A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation

Andrew J. Chalk,[†] Bernd Beck,[‡] and Timothy Clark*,[†]

Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Oxford Molecular Ltd., a subsidiary of Pharmacopeia Inc., Computer-Chemie-Centrum, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received September 26, 2000

We present QSPR models for normal boiling points employing a neural network approach and descriptors calculated using semiempirical MO theory (AM1 and PM3). These models are based on a data set of 6000 compounds with widely varying functionality and should therefore be applicable to a diverse range of systems. We include cross-validation by simultaneously training 10 different networks, each with different training and test sets. The predicted boiling point is given by the mean of the 10 results, and the individual error of each compound is related to the standard deviation of these predictions. For our best model we find that the standard deviation of the training error is 16.5 K for 6000 compounds and the correlation coefficient (R^2) between our prediction and experiment is 0.96. We also examine the effect of different conformations and tautomerism on our calculated results. Large deviations between our predictions and experiment can generally be explained by experimental errors or problems with the semiempirical methods.

I. INTRODUCTION

The traditional approach to prediction of physical properties or QSPR (quantitative structure−property relations) has been to use incremental or group additivity methods. Models for such diverse properties as log P[1−3] (log of the octanol/water partition coefficient), melting points,[4] enthalpies of formation,[5−7] aqueous solubility,[8,9] and boiling points[10,11] have been based on such a strategy. Although these methods can produce good results in many systems, there are some potential problems. It may be impossible to calculate values for some classes of molecules because of lack of increments for specific groups, and in some cases partitioning of groups may be ambiguous.[12] The accuracy of these methods can also fall off sharply for unusual molecules.

More recently, approaches employing linear regression[12−17] or neural networks[12,14,18−20] have been used in the prediction of boiling points and other properties. Here we take a neural network approach with descriptors that correspond to electronic and geometrical features of the molecule with simple physical interpretations. These descriptors also have the advantage that they can be calculated for any molecule of interest. Because the 3D structure is used, these methods also offer the possibility of modeling properties that depend on variable structural features such as conformations.

Previous neural network models of boiling points have been either restricted to certain classes of molecules[14,18] or use relatively small training sets.[12,19] By using a large data set of diverse molecules we hope to create a neural net based model, including cross-validation and estimates of error limits for individual species, that will give accurate predictions for a wide variety of molecules. The accuracy of our model is assessed and compared to others and the failures examined in detail. We also investigate the effect of conformational changes on our results.

II. METHODS

A. Semiempirical MO Calculations. Two-dimensional descriptions of molecules in SMILES[21] format were converted to three-dimensional structures using CORINA.[22] This results in a single conformation which was fully optimized using the AM1[23] and PM3[24] Hamiltonians within VAMP 7.5.[25] The standard parameters for the semiempirical methods for all elements were used throughout. Electrostatic potentials were calculated using the natural atomic orbital/point charge (NAO−PC)[26−28] model. All surface-based descriptors are calculated on the solvent-excluded surface (SES) which is determined by an algorithm based on GEPOL[29] and the Marsili marching cube algorithm[30] with a solvent radius of 1.4 Å. We have examined polarizabilities calculated by both the variational method of Rivail and Rinaldi[31] and the parametrized method of Schürer et al.[32]

B. Descriptors. An initial set of descriptors was chosen by first performing a formal inference-based recursive modeling (FIRM)[33] analysis on the descriptors with the experimental boiling point as the lead variable. Descriptors were then removed from consideration if their deletion did not result in any significant decrease in accuracy. New descriptors were also added if their inclusion resulted in significant improvement. The final 18 descriptors that we have chosen, shown in Table 1, can be divided into two classes, electrostatic and structural. The first category includes the descriptors of Murray et al.,[16,34] which provide information on the likely strength of intermolecular interactions. A previous SCF study has shown that there is a relationship between several of these descriptors and boiling points.[16] In an attempt to take more specific interactions, such as hydrogen bonding, into account, we have also included

* Corresponding author phone: ++49 (0)9131 8522948. E-mail: clark@chemie.uni-erlangen.de.
[†] University Erlangen-Nürnberg.
[‡] Oxford Molecular.

