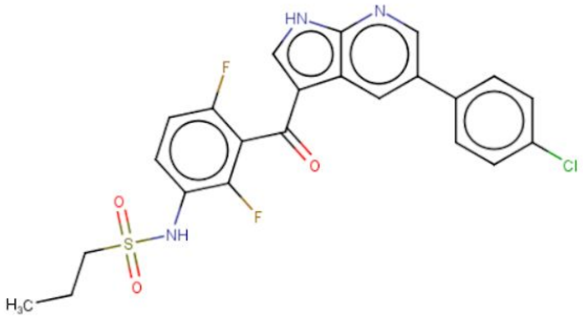
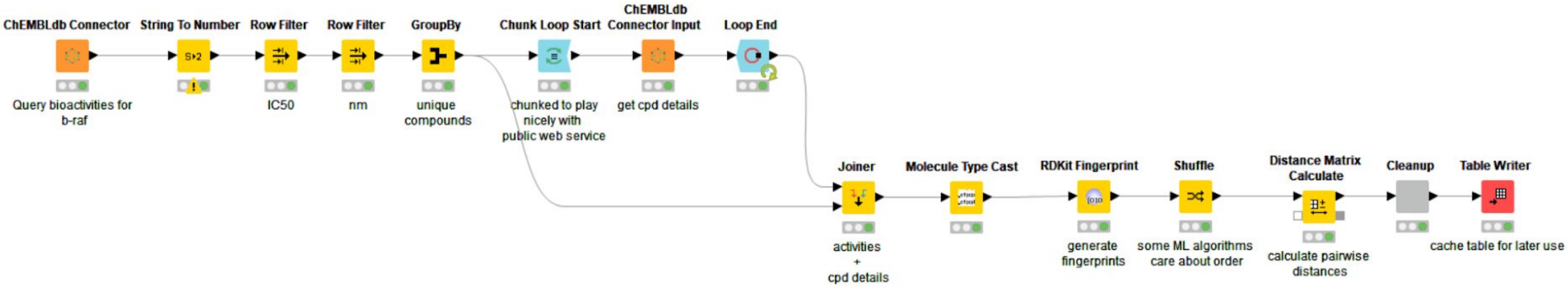
Cheminformatics example workflows

Created by Aaron Hart
(aaron.hart@gmail.com)

