dotmatics
knowledge solutions

Advanced Searching using Chemselector

Rob Brown, Jameed Hussain, Gianpaolo Bravi and Mike Hartshorn

Introduction

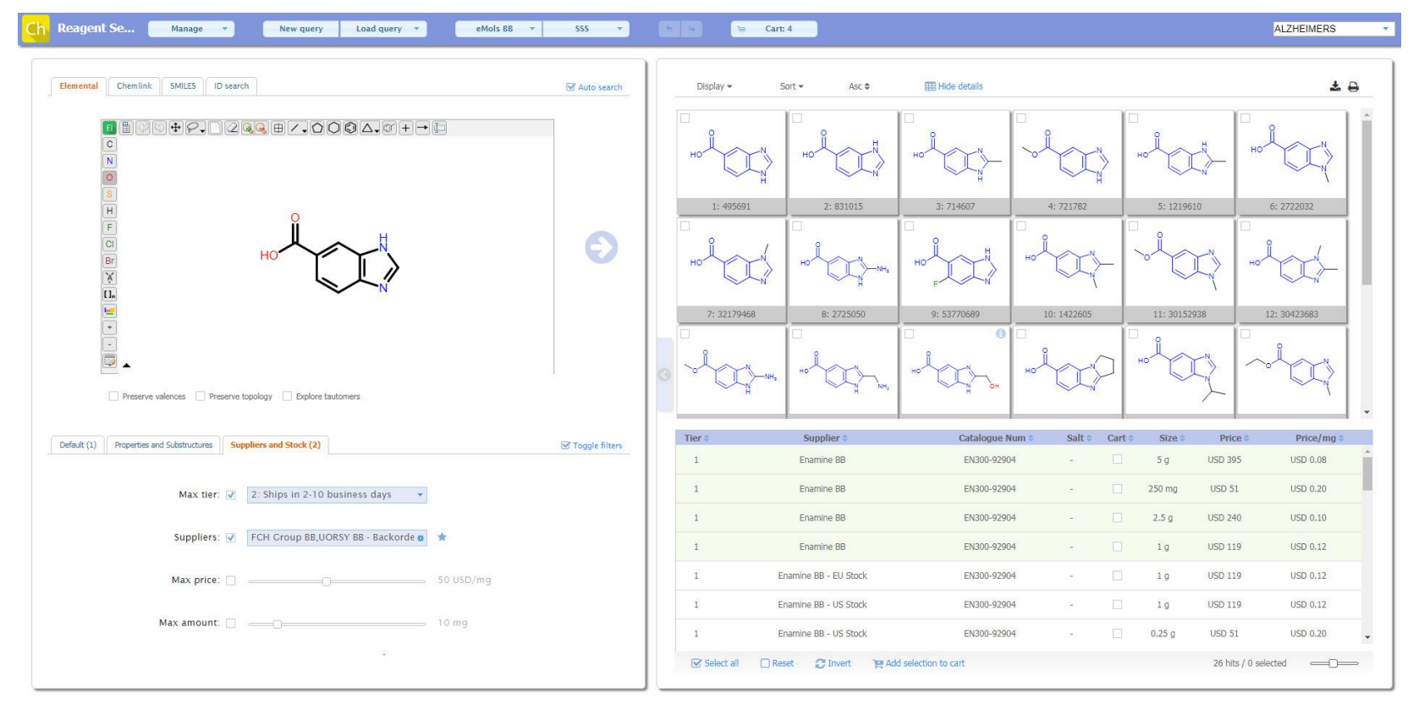
Chemical structure searching is an important technique within small molecule drug discovery. It is used to find available analogues to expand the SAR around a biological active as well as to identify appropriate reagents for compound synthesis. Modern search systems typically provide a few search types including exact structure, substructure and similarity searching and almost all provide these search services through an Oracle cartridge.

There are two challenges inherent in providing these types of workflows to chemistry teams

1. Building, maintaining and distributing very large (10M+) databases of compounds such as screening collections (e.g eMolecules) or public data (e.g. SureChEMBL), with up to date data
2. Providing a range of search types to allow important tasks such as lead hopping or SAR expansion in a user interface appropriate for end user chemists rather than power users such as cheminformaticians and modellers.

Chemselector

Chemselector (Ch) is a new web-based application developed by Dotmatics for chemical searching. The application utilizes a novel chemical search engine (minpoint) and it provides a intuitive interface for bench chemists to perform the powerful (and often complex) techniques used in chemical structure searching.



