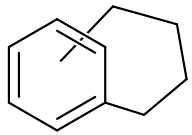
**Attachment Point [Rgroup]**

`M APOnn2 aaa vvv ...`

`vvv`: Indicates whether atom `aaa` of the Rgroup member is the first attachment point (`vvv` = 1), second attachment point (`vvv` = 2), both attachment points (`vvv` = 3); default of 0 = no attachment.

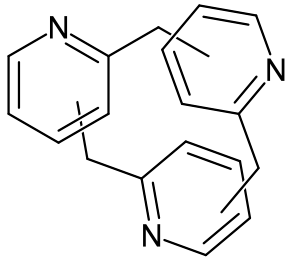
Pros/cons but explicit attachment more options IMO

# QUIRKY POSITIONAL VARIATION



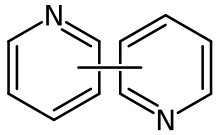
*"ouroboros"*

`c1ccccc1CCCC* |m:10:0.1.2.3.4|`



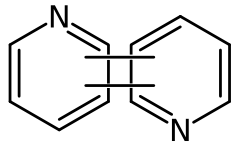

*"triskelion"*

`c1ccnc(c1)CC.c1ccnc(c1)CC.c1ccnc(c1)CC`
`|m:23:0.1.2.3.4.5,7:8.9.10.11.12.13,15:16.17.18.19.20.21|`




*"bridge to nowhere"*

`**.n1ccccc1.n1ccccc1`
`|m:0:2.3.4.5.6.7,1:8.9.10.11.12.13|`

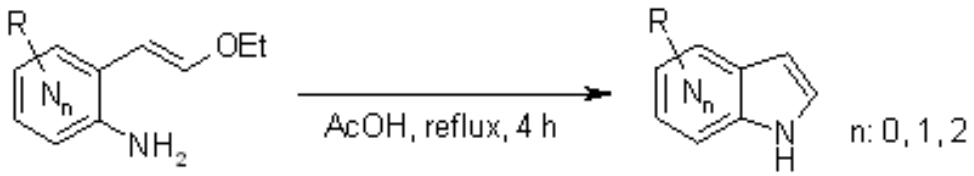



*"double rainbow"*

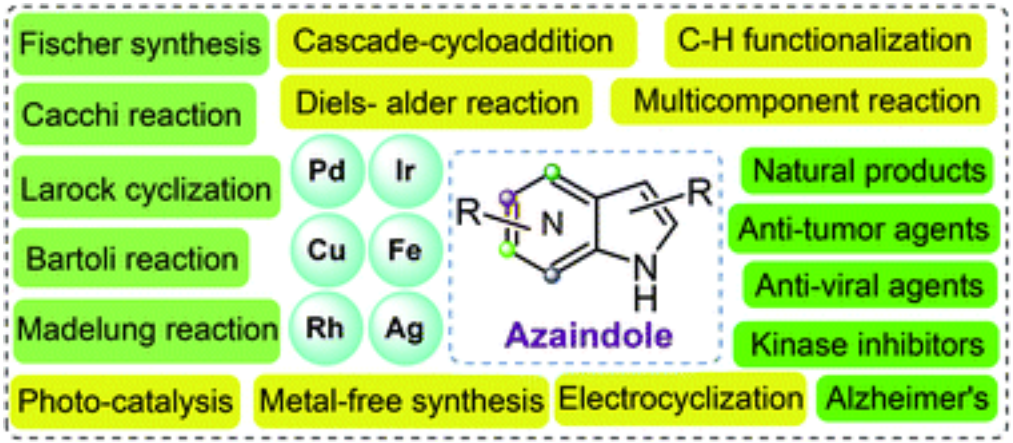
`**.**.n1ccccc1.n1ccccc1`
`|m:0:5.6.7.8.9,1:11.12.13.14.15,2:5.6.8.7.9,3:11.12.13.14.15|`

Examples of non-trivial positional variation queries to handle from NextMove Software's Arthor test suite, also supported by Chemistry Development Kit.

