A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds

David T. Manallack,*,† Benjamin G. Tehan,‡ Emanuela Gancia,† Brian D. Hudson,§ Martyn G. Ford,§ David J. Livingstone,§,# David C. Whitley,§ and Will R. Pitt†

Celltech R&D Ltd., Granta Park, Great Abington, Cambridge, CB1 6GS, United Kingdom, Department of Medicinal Chemistry, Victorian College of Pharmacy, Monash University, 381 Royal Parade, Parkville, Australia, Centre for Molecular Design, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, Hampshire, PO1 2DY, United Kingdom, and ChemQuest, Delamere House, 1 Royal Crescent, Sandown, Isle of Wight, PO36 8LZ, United Kingdom

Received July 20, 2002

BCUT [***B***urden, ***C***AS, and ***U***niversity of ***T***exas] descriptors, defined as eigenvalues of modified connectivity matrices, have traditionally been applied to drug design tasks such as defining receptor relevant subspaces to assist in compound selections. In this paper we present studies of consensus neural networks trained on BCUTs to discriminate compounds with poor aqueous solubility from those with reasonable solubility. This level was set at 0.1 mg/mL on advice from drug formulation and drug discovery scientists. By applying strict criteria to the insolubility predictions, approximately 95% of compounds are classified correctly. For compounds whose predictions have a lower level of confidence, further parameters are examined in order to flag those considered to possess unsuitable biopharmaceutical and physicochemical properties. This approach is not designed to be applied in isolation but is intended to be used as a filter in the selection of screening candidates, compound purchases, and the application of synthetic priorities to combinatorial libraries.

INTRODUCTION

The pharmaceutical industry is currently striving to reduce the costs of researching and developing new medicines. One estimate has put this cost at over $800 million for a single new product.[1] New technologies have been applied to automate the drug discovery process including high throughput screening (HTS)[2] and combinatorial chemistry.[3,4] Unfortunately, the overheads of running HTS and supplying these assays with sufficient compounds means that the cost of each project can be considerable. Avoiding compounds that cause problems in HTS is paramount to improve efficiencies. Computational filters can be applied to lists of compounds to remove those with toxic functional groups, unsuitable physicochemical properties or chromophores.[5−8] Other *in silico* methods that have been applied to compound lists include those attempting to predict biopharmaceutical properties such as absorption, distribution, metabolism, excretion and toxicity (ADMET).[9−11]

Aqueous solubility is a physicochemical property that is currently being studied by numerous researchers.[12] Not only is this a fundamental factor in determining the bioavailability of a compound,[12] the screening, synthesis, or purchase of poorly soluble compounds is a waste of resource and capital. Many different methods have been employed to predict solubility. Examples include multiple linear regression using a variety of descriptors,[13−17] group contributions,[18,19] neural networks (NNs),[15,20−23] Monte Carlo methods,[24] and mobile order theory.[25] The best of these methods typically give correlation coefficients of around 0.9. Given the variation that can be observed in the measurement of solubility from one laboratory to another, one should be wary of results that exceed the accuracy of the experimental data.[12]

This present study has concentrated on developing a computational method to identify those compounds with solubility less than 0.1 mg/mL. The choice of this cutoff was established following discussions with formulation scientists and biologists who struggle to deal with compounds less soluble than 0.1 mg/mL.[6,26] While a number of established drugs have solubility below this level, our intention was to avoid poorly soluble compounds at an early stage in the discovery process in order to save on synthetic efforts, screening slots, and compound acquisition costs. Another aspect of this study was to explore the use of novel descriptors that are relatively quick to calculate. To this end we have investigated BCUT metrics stemming from the work of Burden[27] and Pearlman and Smith,[28,29] which have been implemented in the DiverseSolutions (DVS) software.[30]

BCUT metrics have been used largely for applications involving the use of the biological profile of compounds.[28,31] Typically this would involve defining regions of space known as receptor relevant subspaces, in which active compounds cluster.[28] Further applications using these metrics have emerged,[31−34] highlighting their utility in drug discovery. The study by Pirard and Pickett[31] was noteworthy as it suggested that BCUTs were not merely encoding 2D structural information but contained information relevant to ligand−receptor interactions. Our interest in these parameters has

* Corresponding author phone: +44 1223 238000; e-mail: David.Manallack@denovopharma.com. Current address: De Novo Pharmaceuticals Ltd. Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, UK.
† Celltech R&D Ltd.
‡ Monash University.
§ University of Portsmouth.
# ChemQuest.