Minpoint

Following early innovation in cheminformatics research, cartridges became the norm for database searching in the 1990s and since then the pace of innovation in this field has slowed dramatically. However, cartridges have definite limitations especially when needing to build and distribute very large databases on a regular basis.

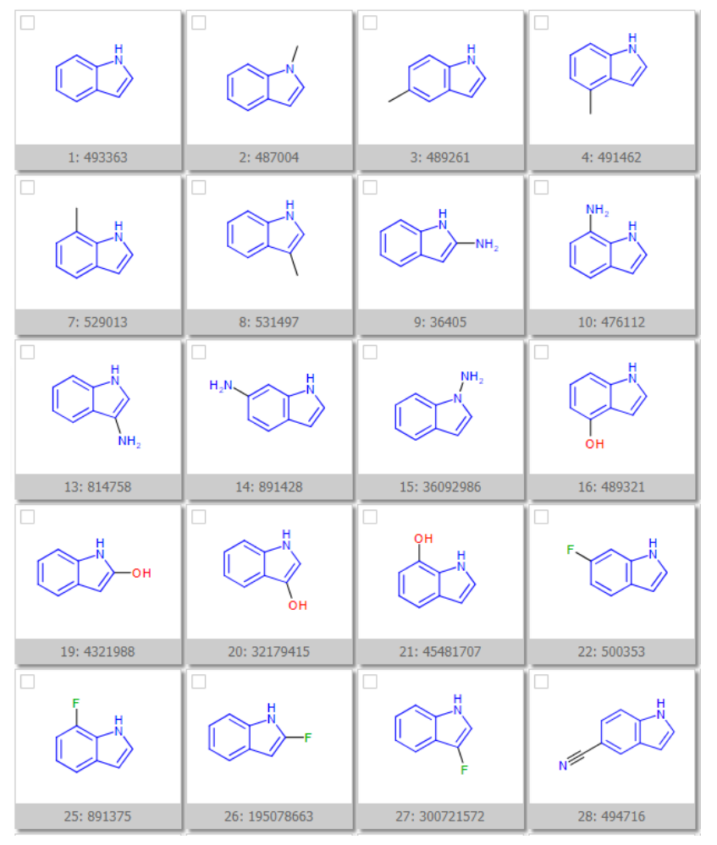
Minpoint is a novel search technology designed for very fast search performance – a substructure search against 10+ million structures can be run on a standard laptop. The search performance means almost all searches can be performed interactively as a structure is being drawn. The high performance can be utilized to help deal with tautomers when searching. Additionally, searching for matched molecular pairs of a query molecule in a large database becomes possible.

Minpoint does not rely on an Oracle cartridge for searching, instead holding structures and indexes as files on an application server. The scalability of this approach to very large databases is important since many tasks such as compound or reagent searching, or searching public datasets of patents, require the datasets that would be prohibitively expensive to build, maintain and distribute as Oracle databases. With Minpoint these large databases can be distributed as files for download. Installation of these dbs with Chemselector simply involves telling the system where the files are accessible.

Searching

The application allows the user to perform most of the typical searches required. The types of querying that can be carried is shown below. Additionally, a number of different dbs relevant for medicinal chemistry are available.

- SSS, Exact, Sim, ID and List searches
- eMols BB & SC, ChEMBL, SureChEMBL or add your own
- Query against chemistry annotations
  - Functional groups, heterocycles, protecting groups
  - Include/exclude options
- Query against phys. chem. properties
  - Within range e.g. 350≤ MW ≤ 450
  - Within delta from specified structure e.g. MW ±100
- Query against stock information
  - Supplier, Amount, Price, Tier etc.



