ELSEVIER

Can we estimate the accuracy of ADME–Tox predictions?

Igor V. Tetko[1], Pierre Bruneau[2], Hans-Werner Mewes[1], Douglas C. Rohrer[3] and Gennadiy I. Poda[3]

[1] Institute for Bioinformatics, GSF – National Research Centre for Environment and Health, Neuherberg, D-85764, Germany
[2] AstraZeneca Centre de Recherche Parc Industriel Pompelle BP 1050 – 51689 Reims, Cedex 2, France
[3] Structural and Computational Chemistry, Pfizer Global Research and Development, 700 Chesterfield Parkway West, Chesterfield, MO 63017, USA

There have recently been developments in the methods used to access the accuracy of the prediction and applicability domain of absorption, distribution, metabolism, excretion and toxicity models, and also in the methods used to predict the physicochemical properties of compounds in the early stages of drug development. The methods are classified into two main groups: those based on the analysis of similarity of molecules, and those based on the analysis of calculated properties. An analysis of octanol-water distribution coefficients is used to exemplify the consistency of estimated and calculated accuracy of the ALOGPS program (http://www.vcclab.org) to predict in-house and publicly available datasets.

Each year an increasing number of computational methods devoted to the development of predictive ADME–Tox models is published. Despite the fact that their importance for the drug discovery process is well recognized [1], the available methods are not yet sufficiently reliable and are limited in their application [2]. For example, recent reviews [3,4] indicate that as many as 50 articles devoted to methodological developments to predict lipophilicity and aqueous solubility were projected to be published in 2005. This is about a fivefold increase compared with 1995. However, the prediction accuracy for proprietary datasets remains disappointingly low [5–8].

One can describe such relative levels of failure in terms of the applicability domain (AD) of the models. In the 'ontology' classification of the model failure, one can distinguish at least two major problems: experimental design and diversity of compounds. The experimental design problems can result from different endpoints of the models [9] (i.e. agreement of protocols used in the development of the models, data consistency and quality, and model applicability).

The second reason for model failure is the difference in chemical space of compounds that were used to develop and apply the models. This problem can also be attributed to experimental design problems: in predictive models, both training and test set compounds have to be from the same chemical space [10–14]. However, there are at least two principal reasons why such a situation is unlikely. First, because of the proprietary nature of research in pharmaceutical firms, one cannot expect a situation to arise in which a sufficient amount of proprietary experimental data will be publicly released to develop specific models [15]. Second, the available chemical space of synthetically feasible chemistry is extremely large. Therefore, it is unrealistic to hope that a 'magic bullet'– that is, a method that would reliably predict physicochemical properties of any possible chemical – will be developed.

The problem of the AD of chemical models has also received considerable attention in the European Union. As a result of a new system for registration evaluation and authorization of chemicals (REACH), the European Union requires a clear estimation of the accuracy of developed QSAR models before they can be used within the REACH system. The European Centre for the Validation of Alternative Methods recently published a comprehensive meeting report with the results of the current status of the field [16]. The AD of a model is defined in this study as: 'the response and chemical structure space in which the model makes predictions with a given reliability'.

In this review, we focus mainly on methods to predict selected physicochemical properties of compounds, particularly

Corresponding author: Tetko, I.V. (itetko@vcclab.org)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.06.013

lipophilicity and aqueous solubility. These two properties are supported with the largest experimental datasets collected by industry and publicly available databases – for example, PHYSPROP (http://esc.syrres.com/interkow/KowwinData.htm) or LOGKOW (http://logkow.cisti.nrc.ca), and, thus, they are frequently used to develop and test new methods to estimate the accuracy of prediction of these properties. Here, we consider two major groups of methods: approaches based on analysis of similarity of molecules ('molecular-based similarities') and approaches based on analysis of calculated ADME–Tox or physicochemical property models ('property-based similarities').

Methods that employ molecular-based similarities

The main hypothesis of this group of methods is based upon an assumption that similar molecules exhibit similar properties [17]. Neighborhood behavior can be expected for additive properties–that is, when addition of each additional group of atoms or a functional group consistently provides a certain increase or decrease of the target property of molecules. The success of fragment-based methods [3,4,18,19] for the prediction of physicochemical properties does confirm that some of them can be considered as additive. Thus, it is not surprising that methods employing molecular-based similarities have been widely developed.