10.1021/ci0202741 CCC: $25.00
Published on Web 01/14/2003

been stimulated by the ability of these metrics to be able to identify bioactive compounds belonging to novel structural classes for particular biological targets.[35] This ability has sometimes been called "scaffold-hopping".[36] We therefore wished to apply these parameters to predicting solubility, particularly if the method was able to predict outside of the structural classes used in the training procedure.

EXPERIMENTAL SECTION

Compound Selections. The Syracuse Research Corporation's Physprop database,[37] which contains information on over 4600 compounds, was used as a source of experimental aqueous solubility data. As many of these compounds contain functional groups or possess physicochemical properties not relevant to the pharmaceutical industry, and our methods are not intended to be applied to compounds of this nature, a number of filters were used in an attempt to remove them. These filters were similar to those described by Hann et al.[5] and included a molecular weight range of 150−700, removal of mixtures, salts, and compounds with toxic functionality (see also ref 38). Finally, only experimental data measured between 10 and 40 °C was used. Additional compounds were removed by inspecting each structure with an experienced medicinal chemist. Included in this set of compounds were measurements where the temperature of the determination was not specified and these were set aside as a test set (set #1). The Physprop database contains many congeneric series of molecules. We decided that our data sets should contain only a few representatives of any of these series in order to prevent the networks converging on overly simple solutions based on highly similar compounds. To this end, the data set was reduced in size by the application of a simple Tanimoto cutoff of 0.85 (based on Unity[30] fingerprints). The compounds removed at this step were also retained as a test set (set #2). Finally, the neural network training method requires approximately equal numbers of positive and negative cases. These were defined as compounds with solubility greater or less than 0.1 mg/mL, respectively, and for the purposes of this paper are referred to as "soluble" and "poorly soluble". As there were more soluble than poorly soluble compounds, a random selection of compounds was removed from the soluble set to maintain a 1:1 ratio for training purposes. Once again the compounds removed were retained as a test set (set #3).

Descriptors. Using the Diverse Solutions (DVS) software as implemented within the Tripos' Sybyl software package,[30] 29 standard 3D H-suppressed BCUT metrics were calculated for each compound. The BCUTs are derived from a series of three classes of matrices. The first is related to atomic charge-related values on the diagonal, the second is related with atomic polarizability-related values on the diagonal, and the third relates to H-bond-abilities on the diagonal. Various definitions may be used for the off-diagonal elements such as interatomic distance (employs 3D structure of the compound). The work of Pearlman and Smith[28,29] demonstrated that both the lowest and highest eigenvalues of these matrices reflected aspects of the molecular structure. Table 1 lists the 29 standard 3D H-suppressed BCUT metrics employed in this study.

Variable Selection. The unsupervised forward selection (UFS) procedure[39] was applied to reduce the level of multicollinearity in the training compounds data set. UFS is a forward stepping algorithm that selects a subset of variables with a minimal amount of multicollinearity, halting when the squared multiple correlation coefficient of each remaining variable with those already chosen exceeds a pre-specified level. In this case a value of 0.99 was used as the cutoff, so that UFS omitted only redundant variables and those with a very high degree of multicollinearity while retaining a large proportion of the information in the original data. This reduced the number of BCUT variables from 29 to 20 for the training data.

Table 1. Twenty-Nine Standard 3D H-Suppressed BCUT Metrics Listing the Diagonal and Off-Diagonal Elements and the Eigenvalue Calculated