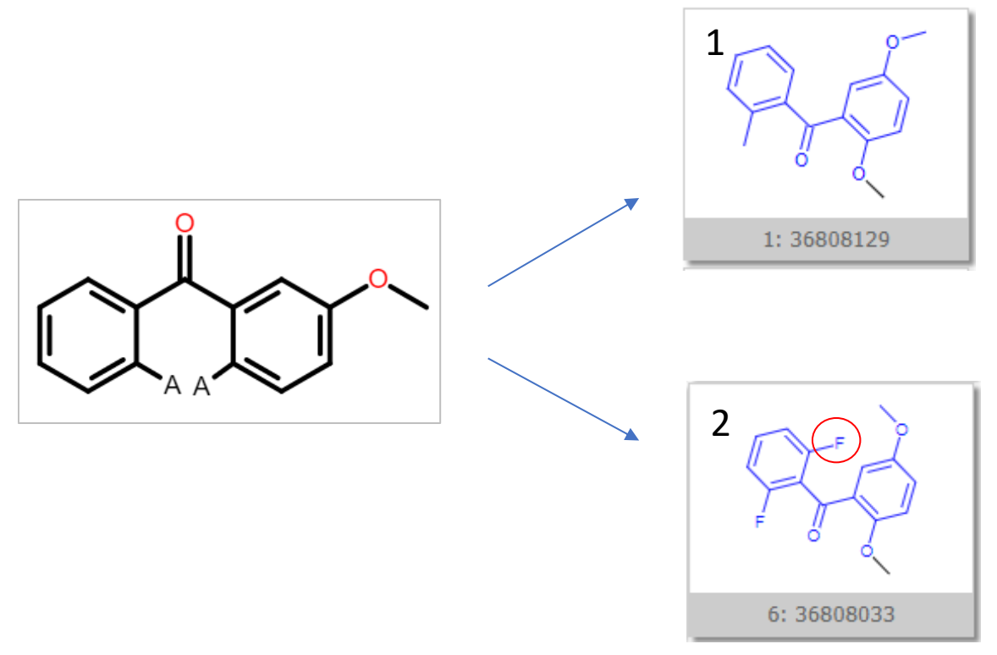
Substructure searching

Substructure searching is a powerful tool when it comes to searching for small molecule analogues that fit the SAR requirements for a given biological target. However, formulating an appropriate substructure query to capture the SAR requirements of your target can be a daunting task.

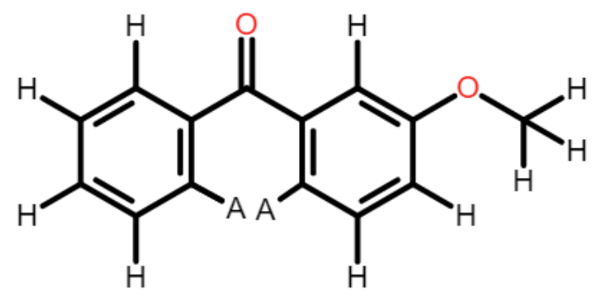
One possible way to overcome the problem of generating a specific substructure query is to use a more general one and remove any unwanted compounds by eye. This can be an onerous task if the search query generates many hits. Hence any features that can be used to make the creation of substructure queries easier and reduce the number unwanted compounds from a substructure search are likely to be very useful for chemists. A couple of these are available in Chemselector.

Preserve valence

Given the substructure query below (where the "A" is any non-hydrogen atom), the search will retrieve compound 1 and compound 2. However, if you want to only allow substitutions at the position marked with an A, compound 2 is not what you want.

