molecular pharmaceutics

articles

On-the-Fly Selection of a Training Set for Aqueous Solubility Prediction

Hongzhou Zhang,[†,‡] Howard Y. Ando,[*,†] Linna Chen,[†,§] and Pil H. Lee[*,||]

Research Formulations and Computer-Assisted Drug Discovery, Pfizer Global Research & Development, Michigan Laboratories, 2800 Plymouth Road, Ann Arbor, Michigan 48105

Received January 24, 2007; Revised Manuscript Received June 6, 2007; Accepted June 7, 2007

Abstract: Training sets are usually chosen so that they represent the database as a whole; random selection helps to maintain this integrity. In this study, the prediction of aqueous solubility was used as a specific example of using the individual molecule for which solubility is desired, the target molecule, as the basis for choosing a training set. Similarity of the training set to the target molecule rather than a random allocation was used as the selection criteria. The Tanimoto coefficients derived from Daylight's binary fingerprints were used as the molecular similarity selection tool. Prediction models derived from this type of customization will be designated as "on-the-fly local" models because a new model is generated for each target molecule which is necessarily local. Such models will be compared with "global" models which are derived from a one-time "preprocessed" partitioning of training and test sets which use fixed fitted parameters for each target molecule prediction. Although both fragment and molecular descriptors were examined, a minimum set of MOE (molecular operating environment) molecular descriptors were found to be more efficient and were use for both on-the-fly local and preprocessed global models. It was found that on-the-fly local predictions were more accurate ($r^2 = 0.87$) than the preprocessed global predictions ($r^2 = 0.74$) for the same test set. In addition, their precision was shown to increase as the degree of similarity increases. Correlation and distribution plots were used to visualize similarity cutoff groupings and their chemical structures. In summary, rapid "on-the-fly" similarity selection can enable the customization of a training set to each target molecule for which solubility is desired. In addition, the similarity information and the model's fitting statistics give the user criteria to judge the validity of the prediction since it is always possible that good prediction cannot be obtained because the database and the target molecule are too dissimilar. Although the rapid processing speed of binary fingerprints enable the "on-the-fly" real time prediction, slower but more feature rich similarity measures may improve follow-up predictions.

Keywords: On-the-fly selection; training set; test set; QSPR; aqueous solubility; descriptors; accuracy; precision; local prediction; global prediction; average similarity; Tanimoto coefficients

Introduction

The proper selection of a training set is one of the most basic operations in quantitative structure property relationships (QSPRs). Small, relevant, and homogeneous data sets have and continue to be the workhorse for structure−activity predictions when the activity for a new analogue is needed for a particular chemical series. For large data sets that have

* Corresponding authors. H.Y.A.: Research Formulations, Pfizer Global Research & Development, Michigan Laboratories, 2800 Plymouth Road, Ann Arbor, MI 48105; tel, (734) 622-1278; fax, (734) 622-3609; e-mail, howard.ando@pfizer.com. P.H.L.: Computer-Assisted Drug Discovery, Pfizer Global Research & Development, Michigan Laboratories, 2800 Plymouth Road, Ann Arbor, MI 48105; tel, (734) 622-4744; e-mail, pil.h.lee@pfizer.com.

[†] Research Formulations.

[‡] Current address: Eli Lilly and Company, Lilly Corporate Center Indianapolis, IN 46285.

[§] Current address: Klarquist Sparkman, LLP, One World Trade Center, 121 S.W. Salmon Street, Suite 1600, Portland, OR 97204.

[||] Computer-Assisted Drug Discovery.

10.1021/mp0700155 CCC: $37.00
Published on Web 07/12/2007

been compiled, however, the selection of a training set is critical since compounds of diverse chemical structure are contained within the chemical space of the database. PHYSPROP (www.syrres.com), for example, is a database of physical chemical properties that contains 13,250 compounds (December 2006). It is, in general, very challenging to build a satisfactory global QSPR for a large database that contains such diverse structural classes. Bergstrom et al.[1] have discussed the advantages of local models over truly global models and the requirements that are placed on the training set. Choosing proper training and test sets is critical for successful predictions. Training sets codify the relationship between the relevant property and chemical structure while test sets validate the predictions obtained from these relationships. Bias in either set will impact the statistical probability that the desired property can be accurately predicted for an unknown compound. Thus a randomized selection of these two sets is most often used to increase the probability that they reflect the database as a whole. Tetko, in selecting test and training sets in associated neural networks, divided the PHYSPROP database randomly into two equal sets of 6454 compounds.[2]

For global modeling, especially, it is important that both sets reflect the database as a whole.[3−6] Although a model can be developed from a large diverse training set as shown in Figure 1, the training set that was used might not be appropriate for a particular target molecule. The target may have properties that are not found in the training set. Even if there are relevant examples in the training set, the global model will be biased toward those examples that are in greatest numbers, leaving sparsely represented substructures to suffer a poor fit that accompanies a minority class.

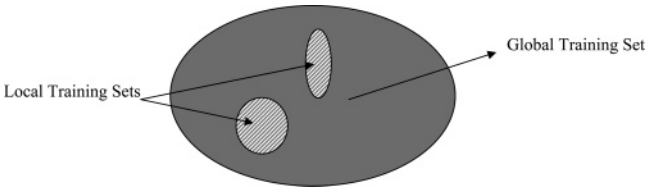

Figure 1. On-the-fly local versus preprocessed global models.

