Predictive Power of Time-Series Based Machine Learning Models for DMPK Measurements in Drug Discovery

Modest von Korff^(✉) , Olivier Corminboeuf , John Gatfield,
Sébastien Jeay , Isabelle Reymond , and Thomas Sander

Idorsia Pharmaceuticals Ltd., 4123 Allschwil, Switzerland
{modest.korff, olivier.corminboeuf, john.gatfield,
sebastien.jeay, isabelle.reymond,
thomas.sander}@idorsia.com

Abstract. Four datasets measuring DMPK (drug metabolism and pharmacokinetics) parameters, and one target protein-specific dataset were analyzed by machine learning methods. Parameters measured for the five compound sets were biological activity data, plasma protein binding, permeability in MDCK I cell layers, intrinsic clearance by human liver microsomes, and plasma exposure in orally dosed rats. The measured data were sorted chronologically, reflecting the order in which they had been obtained in the discovery project. Subsets of the chronologically sorted data that appeared early in the project were used as training datasets to build predictive models for subsequent compounds based on kNN, partial least squares regression (PLSR), nonlinear PLSR, random forest regression, and support vector regression. A median model was used as a baseline to assess the machine learning model prediction quality. Data sets sorted in order of increasing test set prediction error: intrinsic clearance, plasma protein binding, cell layer permeability, biological activity on target protein, and bioavailability as AUC in rats. Our results give a first estimation of the power of machine learning to predict DMPK properties of compounds in an ongoing drug discovery project.

Keywords: Machine learning · DMPK · Drug discovery · Biological processes

1 Introduction

In drug discovery, new molecules undergo clinical trials with human subjects only after passing numerous checks for safety and potency in biological test systems. Often desired is a drug suitable for oral administration, i.e., a molecule that can cross cellular membranes separating the gastrointestinal system from blood vessels. Cell assays with MDCK I cells are used to assess membrane penetration. After absorption, blood vessels distribute the molecule across the organism and bring it to its site of action. Blood contains many proteins that bind a substantial fraction of the compound. This is measured as plasma protein binding (PPB). On its way, the molecule passes through the liver that contains enzymes able to metabolize many types of chemical substances, thus reducing the active drug's concentration (clearance). In drug discovery, a

suspension of human liver microsomes is used to assess intrinsic clearance. An important measure to optimize for a bioactive molecule is its plasma exposure after oral administration that is often expressed as “area under the curve” (AUC), i.e., the concentration of the active molecule in blood plasma integrated over time. Bioavailability depends on multiple properties of the molecule including cell layer permeability and clearance in the liver.

Medicinal chemists need quantitative models that would allow to prioritize the most promising molecules for biological testing. Quantitative structure-activity relationships (QSAR) is a central technology in drug discovery and have been investigated in multiple publications [1, 2]. But the chronological occurrence of information in the project workflows of drug discovery has rarely been investigated. A chemical series in drug discovery starts usually with one or a few molecules, often, with modest activity on the target protein. The starting compounds are modified by medicinal chemists to improve their properties. In a maturing drug discovery project, compounds become “smarter”, because they contain more information from previous measurements. Here, we show how to set up quantitative machine learning models in an industrial drug discovery context in order to predict biological properties.

2 Methods

2.1 Data Sets (Table 1)

Each data set contained molecules from the same chemical series that shared an identical ‘backbone’ chemical substructure. Our first dataset contained measurements of half-maximal effective concentration (EC₅₀) on the target protein for 1400 molecules. Much fewer data points were available for the other four datasets originating from active drug discovery programs. PPB is an important measure to assess the concentration of the free (unbound) molecule in the blood.

Table 1. Data sets used for machine learning models.