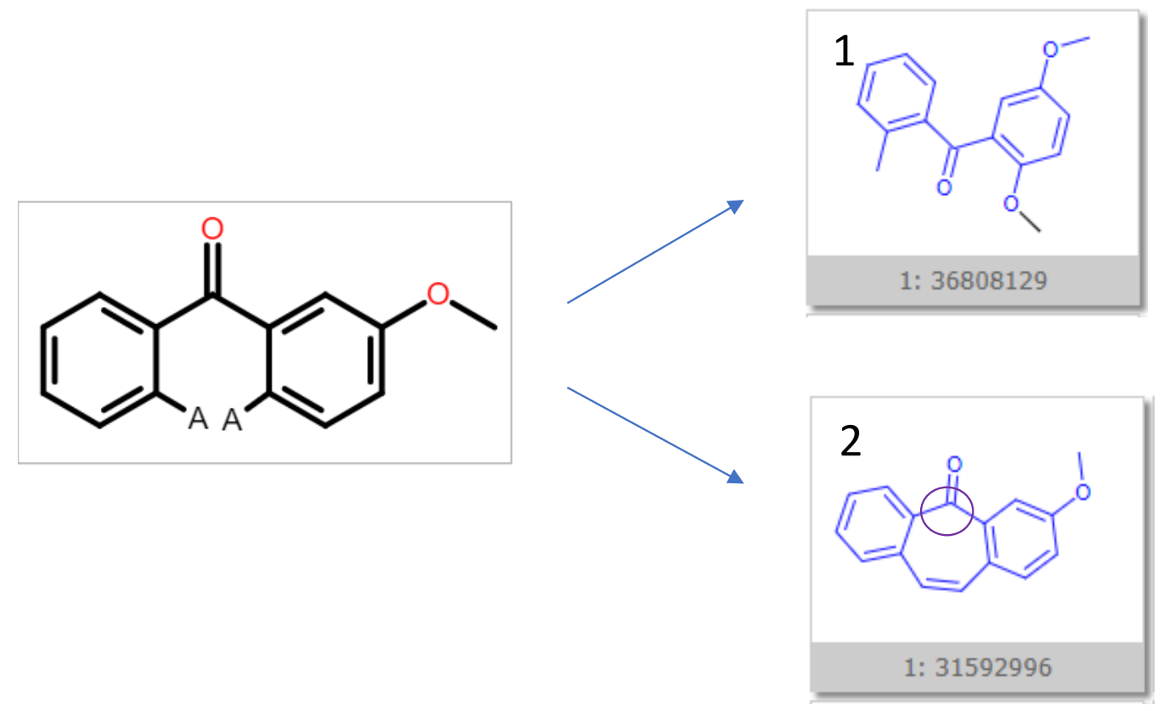


To get the query to only retrieve compounds where only the A is substituted, you need to add explicit hydrogens at every position you do not want a substitution (see below). This can be tedious. To get the desired query search behaviour (without having to add hydrogens everywhere) you can use the "Preserve valences" option in Chemselector.



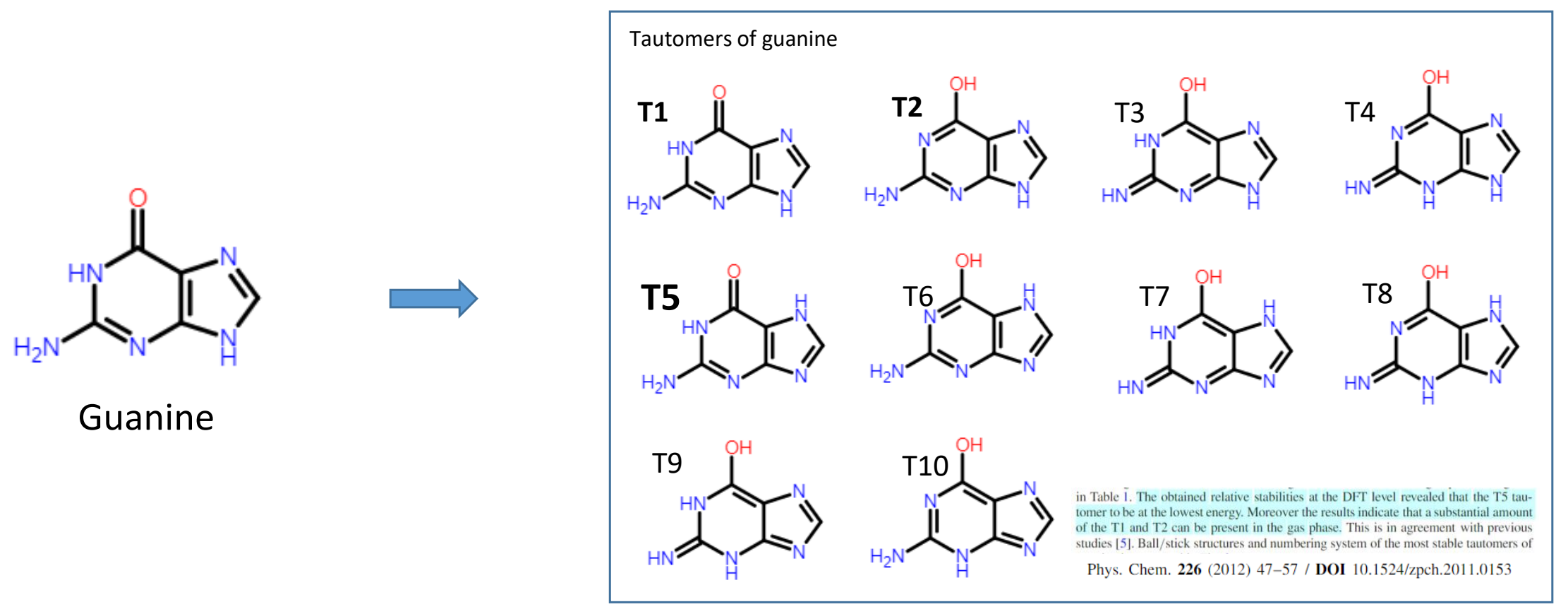
Preserve topology

Another useful feature is the "Preserve topology" option. This preserves the topology or ring / non-ring nature of the query atoms. In the example below, compounds 1 and 2 are both valid hits for the shown substructure query. If you would like the search results not to contain hits where the carbonyl moiety is part of a ring system, the "Preserve topology" option can be used. With this option, compound 2 will not be retrieved as the circled carbonyl atom is in a ring (unlike the query). Note that it is possible to do achieve the same results using the SMARTS language, however very few bench chemists use SMARTS for substructure searching.



