UPPSALA
UNIVERSITET

Machine learning in drug discovery

Jonathan Alvarsson[1], Samuel Lampa[1], Claes Andersson[2],
Jarl E.S. Wikberg[1], and Ola Spjuth[1,3]

1 Department of Pharmaceutical Biosciences, Uppsala University
2 Department of Medical Sciences, Uppsala University
3 Science for Life Laboratory, Uppsala University

Contact: jonathan.alvarsson@farmbio.uu.se

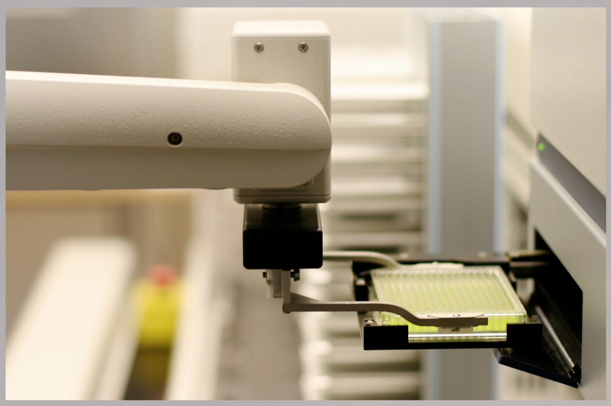


Figure: A robot for testing drugs on large scale

A large chemical space

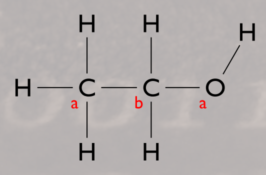
In drug discovery large numbers of drugs are regularly synthesized and tested by robots. The possible chemical space is enormous and to synthesise and test everything is impossible. However computer-based drug screening can be used to pick substances to study further.

Molecular properties are predicted using machine learning methods and decisions are made based on these properties.

Molecular signatures

We often use a molecular descriptor called the molecular signature descriptor to describe the molecules in a way so that the computers can work with them.

A molecular signature consists of the atom signatures for the atom in that molecule. The atom signatures canonically describes the connectivity of the atoms in a tree-like fashion. A height parameter decides the distance, i.e., number of atoms from the central atom to be included in the atom signature.



	Molecular signature		
Height	$_{a}C$	$_{b}C$	$_{a}O$
0	[C]	[C]	[O]
1	[C]([C])	[C]([C][O])	[O]([C])
2	[C]([C]([O]))	[C]([C][O])	[O]([C]([C]))

Figure: Example of molecular signatures for ethanol. Notice that hydrogen atoms are ignored. Ethanol is so small that heights larger than 2 makes no difference.

Cloud computing

We have evaluated the use of cloud computing while doing machine learning for drug discovery. Scaling up computation using cloud computing gives lower up-front costs and resources can scale with computation need. Also, the costs of modeling is easily quantified.

Moghadam *et al.* "Scaling predictive modeling in drug development with cloud computing." *Journal of chemical information and modeling* 55, no. 1 (2015): 19-25.

The signature fingerprint

An effective representation of molecular descriptors are as so called fingerprints. Classic fingerprints are bit vectors but also lists of integers are common. We created and evaluated open source fingerprints based on the molecular signature descriptor using a nearest neighbour approach.

We tested out fingerprints for predicting binding affinities using a simple nearest neighbour approach and found that our open source signature fingerprints were capable of performing on par with commercial industry standard fingerprints.

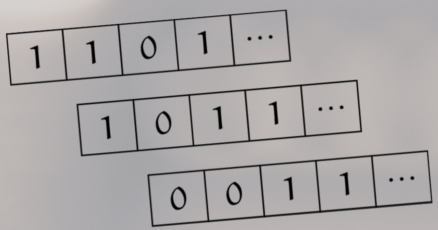


Figure: Classic fingerprints are bit vectors

The resulting signature fingerprints were made available as open source software through the Chemistry Development Kit (CDK) project.

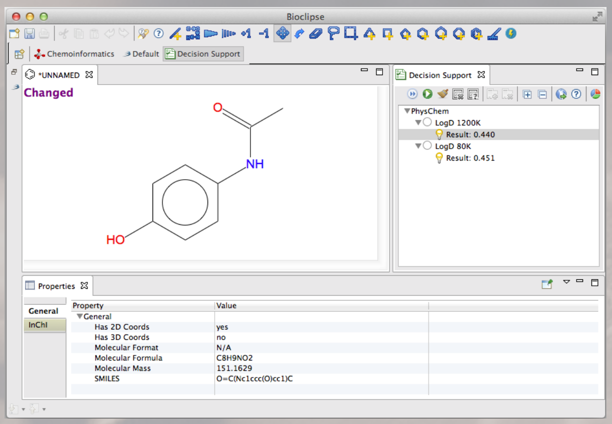
Alvarsson et al. "Ligand-based target prediction with signature fingerprints." Journal of chemical information and modeling 54, no. 10 (2014): 2647-2653.

Molecular signatures and support vector machines

After the nearest neighbour approach we moved on to Support Vector Machines (SVM) and in a benchmarking study found good parameter values to use when using SVM and the Radial Basis Function (RBF) kernel on data sets of a couple of thousands molecules.

Currently we are evaluating the use of linear SVM through the LIBLINEAR software package for building SVM models on large molecular datasets in the size of approximately a million compounds. At these sizes the parameter optimization involved with the RBF kernel becomes infeasible.

However, initial results show that we get similarly good results using LIBLINEAR as we do with an RBF kernel based SVM.



PhysChem
LogD 1200K
Result: 0.440
LogD 80K
Result: 0.451

Property	Value
General	
Has 2D Coords	yes
Has 3D Coords	no
Molecular Format	N/A
Molecular Formula	C8H9NO2
Molecular Mass	151.1629
SMILES	O=C(Nc1ccc(O)cc1)C

Figure: Our final models can be accessed from the Bioclipse workbench and used to get near real time feedback on molecules as a user draws them in a molecular editor.

Alvarsson et al. "Benchmarking study of parameter variation when using signature fingerprints together with support vector machines." Journal of chemical information and modeling 54, no. 11 (2014): 3211-3217.

essence
THE e-SCIENCE COLLABORATION

SciLifeLab