Randomization minimizes, but does not overcome, the potential disparity that occurs for a minority class. Such a model based on a preprocessed training set usually will have no indication of the validity of the prediction for a particular target molecule. Retraining will be needed when new structural features emerge as new data accumulates. Ensembles of many local models overcome some of the issues with global models. Tetko and Tanchuk[2] tested an ensemble of up to 256 associative neutral networks (ASNN), each optimized for a particular domain, to predict solubility and partition coefficient. This ensemble approach was shown to work well,[7−9] but it has the disadvantage of needing to specify the number of ensembles *a priori*.

Use of similarity-based selection of a local training set method is an alternative method to the preprocessed global approach. Lazy learning methods[10] defer the selection of a training set until the target molecule is identified. Local lazy regression (LLR)[11] obtains a prediction using a local neighboring set. Recently, Zhang and co-workers[12] developed a novel automated lazy learning QSAR (ALL-QSAR) using a locally weighted regression technique and applied it to

(1) Bergstrom, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.
(2) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neutral Networks for Prediction of Lipophilicity in AlogPs 2.1 Program. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1136−1145.
(3) Platts, J. A.; Butina, D.; Abraham, M. H; Hersey A. Estimation of Molecular Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.
(4) (a) Yalkowsky, S. H.; Valvani, S. C.; Roseman, T. J. Solubility and Partitioning VI: Octanol Solubility and Octanol-Water Partition Coefficients. *J. Pharm. Sci*. **1983**, *72*, 866−870. (b) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252. (c) Yang, G.; Ran, Y.; Yalkowsky, S. H. Prediction of the Aqueous Solubility: Comparison of the General Solubility Equation and the Method Using an Amended Solvation Energy Relationship. *J. Pharm. Sci*. **2002**, *91*, 517−533.
(5) Yalkowsky, S. H.; Pinal, R.; Banerjee, S. Water Solubility: A Critique of the Solvatochromic Approach. *J. Pharm. Sci*. **1988**, *77*, 74−77.
(6) Bergstrom, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1177−1185.
(7) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D Molecular Similarity Measures and Cluster Analysis. *J. Chem. Inf. Model.* **2004**, *44*, 2083−2098.
(8) Klekota, J.; Brauner, E.; Schreiber, S. Identifying Biologically Active Compound Classes Using Phenotypic Screening Data and Sampling Statistics. *J. Chem. Inf. Model*. **2005**, *45*, 1824−1836.
(9) He, L.; Jurs, P. Assessing the Reliability of a QSAR Model's Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.
(10) (a) Aha, D. W. Lazy Learning. *Artif. Intell. Rev*. **1997**, *11*, 7−10. (b) Armengol, E.; Plaza, E. Discovery of Toxicological Patterns with Lazy Learning. *Knowl.-Based Intell. Inf. Eng. Syst. Part 2, Proc*. **2003**, *2774*, 919−926. (c) Armengol, E.; Plaza, E. Relational Case-Based Reasoning for Carcinogenic Activity Prediction. *Artif. Intell. Rev*. **2003**, *20*, 121−141. (d) Atkeson, C. G.; Moore, A. W.; Schaal, S. Locally Weighted Learning. *Artif. Intell. Rev*. **1997**, *11*, 11−73. (e) Wettschereck, D.; Aha, D. W.; Mohri, T. A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. *Artif. Intell. Rev.* **1997**, *11*, 273−314.
(11) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. *J. Chem. Inf. Model*. **2006**, *46*, 1836−1847.
(12) Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (ALL_QSAR) Approach: Method Developemnt, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J. Chem. Inf. Comput. Sci*. **2006**, *46*, 1984−1995.

virtual screening with reasonable success. The weights with which training set compounds are included in the regression depend on the similarity of those compounds to the target molecule; Euclidean distance in multidimensional descriptor space was used as the similarity metric. Guha et al.[11] investigated the use of local lazy regression (LLR), where the neighborhood of the target compound in the database is determined on-the-fly and is used to build a linear model, which is then used to predict the activity of the target molecule. The neighborhood was determined as the *k* nearest neighbors in the descriptor space where *k* was automatically determined using a leave-one-out (LOO) cross-validation procedure in the lazy package available for R 2.2.0.

Previous similarity based approaches used (1) small databases of 50−75 molecules and (2) the same descriptor set to calculate both similarity (Euclidian distance) and prediction (regression). In this study, we studied a large database of 9433 molecules to increase the probability of being able to select a training set that has highly similar properties to the target molecule. In addition, the training set selection descriptor (Tanimoto coefficients from Daylight fingerprints[13]) was orthogonal to the descriptors (MOE) that were used for regression prediction. We hypothesized that superior results over previous attempts might be possible if the descriptor set reflected as much molecular specificity as possible consistent with high speed on-the-fly evaluations. Thus binary fingerprint evaluations were chosen as the basis for a similarity metric. Similarity was chosen as the selection parameter for choosing training sets since many studies[14] show that prediction accuracy is correlated to the similarity of a test compound to those in the training set. In addition, a similarity-based training set selection can provide for a determination of the relative validity of the training set for the target molecule. In some situations, it is conceivable that good predictions are not possible because sufficiently similar molecules are not available in the database. Knowledge of this fact was deemed to be useful since the user could then disregard appropriate predictions until more relevant molecules are available in the database. This is the advantage of the on-the-fly nature of this procedure compared to preprocessed global models.