# VALID?

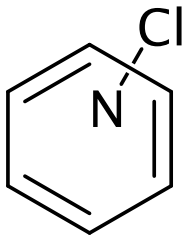



https://www.organic-chemistry.org/synthesis/heterocycles/benzo-fused/azaindoles.shtm

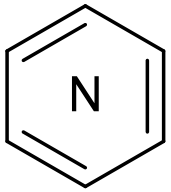



https://pubs.rsc.org/en/content/articlelanding/2021/qo/d0qo01079k

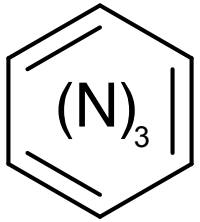
# VALID?




c1ccccc1.[N]Cl |m:6:0.1.2.3.4.5|




c1ccccc1.[N] |m:6:0.1.2.3.4.5|




triazine?

c1ccccc1.[N].[N].[N] |m:6:0.1.2.3.4.5,7:0.1.2.3.4.5,8:0.1.2.3.4.5|

Standardisation committee

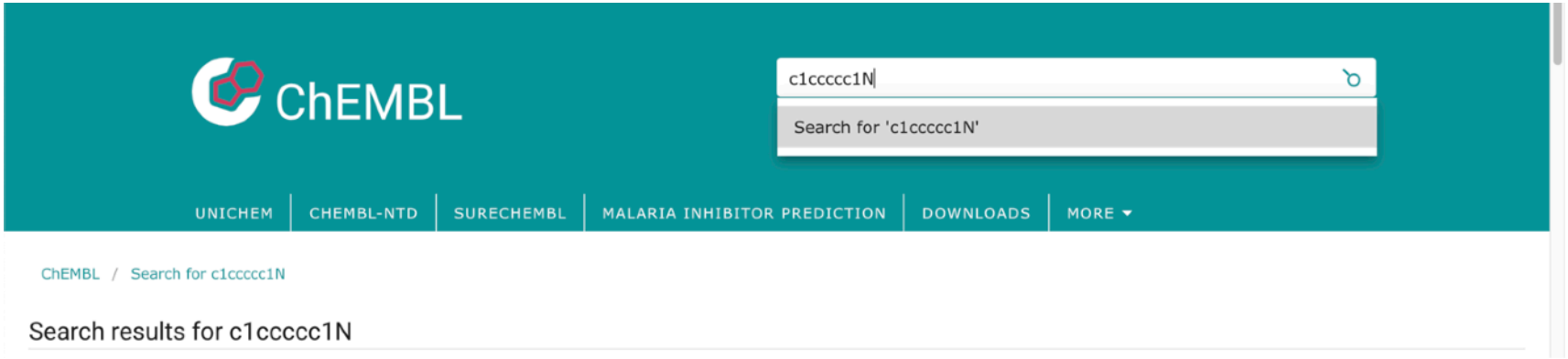
A variable bond-fusion operator would also be useful to reduce enumeration

# EXTRA

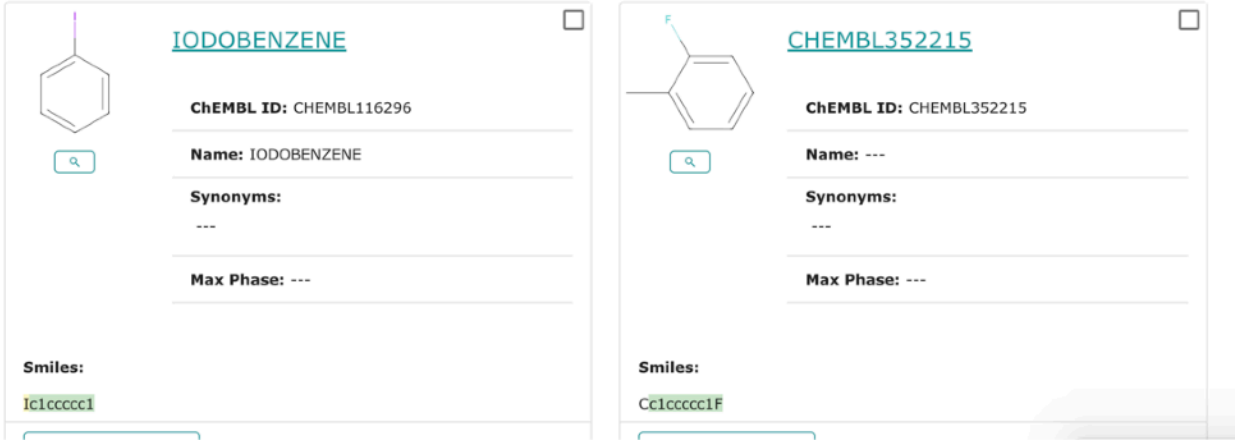


# MOTIVATION

**ChEMBL** possibly only **“text based”**




Hits 5 and 6:



# MOTIVATION

**ChEMBL** possibly only **“text based”**

