Draft accepted for ChemMedChem

For the finished article, see dx.doi.org/10.1002/cmdc.200700041

Predicting Lipophilicity of Drug Discovery Molecules using Gaussian Process Models

Timon S. Schroeter,[a,b] Anton Schwaighofer, [a] Sebastian Mika, [c] Antonius Ter Laak, [d] Detlev Suelzle, [d] Ursula Ganzer, [d] Nikolaus Heinrich, [d] Klaus-Robert Müller*[a,b]

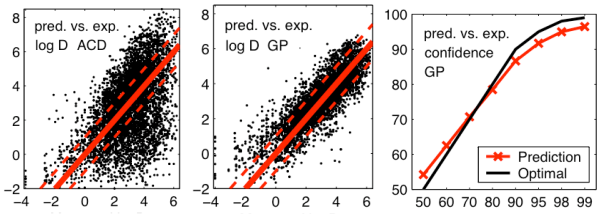

Figure 1. Evaluation on Bayer Schering Pharma in-house data in blind test: Scatterplots for ACDLabs v9 (left) and the GP log D_7 model (center), predicted (black) vs. ideal (red) error bar confidence (right)

Many drug failures are due to an unfavorable ADMET profile (Absorption, Distribution, Metabolism, Excretion & Toxicity). Lipophilicity is intimately connected with ADMET and in today's drug discovery process, the octanol water partition coefficient log P and it's pH dependant counterpart log D have to be taken into account early on in lead discovery. Commercial tools available for 'in silico' prediction of ADMET or lipophilicity parameters usually have been trained on relatively small and mostly neutral molecules, therefore their accuracy on industrial in-house data leaves room for considerable improvement (see Bruneau et al. and references therein).[1] Using modern kernel-based machine learning algorithms – so called Gaussian Processes[2] (GP)– this study constructs different log P and log D_7 models that exhibit excellent predictions which compare favorably to state-of-the-art tools on both benchmark and in-house data sets.

GP models are Bayesian non-linear regression models and it is the Bayesian framework that allows to provide theoretically well founded criteria to automatically choose the "right amount of nonlinearity" for modeling, thereby avoiding to depend on the users experience for choices like the architecture of neural networks.[3] For chemistry applications one of the most interesting virtues of GPs certainly is that they can provide insihts into the relevance of individual descriptors, e.g. like in this work to lipophilicity. During model fitting the GP algorithms automatically assigns weights to each descriptor that enters the model as relevant input. Moreover and equally important, GPs automatically supply the user with an error bar when predicting the outcome of an experiment. In practice, the latter should be valued especially high since the machine will quantify its uncertainty, which allows to reduce the error rate by discarding predictions with large error bars (for a detailed explanation of GPs and the algorithmic approach used see Schwaighofer et al.).[2a]

The machine learning approach to computational chemistry requires a training set from which the underlying statistical properties are inferred and a prediction model is selected.[2,4] Typically, cross-validation or resampling methods help to tune the hyperparameters of this modeling. Once the model has been fixed, an out-of-sample prediction is performed on held-out data (test set) that was not used to tune the model. Ideally the prediction quality should be measured in a blind test, where the predicting team (1) has no knowledge of the labels of the blind test set, (2) has to apply the statistical model to this set and (3) has to provide its predictions to the evaluating team, which only has knowledge of the labels of the blind test data and can therefore assess the prediction error in a more objective manner. The latter setup, as opposed to usual benchmark evaluations, allows a nearly unbiased evaluation where 'cheating', i.e. re-tuning the model on held-out data, becomes unfeasible. Note however that the blind test data needs to surpass certain minimal size criteria otherwise the evaluation results of the blind test will not be statistically significant.

