Open Source Implementation of Astex Fragment Network and Availability as REST Service

Tim Dudgeon[1], Anthony Bradley[3,5], Alan Christie[1], Chris Leeding[2], Matt Segall[2], Rachael Skyner[3], Ric Gillams[3,4], Frank von Delft[3,4]

1. Informatics Matters Ltd. 2. Optibrium Ltd., 3. Diamond Light Source, 4. Structural Genomics Consortium, University of Oxford, 5. Exscientia Ltd.

AIM: *Make the published Astex Fragment Network methodology readily available to a wide audience, applied to commercially available compounds*

Fragment Network Origins

The Fragment Network concept was created at Astex and used in their fragment-based drug design work, primarily for the purpose of identifying follow-up compounds from their initial fragment leads. The work is described in detail in [1].

The XChem project at the Diamond Light Source is performing high-throughput fragment screening and saw the benefit of this approach for following up its fragment screening hits. Whilst the Astex publication contained a thorough description of the methodology, the source code was not released and, as the code was based on the Daylight Toolkit [2], it would not have been suitable for a modern open implementation. It was therefore decided to re-implement this work using the RDKit cheminformatics toolkit [3].

Implementation

The methodology described by Astex was reproduced as a set of Python modules and scripts that are released under an Apache-2.0 license and can be found in GitHub [4]. The methodology described by Astex was followed as closely as possible, including the format of the output files. Results described in the Astex paper are reproduced.

A number of datasets have been processed (Table 1). Processing is split across a small cluster of computers with the nodes and edges result files being de-duplicated and then augmented with additional information from the source files, such as vendor codes, pricing information and activity data. Final files are in csv format data suitable for loading into the Neo4j graph database [5] and queryable using the Cypher query language [6]. Automated pipelines for generating these datasets are being created and are mostly operational. New datasets can easily be added.

Source	Compounds	Nodes	Edges
MolPort (November 2018)	7,486,593	107,899,273 (14)	607,806,848 (81)
Enamine REAL DSI poised library rule of 4	1,394,963	5,380,055 (4)	24,439,600 (17)
Enamine REAL DSI poised library rule of 5	39,765,321	191,230,680 (5)	1,038,917,131 (26)
SENP7 HTS dataset	330,688	6,949,173 (21)	36,188,234 (109)

Table 1: Processing of different datasets.
MolPort: Screening compounds and Building Blocks from Molport.
Enamine REAL DSI poised library - a subset of Enamine REAL database based around the XChem Poised screening library, with Lipinski Rule-of-5 filter applied and with an equivalent rule based on the number 4.
SENP7: an HTS screening dataset provided by Cancer Research UK.
The numbers in brackets in the n-fold increase in size compared to the number of input compounds. The variations reflect the different complexity of the datasets.
Note: filters to the input molecules can be applied to exclude structures with large number of fragments. Typically we use a heavy atom count of 36 (Astex used a limit of 24), but can also filter by the number of direct child fragments of the input structure.

