Rapid and simple Blaze database population and searching using KNIME and Forge

Tim Cheeseright, Giovanna Tedesco[†]

[†] Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, UK

Abstract

Blaze[1] is Cresset's ligand-based virtual screening platform. It uses the shape and electrostatic character of known ligands to rapidly search large chemical collections for molecules with similar properties. In this case study, a Blaze database of approximately 200,000 compounds from ChEMBL[2] was prepared in a seamless manner using a KNIME[3] workflow and standard Blaze database creation routines. The new collection, named 'Chembl20_filtered', is available from the Blaze Demo Server[4]. Blaze searches were launched within Forge[5] and by means of a KNIME workflow to test the ease of use of both workflows. The output of the searches was finally downloaded into Forge and visually inspected.

Background

Blaze, Cresset's ligand-based virtual screening platform, uses the shape and electrostatic character of known ligands (as encoded by Cresset's field technology[6]) to rapidly search large chemical collections for molecules with similar properties. It is excellent for finding novel leadlike hits from known actives, replacing peptides with non-peptides or steroids with non-steroids.

Using Blaze you can increase the diversity of your project's lead compounds and jump into new areas of chemical space giving substantial improvements in the properties of your hits. Cresset have run hundreds of projects through Blaze with an excellent track record: hit rates as high as 30% are reported by our customers.

Blaze

Blaze is a full virtual screening system containing the infrastructure to manage compound collections and the associated conformation populations. It automatically records additions and removals from any collection and handles duplication across collections. New compounds are automatically submitted to a queuing system (typically SGE or Platform LSF) for conformer generation on a Linux cluster.

Database searching is configured through a single webpage, REST call or on the command line. Compounds are automatically triaged through a cascade of increasingly accurate search methods. Blaze automatically manages database searches with differing priorities, submitting them to a queuing system of either a GPU or CPU cluster).

Lastly, Blaze contains a full user and project based permissions system to control the visibility of individual and groups of search results.

Blaze V10.2

This most recent version of Blaze includes:

- A new search algorithm that enables full 3D assessment of molecules at four times the previous speed, enabling the processing of databases of over 10 million compounds.
- A new RESTful web service providing easy integration with Forge, KNIME and Pipeline Pilot[7] and custom software solutions.
- Simplified security features that are easier to unify with corporate authentication servers, in response to customer requests. This makes user management significantly simpler for large installations.
- A free demo server, enabling you to test the performance and functionality of Blaze on a small collection of commercially available compounds.

In this case study a Blaze database of approximately 200,000 compounds from ChEMBL (database of bioactive data for drug discovery) was rapidly prepared and uploaded (added) to the Blaze demo server using the new REST API interface.

Method

Filtering

The full ChEMBL 20 data set (containing approximately 1.5 million compounds) was downloaded as an SDF file.

The set was filtered using a KNIME workflow (Figure 1) applying the following physico-chemical cut-offs to select potential leadlike structures to be used as starting points for medicinal chemistry optimization:

- MW 200-400
- TPSA 40-80
- RotBonds 0-5
- Aryl rings 0-3

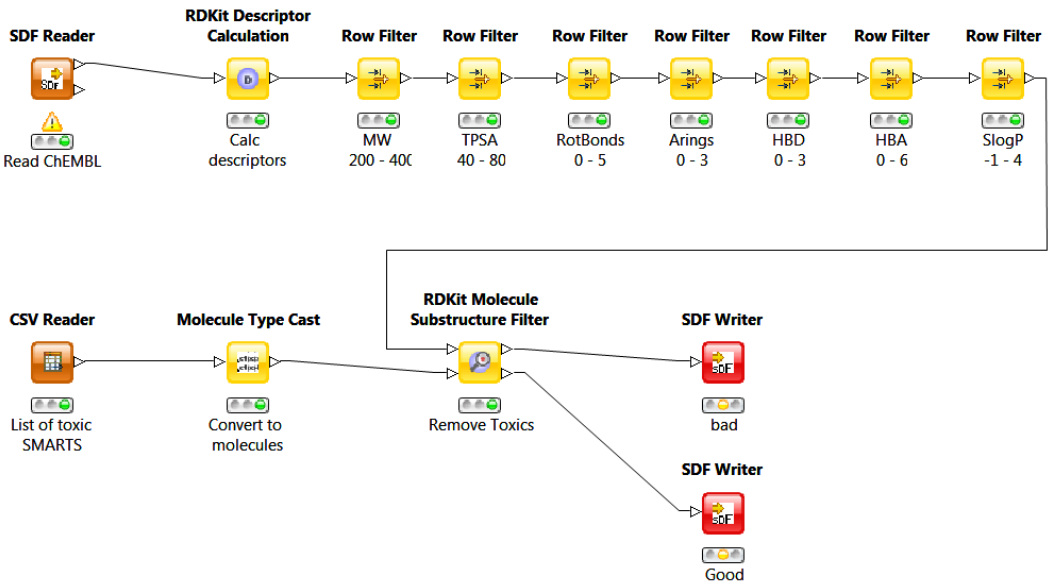


Figure 1. KNIME workflow used to filter the original ChEMBL data set (1.5M compounds).

- HBD 0-3
- HBA 0-6
- SlogP -1-4

The data set was further cleaned with the removal of compounds carrying reactive functional groups (e.g. alkyl halides), potentially toxic groups (e.g. azides) or other unwanted chemical moieties (e.g. heavy metals). After filtering, approximately 208,000 compounds remained for uploading to Blaze.

Upload to Blaze

The upload of the new collection could be achieved using the command line or the web interface. However, as all the compounds exist within KNIME we chose to directly upload to the Blaze free demo server using the Blaze REST API (Figure 2).

The creation of the Blaze Chembl20_filtered collection took a few hours on 150 cores using Cresset's internal Linux cluster.

Using Blaze from Forge/Torch

The introduction of the REST interface has enabled Blaze searching directly from many platforms and scripts including Cresset's desktop applications Forge and Torch. To work with Blaze the applications require the address of the Blaze server and your username and password in the relevant preference setting (Edit menu -> Preferences -> Blaze panel, Figure 3).

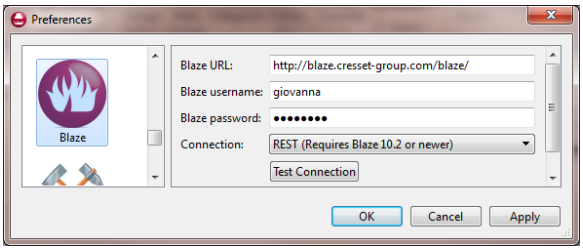


Figure 3. Set up of Forge/Torch connection to Blaze.

The interface enables sending the current molecule, including any field constraints and the current protein excluded volume, to Blaze, configuration of the search options and

