

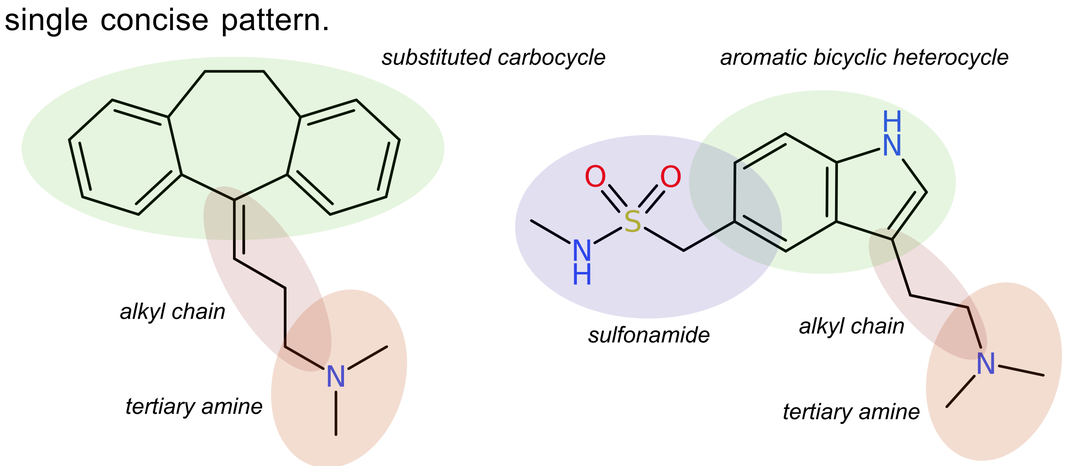
ADVANCED CHEMICAL QUERYING - GENERIC STRUCTURAL SEARCHES FROM IUPAC NOMENCLATURE

Michael Blakey, John Mayfield and Roger Sayle
NextMove Software Ltd, Cambridge, UK.

Introduction

For any search query language writing succinct, accurate queries that do exactly (and only) what you intend is a skill that demands serious practice and knowledge within the language domain. SMARTS is the industry standard for this, but what can be done for queries that do not naturally fit to a singular pattern? The natural language of chemists includes many such generic terms and ontologies. Several largely successful attempts have been made previously to fully classify chemical structures described by these classes[1,2]. Public resources like ChEBI, CHMO, and ChemFOnt, offer tools and databases for searching and exploring these ontologies [3,4,5].

Homology groups are a subset of these terms that define sets of structurally related substituents such as alkyl, aryl, or heterocyclic moieties, commonly used in Markush structures to represent chemical variability. While conceptually useful for capturing chemical diversity, these groups present significant challenges for structure searching and indexing. Existing cheminformatics tools, including SMARTS, often struggle to represent and retrieve compounds containing abstract or generalised substructures with a single concise pattern.



[#6]-[#7](-[#6])-[*]-[*] |$;;;Alkenyl_p;Carboaryl_p$, atomProp:3.TEXTNOTES.E1:3.SIZE.LO:4.TEXTNOTES.NR3|

[#6]-[#7]S(=O)(=O)[#6]-[*]-[*]-[#7](-[#6])-[#6] |$;;;;;;Heteroaryl_p;Alkyl_p;;;$|

Without a well defined ontology, searching chemical databases using these terms can at best return some exact text matches if other synonym identifiers are present, or at worst nothing at all. Typical search queries may include:

? Heterocycles with an even number of oxygen atoms
? Aromatic or aliphatic pyrimidines
? Indole substructure
? Optionally substituted indoles
? Imidazotriazines
? Oxygen containing cyclic compounds
? Metal-containing drugs
? Reactions with phenolphthalein

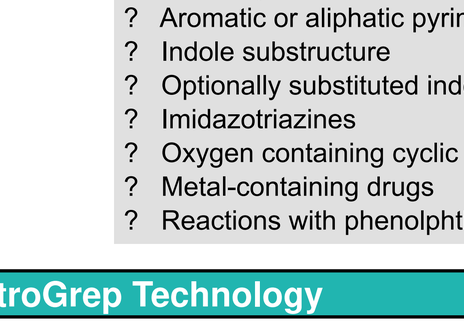
OntroGrep Technology

- **OntoGrep**: natural language and generic chemical nomenclature queries that previously would be too broad to be accepted as valid to perform *fast searches* of chemical databases.

- Query handling by assigning *Part of Speech* (PoS) tags to words. Potential tags include: homology groups, functional groups (as SMARTS) or generic properties.

- *Ambiguity* in tags handled using English language (and potentially special notation). Typically verb phrases e.g. "contained in", "attached to" following naturally after the defined group.

- Build a *parse tree* to decide best strategy - SMARTS (where possible) built using NextMove Software's Patsy.

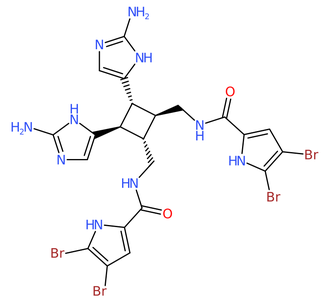


"charged substituted nitrogen spirocycles"

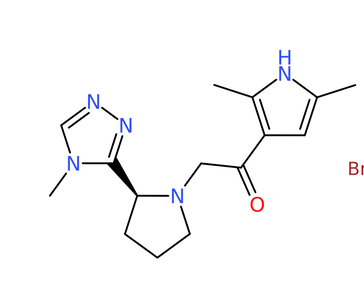
- *Homology groups* use mixture of SMARTS and search functions for the most appropriate match algorithm.

