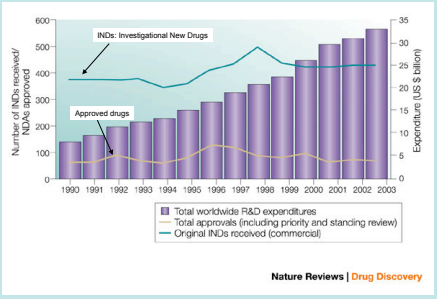
HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Ideen zünden!

Development of novel ADME/T models and estimation of the accuracy of predictions for drug discovery

Igor V. Tetko, Institute of Bioinformatics & Systems Biology, Neuherberg

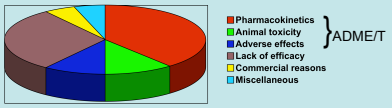
The productivity gap in pharmaceutical industry



Despite an increasing investment in drug discovery over last 15 years, there has been no corresponding increase in the number of approved drugs. One of the bottlenecks causing the low success rate is the failure of drugs in the pre-clinical and clinical studies because of their unfavorable ADME/T (Absorption, Distribution, Metabolism, Excretion and Toxicology) properties. The development of methods that can predict as early as possible a failure of compounds at the late stages of drug testing are very important. The rapid progress of combinatorial chemistry approaches and the possibility to virtually screen millions of compounds enormously raised massive interest in the computational prediction of these properties as the only way to analyze such enormous collections of molecules.

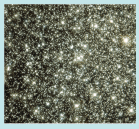
We are developing novel software suites for ADME/T properties predictions. These software tools provide self-learning features to develop local models for "in-house" data and estimate the prediction accuracy. On this poster we highlight some of our recent studies based on applications of local models and we present novel tools to assess the prediction accuracy of different methods.

Reasons for failure in drug development

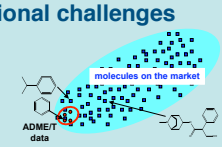


➢60% of drug failures are due to absorption, distribution, metabolism, excretion and toxicology (ADME/T) problems. Selection of compounds with most favorable properties can prevent failures on the later stages of drug development.

Computational challenges



The amount of molecules potentially available for chemist is comparable to the number of stars in a galaxy.

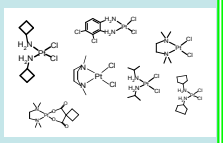


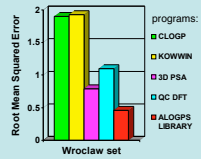
Typical ADME/T datasets have 10^2-10^5 molecules, while prediction of 10^7-10^{24} could be required. **Thus predictions *cannot be* performed with the same accuracy for all possible molecules.**

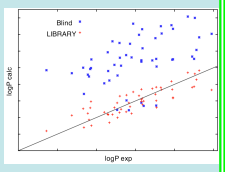
We are developing methods which calculate accurate ADME/T models and estimate the quality of predictions for target scaffolds of molecules

Prediction of lipophilicity (logP) of Pt-complexes

Platinum containing compounds are promising antitumor agents. Their activity is to a large extent determined by their lipophilicity (logP) but existing programs cannot satisfactory predict this property for platinum complexes.



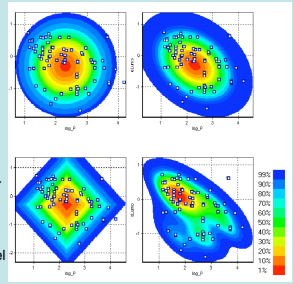




The use of our program in local mode achieved a reasonable accuracy of predictions for a set of 12 new complexes compared to experimentally determined values.[1]

Applicability Domain (AD)

Compounds that are more „near" to the training set (model) are expected to have higher accuracy of predictions, i.e., they are within the model AD. The regions of the chemical space with more confident predictions (small distances to model) are shown in red color. There are many ways to define a distance to a given model and thus its AD (four distances to models are shown). Therefore there is a need to quantify different ADs with respect to their ability to differentiate molecules with reliable and non-reliable predictions.