Materials and Methods

Dataset: Source and Preparation. An internal database of aqueous thermodynamic solubility, collected and compiled by Lipinski,[15] was used for analysis. The set of 11,026 compounds with experimentally measured solubility values was cleaned using the Pipeline Pilot v.4.0 from the Scitegic to remove all salts and keep only organic compounds which contain exclusively C, H, O, N, S, P, F, Cl, Br, and I atoms. The final dataset of 9443 druglike compounds with a molecular weight greater than 100 was obtained after applying Lipinski's Rule of Five.[16]

Partition of the Dataset into Training and Test Sets. The entire dataset of 9443 compounds was partitioned into training and test sets to develop and evaluate the models. To select a training set to be representative of the whole dataset in chemical space, a diversity analysis was performed using the Pipeline Pilot's FCFP_4 (functional class fingerprint to a maximum diameter of 4). Based on the diversity analysis, 30% of the most diverse structures from the entire dataset were selected to be the diverse set. From the remainder of the dataset (70%), approximately 30% was randomly selected and saved as the test set, and the rest were combined with the diverse set to form the training set. The final training set consists of 7543 compounds (80% of the whole dataset); the test set 1900 compounds (20% of the whole dataset). The solubility distributions of both training and test sets were fairly normal. For both data sets, log *S* (μM) values range from −6 to +5 log units and were centered at 0.5 log unit.

Similarity. All molecules were characterized using Daylight's SMILES/SMARTS/FINGERPRINT toolkits.[13] Canonical SMILES strings were then used to represent the whole molecule and SMARTS for the functional fragments. The Tanimoto similarity coefficients based on the Daylight fingerprints were used as a measure of the similarity between two molecules.[17]

Molecular Descriptors. For model building, two sets of descriptors were calculated. (1) MOE 2D descriptors: A set of 146 2D molecular descriptors were calculated using the MOE 2004.03 software[18] from the Chemical Computing Group, Inc. (2) Fragment descriptors: A wide variety of molecular fragments similar to Abraham's[3] were generated and defined as SMART strings. A set of the 60 most common fragments in the training set is shown in Table 1. The Daylight's SMARTS toolkit was used to parse the predefined SMARTS strings to produce pattern objects. For each molecule, a molecule object is created from its SMILES string. Then, a series of pattern objects was tested to see if the molecule contained the specific patterns. If the molecule contained a specific pattern object, the number of occurrences of this fragment was recorded, and if not, zero is recorded. The number of occurrences of these fragments was used as the descriptor.

(13) Daylight Theoretical Manual, Daylight CIS, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691.

(14) Sheridan, R. P.; Bradley Feuston, P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1912−1928.

(15) In-house database.

(16) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug discovery. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(17) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996

(18) MOE (Molecular Operating Environment) from Chemical Computing Group Inc., 1010 Sherbrooke St. West Suite 910, Montreal, Quebec, 3A 2R7, Canada.