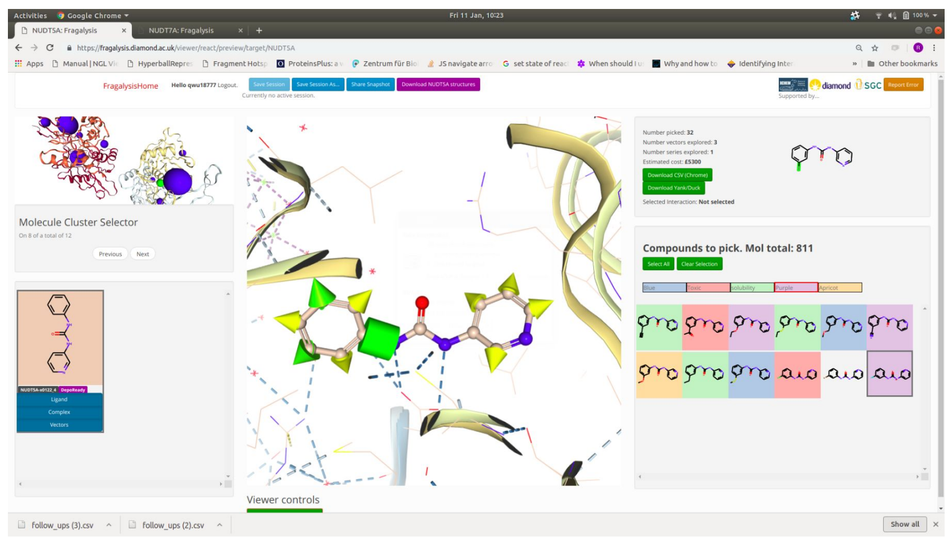
Accessing the Fragment Network Data

The Neo4j database can be queried directly using the Cypher query language. In addition we have created a REST Query API that can be used and is expected to become publicly available. This Query API can also group and classify the compounds based on the type of transformation and can calculate a number molecular properties, including the Tanimoto distance to the query structure base on RDKit and Morgan fingerprints. Results are returned in JSON format.

Query response times are typically less than 1 second.

3D visualisation - Fragalysis

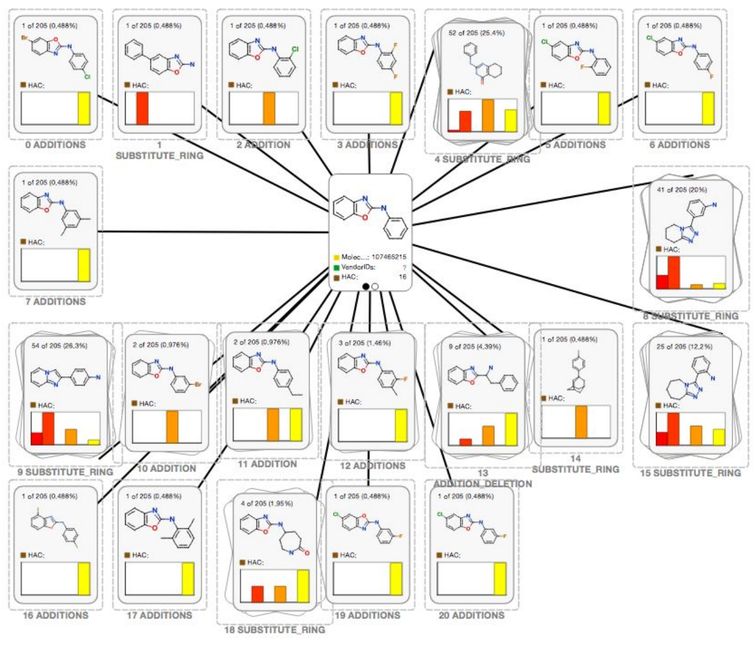
A suite of web based tools, under the collective name of 'Fragalysis' is being created for exploring the follow up potential of XChem screening hits. Data from the Fragment Network is used to propose compounds of interest that could be purchased and tested. The web front end uses NGL Viewer [7] to display the 3D context of the fragment bound to the target active site, together with the 'vectors' along which the fragment can be derivatised and the available compounds that could be purchased.



The fragalysis application is publicly available [8] and can be used to analyse your XChem fragment screening data.

2D visualisation - Optibrium StarDrop

For displaying results in a 2D context, enabling optimisation of multiple drug design properties, a link has been created to Optibrium's StarDrop™ software [9]. StarDrop's Card View™ display [10] can show groups of compounds of same transformation type and associated properties. The link was implemented in StarDrop's Python scripting interface and utilises a REST API for fetching compounds related to the query structure.



Further Information

For additional information about using the Fragment Network data or tooling contact Tim Dudgeon <tdudgeon@informaticsmatters.com>.
For information about fragment screening at Diamond contact Frank von Delft <frank.von-delft@diamond.ac.uk>.
For information on StarDrop, please contact info@optibrium.com.

References

1. Hall *et al.* J. Med. Chem. 2017, 60, 14, 6440-6450
2. http://www.daylight.com/products/toolkit.html
3. http://rdkit.org/
4. https://github.com/xchem/fragalysis
5. https://neo4j.com/
6. https://neo4j.com/developer/cypher/
7. http://nglviewer.org/
8. https://fragalysis.diamond.ac.uk/
9. https://www.optibrium.com/stardrop/
10. Segall *et al.* Drug Discov. Today 2015, 20, 9, 1093-1103