N	diagonal	off-diagonal	eigenvalue
1	gasteiger charge	inverse distance^2 * 0.05	highest
2	gasteiger charge	inverse distance^2 * 0.08	highest
3	gasteiger charge	inverse distance^2 * 1.25	lowest
4	gasteiger charge	inverse distance^2 * 2.75	lowest
5	gasteiger charge	inverse distance^6 * 0.60	highest
6	gasteiger charge	inverse distance^6 * 2.25	lowest
7	gasteiger charge	inverse distance * 0.02	highest
8	gasteiger charge	inverse distance * 1.50	lowest
9	gasteiger charge	inverse distance * 3.00	lowest
10	hacceptor ability	inverse distance^2 * 2.00	highest
11	hacceptor ability	inverse distance^2 * 3.00	highest
12	hacceptor ability	inverse distance^6 * 16.00	highest
13	hacceptor ability	inverse distance * 0.60	highest
14	hacceptor ability	inverse distance * 0.90	highest
15	hdonor ability	inverse distance^2 * 1.20	highest
16	hdonor ability	inverse distance^6 * 8.00	highest
17	hdonor ability	inverse distance * 0.30	highest
18	hdonor ability	inverse distance * 0.45	highest
19	polarizability	inverse distance^2 * 1.00	lowest
20	polarizability	inverse distance^2 * 1.50	highest
21	polarizability	inverse distance^2 * 2.00	highest
22	polarizability	inverse distance^2 * 3.00	lowest
23	polarizability	inverse distance^6 * 1.25	lowest
24	polarizability	inverse distance^6 * 11.00	highest
25	polarizability	inverse distance^6 * 2.75	lowest
26	polarizability	inverse distance^6 * 8.00	highest
27	polarizability	inverse distance * 0.50	highest
28	polarizability	inverse distance * 2.00	lowest
29	polarizability	inverse distance * 4.00	lowest

Neural Network Methods. Feed-forward, back-propagation networks[40] employing a single layer of hidden units were trained using the BCUT data. Targets of 0 and 1 were used for the poorly soluble and soluble cases, respectively. The networks were trained to minimize a cross-entropy error function with a weight decay regularization term. The weight decay introduces bias to improve the generalization ability of the trained networks and produces an effect similar to that obtained by the early-stopping technique commonly used to avoid over-training in applications of neural networks to QSARs.[40] Preliminary experiments indicated that 0.1 was an appropriate value for the weight decay parameter α. The calculations were performed in MATLAB[41] using the NETLAB software library.[42] Training was conducted in two phases. After dividing the compounds into training, validation and test sets in the ratio 50%:25%:25%, preserving the 1:1 ratio between positive and negative cases in each subset, the size of the hidden layer was varied from 2 to 10 hidden units, and 20 networks were trained for each architecture. Plotting the error for the validation set against the number of hidden units allowed the optimal architecture to be determined. The test set was not involved in the training

process and was used only to give an independent estimate of the performance of networks with the optimal architecture. Comparison of the network performance on the training and test sets revealed only a small decrease in accuracy, indicating that both over-fitting and over-training were avoided. In the second stage of training, 1000 networks were trained using the optimal architecture, employing all the data. From these, a set of 100 networks with the smallest training set error was chosen as an ensemble for future predictive work. In addition, a set of 100 networks with the largest *predictive power of a negative test* (PPNT) (*vide infra*) was selected as a second ensemble for predictive purposes. For each ensemble, the mean of the outputs from the 100 networks was used as a consensus prediction, while the standard deviation of the outputs provided a measure of confidence in the results.

Confusion Matrices. Each network has a single output unit, and a compound was predicted to be soluble if the output value was greater than 0.5, while a value less than this indicated a prediction of poor solubility. To demonstrate the performance of the best networks, a confusion matrix can be constructed in the following form:

TP	FN	TP+FN	100*TP/(TP+FN)
FP	TN	FP+TN	100*TN/(TN+FP)
TP+FP	FN+TN	TP+TN+FP+FN	
100*TP/(TP+FP)	100*TN/(FN+TN)		100*(TP+TN)/(TP+TN+FP+FN)

Here TP = true positive, FP = false positive, TN = true negative, and FN = false negative. (For positive/negative read soluble/poorly soluble.) The entries in the last column are the *sensitivity* (the percentage of positives predicted correctly) and the *specificity* (the percentage of negatives predicted correctly). The entries in the last row are the *predictive power of a positive test* (the percentage of cases correctly predicted to be positive) and the *predictive power of a negative test* [PPNT] (the percentage of cases correctly predicted to be negative). The entry at the bottom right is the total percentage predicted correct. This is the quantity maximized by the network training using an output value of 0.5 to determine the classification.