The truth about 'missing fragments'

In QSAR methods (in particular, fragment-based approaches), the accuracy of prediction depends upon the presence of all fragments required to estimate a given property. The accuracy decreases when some fragments are entirely absent in the training set, or have a very low frequency of appearance. Thus, for these fragments, no statistically significant coefficients can be calculated. For example, users of the CLOGP program [20] are familiar with the problem of missing fragments, particularly in the earlier versions of this software. The CLOGP calculator of version 4 and below was unable to predict lipophilicity for molecules containing such fragments. Later on, a 'no missed fragments' version of the program was developed, which included *ab initio* estimation of the contributions of the missed fragments [21]. It was claimed to estimate the accuracy of new compounds with an error below 0.5 log units. However, our studies indicated that ~67% of molecules (376 out of 558) with large prediction errors in logP (>1.5 log units) in the PHYSPROP dataset (12,908 molecules from PHYSPROP database selected as indicated in Ref. [22]) contained fragment values calculated by the *ab initio* method [22]. The ALOGPS program [22] (http://www.vcclab.org), which predicts lipophilicity and aqueous solubility of chemical compounds, flags unreliable predictions if the analyzed molecule contains one or more E-state atom or bond types that were missed in the training set. This simple flag made it possible to indicate 90% (357/394) of outlying molecules with large prediction errors (>1.5 log units) for the same training and test sets.

The ISIDA (*In Silico* Design and Data Analysis) software suite [23] calculates an average model as a combination of $n = 3$–5 of the most statistically sound models developed with up to 49 different types of molecular fragments. When testing a new molecule, the models that contain missed fragments are not considered in the averaging. This makes it possible, on the one hand, to moderate the problem of missed fragments and, on the other hand, to improve the predictive ability of the method as demonstrated by benchmarking studies [24]. By analyzing the number of rejected models, as well as the variance of model predictions, the user can have a qualitative assessment of the reliability of the predicted value.

Thus, although the problem of 'missing fragments' is sometimes considered to be a drawback of the approach, it does have an important quality control role. Experimental measurements of compounds with missing fragments can be used to determine the contribution of such fragments and, thus, increase the predictive power of the method [25].

The use of expert knowledge to define the AD (e.g. the mechanistic analysis of chemical reactions leading to skin sensitization [26]) can also be classified to the group of methods considered in this section.

AD in the descriptor space

The 'missing fragment' approach will work only in cases where a new molecule contains fragments not covered in the training set. The problem with this method lies with the definition of a missing fragment. One can treat the whole molecule as a fragment. Thus, depending on the methods used to generate the fragments, different approaches developed using the same training set will find different molecules as having 'missing fragments'. The 'missing fragments' approach is not directly applicable to methods that rely on descriptors determined for molecules as a whole – for example, topological descriptors [27,28].

The analysis of the range of descriptors was shown to be a very efficient approach for the determination of the AD of models. It is the basis of the optimal prediction space (OPS) used in the TOPKAT package [10,29]. The OPS was initially developed for the prediction of lipophilicity of chemical compounds in the VLOGP program [10]. The authors discussed the importance of the development of a robust model, and provided several characteristics that would be necessary for the model to be robust: all descriptors are significant, there are no compounds with unique variables, no outlying or leverage compounds are left in the training set, residuals are normally distributed and cross-validation results are not statistically different from those calculated for the training set. After the creation of such a model, the range-based cut-offs are used to determine whether the query molecule is inside or outside of the space of the training set of molecules (for more details, see the TOPKAT patent [29]). Using a test set of 113 compounds, the authors demonstrated that the predictive ability for 29 compounds outside the OPS was about five times lower than compounds inside it [10]. Therefore, the use of the OPS enables discrimination between 'bad' and 'good' predictions.

The ranges in the descriptor space were applied to define the ADs for KowWIN [30]. The authors concluded that principal component analysis provided the simplest acceptable solution, and the use of more complex distance-based molecular similarity approaches in the descriptor space did not improve the results.

Because of its simplicity and the ease of interpretation, the analysis of 'missing fragments' and range of descriptors is used in major physicochemical software packages produced by ACD Laboratories LogD/Solubility suite, BioByte CLOGP, Simulation Plus ADME–Tox Predictor, QikProp2.2 and others.

Methods based on similarity of molecules in the descriptor space

A sound classification of methods based on the similarity of molecules in the descriptor space was proposed by Jaworska *et al*. [31]. According to this article, the methods can be attributed to several major categories: (i) range-based methods; (ii) geometric methods, (iii) distance-based methods and (iv) probability density distribution range methods. Some of these methods are implemented in the Ambit software (http://ambit.acad.bg). The methods considered in the previous two sections fall into the first category. The geometric methods determine a convex hull or convex envelope – that is, the smallest convex region enclosing all points from the training set. This convex hull can be used to define the AD of the model (Figure 1). However, the hull might contain data regions with a low density of points and, thus, result in models with low accuracy. In fact, the requirement of the OPS to build a model without outlying or leveraging compounds implicitly tries to avoid such situations.