10.1021/ci0004614 CCC: $20.00
Published on Web 01/26/2001

Table 1. Descriptor Set Used for Boiling Point Predictions

	property
α	mean polarizability
Descriptors Introduced by Murray et al..[16,34]	
MEP_+	mean positive electrostatic potential
MEP−	mean negative electrostatic potential
V_{max}	max electrostatic potential
V_{min}	min electrostatic potential
σ_+^2	variance of the positive electrostatic potential
σ_-^2	variance of the negative electrostatic potential
σ_{tot}^2	variance of the total electrostatic potential
$\nu\sigma_{tot}^2$	balance × variance
Sum of the Electrostatic Potential Derived Charges on:	
NSUM	nitrogen
OSUM	oxygen
HalSUM	all halogen atoms
nAcc	no. of hydrogen bond acceptor groups
nDon	no. of hydrogen bond donor groups
nAryl	no. of aryl groups
MW	molecular weight
SA	surface area
Glob.	globularity

descriptors such as the number of hydrogen bonding donor and acceptor sites and the sum of charges on nitrogen, oxygen, and halogens. Structural descriptors include molecular weight, surface area, and globularity,[35] which is a measure of the deviation of the solvent-excluded surface from a sphere. A perfectly spherical species such as an atom will have a globularity of 1 while a for long thin molecule, for example a long chain hydrocarbon, the globularity approaches a value of zero in the limit of an infinitely long chain.

C. Training/Test Data. We have taken 6629 experimental boiling points, ranging from 112 to 824 K, and 2D structures in SMILES[21] format from the Registry of Physiochemical Data.[36] This data set contains molecules of very diverse functionality, containing elements H, B, C, N, O, F, Al, Si, P, S, Cl, Zn, Ge, Br, Sn, I, and Hg. We have divided the available 6629 molecules into two sets, the training/test set of 6000 molecules and a validation set of 629 molecules. The latter set was selected in such a manner that it contains molecules spanning the entire range of boiling points examined and is used to test the predictive power of a model on unseen data. There is no information available on the accuracy of these boiling points.

D. Feedforward Neural Nets. Three-layer feedforward neural networks using sigmoid $(1/1 + \exp(-x))$ transfer functions and trained with the back-propagation of errors algorithm[37−39] have been employed in this work. For the sake of brevity, we refer to these simply as neural nets. After examining neural nets with various numbers of hidden nodes, we found that 10 offered the most accurate results, and we therefore employ a 18:10:1 network architecture. Ten separate networks with random starting weights were trained with different training and testing sets, chosen such that each molecule appears in the test set for one and only one network. The cross-validated result for a given molecule is then the prediction of the net in which the molecule does not appear in the training set and should therefore give a worst case prediction for the given model. Training of the nets is halted when the root-mean-square (rms) error of the cross-validated predictions for the entire data set reaches a minimum. This differs from our previous approach,[40] where each network was trained until a minimum in its own rms test set error is reached. The prediction of a model is given by the mean of the results for the 10 nets. Using our method for prediction of individual errors discussed elsewhere,[40] we find that a reliable estimate of the prediction error of an individual molecule can be given by 2.15 times the standard deviation of the results of the 10 neural nets for that species.

Table 2. Comparisons of Predicted and Experimental Results for the AM1 Variational Polarizability Model

	corr coeff (R^2)	std dev	mean unsigned error	largest error[a]
training[b]	0.959	16.54	11.76	−119.4
cross-validation[b]	0.941	19.89	13.82	−171.6
validation[c]	0.947	19.02	12.99	−94.0

[a] Experiment − prediction. [b] $N = 6000$. [c] $N = 629$.

Table 3. Comparisons of Predicted and Experimental Results for the PM3 Model

	corr coeff (R^2)	std dev	mean unsigned error	largest error[a]
training[b]	0.950	18.33	13.00	−107.5
cross-validation[b]	0.931	21.38	14.78	−155.6
validation[c]	0.940	20.27	14.12	−102.0

[a] Experiment − prediction. [b] N = 6000. [c] N = 629.