Table 1. Fragment Definition and the Descriptions

fragment smarts	description
[$([CH3X4]C),$([CH2](C)C)]	# 1′ and 2′ carbon
[$([CH1X4](C)(C)C)]	# 3′ and
[$([CH0X4](C)(C)(C)C)]	# 4′ carbon
[CX4]([A,H])([A,H])([A,H])a	# carbon connected to one aromatic
[CX4]([A,H])([A,H])(a)a	# carbon connected to two aromatic
[CX3]=[CX3]	# C=C
[CX3]=[!C]	# C double bond with other atoms
[cH](c)c	# aromatic carbon CH
[cX3H0;R1][C]	# aromatic C with carbon substitution
[cX3H0;R1][N,O,S,P]	# substituted aromatic carbon
[cX3H0;R1][Cl,F,Br,I]	# halide substituted aromatic carbon
[c&!R1]	# bridge aromatic carbon
[$(C#CA),$(C=C=C)]	# triple bond and C=C=C
A=A-A=A	# resonant structure C=C−C=C
[$([OX2H1]C)]	# OH−C
[$([OX2H1]c)]	# OH−c
[$(C-O-C)]	# ether, aliphatic
[$(C-O-c),$(c-O-c)]	# ether, aromatic″
[$(c[CX3;!R1](=[OX1])[OX2H])]	# aromatic carboxylic acid −COOH
[$(C[CX3;R1](=[OX1])[OX2,NX3])]	# lactam or lactone
[#6][#16][#6]	# thio ether
[$([NX3H2]C),$([NX3H1](C)C)]	# 1′ and 2′ amine attached to aliphatic
[NX3H2,NX3H1]c	# 1′ or 2′ amine attached to aromatic
[nX2H0]	# pyridine nitrogen
[nX3H0&R2]	# bridge nitrogen
[O-O,S-S]	# O−O, S−S″
N-O	# N−O oxide
N-N	# hydrozine
N=N	# N=N
S=C	# S=C
[$(N#CA)]	# CN with aliphatic
[$(N#Ca)]	# CN with aromatic
[$([CX3R1](=)[OX2,NX3][CX3R1](=))]	# phthalimide or anhydrous acid
[#9]	# Fluorine
[#17]	# chlorine
[#35]	# bromide
[#53]	# iodine
[#15]	# phosphorus
[$([NX2]=O)]	# NO
[$([#7](=)(~[#8]))	# NO2
[$([CX3H0;!R1](=)([NX3H2,NX3H1])C)]	# aliphatic amide
[$([CX3H0;!R1](=)([NX3H2,NX3H1])c)]	# 1′ and 2′ amide, aromatic
[$([CX3H0;!R1](=)([NX3H0])C)]	# 1′ and 2′ amide, aliphatic
[$([CX3H0;!R1](=)([NX3H0])c)]	# 3′ amide, aromatic″
[$([PX4](=))]	# PO3
[$([PX4](=))]	# P(=)O2
[$([SX4](=)(=))]	# SO2
[$([SX4](=)(=)[NX3H0,OX2H0])]	# SO3
[$([SX4](=)(=)[NX3H1,NX3H2,OX2H1])]	# SO2N
[$([OX2H][AA,aa][OX2&!R,NX3H0&!R])]	# 5-member hbond
[$([OX2H1]Caa~[O,NH0])]	# 6-member hbond
[$([OH1,NX3H1&!R,NX3H2][aR1][aR2][aR1]~[O,N])]	# 6-member hbond
[$([NX3H1&!R,NX3H2][AA,aa][O&!R,NH0&!R])]	# 6-member hbond
n:n	# 1,2-aromatic nitrogen
n:c:n	# 1,3-aromatic nitrogen
n:c:s	# 1,3-thiozine
F,Cl,Br,I]-cc-[F,Cl,Br,I]	# neighboring halide
F,Cl,Br,I]-CC-[F,Cl,Br,I]	# neighboring halide
F,Cl,Br,I]-C-[F,Cl,Br,I]	# multiple halide
[r3,r4]	# 3, 4-member rings

Statistics Analysis. A statistical package CoStat[19] was used to dynamically build multiple linear regression models and to make prediction from the models. The results from CoStat were verified from Minitab[20] as well as SPlus 2000 (Professional release 3, MathSoft Inc, Seattle, WA). We used an independent test set to evaluate the predictability of the models rather than other popular cross-validation techniques such as leave-one-out (LOO) or leave-group-out (LGO).[21] To measure the performance of the model on the test set, we used the correlation between experimental and the predicted values, r^2, as well as the absolute prediction error (APE), which is the difference between predicted and experimental values.

Preprocessed Global Model. Global models were developed from the entire training set. The multiple linear regression (MLR) models for log *S* were generated using the above-mentioned two sets of descriptors. Variable selections were performed with both descriptor sets using the subset selection available in Minitab and stepwise regression in JMP.[22] The final models were selected with the highest r^2 with the least number of descriptors.

On-the-Fly Local Models. For each molecule in the test set of 1900 molecules, an on-the-fly local model was developed from a customized training set using the same set of descriptors in the preprocessed global model. The customized training set was selected based on the molecular similarity from the entire training set (7543 molecules). For a model with the MOE descriptors, a set of the 50 most similar molecules was selected from the entire training set to build a MLR model. For a MLR model with fragment-based descriptors, a set of 100 of the most similar molecules was selected from the entire training set. For the test set of 1900 molecules, 1900 predictions were made from 1900 on-the-fly local models for each of the two descriptor sets. The performances of on-the-fly local models and preprocessed global models were compared in terms of r^2 and the absolute deviation (AD).

Application Work Flow. The application had client and server components. The server, running on a SGI Octane 2 workstation, accessed the data stored in an Oracle database and calculated the similarity between molecules using Daylight fingerprints; on the other hand, the client, which was deployed with Java Web Start technology, communicated with the server using Java RMI (remote method invocation). The overall work flow of this application is shown in Figure 2. When test set molecules have been loaded into the application, the user has options to select the descriptor sets, the size of the customized training sets, or the similarity cutoff. Then, for each test molecule, the algorithm calculates and sorts the molecular similarity coefficients for the training set. The most similar molecules (based on the Tanimoto similarity coefficient and user's option) make up the customized training set for the test molecule. An on-the-fly local MLR model is created and the prediction for the test molecule from the model is made along with statistics, such as r^2, r^2_{adj}, the number of training molecules, the number of descriptors, and the molecular similarity distribution. The chemical structures and properties of the associated training set are also provided to the end user for visualization.