The distance-based methods calculate a distance from the test set compounds to the training set compounds. Different measures, such as Euclidian (Equation 1), city block, as well as three other interrelated measures such as Mahalanobis, hotelling T^2 and leverage (Equation 2), were used to assess the quality of predictions [31–35].

$$D_{ij} = \left(\Sigma\left(x_k^i - x_k^j\right)^2\right)^{1/2} \quad \text{[Eqn 1]}$$

The leverage is defined as:

$$H = x^T\left(X^T X\right)^{-1} x \quad \text{[Eqn 2]}$$

Where x is the vector of descriptors of a query compound and X is the matrix formed with descriptors from the training set. High h values indicate that the analyzed compound stands out from the training set and might involve extrapolation rather than interpolation. Leverage was recommended for assessing AD in several studies [36,37].

The more complex probability density distribution-based methods (Figure 1c) can also be used to detect dense and low populated regions of the structural space [16]. However, these methods are computationally intensive and, thus, cannot be efficiently used in models involving a large number of descriptors or molecules [16,30]. Of course, one can also attempt to correlate the accuracy of prediction with several other molecular similarities – for example, based on the shape of molecules, their electron densities, molecular holograms and others, as reviewed by Nikolova and Jaworska [38] and Bender and Glen [39].

To use the methods described in this section mostly relies upon the establishment of a threshold value for compounds lying inside and outside of the AD. Some authors connected the predictive accuracy for molecules with their distance measures from the training set of compounds. Such distance measures include a dimension-related distance, a combined Euclidian distance to the mass center of the convex hull and a distance to the nearest vertex, among others [11]. Those compounds outside of the convex hull were additionally penalized. This measure enabled the authors to calculate confidence levels for the prediction of the aqueous solubility of molecules.

A comprehensive study based on 20 diverse, in-house, activity datasets, including logD, aqueous solubility, pK_a and biological activities was performed to correlate the accuracy of prediction with molecular similarity to the training sets [12]. The authors explored a variety of different machine-learning methods using five sets of descriptors, ranging from regular atom pair to 3D structure based. The accuracy of prediction correlated best with its similarity to the nearest molecule in the training set or with the number of neighbors in the training set. Moreover, according to the authors, this trend did not depend on the nature of used descriptors or on the employed QSAR method.

Structural similarity can also be used to guide the model selection. The accuracy of the aqueous solubility prediction of the most structurally similar molecules to the target compound was estimated using several published models [40], and the model with the lowest bias was selected to predict the target compound.

AD based on predicted property

The previous approaches were mainly focused on molecular descriptors and actually ignored the most important descriptor, the predicted property itself. In fact, the target property was implicitly included in the similarity measures, because it guided the selection of sets of descriptors to optimize the target property. This kind of molecular similarity, based on relevant descriptors only, is known as tailored similarity [41], and its applications have been reviewed elsewhere [42]. One way to enhance the influence of the target property on the AD determination is to weight the variables for similarity distances measures using, for example, the importance of descriptors in the model, such as:

$$D_{ij} = \left(\Sigma w_k\left(x_k^i - x_k^j\right)^2\right)^{1/2} \quad \text{[Eqn 3]}$$

where weights w_k correspond to the importance of the k^{th} descriptor in the model calculated using autoscaled descriptors [16]. The weighting (Equation 3) makes it possible to account for the relative contribution of each variable to the similarity and improves the detection of the AD of the model.

Several methods explore variations in the model residuals as a measure of accuracy and, thus, estimate the AD of the models. In this type of analysis, not one but a set (ensemble) of models is usually calculated (e.g. generating models using different subsets of the data [43], different variables [44] or using simulated annealing [45]). The residuals and/or confidence values of predictions are analyzed to derive the AD of the models.

Significant variation in the predictions of some molecules that could indicate low confidence in their classification was reported some time ago [46]. The use of statistical tests and a large number of models was proposed to improve the accuracy of prediction for such cases. Other studies indicated that, even with as many as 10,000 models, no significant consensus predictions could be derived for some data compounds [47]. Thus, molecules with low predictive confidence might have low prediction accuracy (i.e. they are not covered by the AD). Recently, several approaches have been derived to provide qualitative and quantitative estimates.

The decision forest method builds multiple models by combining, in one predictor, the results of multiple decision trees [44]. The decision trees are constructed to be as heterogeneous as possible, using each variable a maximum of one time in the models. Using the example of the analysis of estrogen receptor