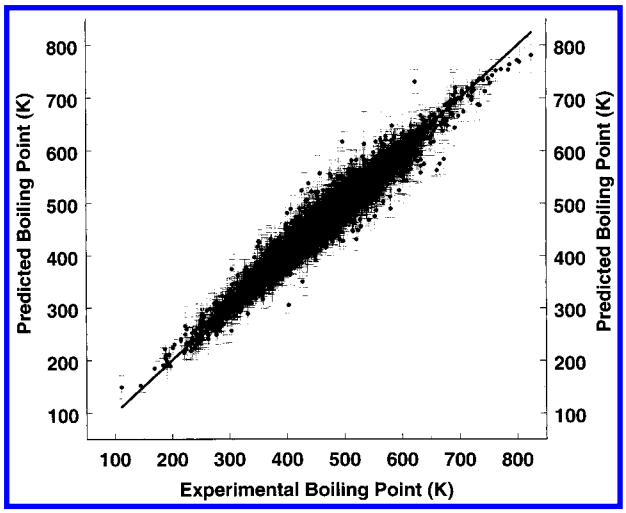


Figure 1. Experimental vs predicted boiling points for the training set ($N = 6000$).

III. RESULTS AND DISCUSSION

A. Performance of the Models. Summaries of the results for the AM1 variational polarizability and PM3 models, all employing a 18:10:1 net architecture and the descriptors discussed above, are shown in Tables 2 and 3. We find little difference between the variational and parametrized polarizabilities so only results for the variational method are reported. The PM3 standard deviations for the training, cross-validation, and validation sets are all around 1.5−2 K higher than the AM1 results. Since the AM1 models are significantly more accurate than PM3, we will restrict the following detailed discussion to the former unless otherwise noted.

Plots of the predicted versus experimental boiling points for the training/test data, cross-validated training/test data, and validation data are shown in Figures 1−3, with error bars are derived as in Beck et al.[40] The training error of 16.5

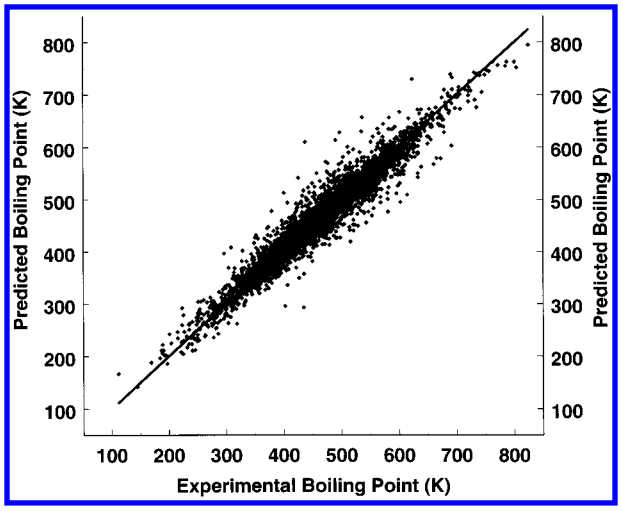


Figure 2. Experimental vs cross-validated predicted boiling points for the training set ($N = 6000$).

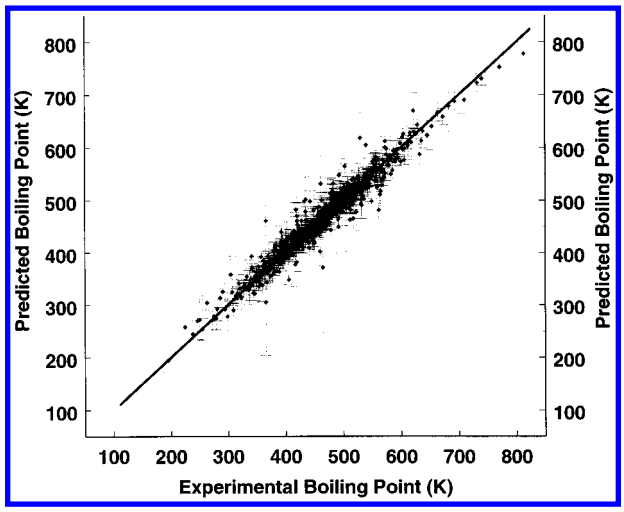


Figure 3. Experimental vs predicted boiling points for the validation set ($N = 629$).