Tautomers

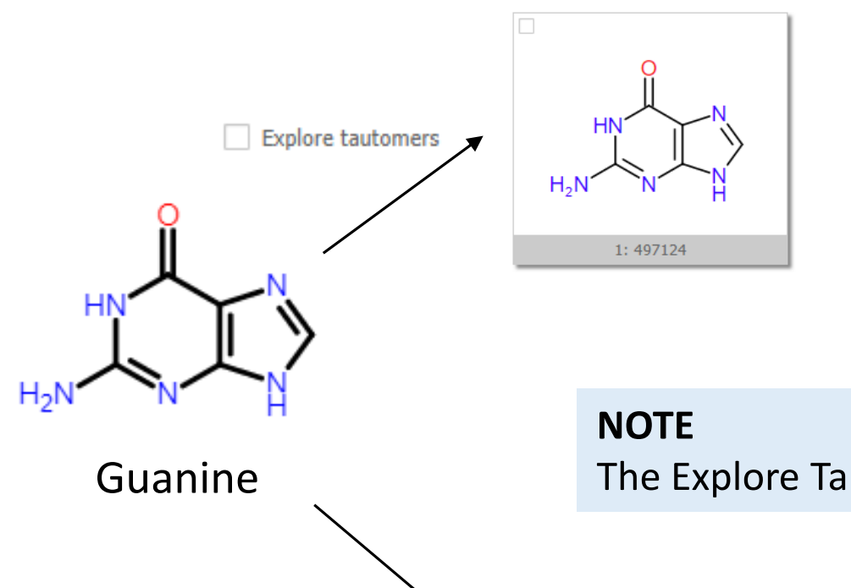
Tautomers are isomers of a compound that can exist in equilibrium and are readily interconverted. One of the more famous examples, guanine, is shown below. When you are in the situation where you need to search for a tautomeric compound in a chemical database, care needs to be taken to make sure it is drawn as it appears in the database. This is likely to be straight-forward in cases where you are familiar with how it is stored in the database. However, when this is not the case (e.g. a 3rd party database) you probably need to try several tautomeric forms.



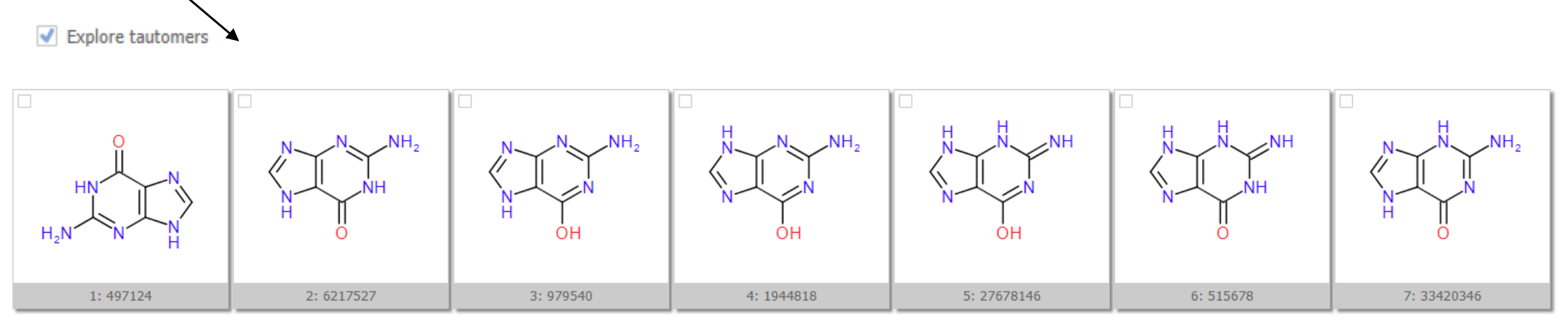
in Table 1. The obtained relative stabilities at the DFT level revealed that the T5 tautomer to be at the lowest energy. Moreover the results indicate that a substantial amount of the T1 and T2 can be present in the gas phase. This is in agreement with previous studies [5]. Ball/stick structures and numbering system of the most stable tautomers of