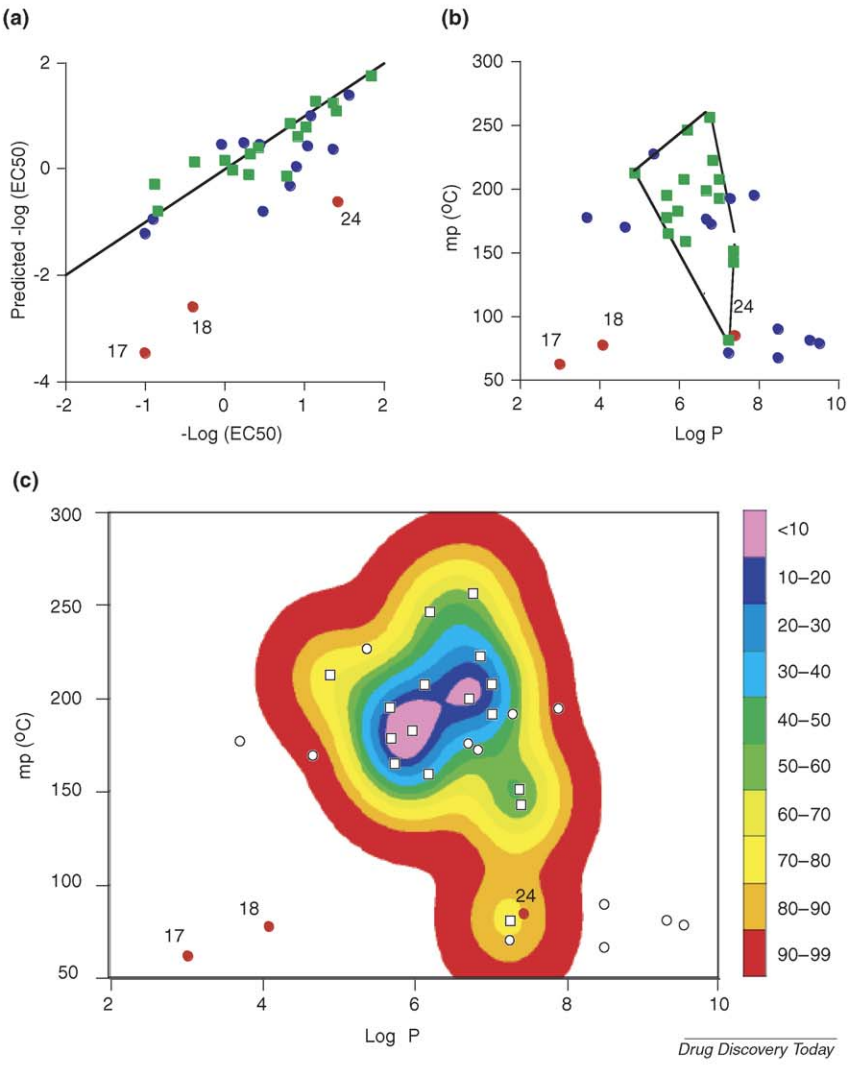

FIGURE 1

(a) The regression line for *in vitro* activity −log (EC_{50}) = 0.016 mp + 0.56 logP − 6.14 was calculated using a training set (squares) for the Selwood dataset [59]. **(b)** Only six test set compounds (circles) are within or near to the AD identified as a convex hull for the training set. **(c)** Probability density distribution estimated using the kde2d{MASS} function of *R* (http://www.r-project.org). The high-density regions covering 60% (very reliable), 90% (reliable) or 99% (less reliable) of total density can be defined as an AD. Only three, six and eight test set compounds are within these regions, respectively. Three outlying compounds from the test set are shown as red circles. The large prediction errors for compounds 17 and 18 (but not for compound 24) can be explained by their out-of-the-domain position. The reported *in vitro* activity for compound 24 could be an experimental error. Indeed, although this compound was reported to be active *in vitro* [−log (EC_{50}) = 1.41], it did not have any *in vivo* activity [59].

binding, it was demonstrated that the prediction accuracy of molecules increased as the confidence level of the prediction increased [16,48].

A similar effect was observed in methods developed to discriminate soluble from poorly soluble molecules [49]. Manallack *et al.* [49] applied an ensemble of neural network models and demonstrated that molecules with small standard deviations of predictions (<0.01) had 2–3 times lower errors compared with the rest of the dataset. Thus, predictions with high standard deviations are outside of the AD of models.

In another study [32], the standard deviation of predictions issued from an ensemble of Bayesian regularized neural nets was shown to be positively correlated with the distance to the model, and both metrics correlate well with the errors of predictions. Both distances were combined [50] to determine a combined distance measure as:

$$CD_{i,model} = (sd_i * D_{i,model})^{1/2} \quad \text{[Eqn 4]}$$

Where sd_i is the dispersion of predictions of the compound *i*, and $D_{i,mod}$ is the minimum Mahalanobis distance from the analyzed compound to all compounds in the training set. The use of the combined distance (Equation 4) provided a better estimation of the accuracy of prediction of a new compound compared with the Mahalanobis distance itself.

The associative neural networks [43] use residuals calculated from an ensemble of models to define a new representation of

molecules. In this approach, each molecule is represented as a vector of residuals. A similarity function between molecules is then introduced as a rank correlation between these vectors (see also the study by Tetko and Tanchuk [51] for other measures). A property-based similarity, R, of a given molecule to a dataset is identified as a square of a maximum correlation of a vector of residuals of the query molecule to vectors of residuals of all molecules in the training dataset [52]. Using an example of a lipophilicity prediction from a PHYSPROP dataset using the ALOGPS program [22,51], we have shown that molecules with $R > 0.8$ and $R < 0.3$ had mean absolute errors (MAE) of ~0.3 and 0.7 log units, respectively [52].