Earlier studies have already shown the applicability of Gaussian Process models to problems in computational chemistry, however mainly on comparatively small data sets and typically without blind test.[5] Note that until recent improvements in GP algorithms, it was unfeasible to learn on larger data sets and it is due to elegant approximations and advances in sampling techniques that large systems can now be analysed.[6] While Burden predicted activity of compounds with respect to benzodiazepine and muscarinic receptors and their toxicity,[5a] the largest data set used contained only 277 compounds (no blind validation). Enot et al. used GP models to predict log P on a set of 1,2-dithiole-3-one molecules; only 44 compounds were

[a] *Timon S. Schroeter, Anton Schwaighofer, Klaus-Robert Müller*
Intelligent Data Analysis Group
Fraunhofer FIRST
Kekulestraße 7, 12489 Berlin, Germany
Fax: (+49)30-6392-1805
E-mail: pcadmet@first.fraunhofer.de
[b] *Klaus-Robert Müller, Timon S. Schroeter*
Computer Science
Technical University of Berlin
Franklinstraße 28/29, 10587 Berlin, Germany
[c] *Sebastian Mika*
idalab GmbH
Sophienstraße 24, 10178 Berlin, Germany
[d] *Antonius Ter Laak, Detlev Suelzle, Ursula Ganzer, Nikolaus Heinrich*
Research Laboratories
Bayer Schering Pharma
Müllerstraße 178, 13342 Berlin, Germany

Supporting information for this article is available on the WWW under http://www.chemmedchem.org/ or from the author.

employed (no blind validation).[5b] Tino et al. built GP models for log P on a public data set of 6912 compounds. They performed a blind evaluation, however, with a validation set (from Pfizer) that contained only 226 compounds.[5c]

The present study goes beyond this prior work as our model was trained and evaluated on large sets of public and in-house data, furthermore a blind test was performed on a large set of 7013 recent drug discovery molecules at Bayer Schering Pharma that have not been available to the modeling team. The complete list of compounds in the public data set is included in the supporting information to facilitate reproduction of our results by other researchers.

Modeling was performed as follows: For each molecule, the 3D structure of one conformation is predicted using the program Corina.[7] From this 3D structure, 1,664 Dragon descriptors are generated.[8] We inspected the relative weighting of descriptors as computed by the GP model. Among the descriptors with highest weight, the following set with a clear link to lipophilicity was identified automatically: Number of hydroxy groups, carboxylic acid groups, keto groups, nitrogen atoms, oxygen atoms and total polar surface area. This information can be used to select a subset of features for model building. For all three models employed in this study, reducing the number of descriptors resulted only in a slight performance decrease, even when less than 100 features were retained. The quality of the predicted error bars of the GP model, however, was significantly decreased. Therefore, the full set of descriptors was retained.

Based on consensus values of log P / log D_7 measurements and molecular descriptors of a large set of compounds, a Gaussian Process model is fitted to infer the relationship between the descriptors and the log P / log D_7 for two data sets:

Table 1. Public data, N=7926, results for Gaussian Process model, Support Vector regression, Ridge regression and several commercial tools

log P	MAE	RMSE	% ±1
ACDLabs v9	0.43	0.90	89.2
Wskowwin v1.41	0.25	0.90	91.6
AdmetPredictor v1.2.3	0.65	1.32	86.9
QikProp v2.2	0.76	1.23	79.6
this study GP (trained on in-house data) [a]	1.21	1.68	56.4
this study GP (trained on in-house data, pred. err. bar < 0.3, N=179) [a,b]	0.41	0.69	92.2
this study RR	0.59	0.89	84.4
this study SVM	0.40	0.71	91.8
this study GP	0.38	0.66	92.6
this study GP (pred. err. bar < 0.7, N= 7072) [b]	0.33	0.53	96.0
this study GP (pred. err. bar < 0.3, N= 5802) [b]	0.28	0.45	96.8

[a] Predicting public compounds with a GP model trained on in-house data results in low performance. [b] Focusing on confident predictions (small predicted error bars) results in increased performance