Phys. Chem. **226** (2012) 47–57 / **DOI** 10.1524/zpch.2011.0153

ChemSelector has an option called "Explore Tautomers" which takes your drawn structure, enumerates most (if not all) of the tautomeric states it can exist in and searches using each one. Therefore, even if the compound exists in the database in a different tautomeric state than the structure you have drawn, it should be found. The results of an exact search for guanine in the eMolecules database (May 2017) with and without the "Explore Tautomers" option are shown below. As can be seen all the different tautomeric forms of guanine in the database are found.

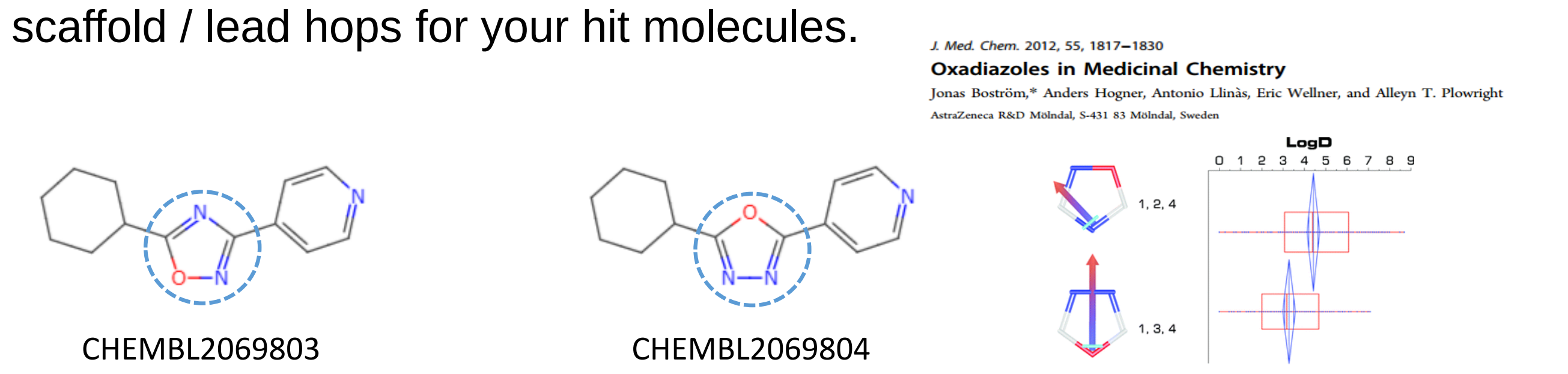


NOTE
The Explore Tautomers option is available for all structure searching methods



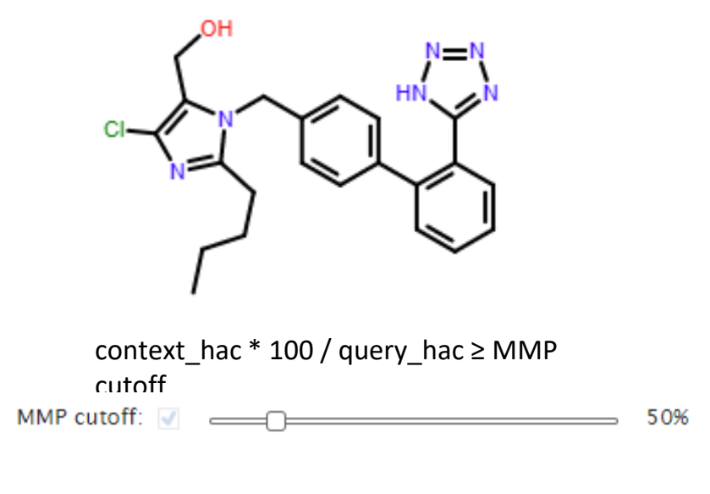
Matched molecular pair search

Matched molecular pairs (MMPs) are pairs of compounds that differ by a single local change. They can be a useful source of analogues or scaffold / lead hops for your hit molecules.