Linear Discriminant Analysis. The classification into soluble and poorly soluble compounds provided by the network is a nonlinear form of discriminant analysis. To compare this with traditional linear methods, a linear discriminant analysis (LDA) was performed using the network training data, and the resulting discriminant function was applied to the remaining data sets. This analysis was carried out in SYSTAT,[43] using the forward stepping LDA with default parameters.

COMPOUND FILTERING

For compound filtering a number of properties were taken into consideration. First the four properties used by Lipinski and co-workers[6] to define their rule of five were calculated (ClogP,[44] molecular weight, the number of OH and NH groups, and the number of nitrogen and oxygen atoms was calculated using Sybyl[30]). The study by Lipinski[6] concluded that compounds would be predicted to have poor absorption or permeation if two or more rules were broken (MW > 500, ClogP > 5.0, NH + OH > 5, N + O > 10). In addition to these properties, the number of rotatable bonds was calculated using Sybyl,[30] and the results from the consensus neural networks were divided into three classes: 1. mean above 0.5 (predicted soluble), 2. mean below 0.5 and standard deviation below 0.1 (consistently predicted poorly soluble), and 3. mean below 0.5 and standard deviation greater than 0.1 (predicted poorly soluble with less confidence). Compounds failing two or more of Lipinski's rules were discarded. Compounds falling into class 2 from the network results were flagged for inspection; however, these were generally discarded without being viewed. Those compounds failing a single Lipinski rule and belonging to class 3 network results were flagged for further inspection to allow examination by experienced scientists. The number of rotatable bonds, a factor known to influence oral bioavailability,[45] was employed as a filter, if requested, to discard additional compounds deemed unsuitable for screening.

Table 2. Confusion Matrix for the Ensemble of 100 Networks

	soluble	poorly soluble	total	% correct
Solubility Consensus Networks − BCUT				
soluble	344	50	394	87.31
poorly soluble	51	343	394	87.06
total	395	393	788	
% correct	87.09	87.28		87.18

RESULTS AND DISCUSSION

Following the filtration of compounds deemed unsuitable for training, the application of a 0.85 Tanimoto and trimming the final list to maintain a 1:1 ratio of soluble and poorly soluble compounds, we were left with 788 compounds. Training of a series of neural networks showed that the lowest error occurred when two hidden layer units were used. Past experience has shown that when a set of 1000 two-hidden unit networks are trained, many of these converge to the same local minima. To avoid this problem we chose to use three-hidden-unit networks for training. The confusion matrix following the second phase of network training is shown in Table 2. This shows that 87% of the compounds were classified correctly.

Three test sets were retained and used to help validate the NN ensemble. The results for test set #1 found that 82.6% of the soluble and 72.6% of the poorly soluble compounds were classified correctly (Table 3). The second test set performed well with 81.8% of the soluble and 82.9% of the poorly soluble compounds being correctly assigned (Table 3). Good predictions should perhaps be expected with this set, as representatives of these compounds are present in the training sets. The final test set (#3) consisted of compounds left out to maintain the 1:1 ratio of compounds for training. Once again these were predicted with reasonable accuracy [79.8% correct] (Table 3).

The distribution of mean and standard deviation values for the NN ensemble is shown in Figures 1 and 2. The mean values show maximum frequencies near 0 and 1 suggesting a bimodal distribution (Figure 1), while the standard deviation frequencies fall toward zero (Figure 2). Figure 3 shows the mean values for the 314 compounds with a standard deviation less than 0.1, clearly demonstrating the bimodal distribution. Of these 314 compounds, only 3.2% were

Table 3. Results Showing the Predictive Capabilities of the NN Ensemble, the PPNT Ensemble, and ALOGPS Expressed as Percentages for the Training and Test Sets

data set name	N^a	% correct NN ensemble	% correct PPNT ensemble	% correct LDA	% correct ALOGPS	% incorrect ALOGPS	% unclassified ALOGPS
training soluble	394	87.3	83.0	73.1	83.0	11.9	5.1
training poorly soluble	394	87.1	90.6	76.7	69.3	23.4	7.3
set #1 soluble	298	82.6	77.2	72.5	83.6	12.1	4.3
set #1 poorly soluble	73	72.6	83.6	78.1	67.1	21.9	11.0
set #2 soluble	868	81.8	78.3	76.7	89.6	8.8	1.6
set #2 poorly soluble	680	82.9	83.2	82.8	84.3	14.6	1.1
set #3 soluble	461	79.8	75.3	70.1	80.0	15.2	4.8

[a] *N* indicates the number of compounds in the set.