The first set of data contains 7926 log P measurements of neutral (between pH 2 and 13) molecules that were extracted from the PhysProp and Beilstein databases (supporting material). Different machine learning methods were validated on this set of data in leave 50 % out cross-validation. Achieved accuracies are given in Table 1, along with results of four commercial tools, evaluated on the same dataset (plots are included in the supporting information). The two best performing commercial tools, Wskowwin and ACDLabs and our own SVM and GP models perform equally well (89 to 92 % correct within one log unit) when applied to the whole set of data. The accuracy of the linear Ridge Regression model being much lower (60 %), we conclude that modeling Log P based on the given data and descriptors requires non-linear regression models like GP, SVM, or neural networks. Note that all four commercial tools have been constructed using some measurements that are also included in the PhysProp and Beilstein databases. Predictions for measurements that have been used to train the model are clearly not 'out-of-sample' predictions and thus in a sense trivial, therefore these results are somewhat biased towards better performance. Our own validation procedure is based on repeatedly leaving out 50% of the data from training and then only evaluating predictions for truly "unseen" compounds. The compounds to leave out were picked at random, so the distribution across different compound classes is similar for test and training data. In drug discovery practice, this idealized statistical assessment does typically not hold: In new projects, new compound classes may be investigated, resulting in less accurate predictions. To get a realistic estimate of the performance on unseen data, a blind evaluation of models using data from new projects is crucial for a real-life out of sample estimate.

Table 2. Bayer Schering Pharma in-house data in blind test, N_{blind}=7013, N_{train}=14556, results for Gaussian Process model, Support Vector regression, Ridge regression and ACDLabs v9

Log D, blind test	MAE	RMSE	% ±1
ACDLabs v9	1.40	1.79	44.2
this study GP (trained on public data) [a]	1.21	1.68	56.4
this study GP (trained on public data, pred. err. bar < 0.3, N=339) [a,b]	0.66	0.86	79.4
this study RR	0.60	0.83	82.2
this study SVM	0.58	0.81	81.6
this study GP	0.60	0.82	81.2
this study GP (pred. err. bar < 0.7, N=5398) [b]	0.51	0.70	86.8
this study GP (pred. err. bar < 0.3, N=2603) [b]	0.40	0.55	91.3

[a] Predicting in-house compounds with a GP model trained on public data results in low performance. [b] Focusing on confident predictions (small predicted error bars) results in increased performance

We independently constructed models based on an in-house set of 14556[9] HPLC log D_7 measurements from Bayer Schering Pharma. The GP model was validated in blind evaluation by our colleagues at Bayer Schering Pharma on a set of 7013 new measurements of drug discovery molecules from the last months. Afterwards, the new data was made available to the modelers and used to validate to remaining models. See Table 2 and figure 1 for results. All three models constructed in this study exhibit reasonable overall performance (81.2 – 82.2 %). The advantage of the GP model becomes obvious when the predicted error bars are used to focus on reliable predictions: 5398 compounds are predicted with error bars smaller than 0.7 and 87 % of these predictions are correct within one log unit. Focusing on the 2603 compounds with error bars below 0.3 results in 91 % of these predictions being correct within one log unit.

Out of the four commercial tools available to us, only ACDLabs v9 can calculate log D_7. It predicts 44.2% of all compounds correct within one log unit. One has to keep in mind that ACDLabs predicts log D based on shake-flask measurements, while the measurements used in the blind-test scenario were done using the HPLC methodology described in

the supporting information. Moreover, in-house compounds are structurally quite different from publicly available data: When applying GP models trained on in-house data to public data or vice versa, only a small subset of all predictions is made with high confidence (i.e. small predicted error bars, see Table 1,2, rows labeled [b]). Nevertheless evaluating *all* predictions results in low performance (see Table 1,2, rows labeled [a]). This is consistent with results of Bruneau[10] and others.