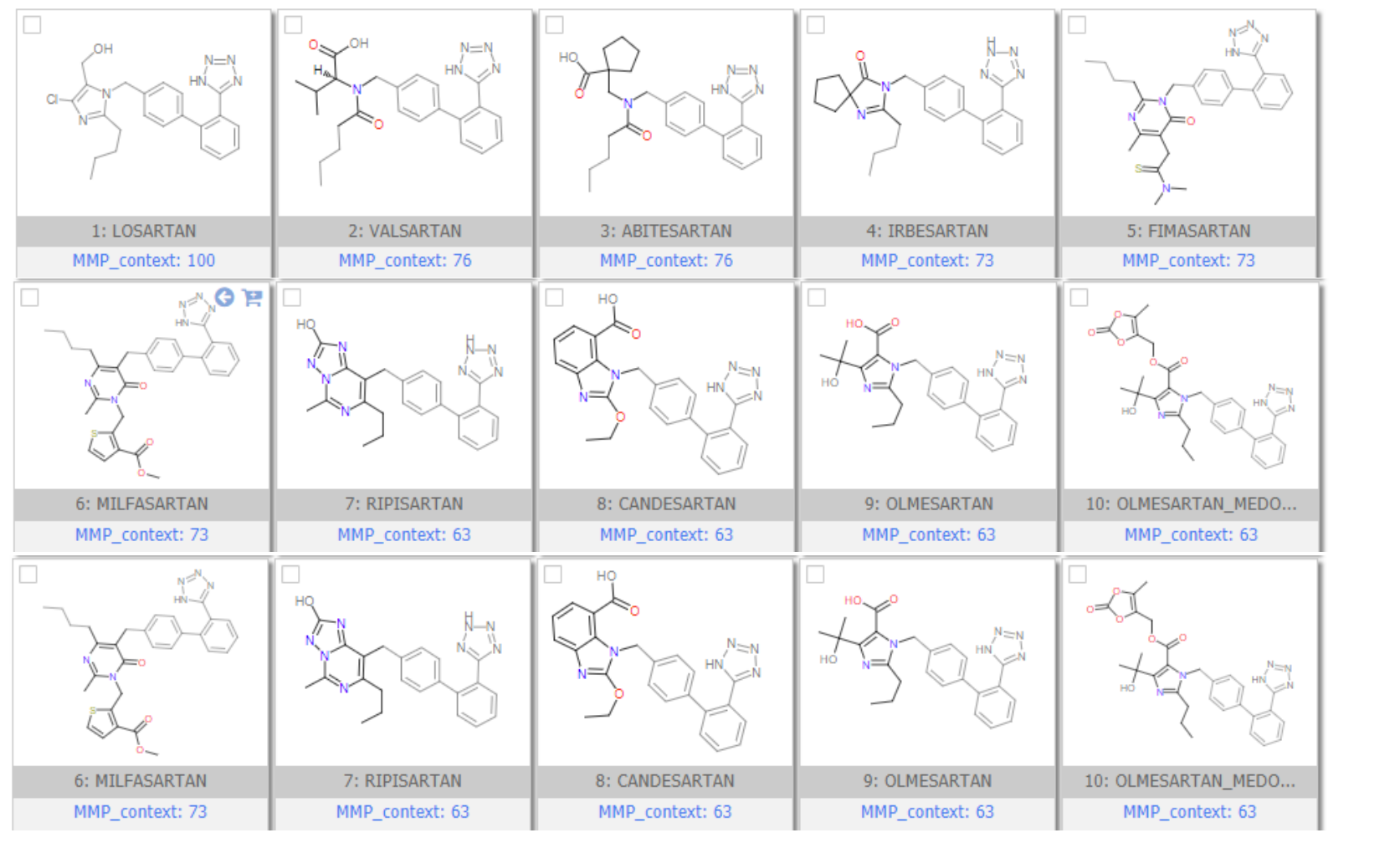
J. Med. Chem. 2012, 55, 1817−1830
Oxadiazoles in Medicinal Chemistry
Jonas Boström,* Anders Hogner, Antonio Llinàs, Eric Wellner, and Alleyn T. Plowright
AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden

Chemselector can be used to search for MMPs of an input compound within any of the databases selected. A MMP cutoff is used to define the minimum percentage of heavy atoms in common between the query and the retrieved compounds and this can be used to control the number of hits retrieved.