Data set ID	Dataset name	Data response (min, median, max)	Unit	Number of molecules
1	Activity on target protein	45, 1100, 20000	nmol/l	1400
2	Plasma protein binding	57, 94, 100	%	129
3	Permeability in MDCK I	0.06, 2.7, 39.4	10–6 cm/s	89
4	Intrinsic clearance HLM	65, 313, 1250	l/(min mg)	179
5	Bioavailability AUC	0, 540, 31600	ng h/ml	182

Dataset 2 contained 129 molecules. The permeability MDCK I (dataset 3) contained data from 89 molecules that were tested in a cell permeability assay. High permeability values are desired when the compound should pass the intestinal membranes in humans. The intrinsic clearance (dataset 4) contained data for 179 molecules. To determine the stability of the molecules they were measured in a human liver

microsomes (HLM) assay. Dataset 5 contained area under the concentration vs time curve (AUC) values from rats for 182 compounds representing the overall bioavailability of the compound. After the compound was administered to rats, blood samples were taken and the concentration of the test compound in the blood was measured.

To create quantitative computer models for molecules it is necessary to encode the molecules as vectors $\mathbf{x}$. We decided to use the Skeleton Spheres descriptor [3], where one row in the matrix $\mathbf{X}$ represents one molecule. For each molecule in a data set, there is a single response value $\mathbf{y}_i$. In every data set, the molecules were highly similar in descriptor space as they were in chemical space.

2.2 Machine Learning Techniques

Five modeling techniques were applied to construct regression models for the data sets: KNN regression, PLSR, PLSR with power transformation, random forest regression, and support vector (SVM) regression. All parameters for these machine learning models were optimized by an exhaustive search. The median model was used as a baseline model. Any successful machine learning model should be significantly better than the baseline model. Almost as simple was the k next neighbor model for regression. Partial least square regression (PLSR) is a multivariate linear regression technique [4] that requires the number of factors as the only input parameter. PLSR with power transformation includes a Box Cox transformation. It is often used to model biological data, which are notorious to be not normally distributed [5]. For random forests, we used the implementation from Li [6]. The Java program library libsvm was used for the support vector machine regression [7].

2.3 Successive Regression

To assess the predictive power of a machine learning tool in a drug discovery project it is necessary to consider the point in time a compound was made. Therefore, we ordered all molecules in a dataset according the point in time it was synthesized. A two-step process was implemented to ensure an unbiased estimation for the predictive power of a model. The first step was the selection of one meta-parameter set for every machine learning technique. The algorithm started with the first 20% of the molecule descriptors $\mathbf{X}_{0,0.2}$, $\mathbf{y}_{0,0.2}$ together with the measured response values to determine the meta parameters of the machine learning models via an exhaustive search. An eleven-fold Monte Carlo cross validation was employed to split all data into the training and validation datasets [8]. A left out of 10% was chosen as the size of the validation dataset. With this setup, the average error for all meta-parameter sets was calculated. For each machine learning technique t , the meta-parameter set $\mathbf{M}_{\min,t}$ was chosen that showed the minimum average error. This meta-parameter set was used to construct a model from all data in $\mathbf{X}_{0,0.2}$, $\mathbf{y}_{0,0.2}$. In the second step, an independent test set was compiled from the next 10% of data, $\mathbf{X}_{0,3}$, $\mathbf{y}_{0,3}$. The average prediction error of $\widehat{\mathbf{y}}_{0,3}$ gave an unbiased estimator for the model, because the machine learning algorithm $\mathbf{M}_{\min,t,0.2}$ had not seen these data before prediction. Subsequently, step one was repeated, this time with the data set $\mathbf{X}_{0,3}$, $\mathbf{y}_{0,3}$. So, the former test data were added to $\mathbf{X}_{0,0.2}$, $\mathbf{y}_{0,0.2}$. The meta parameter for the machine learning algorithms $\mathbf{M}_{\min,t,0.3}$ were

now determined with $\mathbf{X}_{0,0.3}$, $\mathbf{y}_{0,0.3}$. So, the prediction was done for $\mathbf{y}_{0.4}$. This process was repeated eight times, up to a model size with $\mathbf{X}_{0,0.9}$, $\mathbf{y}_{0,0.9}$ and a prediction for $\mathbf{y}_{1,0}$. With this method, we assessed how the predictive power depends on the time point when the data were obtained in a drug discovery project. The 10% test set, next in time, was an unbiased estimator of the model's quality.