K (Table 2) corresponds to a percentage error of 2.3%, and the standard deviation of the cross-validated error, which represents the worst case performance, is only a few degrees higher than this. The validation set error is also several degrees higher than the training error but is still less than worst case error of the cross-validated result.

For the training set we find that 56.0% of the molecules have errors below 10 K, 84.9% below 20 K, and 93.4% below 30 K. This is in fairly close agreement with the corresponding values for the validation set of 56.9% below 10 K, 79.2% below 20 K, and 90.3% below 30 K. It is also found for the training set that 37% of the compounds are outside their predicted error limits while for the validation set, a value of 38% is observed. This corresponds roughly to the expected error limits for ± 1 standard deviation.[40]

On closer examination of Figure 1, a slight "S" shape is evident with the lowest boiling points being overestimated and the higher end underestimated. This is most likely due to the lack of training data at the extremes of the boiling point range. However, this effect is adequately compensated for by the error bars which become correspondingly larger at these extremes.

The performance of several models trained with the 298 molecule data set from Egolf *et al.*[12] has been examined previously, allowing us to compare the accuracy of our predictions to other models. Our model is actually at a disadvantage in this comparison since, except for the group additivity model, all models were trained using only the Egolf data set and 51 of these molecules were not included in our data set. We have compared our model with the 4 parameter regression model of Katritzky *et al.*,[13] the 8 parameter regression model of Egolf *et al.*,[12] the 19 descriptor neural net model of Hall and Story,[19] and the group additivity model of Joback.[10,16] The results of this comparison are shown in Table 4. Our model performs slightly better than the regression models, despite the disadvantage mentioned above, and much better than the Joback method, which has a very large maximum error. Although these regression models used fewer descriptors than here, the number of training molecules/descriptor for our model is 1 order of magnitude greater. The performance of the neural net model of Hall and Story is significantly better than ours but at the expense of much less economical use of descriptors. In fact, this model performs better than the estimated experimental error for this set of around 9 K.[12] This and the low number of training molecules per descriptor suggest that this model is overtrained and is unlikely to perform well on more diverse data sets such as that examined in this work. The standard deviation and maximum error for the 51 molecules not present in our training set are found to be 10.0 and 33 K, respectively.

Table 4. Comparison of Results for Various Methods with the Data Set of Egolf *et al.*[12] (N = 298)

method	R^2	std dev	largest error	molecules/descriptor
present work	0.980	10.8	33	333.3
Hall and Story[19]	0.995	5.36	−27.4	14.1
Egolf *et al.*[12]	0.975	11.9	−48	33.5
Katritzky *et al.*[13]	0.973	12.4		74.5
Jobak[10,12]	0.943	19.5	−138	

B. Analysis of Large Outliers. In order for a model to be robust, it is important that it be free of extremely large errors. It can be seen in Figures 1 and 3 and in Table 2 that some significant deviations from experiment are produced by this model. In an effort to find possible explanations for these large errors, we have examined in detail the 20 species having the greatest errors from the training set and 15 from the validation set. The predicted and experimental results for these molecules in the training and validation sets are shown in Tables 5 and 6, respectively.

During the training of this model we identified 28 large outliers where the experimental values, stated to be at standard pressure, were in fact measured at reduced pressure. This resulted in a reduced experimental boiling point and hence a prediction that was significantly higher than the incorrect experimental result. Hence, before looking for sources of error in our model, we have sought to confirm the experimental values of the largest outliers by comparing them with those in the Beilstein[41] database. We find that our values for species **7** and **13** (Table 5) were indeed measured at reduced pressure, explaining the apparent error of our prediction. We were also unable to locate further reliable data at standard pressure for species **1**−**3**, **12**, **14**, **15**, and **18**, for which our predictions are higher than experiment, consistent with low-pressure experimental measurements. Although our predictions are below the experi-