Premise:
Explore the chemical information around Vemurafenib

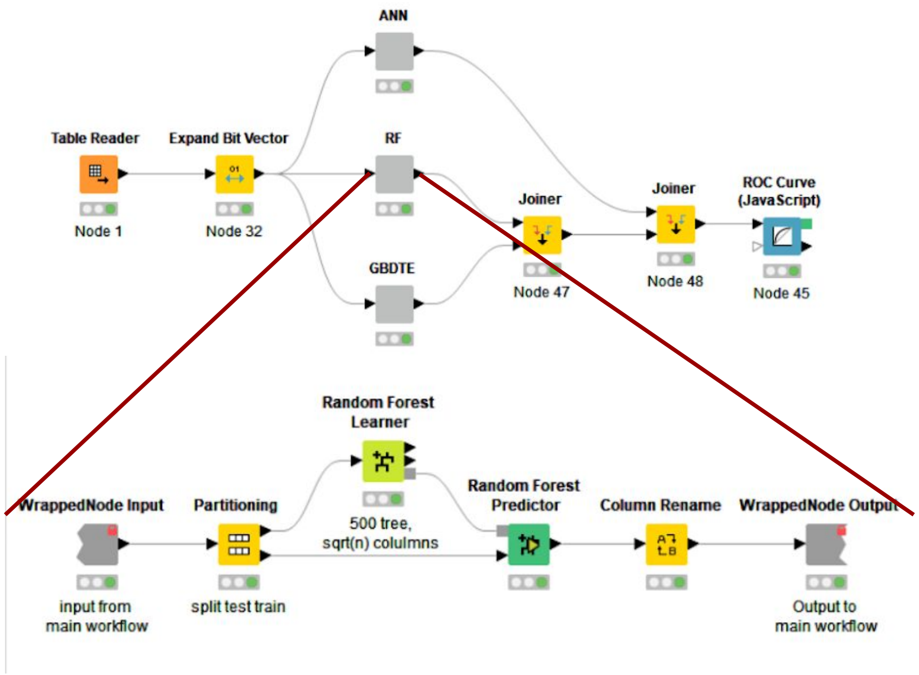


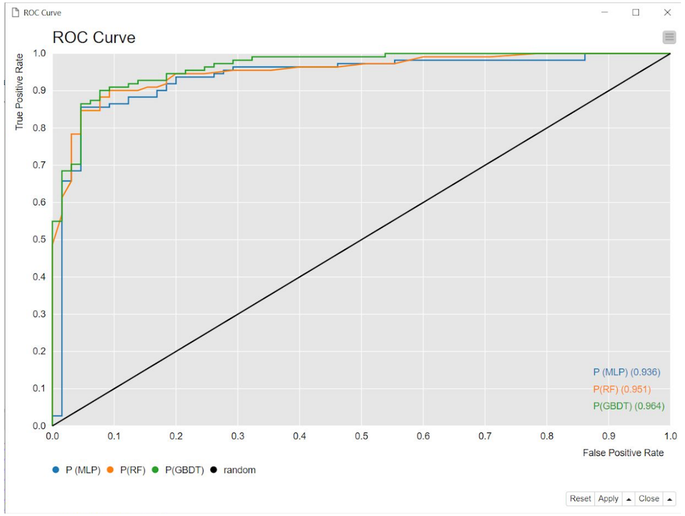
Bioactivity retrieval from ChEMBL



1. Fetch relevant data for the target from ChEMBL
2. Prepare molecular structures, fingerprints and distance matrix (Tanimoto similarity)
3. Clean up column names, tag activity and write to cache

Comparing Activity Models



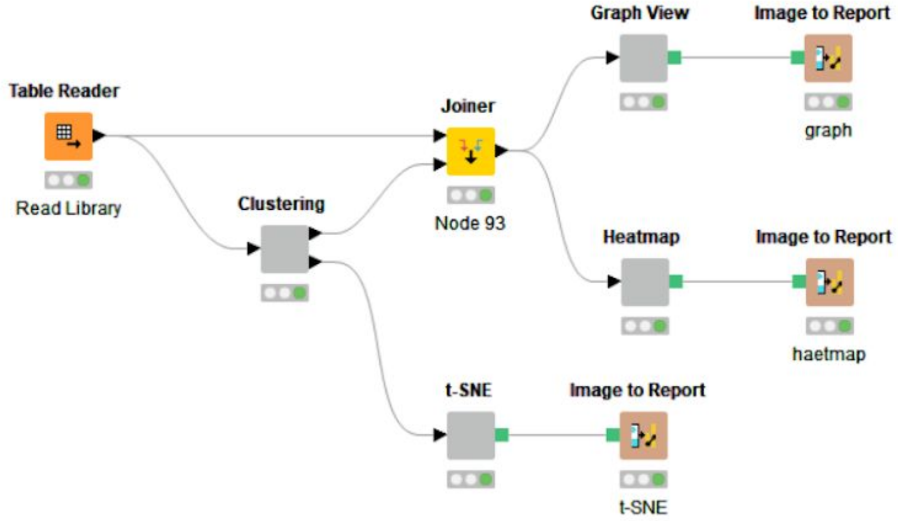


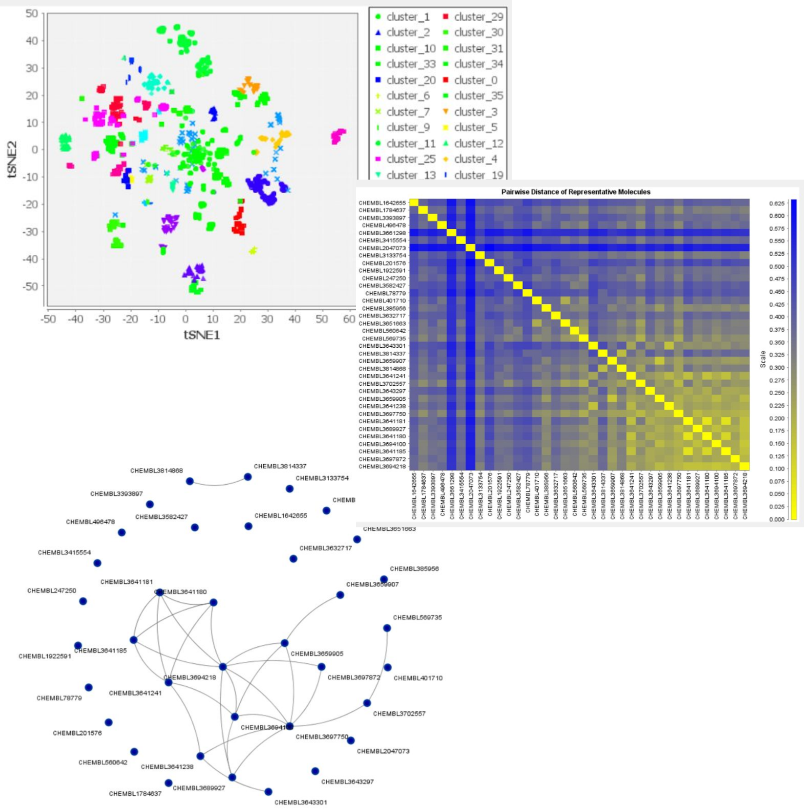
Compare activity models based on the fingerprint:

1. A deep MLP neural network
2. A classical random forest
3. A gradient boosted decision tree ensemble

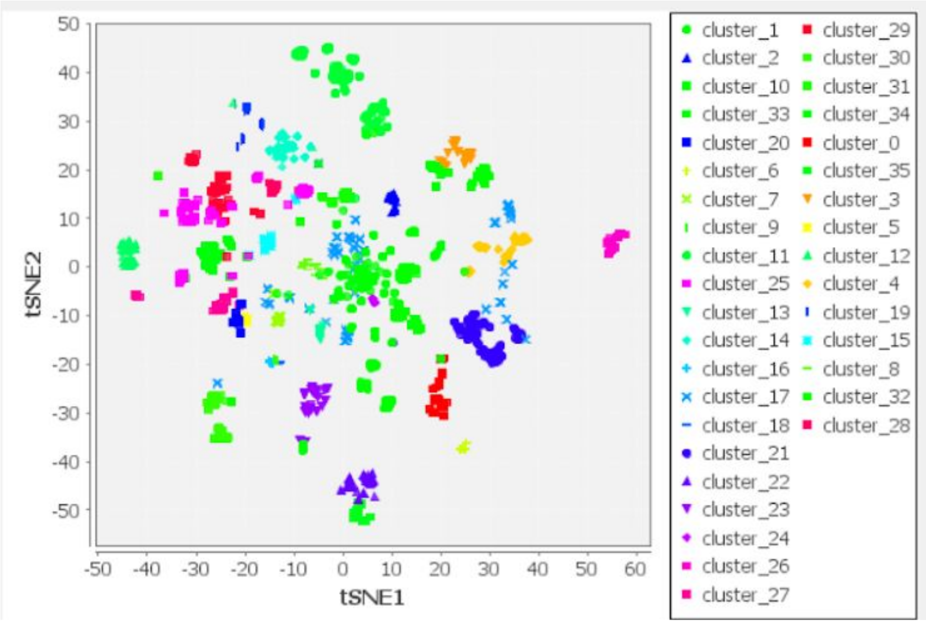
Note: no work was put into optimization of any of the models but that is also something I can do.

Library visualization I: Representative compounds





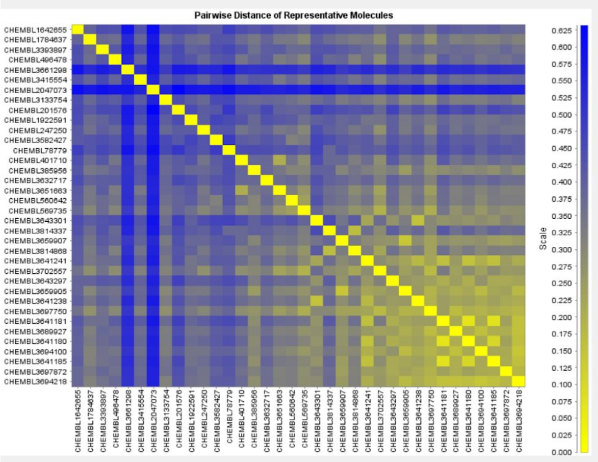
Library visualization



1. PCA of the fingerprints
   (should probably be truncated SVD)
2. k-Means on the first ~200 components
3. Visualization via t-SNE

Note: some clusters could be merged for a more compact representation of the library