It follows from the definition of the error bar (σ), that 68,7 %, 95 % and 99,8 % of all predictions have to be within σ, 2σ and 3σ intervals of the experimental values, respectively. The quality of predicted error bars can therefore be evaluated by counting how many of the predictions are actually within the respective σ, 2σ etc. intervals of the experimental values. Figure 1 (right) shows that the predicted errors indeed exhibit the correct statistical properties: Results on the blind test data (black line) are close to the ideal run of the curve (red line). In addition, predicted error bars can be used to identify reliable predictions. Focusing on predictions with small predicted error bars results in significantly increased performance (see Table 1,2, rows labeled [b]).

In conclusion, we presented results of modeling lipophilicity using the Gaussian process methodology, Support Vector Machines and linear Ridge Regression. On public data the prediction quality of our models compares favorably with four commercial tools, with the non-linear models performing better than the linear model. On in-house data of Bayer Schering Pharma, all three models perform better than commercial software. If predicted error bars from the GP model are used to focus on compounds inside its domain of applicability, it clearly outperforms all remaining models. This is furthermore underlined by a blind evaluation on a large set of measurements from new drug discovery projects. Finally we would like to stress that machine learning techniques (in particular GP models) are not only capable to contribute good predictions, but they can provide automatized tools to gain insight in what descriptors are most important for the modeling task and even more important for drug discovery practice: GPs quantify the trust in a given prediction in a statistically very well founded manner. Future research will therefore strive for a continuous improvement of modeling for computational chemistry using machine learning methods.

Acknowledgements

The authors gratefully acknowledge partial support from the PASCAL Network of Excellence (EU #506778). We thank Vincent Schütz and Carsten Jahn for maintaining the PCADMET database, and two anonymous reviewers for detailed suggestions that helped to improve the paper.

Keywords: lipophilicity· drug design · machine learning · Gaussian process · domain of applicability

[1] P. Bruneau, N. R. McElroy, *J. Chem. Inf. Model* **2006**, *46*,1379–1387

[2] (a) A. Schwaighofer, T. S. Schroeter, S. Mika, J. Laub, A. Ter Laak, Detlev Suelzle; Ursula Ganzer; Nikolaus Heinrich; Klaus-Robert Müller. *J. Chem. Inf. Model*, **2007**; (b) K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, *IEEE Transactions on Neural Networks* **2001**, *12*, 81–201; (c) C. E. Rasmussen, C. K. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge MA, **2005**

[3] (a) G. Orr, K.-R. Müller, *Neural Networks: Tricks of the Trade*, LNCS, Springer, Berlin, **1998** (b) C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, **1995**

[4] K.-R. Müller, G. Rätsch, S. Sonnenburg, S. Mika, M. Grimm, N. Heinrich, *J. Chem. Inf. Model*, **2005**, *45*, 249-253

[5] (a) F. R. Burden, *J. Chem. Inf. Comput. Sci* **2000**, *41*, 830–835 (b) D. P. Enot, R. Gautier, J. Le Marouille, *SAR QSAR Environ. Res.* **2001**, *12*, 461–469 (c) P. Tino, I. Nabney; B. S. Williams, J. Lösel, Y Sun, *J. Chem. Inf. Comput. Sci* **2004**, *44*, 1647–1653

[6] J. Quionero-Candela, C. E. Rasmussen, *Journal of Machine Learning Research*, **2005**, *6*, 1939–1959

[7] J. Sadowski, C. H. Schwab; J. Gasteiger. *Corina v3.1*, Molecular Networks GmbH Computerchemie, Erlangen, **2005**

[8] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley, New York, **2000**

[9] To speed up model training and reduce the memory demand, we employed a wrapper script to perform a k-means clustering based on descriptors and train one GP model for each cluster of up to 5000 compounds. When applying this model, the wrapper considers each GP and chooses the prediction with the highest confidence (smallest predicted error bar).

[10] P. Bruneau, *J. Chem. Inf. Comput. Sci.* **2001**, *41*,1605–1616