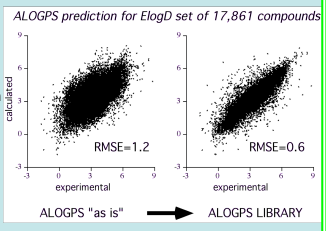
Toxicity against *T. pyriformis*

The growth inhibition of the ciliated protozoan *T. pyriformis* is a commonly accepted toxicity screening tool that has been under development for many years. We used these data to benchmark fourteen different applicability domains developed by different laboratories. These data are of high quality and thus represent valuable resource for comparison of data analysis methods.

Twelve QSAR (quantitative structure-activity relationship) models to models of aqueous toxicity against *T. pyriformis* were calculated with different machine-learning methods and various types of descriptors.[3] Following this analysis we compared fourteen distances to models with respect to their ability to predict errors of molecules.[4]

Prediction of in-house data (Pfizer)

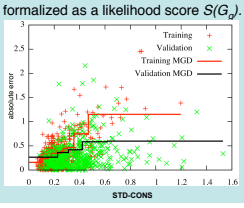
Models developed using data available in literature may not work accurately enough for prediction of „in-house" data of pharma companies. This can be due to chemical diversity of molecules in both sets.

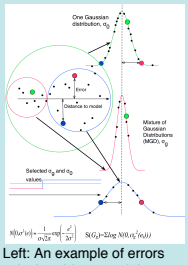


The left panel shows a poor performance of the ALOGPS program to predict private data. The use of our self-learning methodology (local models) dramatically decreased the errors (right panel). The calculations were performed in just a few minutes.[2] The re-development of models using all data calculated similar results.[1,2]

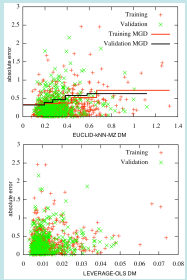
Mixture of Gaussian Distributions

Right: A probability of an error to be generated with a Gaussian distribution is proportional to the height of the corresponding curve on the graph. The use of a mixture of Gaussian distributions provides better description of errors for three colored molecules. This intuitive observation can be formalized as a likelihood score $S(G_\sigma)$.



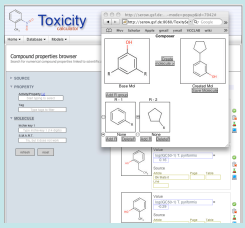


Left: An example of errors calculated for a model as function of standard deviation (STD-CONS) distance to model. There is a good correlation of observed errors with this distance to model.



Up: Examples of two other distances to models. The lower one does not correlate with the observed errors.

WWW interface



The developing user interface allows to introduce molecules and their biological properties, or/and predict new molecules using the developed models. The accuracy of predictions is estimated using the standard deviation of Associative Neural Networks models.

Conclusion and References

The use of our "self-learning" methodology develops very accurate focused local models both for small and large datasets of molecules. Currently we are benchmarking our program with other commercial programs available on the market using commercial data from pharma companies.

1) Tetko et al. *J. Inorg Biochem.*, 2008, in press.
2) Tetko et al. *J. Med. Chem.*, 2004, 47(23), 5601-4

Conclusion and References

A statistical test to compare the different distances to models was developed. We showed that a standard deviation across ensemble of models provided the best correlation between the predicted and estimated errors. We also addressed the problem of incorrect validation of models when using variable selection (overfitting) and analyzed the quality of prediction in a large dataset of about 210,000 molecules.

3) Zhu et al, Combinational QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*, *J. Chem. Info. Model.*, 2008, in press.
4) Tetko et al, A critical assessment of QSAR Models to predict environmental toxicity against *Tetrahymena pyriformis*: An issue of applicability domain and overfitting by variable selection, *in preparation*.

Supported by BMBF GO-Bio Grant 0313883

HELMHOLTZ GEMEINSCHAFT