3 Results and Conclusions

For all five data sets, increasing portions of the chronologically sorted biological data were used as training data to build models that predicted the next 10% of the data. For the largest data set (Fig. 1) already all 'first-step models' had more predictive power than the median model. As the project developed in time, it could be observed that the variance in the EC50 values declined. This was indicated by the smaller median error.

Fig. 1. Results for the successive prediction of EC50 values for dataset 1. The x-axis specifies the fraction of the full dataset used as training data. The y-axis indicates the average error in nmol/l for the prediction of EC50.

For the PPB dataset, the first training data set with $\mathbf{X}_{0,0.2}$, and $\mathbf{y}_{0,0.2}$ contained only 26 molecules. All 'first-step models' predicted the test data better than the median model. From $\mathbf{X}_{0,0.5}$, and $\mathbf{y}_{0,0.5}$ all predictions were superior compared to the median model.

The successive prediction of intrinsic clearance was more successful than the prediction of the permeability MDCK I dataset. The machine learning models outperformed the median model (data not shown). On the AUC (bioavailability) dataset 5, no machine learning technique performed better than the median model (data not shown).

Summary. All vv inner similarity in the SkeletonSpheres descriptor domain. No time-dependent learning curve was observed for the five biological datasets. The introduction of newly designed compounds increased the model error in several cases, even though the added compounds shared larger parts of their molecular structure with the training dataset compounds.

Conclusions. The uncertainty in prediction is correlated with the underlying biological complexity of the modeled parameter. Meaningful models for most of the PPB test datasets were created by all techniques. The permeability MDCK I (dataset 3) can be partially explained by diffusion that is relatively easy to model. However, MDCK I cell membranes contain active transporter proteins. Whether a molecule will be substrate for a transporter is hard to predict. The intrinsic clearance (data set 4) depends on the activity of approximately 20 enzymes, which makes the modeling more challenging than activity prediction for a single target enzyme. The bioavailability measured as AUC (dataset 5), is the result of multiple processes in animals, including cell layer permeability and intrinsic clearance. Consequently, the bioavailability AUC model had the highest uncertainty, followed by the target protein bioactivity data, and the intrinsic clearance model, while the most reliable models were created for the PPB and the permeability dataset. Naturally, all predictions include the uncertainty of the measurements of the response data. We presented time-series prediction results for four important measurements in drug discovery: PPB, permeability MDCK I, intrinsic clearance for human liver microsomes, and oral bioavailability as AUC in rats. The chronology-based prediction gives a good estimation of how well results of biological tests can be modeled in drug discovery projects.

References

1. Cherkasov, A., et al.: Qsar modeling: where have you been? where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014). <https://doi.org/10.1021/jm4004285>
2. Gramatica, P.: Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26**, 694–701 (2007). <https://doi.org/10.1002/qsar.200610151>
3. Boss, C., et al.: The screening compound collection: a key asset for drug discovery. *Chimia (Aarau)* **71**, 667–677 (2017). <https://doi.org/10.2533/chimia.2017.667>
4. de Jong, S.: SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.* **18**, 251–263 (1993). [https://doi.org/10.1016/0169-7439\(93\)85002-X](https://doi.org/10.1016/0169-7439(93)85002-X)
5. Sakia, R.: The box-cox transformation technique: a review. *J. Roy. Stat. Soc.: Ser. D (Stat.)* **41**, 169–178 (1992). <https://doi.org/10.2307/2348250>
6. <https://haifengl.github.io/smile/>. Accessed 02 July 2019

7. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**, 27 (2011). <https://doi.org/10.1145/1961189.1961199>
8. Xu, Q.-S., Liang, Y.-Z.: Monte carlo cross validation. *Chemometr. Intell. Lab. Syst.* **56**, 1–11 (2001). [https://doi.org/10.1016/S0169-7439\(00\)00122-2](https://doi.org/10.1016/S0169-7439(00)00122-2)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.