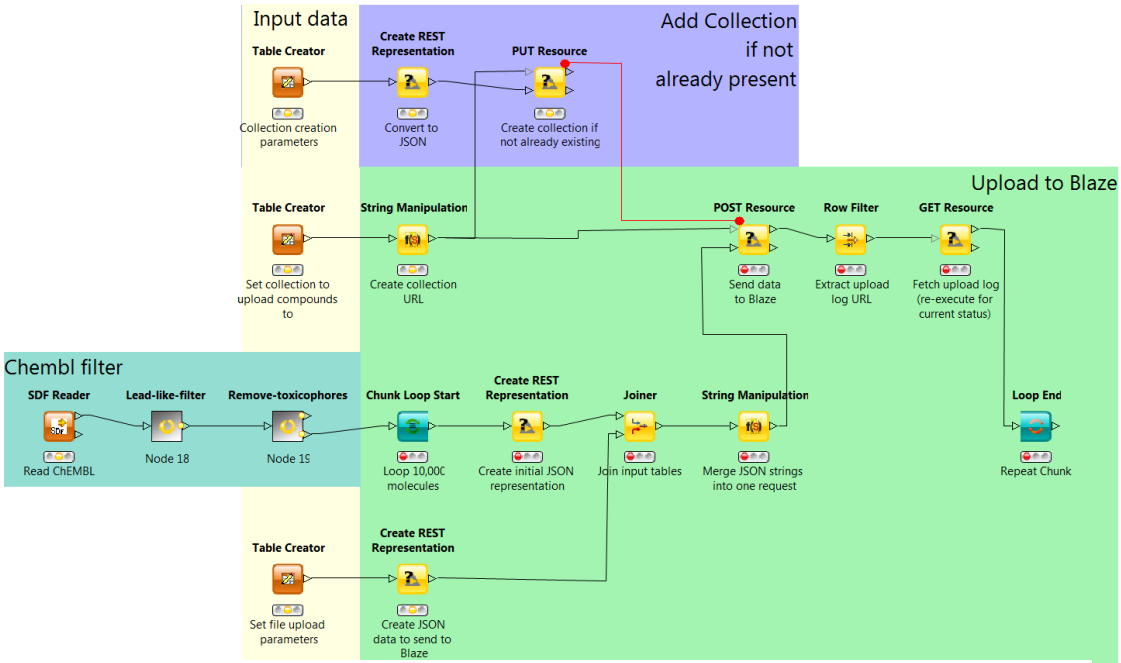


Figure 2. Blaze compound upload protocol.

download of results directly into the application.

To test the new ChEMBL collection and further demonstrate the usefulness of the Blaze REST interface a search was performed using Nevirapine[8], one of the first round of HIV NNRTI inhibitors. The search was submitted using Cresset's Forge and also using a KNIME protocol.

Searching Blaze from Forge

The crystal structure of the Y181C mutant HIV-1 reverse transcriptase in complex with the inhibitor Nevirapine (PDB code 1jlb) was downloaded in Forge (an identical procedure is applied when working with Torch). The workflow is summarized in Figure 4.

Nevirapine was selected as the reference structure and imported into Forge together with the HIV-1 reverse transcriptase protein. Cresset's rules were used to define the protonation state of Nevirapine and the protein. After visual inspection the reference structure was minimized to improve the bond angles.

To initiate the Blaze search, the reference molecule was selected in the main 'Molecules' table then 'Sent to Blaze' using the right click menu. The resulting Blaze search configuration menu was used to name the search '1jlb', select the 'Chembl20_filtered' collection and accept the default search parameters (Figure 4).

Once complete, the search results were imported into Forge (Torch would work identically) for visual inspection and further analysis.