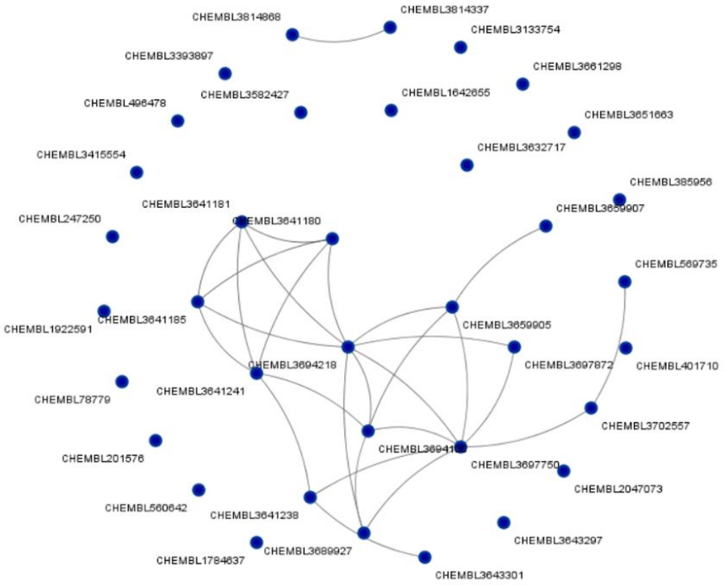
Library visualization II: Visualizing cluster similarity



1. Choose cluster representatives
2. Sort by principal component
3. Generate heatmap based on pairwise distances

Note: A more even distribution of pairwise distances would indicate a more representative sampling of the chemical space captured within this library.

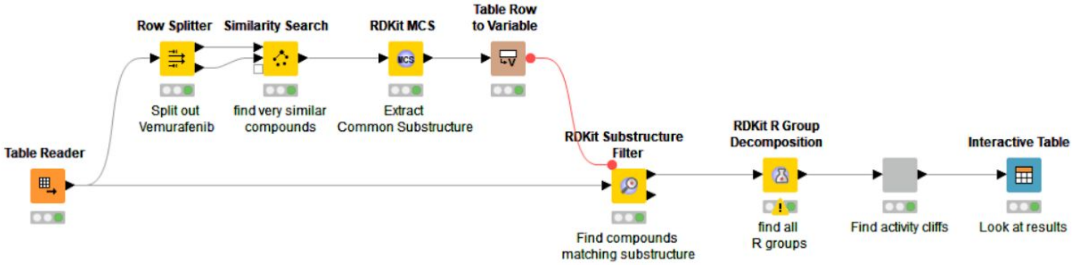
Library visualization III: Distance based network view

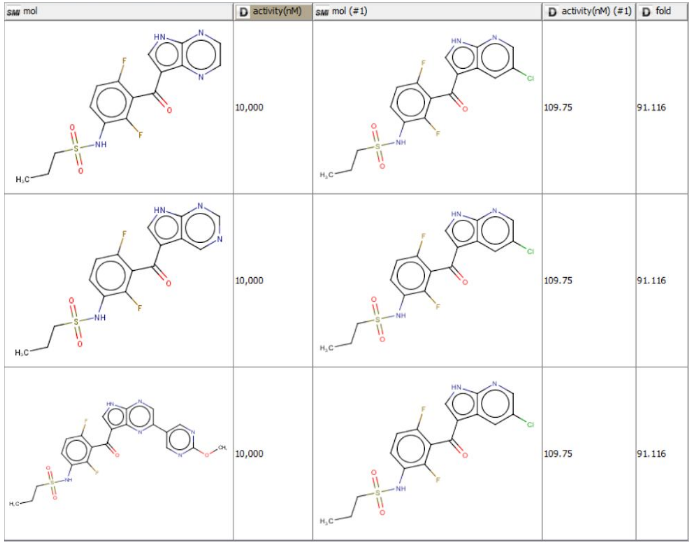


1. Calculate distances for cluster reps.
2. Filter edges by distance threshold
3. Visualization using network view

Note: no real purpose here, just showing an example of a graph visualization.

Matched molecular pairs





SMI mol	D activity(nM)	SMI mol (#1)	D activity(nM) (#1)	D fold
	10,000		109.75	91.116
	10,000		109.75	91.116
	10,000		109.75	91.116

1. Find all compounds with a given substructure
2. Pull out the R groups and find pairs with very small differences
3. Look for activity cliffs where a single change in R causes big change in activity