The estimations based on model analysis are pertinent to the target property and might provide more accurate results [42,43]. For example, only two out of five nearest neighbors of biphenyl in the lipophilicity and aqueous solubility spaces were the same (Figure 2), in spite of both models using the same descriptors. The neighbors of biphenyl in the logS space were symmetrical and all contained two phenyl rings. This reflected the widely known importance of crystal packing of compounds, and thus the symmetry of molecules for their aqueous solubility. The symmetry was not important in the logP space, as exemplified by the nonsymmetrical nearest neighbors of biphenyl detected in this space.

The prediction of complex properties (e.g. biodegradation) might also benefit from hierarchical analysis of the reliability of predictions at different stages of the simulated metabolism [53].

Are assessments of the AD robust?

The methods described in this review have usually been applied to a limited number of compounds or series of molecules measured by one company or experimental group. There is some skepticism as to whether these methods would be sufficiently robust for practical applications – that is, to give consistent predictions for data measured by different pharmaceutical firms. In the next section, we focus on a practical application of the AD estimation for logD experimental data measured by two major pharmaceutical companies.

Estimation of the accuracy of the logP prediction using the ALOGPS program

A PHYSPROP dataset containing 12,908 experimental logP measurements (training set), and two datasets with experimental logD measurements for 7498 neutral compounds from AstraZeneca

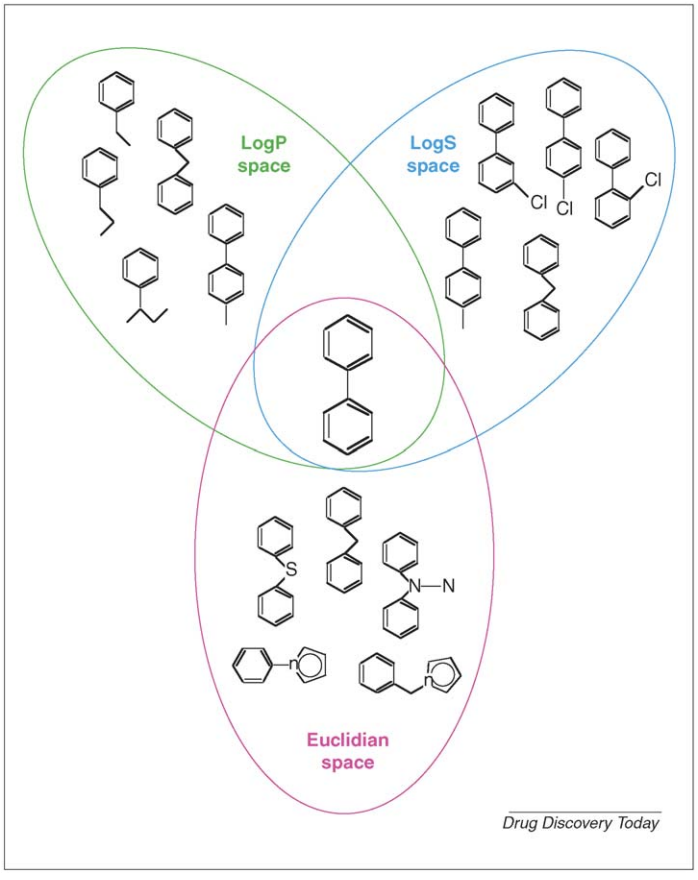

FIGURE 2

The five nearest neighbors of biphenyl in the lipophilicity (logP), aqueous solubility (logS) and Euclidian (Equation 1) spaces. Exactly the same set of 75 descriptors originally used to develop the logP module of the ALOGPS program [22,51] was used by all methods. The neighbors in the logP and logS property-based spaces were calculated as a correlation of vectors of model residuals [43,51]. Neighbors in the logP space were calculated with an interactive version of the ALOGPS 2.1 [60]. A new model was developed to predict water solubility. The nearest neighbors were searched among 12,908 PHYSPROP molecules from the studies by Tetko *et al.* [22] and Tetko and Tanchuk [51]. Only one molecule, diphenylmethane, is common to all three spaces.