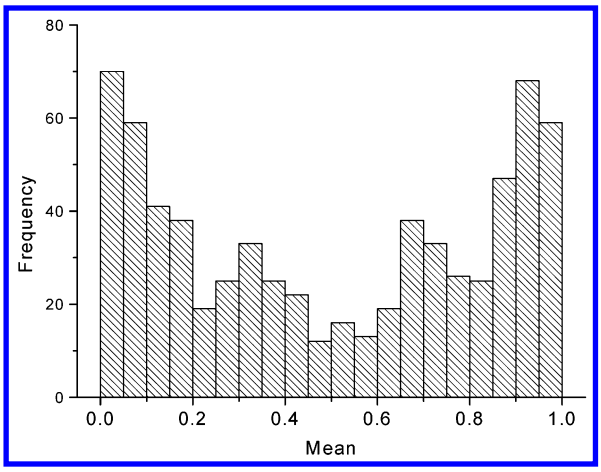

Figure 1. Distribution of mean output values for the NN ensemble.

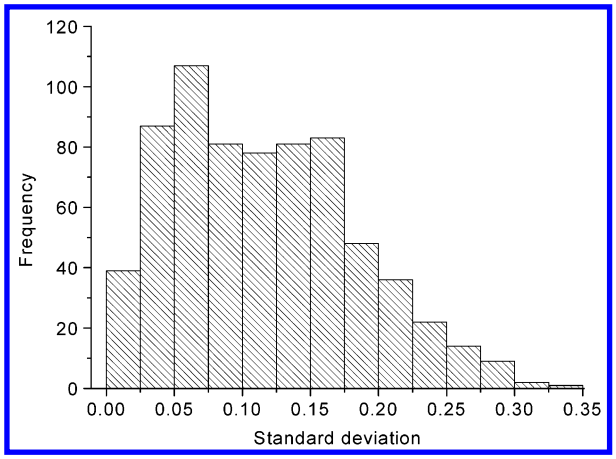

Figure 2. Distribution of standard deviation output values for the NN ensemble.

incorrectly classified. This indicates that applying the standard deviation cutoff results in improved predictions. Using the same cutoff for the standard deviation on the three test sets resulted in error rates of 7.5, 4.6 and 7.6%, respectively. Compounds classified as poorly soluble on this basis can be discarded with about a 1 in 13 chance or better of being predicted incorrectly. While this cutoff improves the predictive ability, this leaves additional compounds with lower levels of confidence in the prediction of insolubility.

One of the aims of this research was to use the consensus neural networks as a filtering tool to highlight poorly soluble compounds. Another way of selecting the ensemble of networks was to base it on the percentage of cases correctly predicted to be poorly soluble (i.e. PPNT). These results are shown in Table 3 and demonstrate that there are improvements in the prediction of the poorly soluble compounds.

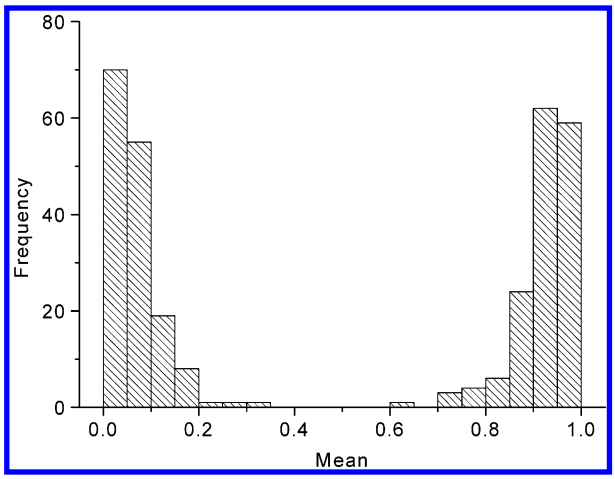

Figure 3. Distribution of mean output values for the NN ensemble using a cutoff of 0.1 for the standard deviation.