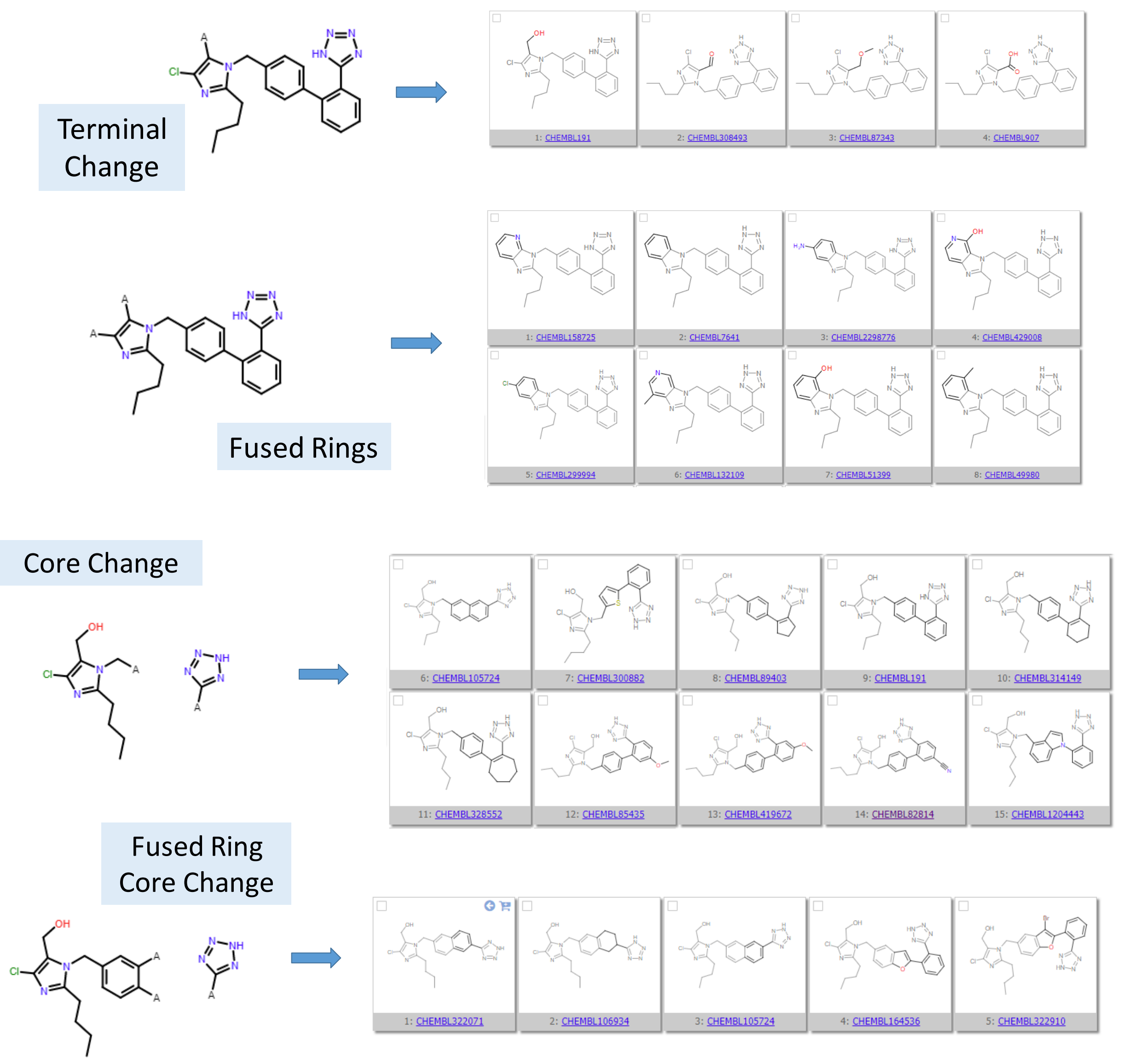
context_hac * 100 / query_hac ≥ MMP cutoff

MMP cutoff: ☑ 50%



NOTE
You can use the MMP cutoff in combination with other property filters e.g. Δ-HAC to ensure that the structure hits are within the desired limit, e.g. that the structure change is not too big.

The MMP search can also be directed to identify MMPs with terminal or core changes at a specific position. Example of each of these searches are shown below.



Conclusion

The powerful and often complex techniques employed when performing a chemical structure search mean designing an appropriate interface is paramount if the system is to be used by non-computational experts such as bench chemists. ChemSelector pulls together several powerful search methods in an interface appropriate for most bench chemists to use.

Contact: jameed.hussain@dotmatics.com