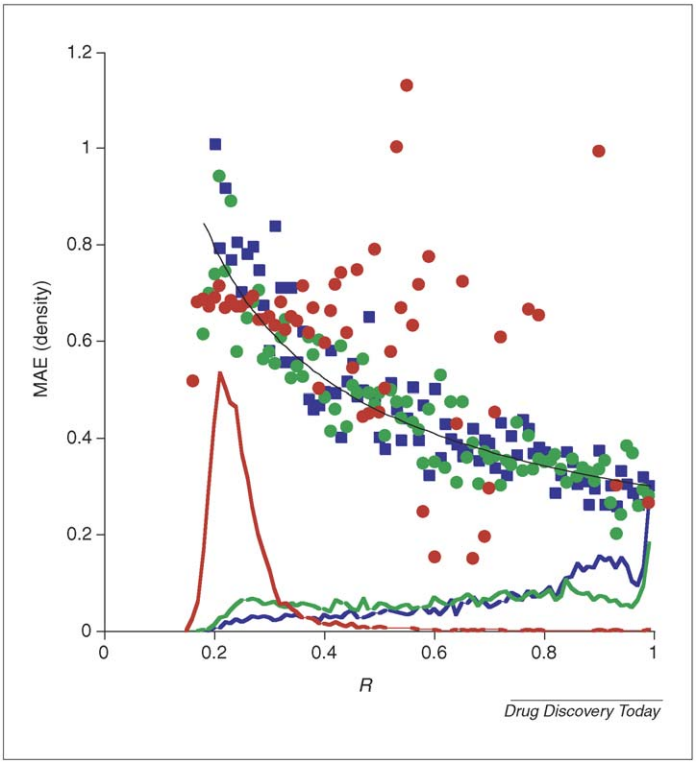

FIGURE 3

MAE (shown as dots) and density of molecules (lines) as a function of property-based similarity, R. Red, green and blue coloring correspond to the AZ blind, AZ LIBRARY and PFE LIBRARY analyses using the ALOGPS program [22,51]. The black line, $y = 0.302 * R^{-0.6}$, indicates the analytical dependency of calculated MAE as a function of property-based similarity. The PHYSPROP training set contains very few compounds that are similar to those of the AZ set with $R > 0.5$, as shown by the density plot (red line). This explains the high variance of the MAE results and the few outlying points observed for the blind prediction of the AZ set with $R > 0.5$ (red dots).

TABLE 1

Estimated and calculated MAE for ALOGPS lipophilicity prediction for different datasets[a]

Dataset	Size	Training or LIBRARY set[b]	Estimated	Calculated
AZ	7498	PHYSPROP	0.69	0.67
AZ	7498	AZ + PHYSPROP	0.42	0.42
PFE	8750	PHYSPROP	0.72	0.74
PFE	8750	PFE + PHYSPROP	0.37	0.37
iResearch Library	13,333,629	PHYSPROP	0.70	n/d[d]
iResearch Library	13,333,629	PHYSPROP + AZ	0.63	n/d
iResearch Library	13,333,629	PHYSPROP + PFE	0.64	n/d
iResearch Library	13,333,629	PHYSPROP + AZ + PFE[c]	0.60	n/d

[a] The PHYSPROP dataset contained 12,908 compounds with logP values.
[b] Datasets used to correct the prediction of the global model [5,6,22,51] and estimate prediction error using Equation 5.
[c] Owing to the confidential nature of in-house datasets, the combined AZ + PFE dataset could not be created. In this analysis, for each molecule we selected the highest correlation coefficient to compounds from the AZ or PFE datasets.
[d] n/d, not determined.

(AZ) and 8750 neutral compounds from Pfizer (PFE) were used and were as described in previous publications [5,6,22,32]. Owing to the confidential nature of proprietary datasets, testing of the AZ and PFE datasets was done by each company independently. Two types of analysis were performed. In the first blind prediction analysis, all molecules were predicted with the 'as is' version of the ALOGPS program [51]. In the second analysis, LIBRARY mode, the corresponding in-house datasets were added to the training set of ALOGPS to extend its applicability domain. LIBRARY mode runs very quickly (it takes ~10 minutes to make the calculation for 17,000 compounds) and significantly improves the predictive power of the models, providing results similar to those obtained from models using an extended training set domain [5,6,32,54].

Figure 3 shows that the accuracy of prediction of compounds increases as R increases for both blind prediction and LIBRARY mode analyses. Moreover, the plots are very similar for data from both AZ and PFE sets. The maximal MAE is observed for molecules with $R < 0.25$ and this value is ~0.8 log units, which is in agreement with previous results [52]. The power fit:

$$MAE_{pred} = 0.302 * R^{-0.6}, \quad \text{[Eqn 5]}$$

where MAE_{pred} is the predicted error, was used as an analytical approximation of the observed dependency of accuracy of prediction on the property-based similarity, R. Using this formula, the expected and predicted error of the ALOGPS program in both blind and LIBRARY prediction modes were in good agreement with the experimental values (Table 1). Equation 5 was also used to estimate the predicted logP errors for the iResearch Library of a collection of 13,333,629 unique SMILES, analyzed in a previous study [55]. The use of in-house data from AZ or PFE decreased the estimated MAE error by ~0.06–0.07 log units. Although these numbers may appear small, the increase in accuracy could be dramatic for some specific subseries of compounds. For example, 514,000 compounds in the iResearch Library that had logP > 5 according to blind prediction with the ALOGPS program changed their values to logP < 5 when the PFE set was used to refine the program predictions. These compounds would not be considered drug-like, according to Lipinski's Rule-of-5 [56,57], using a version of the program developed with the PHYSPROP set only. Moreover, ~495,000 compounds changed their logP values by more than 1 log unit owing to the LIBRARY correction with the PFE set. Thus, the use of in-house data could lead to the development of a program with a higher predictive ability for drug-like compounds than that obtained using public datasets only. It can also be seen that the two companies both explore drug-like, but still different, chemical spaces. The simultaneous use of both sets would decrease the MAE by 0.03–0.04 log units compared with the model developed using the data from only one firm.

Conclusions

Quantitative estimation of prediction accuracy for new compounds can be a very powerful feature in the development of ADME–Tox models and methods for the prediction of physicochemical properties. The results shown in the previous sections illustrate that the estimation of prediction accuracy is now possible. The estimated prediction accuracy might guide the user to apply one or another software package for the analysis of their data. Another way to assess prediction accuracy is to estimate the AD of the model qualitatively and to classify new compounds as being within or outside of the AD.

Incorporation of prediction accuracy in the predicted ADME–Tox properties might significantly improve the quality of compound selection for high-throughput subset screening [58], high-throughput screening hits triage, hit-to-lead stages of drug development and parallel library design for in-house and outsourced chemistry. Indeed, there is a danger that interesting and promising series of compounds could be filtered out from planning as a result of wrongly predicted logP or aqueous solubility. This danger is particularly great for compounds standing far from the investigated chemical space that can be a basis for a new intellectual property. The use of confidence values enables such compounds to be considered for analysis.

Acknowledgements

This work was partially supported by the INTAS-INFO 00–0363 project 'Virtual Computational Chemistry Laboratory' (http://www.vcclab.org). The authors thank Dr Scott Hutton for providing compounds from the iResearch Library (ChemNavigator) used in this study.