This comes of course at a cost, as the percentage of soluble compounds predicted correctly was reduced (Table 3). Surprisingly, selecting compounds with a standard deviation of 0.1 or below resulted in reduced errors of 2.8, 5.8, 3.8, and 5.4% for the training compounds and the three test sets, respectively. This result is improved over the ensemble of networks selected on the overall training error. To our knowledge, ensemble networks have not been selected on this basis before, and the improvements we have seen suggest that this is a useful method and will be the subject of future research.

The neural networks in this study are performing a nonlinear variant of LDA, with the added benefit that an ensemble of networks is able to detect those compounds that are consistently predicted to lie within each class. To compare the ability of the networks to discriminate soluble and poorly soluble compounds we used classical LDA to examine the same data sets. LDA produced a discriminant function based on five BCUT descriptors, and in all but one data set, the ensemble of networks was able to outperform the traditional statistical technique by 6.7% on average (Table 3). This indicates that the nonlinearity inherent in the neural networks results in improved predictive capabilities over the standard linear technique.

While the PPNT results we have presented above seem reasonable, for comparative purposes we applied the same data sets to the ALOGPS program developed by Tetko and co-workers.[46] The results are shown in Table 3, and the overall percentage correct for the training and test sets combined was 82.64%. The PPNT ensemble predicted 82.95% correct for the training and test sets, which compares

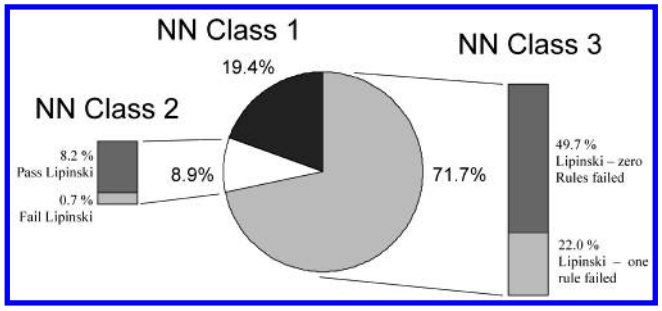

Figure 4. Chart showing the proportion of compounds in the MDDR database belonging to each of the consensus NN prediction classes.

well with ALOGPS. One advantage of using the BCUT parameters was that fewer compounds were unclassified due to missing molecular indices (Table 3). One should keep in mind that the ALOGPS program was not designed to discriminate compounds in this manner and that many of the compounds may have been used for training the ALOGPS system. ALOGPS has, however, the facility to learn using novel data thus enhancing its knowledge base.

To demonstrate the use of our solubility prediction system, it was applied to the MDDR database.[47] Calculation of the BCUT metrics, processing the data, calculation of the consensus network results, and calculation of the physicochemical properties took under 2 h for the MDDR. While this may appear slow, the advantages of not buying and screening poorly soluble compounds would mean that these timelines are considered acceptable for databases of this magnitude (MDDR ~ 120 000 compounds). Using standard Lipinski filters, 17.8% compounds were predicted to have poor biopharmaceutical properties. The consensus neural networks found that 8.9% of compounds belonged to class 2 (i.e. mean < 0.5, standard deviation < 0.1). Interestingly, only 0.7% of compounds were common to both of these lists (Figure 4). Perhaps surprisingly, 71.7% of the database was predicted to be poorly soluble with a standard deviation greater than 0.1 (class 3). Intersection of this 71.7% list of compounds with those that fail a single Lipinski rule highlights 22% of the database (Figure 4). If this were an exercise involving selecting compounds for screening, combinatorial synthesis or purchasing, then the "22% list" would come under additional scrutiny. However, one could argue that for early screening efforts it may be inappropriate to remove these compounds. Indeed, it should be emphasized that our filtration scheme should not be used in isolation but should be a way of flagging compounds that may require additional work to address poor physicochemical characteristics if they are to be the subject of any future lead optimization campaign.