Table 5. Largest Errors for the Training Set

	species	expt	prediction	error[a]
1	3,3-dimethylhydrazobenzene	497	616	−119
2	2,6,10,15,19,23-hexamethyltetracosane	623	730	−107
3	1,3,5-triacetyl-1,3,5-triazacyclohexane	438	538	−100
4	3-pentadecanol	661	563	98
5	tris(trimethylsilyl) borate	458	556	−98
6	trimethylaluminum	403	305	98
7	1-(ethylsulfonyl)2-propanone	426	524	−98
8	2-hydroxy-4-methylpyridine	581	489	92
9	*N*-benzylsuccinimide	666	575	91
10	1,3-dioxolan-2-one	521	431	90
11	*N*-phenylsuccinimide	673	584	89
12	bis(trimethylsilyl)ethyne	407	489	−82
13	allyl sulfone	401	481	−80
14	1-acetyl-3-nitrobenzene	475	555	−80
15	phenylhydroxyacetonitrile	443	522	−79
16	2-hydroxypyridine	553	475	78
17	dimethylnitrosamine	427	350	77
18	1,3-diphenylurea	535	612	−77
19	2-pyrrolidone	524	448	76
20	tetramethyltin	351	427	−76

[a] Experiment − prediction.

Table 6. Largest Errors for the Validation Set

	species	expt	prediction	error[a]
21	chlorodiethoxyphosphine oxide	366	461	−95
22	2,4,6-trichloro-1,3,5-triazine	465	371	94
23	2,4,6-triaminophenol	530	618	−88
24	2-hydroxy-3-methylpyridine	562	482	80
25	bis(1,2,2,2-tetrachloroethyl) ether	461	531	−70
26	hexamethyldisilathiane	436	502	−66
27	tetravinyltin	433	498	−65
28	4-acetamidoaniline	540	605	−65
29	*trans*-1,2-bis(trimethylsilyl)ethylene	419	482	−63
30	5-ethoxy-2-hydroxybenzaldehyde	503	565	−62
31	dimethylmercury	366	305	61
32	*N*-nitrodimethylamine	460	402	58
33	3-chlorooxetane	406	348	58
34	3,6-diethyl-2,6-octadien-4-yne	443	498	−55
35	phenylmethanediol diacetate	493	548	−55

[a] Experiment − prediction.

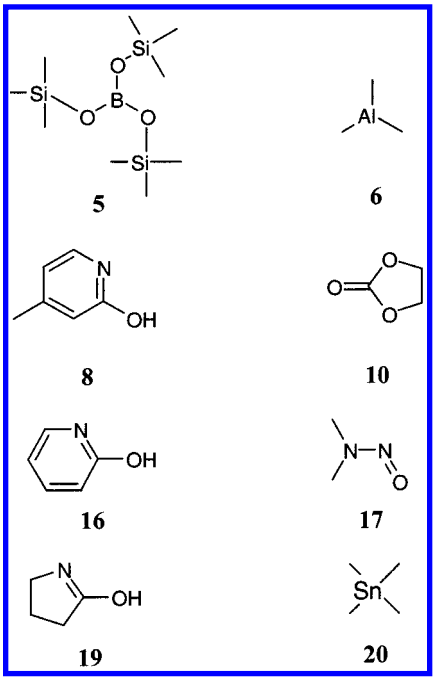


Figure 4. Large outliers from the training set.

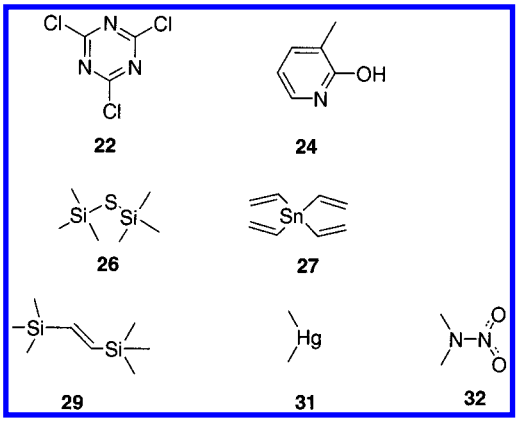


Figure 5. Large outliers from the validation set.

mental values for **4**, **9**, and **11**, we also find no reliable data for these compounds. The questionable accuracy of our database values for these compounds suggests that experiment is at fault in these cases.