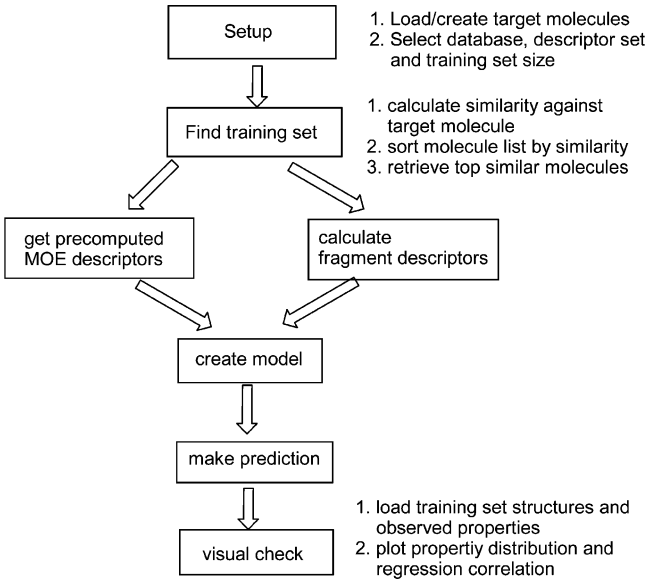

Figure 2. Computational flow from target to prediction and visual check.

The local model implementation provides three possible ways to select the training set: (1) the size of the training set, (2) a molecular similarity cutoff, and (3) both the training set size and a similarity cutoff. If the size of the training set is provided, this application will return the specified number of training molecules regardless of the molecular similarities. If the molecular similarity cutoff was supplied, every molecule with molecular similarity equal to or higher than the cutoff, regardless of the size of the training set, was used for model creation. Finally, if both the size of the training set and the molecular similarity cutoff were selected, the molecular similarity cutoff was applied before the training set size. For example, if there were fewer molecules with similarity cutoffs than specified with the size of the training

(19) CoStat 6.2, CoHort Software, 798 Lighthouse Ave. PMB 320, Monterey, CA 93940.
(20) Release 13.31, Minitab Inc, State College, PA.
(21) Breiman, L.; Spector, P. Submodel Selection and Evaluation in Regression: The X-Random Case. *Int. Stat. Rev.* **1992**, 60, 291−319
(22) Release 5.1.1, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
(23) *CRC Handbook of Chemistry and Physics*; CRC Press: Boca Raton, 1994.
(24) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 868−873.
(25) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem*. **2000**, *43*, 3714−3717.
(26) Labute, P. MOE LogP(Octanol/Water) Model. unpublished. Source code in $MOE/lib/svl/quasar.svl/q_logp.svl (1998).

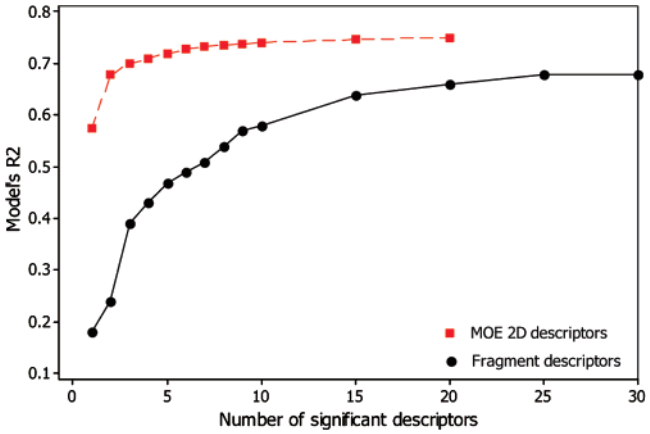

Figure 3. Descriptor efficiency for preprocessed global models.

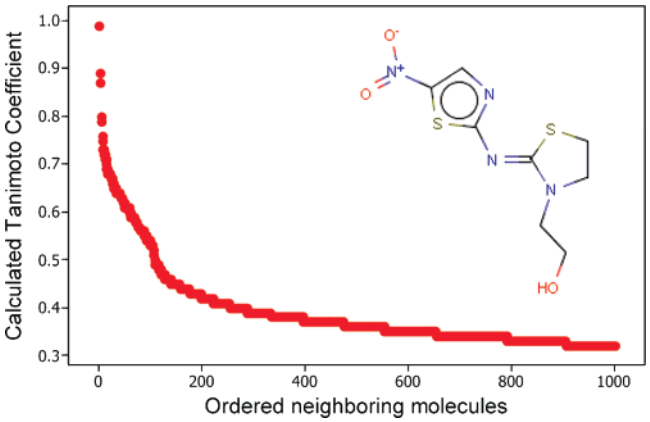

Figure 4. The similarity profile for a given test molecule.

set, then only those with the given similarity cutoff were used. On the other hand, even if more molecules met the given similarity cutoff, only the top molecules were kept. In this paper, we report the results using the fixed training size (case 1) to compare the performance of the on-the-fly local with the preprocessed global model.

Results and Discussion

Since a good model involves an efficient descriptor set as well as a good training set, performances were compared on the same test set to examine the differences between the descriptors (fragment and MOE-molecular) and the models (on-the fly local and preprocessed global).