The results we obtained with the test sets applied to the consensus NNs gave us confidence that this method would be applicable as an additional filter in our screening compound selection procedures. It is intended, of course, to test a number of these predictions in the lab to determine whether the networks have performed adequately. A criticism that may be directed at the current study regards the nature of the Physprop database and its suitability for predicting the solubilities of "drug-like' compounds. Figure 5a,b shows the distribution of molecular weights and ClogP values for the 788 training compounds against the MDDR database. As can be seen from Figure 5b the distribution of ClogP

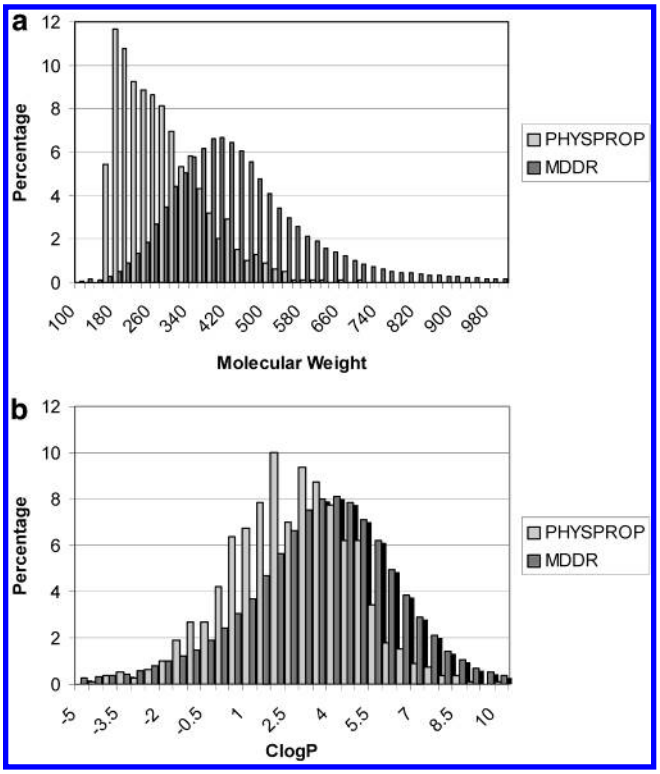

Figure 5. (a) Distribution of the molecular weights for the 788 compound NN training set and the MDDR database. (b) Distribution of the ClogP values for the 788 compound NN training set and the MDDR database.

values for the training compounds is shifted toward lower values. The molecular weight distribution, on the other hand, is considerably different to the MDDR (Figure 5a). It is quite clear that the Physprop database is populated with smaller compounds which is quite understandable perhaps, when one considers the origins of the data. What this does mean, however, is that fewer drug-sized compounds are contained within the training set, and this may have implications given that the consensus networks are applied to predict the solubilities of drug-like compounds. Clearly, there is a need for further solubility data on drug-sized molecules in addition to other needs such as accurate measurements before we will be able to apply these models with full confidence.

CONCLUSIONS

We have demonstrated the use of BCUT metrics via an ensemble neural network technique to adequately discriminate soluble and poorly soluble compounds. This approach is in the process of being implemented as an *in silico* filtering procedure (to be known as BRIKDUST) applied to lists of compounds selected for screening, purchasing, or synthesis via combinatorial methods. When compounds are predicted to be poorly soluble and the standard deviation of the mean value is below 0.1 using the PPNT networks, we estimate there is, at best, a 1 in 20 chance of misclassification. For the other compounds classified as poorly soluble, additional parameters are used to flag compounds that have potentially poor biopharmaceutical and physicochemical characteristics. The advantages of this method are the use of consensus network methods that provide some degree of confidence in the predictive results and the use of BCUT metrics to potentially allow predictions on chemical series not repre-

sented in the training set. Of course, the true test of this system will be to compare the results obtained with our consensus neural network predictions to actual measurements of solubility on compounds of pharmaceutical interest.

ACKNOWLEDGMENT

The authors would like to thank John Cooper for his valuable discussions.