- Speed remains inline with highly optimised SMARTS searching.

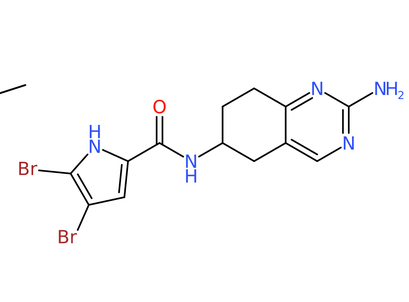
"TRI-SUBSTITUTED PYRROLES WITH ALPHA CARBONYL"



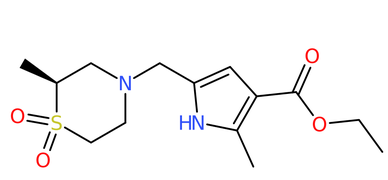
CHEMBL3926591



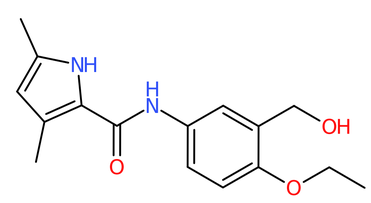
CHEMBL4967010



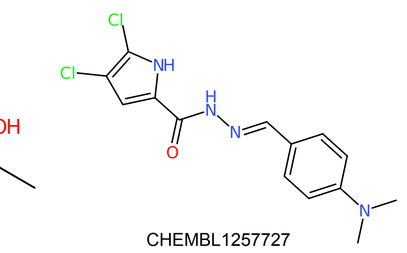
CHEMBL3589523



CHEMBL4993130

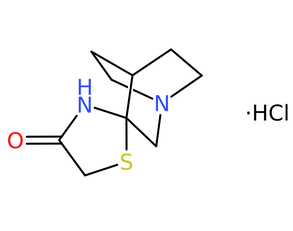


CHEMBL3452579

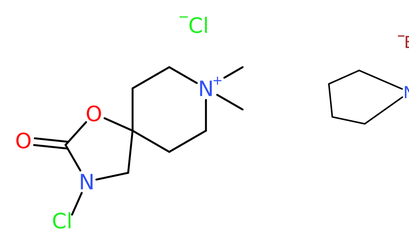


CHEMBL1257727

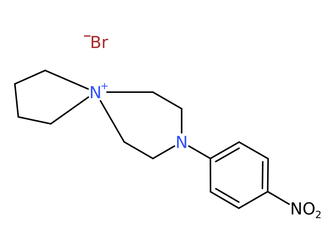
"SPIROCYCLIC COMPOUNDS WITH HALOGEN SALTS"
"HALOGEN SALTS OF SPIRO-SYSTEMS"



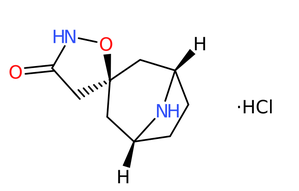
CHEMBL1383946



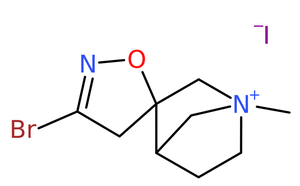
CHEMBL1771399



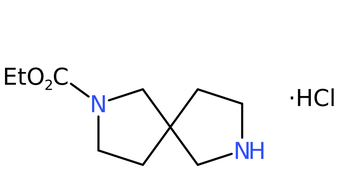
CHEMBL251191



CHEMBL2158872



CHEMBL1917554



CHEMBL3497417

ONTOGREP QUERIES INTERGRATED INTO PISTACHIO



Conclusion & Next Steps

OntoGrep is a natural language input system aimed to enable search of homology group style queries on pre-existing chemical databases. With integration into NextMove Software's Pistachio (due in the **2025Q3** release), reaction databases can be searched with queries such as "substituted pyrrole reactants and intermediates". Homology matching is very natural for those versed in IUPAC style nomenclature, and efficient search based off this language is a move in the right direction.

Acknowledgements

The whole team at NextMove Software!

Bibliography

[1] Greiner, R., & Wishart, D. S. (2016). ClassyFire: automated chemical classification with a comprehensive, computable taxonomy.
[2] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res.
[3] McLaughlin, J. A., Lagrimas, S. J., Iqbal, H., Parkinson, H., & Harmse, H. (2025). The Ontology Lookup Service. Zenodo.
[4] Wishart DS, Girod S, Peters H, Oler E, Jovel J, Budinski Z, Milford R, Lui VW, Sayeeda Z, Mah R, Wei W, Badran H, Lo E, Yamamoto M, Djoumbou-Feunang Y, Karu N, Gautam V. (2023) ChemFOnt: the chemical functional ontology resource. Nucleic Acids Res.
[5] Pascazio L, Rihm S, Naseri A, Mosbach S, Akroyd J, Kraft M. (2023) Chemical Species Ontology for Data Integration and Knowledge Discovery.

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