References

1 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204
2 Wess, G. (2002) How to escape the bottleneck of medicinal chemistry. *Drug Discov. Today* 7, 533–535
3 Tetko, I.V. and Livingstone, D.J. Rule-based systems to predict lipophilicity. In *Comprehensive Medicinal Chemistry II:* In Silico *Tools in ADMET* (Vol. 5) (Testa, B. and van de Waterbeemd, H., eds), Elsevier (in press)
4 Balakin, K.V. *et al.* (2006) *In silico* approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions. *Curr. Med. Chem.* 13, 223–241
5 Tetko, I.V. and Poda, G.I. (2004) Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *J. Med. Chem.* 47, 5601–5604
6 Tetko, I.V. and Bruneau, P. (2004) Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J. Pharm. Sci.* 93, 3103–3110
7 Morris, J.J. and Bruneau, P.P. (2000) Prediction of physicochemical properties. In *Virtual Screening for Bioactive Molecules* (Bohm, H.J. and Schneider, G., eds), pp. 33–58, Wiley–VCH
8 Walker, M.J. (2004) Training ACD/LogP with experimental data. *QSAR Comb. Sci.* 23, 515–520
9 Stouch, T.R. *et al.* (2003) *In silico* ADME/Tox: why models fail. *J. Comput. Aided Mol. Des.* 17, 83–92
10 Gombar, V.K. and Enslein, K. (1996) Assessment of n-octanol/water partition coefficient: when is the assessment reliable? *J. Chem. Inf. Comput. Sci.* 36, 1127–1134
11 Xu, Y.J. and Gao, H. (2003) Dimension related distance and its application in QSAR/QSPR model error estimation. *QSAR Comb. Sci.* 22, 422–429
12 Sheridan, R.P. *et al.* (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* 44, 1912–1928
13 Yan, A. *et al.* (2004) Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput. Aided Mol. Des.* 18, 75–87
14 Japertas, P. *et al.* (2003) Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini Rev. Med. Chem.* 3, 797–808
15 Wilson, E.K. (2005) Is safe exchange of data possible? *Chem Eng. News* 83, 24–29
16 Netzeva, T.I. *et al.* (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* 33, 155–173
17 Patterson, D.E. *et al.* (1996) Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors. *J. Med. Chem.* 39, 3049–3059
18 Japertas, P. *et al.* (2003) Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini Rev. Med. Chem.* 3, 797–808
19 Mannhold, R. and van de Waterbeemd, H. (2001) Substructure and whole molecule approaches for calculating log P. *J. Comput. Aided Mol. Des.* 15, 337–354
20 Leo, A.J. (1993) Calculating log Poct from structures. *Chem. Rev.* 93, 1281–1306
21 Leo, A.J. and Hoekman, D. (2000) Calculating log P(oct) with no missing fragments; the problem of estimating new interaction parameters. *Perspect. Drug Discov. Des.* 18, 19–38
22 Tetko, I.V. *et al.* (2001) Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* 41, 1407–1421
23 Varnek, A. *et al.* (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* 19, 693–703
24 Tetko, I.V. *et al.* (2006) Benchmarking of linear and non-linear approaches for quantitative structure–property relationship studies of metal complexation with organic ligands. *J. Chem. Inf. Model.* 46, 808–819
25 DeWitte, R. *et al.* (2004) Using targeted measurements to improve the accuracy of physical property prediction. ADMET-1 Conference 2004 (http://support.acdlabs.com/um/admet04), accessed 29 March 2006
26 Aptula, A.O. *et al.* (2005) Skin sensitization: reaction mechanistic applicability domains for structure–activity relationships. *Chem. Res. Toxicol.* 18, 1420–1426
27 Devillers, J. and Balaban, A.T. (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach
28 Todeschini, R. and Consonni, V. (2000) *Handbook of Molecular Descriptors*. Wiley–VCH
29 Gombar, V.K. (2000) Method and apparatus for validation of model-based predictions. US Patent 6,036,349
30 Meylan, W.M. and Howard, P.H. (1995) Atom/fragment contribution method for estimating octanol–water partition coefficients. *J. Pharm. Sci.* 84, 83–92
31 Jaworska, J. *et al.* (2005) QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* 33, 445–459
32 Bruneau, P. and McElroy, N.R. (2006) LogD7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *J. Chem. Inf. Model.* 46, 1379–1387
33 Shen, M. *et al.* (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* 46, 3013–3020
34 Papa, E. *et al.* (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). *J. Chem. Inf. Model.* 45, 1256–1266
35 Gramatica, P. *et al.* (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.* 44, 1794–1802
36 Tropsha, A. *et al.* (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* 22, 69–77
37 Eriksson, L. *et al.* (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* 111, 1361–1375
38 Nikolova, N. and Jaworska, J. (2004) Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.* 22, 1006–1026
39 Bender, A. and Glen, R.C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* 2, 3204–3218
40 Kühne, R. *et al.* (2006) Model selection based on structural similarity-method description and application to water solubility prediction. *J. Chem. Inf. Model.* 46, 636–641
41 Basak, S.C. and Grunwald, G.D. (1995) Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. *Chemosphere* 31, 2529–2546
42 Basak, S.C. *et al.* (2006) Predicting bioactivity and toxicity of chemicals from mathematical descriptors: a chemical-cum-biochemical approach. In *Advances in Quantum Chemistry: Chemical Graph Theory: Wherefrom, Wherefor, & Whereto* (Klein, D.J. and Brandas, E., eds), Elsevier–Academic Press (in press)
43 Tetko, I.V. (2002) Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* 42, 717–728
44 Tong, W. *et al.* (2003) Decision forest: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* 43, 525–531
45 Guha, R. and Jurs, P.C. (2005) Determining the validity of a QSAR model – a classification approach. *J. Chem. Inf. Model.* 45, 65–73
46 Tetko, I.V. *et al.* (1993) Applications of neural networks in structure–activity relationships of a small number of molecules. *J. Med. Chem.* 36, 811–814
47 Tetko, I.V. *et al.* (1998) Application of a pruning algorithm to optimize artificial neural networks for pharmaceutical fingerprinting. *J. Chem. Inf. Comput. Sci.* 38, 660–668
48 Tong, W. *et al.* (2004) Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* 112, 1249–1254
49 Manallack, D.T. *et al.* (2003) A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* 43, 674–679
50 Bruneau, P. (2001) Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* 41, 1605–1616
51 Tetko, I.V. and Tanchuk, V.Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* 42, 1136–1145
52 Tetko, I.V. (2005) Encoding molecular structures as ranks of models: A new secure way for sharing chemical data and development of ADME/T models. Abstracts of Papers of the American Chemical Society 229, CINF, U602, ACS (http://vcclab.org/acs), accessed 29 March 2006
53 Dimitrov, S. *et al.* (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* 45, 839–849
54 Poda, G.I. *et al.* (2005) Towards predictive ADME profiling of drug candidates: Lipophilicity and solubility. Abstracts of Papers of the American Chemical Society 229, MEDI 514, ACS (http://vcclab.org/acs), accessed 29 March 2006
55 Tetko, I.V. *et al.* (2005) Surrogate data – a secure way to share corporate data. *J. Comput. Aided Mol. Des.* 19, 749–764
56 Lipinski, C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249

57 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26
58 Oprea, T.I. and Matter, H. (2004) Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* 8, 349–358
59 Selwood, D.L. *et al.* (1990) Structure–activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study. *J. Med. Chem.* 33, 136–142
60 Tetko, I.V. (2005) Computing chemistry on the web. *Drug Discov. Today* 10, 1497–1500