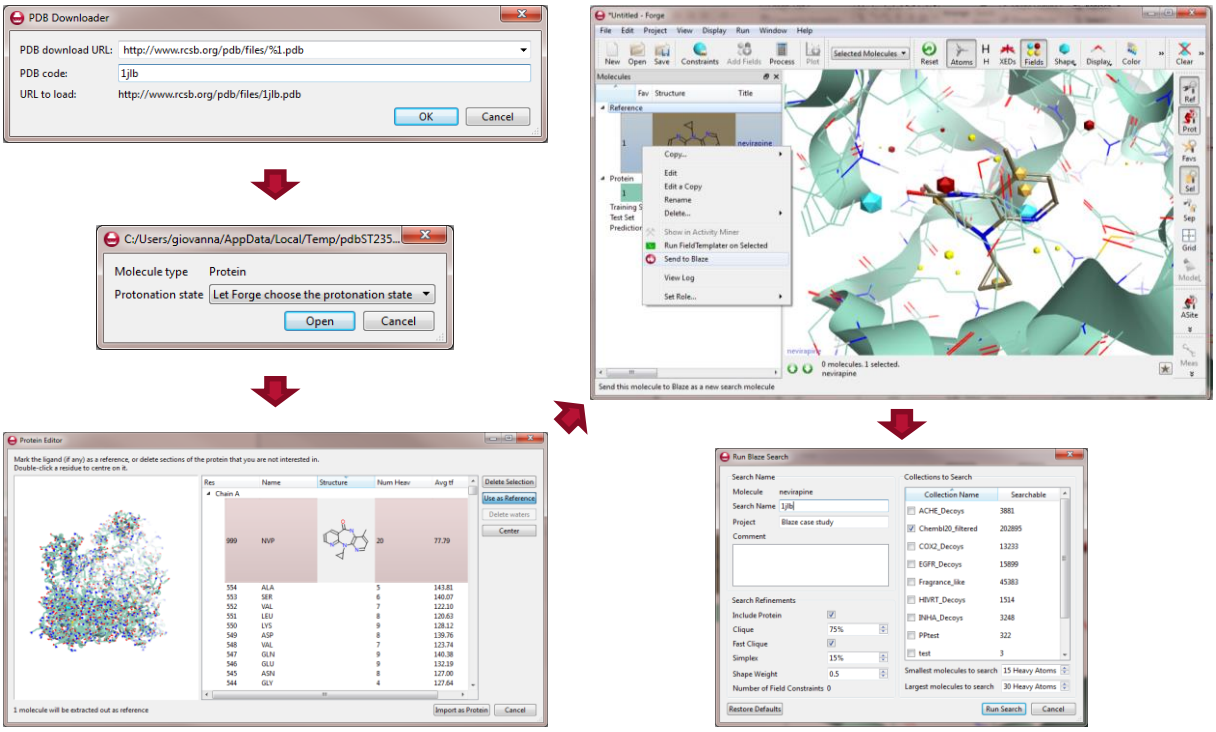


Figure 4. PDB download, selection of reference structure and start of Blaze search in Forge.

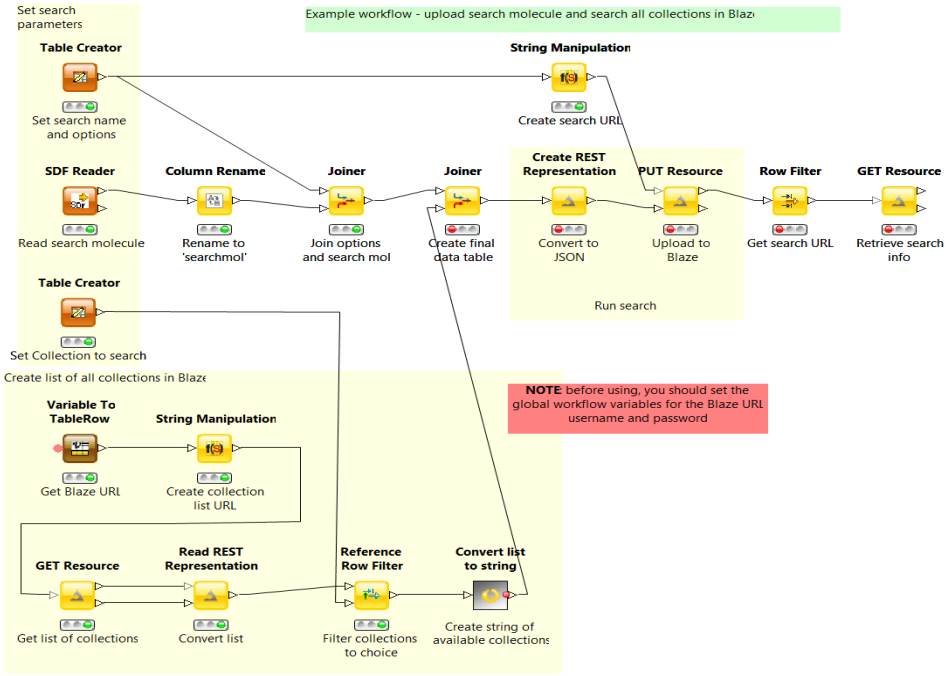


Figure 5. Blaze search protocol.

Blaze Searching from KNIME

A KNIME Blaze search workflow (see Figure 5) was also tested for user friendliness.

The protocol requires the manual setting of a small number of workflow variables (Blaze URL, username and password) and the configuration of 3 input nodes to:

- define the name and conditions of the search (Table creator node),
- load the reference structure as an SDF (using SDF reader node),
- define the name of the Blaze collection to search (Chembl20_filtered, Table creator node).

Download of results to Forge/Torch

The results of the Blaze search on Chembl20_filtered using Nevirapine as the query were downloaded into Forge (Figure 6).