The remainder of these 20 molecules, for which good experimental data exist, are shown in Figure 4. One possibility for our model's poor performance with certain cases is that the AM1 method is not correctly describing the electronic structure of the molecules in question. This would be most likely to occur in species containing unusual bonding or elements. The poor predictions of our model for species **5**, **6**, **17**, and **20** are possibly due to such effects. The molecules **8**, **16,** and **19** can exist in different tautomeric forms, and it is therefore possible that the wrong structure is being examined. This is discussed in more detail in section IIIC below. The experimental result for molecule **10** appears to be accurate, and AM1 should perform adequately for this species, so we have no explanation for the failure in this case. We note however that a similar prediction is also observed for the PM3 model.

Our findings regarding the validation set are similar. The experimental values for **21** and **25** (Table 6) are found to correspond to low-pressure measurements and **35** is found to decompose rather than boil at the given temperature. We also find no reliable data for **23**, **28**, and **30**, whose predictions are also consistent with low-pressure measurements. For **33** and **34** we find that Beilstein has two values, the first of which corresponds closely to our database value. When our predictions are compared to the second value, the errors are reduced from −54 to −25 K for **33** and from −55 to −25 K for **34**.

The remaining species from the validation set are shown in Figure 5. Given the structures of **26**, **27**, **29**, **31**, and **32**, a likely explanation for their large errors is that they are described poorly by AM1. Molecule **24** has the possibility of tautomerism and is examined in section IIIC. We have no explanation for the cause of the large error for **22**, but we note that the error bars for this prediction are extremely large (±124 K), placing the experimental result well within the predicted error range. By removal of these 15 suspect molecules from the validation set the standard deviation of the error is reduced from 19.0 to 15.8 K

We have also examined the results for these outliers with the PM3 model. For the species whose large errors we have claimed are due to experimental errors, large deviations are again observed, providing further evidence for this proposal.

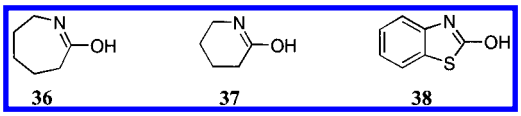
Table 7. Comparison of Boiling Points of Various Tautomers and Their Energy Difference

species	−N=C(OH)−	−NHC(=O)−	expt	ΔE^a
8	489	519	581	0.5
16	475	501	553	0.5
19	447	487	524	−13.2
24	482	519	562	0.2
36	467	510	543	−9.1
37	558	568	633	−11.2
38	456	504	529	−8.8

[a] ΔE = E[−NH−C(=O)−] − E[−N=C(−OH)−] (kcal mol^{-1}).

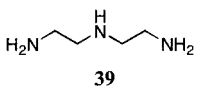
For the molecules for which we suggest AM1 is at fault, significant improvements are generally observed with the PM3 model, although the errors still remain fairly high.

C. Tautomers. As discussed earlier, the possible cause of a number of large prediction errors is that the wrong tautomeric structure was examined. We have investigated the effect of differing tautomers on boiling points for a number of species, the results of which are summarized in Table 7. First, it can be seen that the value for the −NH−C(=O)− tautomers differs significantly from the −N=C(−OH)− tautomers with which the neural network was trained. Furthermore, it is not only found that the −NH−C(=O)− tautomers are closer to the experiment value, but in most cases they are also significantly more stable at the AM1 level. For the cases where they are found to be less stable, the value is sufficiently close to zero that it is not possible to make a definitive statement as to the relative stability of the two structures. It thus appears that the database used does indeed contain incorrect structures. Work is currently underway to develop a method of ensuring that tautomers are treated correctly and consistently.



36 **37** **38**

D. Conformational Effects. Since each molecule in the training data is given as only a single conformation, it is important to examine the possible effects of differing conformations on our predictions. We have therefore performed a conformational search on **39** and examined the predicted boiling points of the resulting optimized conformers. Because of the possibility of hydrogen bonding between the amine groups, differing conformations should result in a significant variation in electronic as well as topological descriptors. This, and the fact that CORINA produces the highest energy gas-phase conformer of this molecule, should mean that any conformational effects observed here should be exaggerated.