Comparison of Descriptor Types in Preprocessed Global Model. Two types of preprocessed global models for log *S* were developed using the fragment and MOE-molecular descriptors. Figure 3 shows the performance of these two descriptor types in terms of r^2. After the rapid rise in r^2 for two to three descriptors, a graded cumulative improvement is seen as the number of descriptors increases to 30. For fragment-based descriptors, 10 descriptors gave an $r^2 = 0.58$; all 60 descriptors, an $r^2 = 0.70$. The MOE descriptors, on the other hand, gave an $r^2 = 0.74$ with the 9 descriptors shown in Table 2; an $r^2 = 0.77$ was obtained with all 146 MOE descriptors. Some of the descriptors are highly correlated. The Pearson product-moment correlations between *S* log *P* and log *P*(o/w) are 0.9 and between TPSA and vsa_pol are 0.93. Every other pair has correlation below 0.6. The relative importance of the descriptors in decreasing order is log *P*(o/w), *S* log *P*, SMR_VSA5, weight, a_ICM, TPSA, vsa_pol, PEOE_RPC-, and PEOE_VSA_POS. The descriptors were selected using the forward selection of the stepwise regression tool in the JMP program. Although log *P*(o/w) and *S* log *P* are correlated, each made a correlation contribution to the minimal set to justify retaining both. Figure 3 shows that MOE molecular descriptors were more efficient than the fragment-based descriptors. Apparently, the MOE descriptors more efficiently capture a broad solubility-chemical space than molecular fragments. Similarly, MOE descriptors outperformed fragment descriptors in absolute prediction error (APE). For absolute predicted deviations of less than 0.5 log unit, the full 60 fragment set gave an APE of 56% whereas the 9 MOE descriptors gave an APE of 51%.

Table 2. The Top Nine MOE Descriptors

descriptor	description
a_ICM	Atom information content (mean). This is the entropy of the element distribution in the molecule (including implicit hydrogens but not lone pair pseudo-atoms). Let n_i be the number of occurrences of atomic number i in the molecule. Let $p_i = n_i/n$ where n is the sum of the n_i. The value of a_ICM is the negative of the sum over all i of $p_i \log p_i$.
weight	Molecular weight (including implicit hydrogens) with atomic weights taken from ref 23.
PEOE_RPC-	Relative negative partial charge: the smallest negative charge divided by the sum of the negative charge.
PEOE_VSA_POS	Total positive van der Waals surface area.
vsa_pol	Approximation to the sum of VDW surface areas of polar atoms (atoms that are both hydrogen bond donors and acceptors), such as −OH.
S log *P*	Log of the octanol/water partition coefficient (including implicit hydrogens). This property is an atomic contribution model[24] that calculates log *P* from the given structure, i.e., the correct protonation state (washed structures). Results may vary from the log *P*(o/w) descriptor. The training set for *S* log *P* was ~7000 structures.
SMR_VSA5	Subdivided molecular refractivity (sum of v_i such that R_i is in 0.44−0.485).
TPSA	Polar surface area calculated using group contributions to approximate the polar surface area from connection table information only. The parametrization is that of Ertl et al.[25]
log *P*(o/w)	Log of the octanol/water partition coefficient (including implicit hydrogens). This property is calculated from a linear atom type model[26] with $r^2 = 0.931$, RMSE = 0.393 on 1827 molecules.

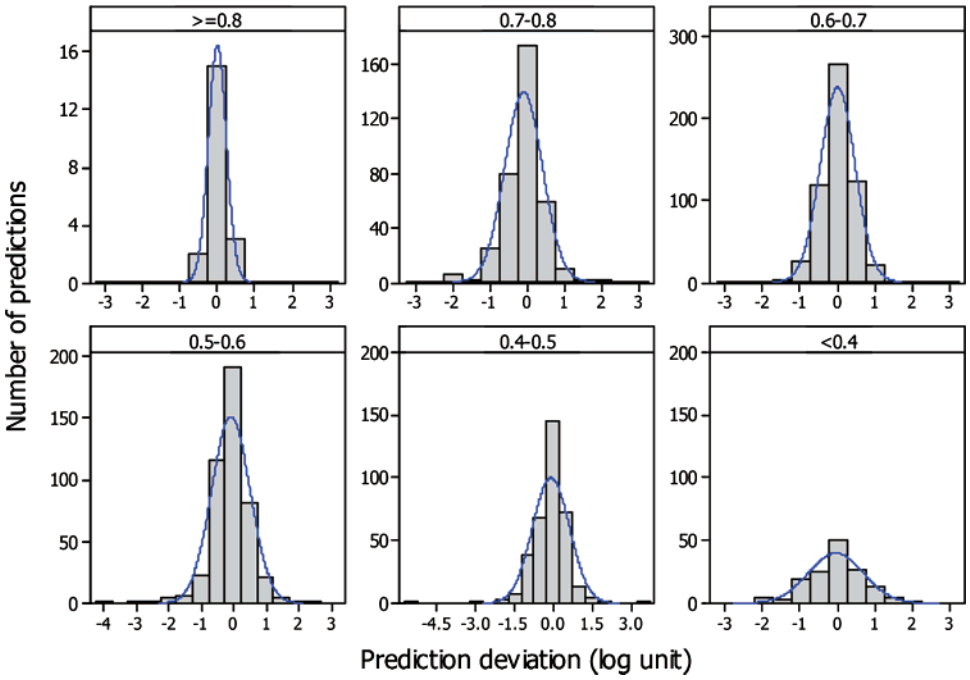

Figure 5. Precision and accuracy increase with greater similarity. Each histogram graph corresponds to a specific average similarity (AS) bin, for example, the first (upper left) graph shows the distribution of prediction deviations with AS ≥ 0.8 bin while the second is for AS bin = 0.7−0.8.