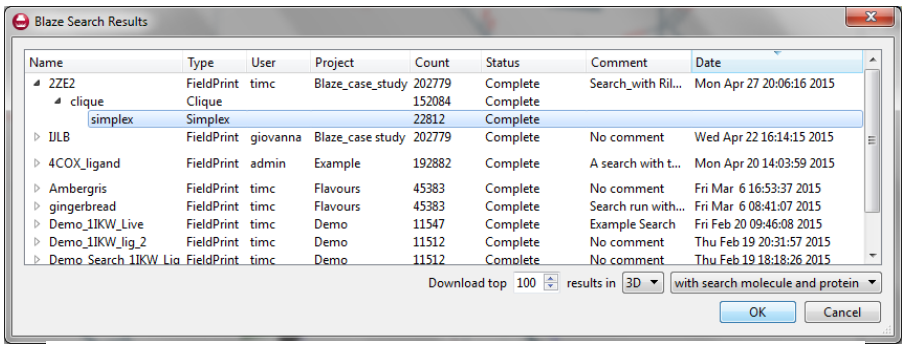
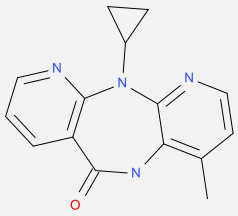
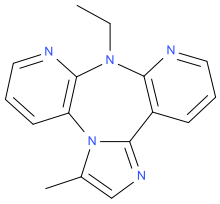
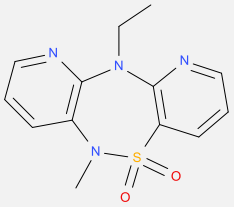
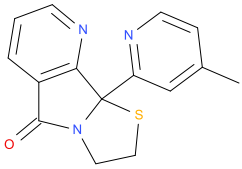


Figure 6. Download of Blaze results into Forge.

While a thorough evaluation of the results of the Blaze search is beyond the scope of this case study, a qualitative analysis of the 200 top scoring results shows that Blaze was able to identify some chemically diverse potential hit compounds. As expected a large fraction of the top scoring compounds belong to the same (widely explored) chemical class of Nevirapine: however a few top scoring molecules (see examples in Table 1, Figure 7) are structurally different and are reported in ChEMBL to have been tested for HIV-1 reverse transcriptase activity.

Table 1. Interesting hits retrieved by the Blaze search on Nevirapine.

ChEMBL ID	Structure
Nevirapine	
45912	
430920	
314103	

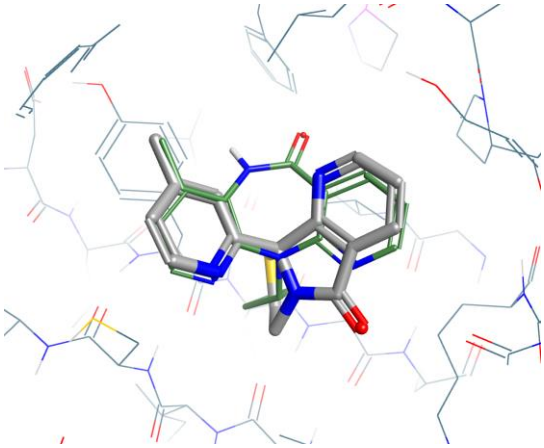


Figure 7. CHEMBL314103 overlaid (grey) with Nevirapine (green).

Conclusion

A Blaze database of approximately 200,000 compounds from ChEMBL was prepared in a seamless manner using a KNIME workflow. Using the Blaze REST interface the dataset could be uploaded to Blaze from within KNIME and was available for searching within a few hours.

To test ease of use of the search workflows available in Forge (Torch) and KNIME, the same search was run on each platform. While both protocols are relatively straightforward the Forge guided interface is definitely simpler to set-up for the end user. The KNIME workflow offers a higher flexibility, however, and allows the integration of Blaze searches into more customized protocols with complex post-processing of results. Using the Torch or Forge viewers within KNIME enables viewing of the 3D alignment of the returned compounds within that platform.

The new Chembl20_filtered collection is available for searching by all users of the Blaze demo server – register for free access by visiting http://blaze.cresset-group.com/blaze/

References and Links

1. http://www.cresset-group.com/products/blaze/
2. https://www.ebi.ac.uk/chembl/
3. KNIME: https://www.knime.org/
4. Blaze free demo server: Register for your username and password at the Blaze demo signup page http://blaze.cresset-group.com/blaze/
5. http://www.cresset-group.com/products/forge/
6. http://www.cresset-group.com/science/field-technology/
7. Pipeline Pilot: http://accelrys.com/products/pipeline-pilot/
8. US5366972 (A) - 5,11-dihydro-6H-dipyrido(3,2-B:2',3'-E)(1,4)diazepines and their use in the prevention or treatment of HIV infection