39

The predicted boiling points for the unique optimized conformers are shown in Figure 6. Although the conformer produced by CORINA (Figure 6, far right) is high in energy, we nevertheless find that the prediction for this conformer is quite accurate. We believe this is due to the fact that conformers are produced in a consistent (via CORINA) manner. It can also be seen in Figure 6 that, although the individual predictions for the conformers of **39** can differ significantly,

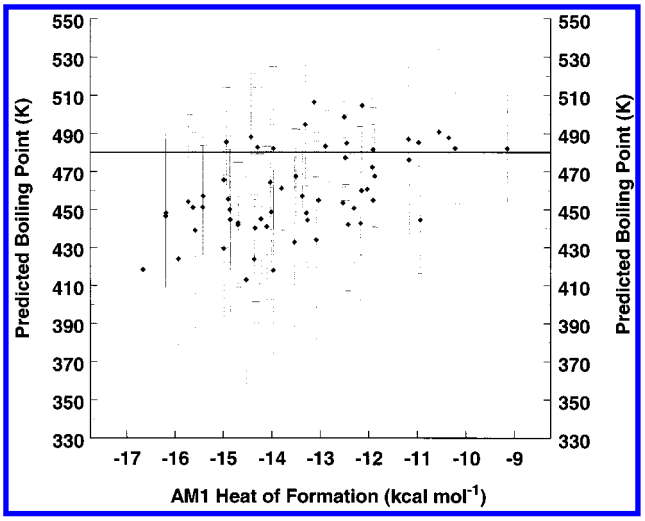


Figure 6. Predicted boiling points vs conformer heat of formation. The experimental value is shown as a horizontal line.

the experimental result of 480 K lies within the error bars for the majority of the results. We have also calculated a Boltzmann average weighted by the heat of formation of each conformer which gives a value of 444 ± 36 K at the temperature of the experimental boiling point. The experimental result is also just within the error bars of this prediction.

IV. CONCLUSIONS

We have developed a model for normal boiling points that is robust and produces good results for a wide range of systems. The largest deviations can generally be attributed to either problems with experimental results or to AM1 poorly describing the electronic structure of the molecule in question. We also have examined the question of conformational dependence and show that, when the error of the predictions is taken into account, the use of a single conformation should not be a major problem.

The fact that many large outliers seem to be caused by incorrect experimental boiling points or structures suggest that our model is sufficiently general that it can recognize these inconsistencies, despite the fact that it was trained with these data. In any study such as this, the quality of the model is limited by the experimental data, but it is obviously impractical to check every value by hand. By focusing only on the data causing the largest deviations, we have been able to identify and eliminate the most serious erroneous experimental results. The large number of training molecules used helps to minimize the effect of erroneous data points.

ACKNOWLEDGMENT

We thank Oxford Molecular for financial support of this work.

REFERENCES AND NOTES

(1) Leo, A. J.; ClogP; Daylight Chemical Information Systems, Irvine, CA, 1991.
(2) Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, D. M. Assessment of Methods Used for Predicting Lipophilicity: Application to Nucleosides and Nucleoside bases. *J. Comput. Chem.* **1993**, *9*, 1019−1026.
(3) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log P Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.