Although we initially explored both descriptor sets, further analysis will focus on models based on the MOE descriptors.

Choosing a Training Set for the On-the-Fly Local Models. The Tanimoto coefficient was used as a measure of molecular similarity for bitwise fingerprints. For a given molecule in a given database, there will be subgroups of molecules, nearest neighbors, that have the same range of Tanimoto coefficients in the solubility database. Figure 4 shows the Tanimoto coefficient for the illustrated compound (2-{2-[(*Z*)-5-nitro-thiazol-2-ylimino]thiazolidin-3-yl}-ethanol) plotted against the number of molecules that are in the top 1000 nearest neighboring subgroups. As similarity declines, the number of nearest neighbors drops first rapidly and then more slowly as the structures become less related. Since molecular structures are not evenly distributed across chemical space, some molecules have more very similar neighbors and some have fewer; however, for the dynamic local model based on similarity, the most similar set of molecules is always used as the training set to construct the model and make the prediction.

On-the-Fly Local Models. A local model was developed with the MLR method for each of the 1900 molecules in the test set using the same descriptor set as the corresponding global models. As discussed above, because of the high efficiency of MOE descriptors, fewer training molecules were sufficient to create a local model and the direct benefit was that the overall similarity of the training set to the target molecule was higher for the MOE descriptor based model. The result was apparent and convincing: for the prediction of 1900 test molecules, the global model with MOE descriptors gave r^2 of 0.74 and the global model with the fragment descriptors 0.7. The local model with the MOE descriptors rendered r^2 of 0.87 and the local model with fragment descriptors 0.82. In both cases, the local model approach delivered a 17% of prediction improvement.

Similarity Impact on Precision and Accuracy. In Figure 4, the Tanimoto coefficients were shown to have a monotonic relationship to the number of molecules that were in subsets with similar coefficients. For the on-the-fly local MOE models, a new parameter, the average similarity, AS, was defined as in eq 1,

$$\mathrm{AS} = \frac{1}{N}\left(\sum_{1}^{N}\mathrm{TC}\right) \qquad (1)$$

where TC is the Tanimoto coefficient and N is the size of training set. We tried to use AS to measure and compare quantitatively the similarities of the training sets. In Figure 5, for each compound in the test set, the AS was calculated for the training set. The 1900 compounds were divided into 6 AS groups: [≥0.8], [0.7, 0.8], [0.6, 0.7], [0.5, 0.6], [0.4, 0.5], [<0.4]. Figure 5 shows the distributions of the absolute prediction deviation (APD) for these six groups. The histograms show that molecular similarity dictates accuracy and precision as measured by the centering and the spread of the diagrams.

For a given test molecule, structurally similar neighbors share a core structure in such a way that their solubilities can be modeled. This is the essence of local modeling in which variations can be handled in an environment of low diversity and reduced prediction error. In addition, average similarity (AS) can be used as a confidence index for the