REFERENCES AND NOTES

(1) A report on the costs of developing new drugs from Tufts University, 192 South Street, Suite 550, Boston, MA, 02111 U.S.A. http://www.tufts.edu/med/csdd/Nov30CostStudyPressRelease.html.
(2) Wolcke, J.; Ullmann, D. Miniaturized HTS Technologies − uHTS. *Drug Discov. Today* **2001**, *6*, 637−646.
(3) Furka, A. A. Combinatorial Chemistry: 20 Years On ... *Drug Discov. Today* **2002**, *7*, 1−4.
(4) Kassel, D. B. Combinatonial Chemistry and Mass Spectrometry in the 21st Century Drug Discovery Laboratory. *Chem. Rev* **2001**, *101*, 255−267.
(5) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(6) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(7) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − An Overview. *Drug Discov. Today* **1998**, *3*, 160−178.
(8) Rishton, G. M. Reactive Compounds and In Vitro False Positives in HTS. *Drug Discov. Today* **1997**, *2*, 382−384.
(9) Clark, D. E. Prediction of Intestinal Absorption and Blood-Brain Barrier Penetration by Computational Methods. *Comb. Chem. High Throughput Screen.* **2001**, *4*, 477−496.
(10) Johnson, D. E.; Wolfgang, G. H. Predicting Human Safety: Screening and Computational Approaches. *Drug Discov. Today* **2000**, *5*, 445−454.
(11) Stewart, B. H.; Wang, Y.; Surendran, N. Ex Vivo Approaches to Predicting Oral Pharmacokinetics in Humans In *Annual Reports in Medicinal Chemistry*, Doherty, A. M.; Academic Press: San Diego, CA, 2002; Vol. 35, pp 299−307.
(12) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility From Structure, in *Advanced Drug Delivery Reviews*, Clark, D. E.; Elsevier: Amsterdam, 2002; Vol. 54, pp 355−366.
(13) Huibers, P. D. T.; Katritzky, A. R. Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.
(14) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(15) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds From Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.
(16) Makino, M. Predictions of Aqueous Solubility Coefficients of Poly-Chlorinated Biphenyls by use of Computer-Calculated Molecular Properties. *Environ. Int.* **1998**, *24*, 653−663.
(17) Meylan, W. M.; Howard, P. H. Estimating Log P With Atom/Fragments and Water Solubility With Log P. *Persp. Drug Discov. Des.* **2000**, *19*, 67−84.
(18) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.
(19) Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, 2061−2077.
(20) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(21) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(22) Liu, R.; So, S-.S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.
(23) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.
(24) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility From Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.
(25) Ruelle, P.; Kesserling, U. W. Solubility Prediction of n-Fatty Alcohols and Sterols in Complexing and Non-Complexing Solvents According to the Mobile Order Theory. *Can. J. Chem.* **1998**, *76*, 553−565.
(26) Suzuki, T. Development of an Automatic Estimation System for Both the Partition Coefficient and Aqueous Solubility. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149−166.
(27) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(28) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(29) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Persp. Drug Discov. Des.* **1998**, *9*, 339−353.
(30) DiverseSolutions, v4.0.6; University of Texas, Austin, USA.; Sybyl and UNITY are distributed through Tripos, Inc.: 1669 S. Hanley Rd., Suite 303, St. Louis, MO 63144, U.S.A.
(31) Pirard, B.; Pickett, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(32) Gao, H. Application of BCUT Metrics and Genetic Algorithm in Binary QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402−407.
(33) Mason, J. S.; Beno, B. R. Library Design Using BCUT Chemistry-Space Descriptors and Multiple Four-Point Pharmacophore Fingerprints: Simultaneous Optimization and Structure-Based Diversity. *J. Mol. Graph. Model.* **2000**, *18*, 438−51, 538.
(34) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.
(35) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Targets Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1256−1262.
(36) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: a Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.
(37) The Physical/Chemical Property Database (PHYSPROP) is available from the Syracuse Research Corporation, Environmental Science Center: North Syracuse, NY.
(38) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(39) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: a Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.
(40) Bishop, C. M. *Neural Networks for Pattern Recognition;* Oxford University Press: 1995.
(41) MATLAB is available from The MathWorks, Inc., Natick, MA.
(42) Nabney, I. T. *NETLAB: Algorithms for Pattern Recognition;* Springer, NY, 2002.
(43) SYSTAT is available from SPSS Inc., 444 N. Michigan Avenue, Chicago, IL 60611 U.S.A.
(44) Daylight Corporate Office 27401 Los Altos − Suite 360 − Mission Viejo, CA 92691 U.S.A.
(45) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.
(46) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(47) The MDDR database is available from MDL Information Systems Inc., San Leandro, CA 94577 U.S.A.

CI0202741