(4) Constantinou, L.; Ganu, R. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 237−244.
(5) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.
(6) Clark, T.; McKervey, M. A. Saturated Hydrocarbons. In *Comprehensive Organic Chemistry*; Barton, D. H. R., Ollis, W. D., Eds.; Pergamon Press: Oxford, U.K., 1979; Vol. 1; p 37.
(7) Cohen, N.; Benson, S. W. In *Chemistry of Akanes and Cycloalkanes*; Patai, S., Rappoport, Z., Eds.; Wiley: Chichester, U.K., 1992; p 215.
(8) Kuhne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubilty of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.
(9) Lipinski, C. A.; Lomardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Water Solubility and Permiability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(10) Joback, K. G. M. S. Dissertation, The Massachusettes Institute of Technology, 1984.
(11) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.* **1994**, *24*, 581−587.
(12) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947−956.
(13) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a test set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(14) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Hetrocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616−625.
(15) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point of Alkanes Based on a Novel Molecular Distance-Edge Vector. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(16) Murray, J. S.; Lane, P.; Brinck, T.; Paulsen, K.; Grice, M. E.; Politzer, P. Relationships of Critical Constants and Boiling Points to Computed Molecular Surface Properties. *J. Phys. Chem.* **1993**, *97*, 9369−9373.
(17) Stanton, D. T.; Egolf, L. M.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *1992*, 306−316.
(18) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *25*, 68−76.
(19) Hall, L. H.; Story, C. T. Boiling Point and Critical Temperature of a Hetrogeneous Data Set: QSAR with Atom Type Electrotopological State Indicies Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004−1014.
(20) Homer, J.; Generalis, S. C.; Robson, J. H. Artificial Neural Networks for the Prediction of Liquid Viscosity, Density, Heat of Vaporization, Boiling-Point and Pitzers Acentric Factor−Part I−Hydrocarbons. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4075−4081.
(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(22) Sadowski, J.; Gasteiger, J. Corina 1.8; Oxford Molecular: Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, U.K.
(23) Dewer, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(24) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Methodol. *J. Comput. Chem.* **1989**, *10*, 209.
(25) Clark, T.; Alex, A.; Beck, B.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Rauhut, G.; Sauer, W.; Steinke, T.; VAMP 7.0; Oxford Molecular Ltd.: Medawar Centre, Oxford Science Park, Standford-on-Thames, Oxford, OX4 4GA, U.K., 1998.
(26) Beck, B.; Clark, T. Some Biological Applications of Semiemprical MO Theory. *Perspect. Drug Discovery Des.***1998**, *9/10/11*, 131−159.
(27) Rauhut, G.; Clark, T. Multicenter Point Charge Model for High Quality Molecular Electrostatic Potentials from AM1 Calculations. *J. Comput. Chem.* **1993**, *14*, 503−509.
(28) Beck, B.; Rauhut, G.; Clark, T. The Natural Atomic Orbital Point Charge Model for PM3: Multiple Moments and Molecular Electrostatic Potentials. *J. Comput. Chem.* **1995**, *15*, 1064−1073.
(29) Pascual-Ahuir, J. L.; Silla, E.; Tuñon, I. GEPOL: An improved Description of Molecular Surfaces III. A New Algorithm for the Computation of a Solvent-Excluded Surface. *J. Comput. Chem.* **1994**, *15*, 1127−1138.
(30) Marsili, M. In *Physical Property Prediction in Organic Chemistry*; Jonchum, C., Hicks, M. G., Sunkel, J., Eds.; Springer: Berlin, Heidelberg, **1988**; p 249.
(31) Rinaldi, D.; Rivail, J. Calculation of Molecular Electronic Polarizabilities. Comparison of Different Methods. *Theor. Chim. Acta* **1974a**, *32*, 243−251.
(32) Schürer, G.; Gedeck, P.; Gottschalk, M.; Clark, T. Accurate Parametrized Variational Calculations of the Molecular Electronic Polarizability by NDDO-Based Methods. *Int. J. Quantum Chem.* **1999**, *75*, 17−31.
(33) Hawkins, D. M. FIRM. http://www.stat.umn.edu/users/FIRM/index.html.
(34) Murray, J. S.; Politzer, P. Statistical Analysis of the Molecular Surface Electrostatic Potential: An Approach to Describing Non-covalent Interaction in Condensed Phases. *J. Mol. Struct (THEOCHEM)* **1998**, *425*, 107−114.
(35) Meyer, A. Y. The Size of Molecules. *Chem. Soc. Rev.* **1986**, *15*, 449−475.
(36) Bavicic, P.; Mackov, M. Registry of PhysioChemical Data, 1.2.2 ed.; Synexchem Consulting Services International, LLC: Alexandria, VA, 1999.
(37) Müller, B.; Reinhardt, J.; Strickland, M. T. *Neural Networks−An Introduction*, 2nd ed.; Springer-Verlag: Berlin, Heidelberg, 1995.
(38) Pao, Y.-H. *Adaptive Pattern Recognition and Neural Networks*; Addison-Wesley Publishing Co.: Reading, MA, 1989.
(39) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Verlag: Weinheim, Germany, 1993.
(40) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and log P. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.
(41) Beilstein. Beilstein Informationsysteme GmbH. <<

CI0004614