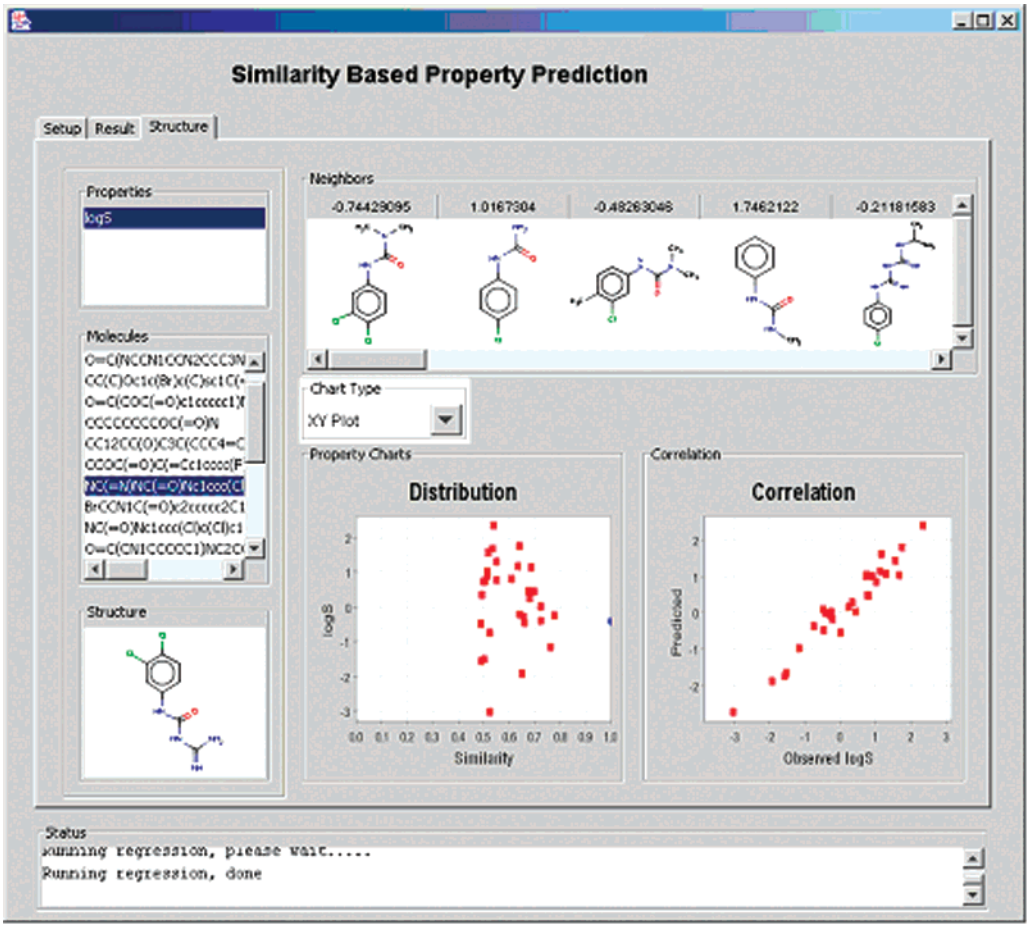

Figure 6. Visual evaluation of the on-the-fly local model's properties.

specific target molecule. The global model, on the other hand, only provides an average level of confidence over the training set.

Simpler and more accurate quantitative structure activity relationships (QSAR) are then possible because the training sets are closer core analogues.[27] This is similar to clustering algorithms[28,29] except for the training set. In contrast, global static models generally employ a large number of descriptors and predictions represent an average response based on global property variations. With the dynamic selection of training sets, the prediction error is small if suitable training molecules are available in the database. For MOE descriptors, this was found to be the case for the aqueous solubility database that was investigated. The average similarity, in turn, can be used as an indicator of the prediction confidence. Although this methodology was applied to aqueous solubility in this study, it should be applicable to many QSAR problems and some ADMET end points. However, because the technique carries out on-the-fly selection of training sets from entire databases, it is not as rapid as pretrained computational routines. A study with a goal similar to that of this study used a technique called local lazy regression.[30]

Visualization of Molecular Similarity. Since each on-the-fly local model was built from a small number of similar training molecules, their experimental property values would be of great interest to the model user. The training set information, readily available from this application, may provide the user further confidence in the prediction through the visual verification of the structures of the training set. The SAR information in the training set can also provide guidance for optimizing the lead compound for desired properties. For each prediction, the training set for the model compound is shown. Structures of the molecules in the training set can be scrolled through to verify the similarity to the target molecule. The plot of log *S* versus similarity for the training set is also shown with the molecules in the training set in red squares and the target molecule in blue (see Figure 6). These plots make it easy for the user to see whether the prediction is a more reliable interpolation or a

(27) Hammett, L. P. *Physical Organic Chemistry*, 2nd ed.; McGraw-Hill: New York, 1970.

(28) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572−584.

(29) Fan Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based studies. *J. Med. Chem*. **2001**, *44*, 3254−3263.

(30) Bontempi, G.; Birattari, M.; Bersini H. Lazy learning for modeling and control design. *Int. J. Control* **1999**, *72*, 643−658.

less reliable extrapolation. The correlation diagram of the experimental property with the predicted property for the training model also allows for a visual assessment of the fitness of the dynamically built local model. Users can ask and evaluate issues such as the following: (1) Was the solubility prediction made with structurally similar training neighbors? (2) How does the training set's solubility vary? (3) How good is the model's fit to the training set?

Because the local on-the-fly models give more information than the preprocessed global model, virtual screening might well be used before chemical synthesis in lead optimization.

Future Work. The current study utilized the Tanimoto coefficient with Daylight fingerprints to select the customized training set for a target molecule from a large dataset and a set of MOE molecular descriptors for the generation of a local on-the-fly model. While this approach clearly delivered a good performance with reasonable computational speed, a set of the same MOE descriptors could be used for the training set selection and model building as was done by others.[11] However, this approach assumes that the same set of descriptors identified from the global model is also appropriate and efficient for selecting the local training set for each target molecule, which might not be always right, as discussed by the authors in the literature. Another potential approach is to select a training set by merging all the compounds identified and apply multiple training set selection strategies to build on-the-fly local models. The closest molecules in the descriptor space could then be combined with the molecules identified by the fingerprints to build models.

Summary

For any modeling effort, quality data is needed. Quality means that data not only is experimentally accurate but also is appropriate for the target molecule. An on-the-fly selection of the training set was developed which enabled the prediction of local QSPR models based on training cohorts that are similar to the target molecule. In this study, the very rapid Daylight fingerprints were used as the basis for similarity. This makes an on-the-fly algorithm practical while enabling the training set to be customized to the target molecule. This has advantages over a prediction that is based on a preprocessed regression model built on the entire dataset. However, further improvements are still possible. If a rapid measure of similarity is used as a screen, then some of the newer feature-rich descriptors like circular fingerprints,[31] COSMOfrag, shape signature, LINGO SMILE substring, and 3D similarity and graphs might be used to further enhance the prediction accuracy. One of the main advantages of a similarity based paradigm is that the prediction can be assessed with respect to a confidence metric based on the similarity of the target molecule to the training set. Predictions that have poor similarity might be disregarded.

MP0700155

(31) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, 9, 199−204.