Melting Point Prediction Employing *k*-Nearest Neighbor Algorithms and Genetic Parameter Optimization

Florian Nigsch,[†] Andreas Bender,[†,‡] Bernd van Buuren,[§] Jos Tissen,[§] Eduard Nigsch,[||] and John B. O. Mitchell*[,†]

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom, Lead Discovery Informatics, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Ave., Cambridge, Massachusetts 02139, Unilever R&D Vlaardingen, Food and Health Research Institute, Olivier van Noortlaan 120, 3133 AC Vlaardingen, The Netherlands, and Faculty of Mathematics and Geoinformation, Vienna University of Technology, Karlsplatz 13, A-1040 Wien, Austria

Received April 24, 2006

We have applied the *k*-nearest neighbor (*k*NN) modeling technique to the prediction of melting points. A data set of 4119 diverse organic molecules (data set 1) and an additional set of 277 drugs (data set 2) were used to compare performance in different regions of chemical space, and we investigated the influence of the number of nearest neighbors using different types of molecular descriptors. To compute the prediction on the basis of the melting temperatures of the nearest neighbors, we used four different methods (arithmetic and geometric average, inverse distance weighting, and exponential weighting), of which the exponential weighting scheme yielded the best results. We assessed our model via a 25-fold Monte Carlo cross-validation (with approximately 30% of the total data as a test set) and optimized it using a genetic algorithm. Predictions for drugs based on drugs (separate training and test sets each taken from data set 2) were found to be considerably better [root-mean-squared error (RMSE) = 46.3 °C, $r^2 = 0.30$] than those based on nondrugs (prediction of data set 2 based on the training set from data set 1, RMSE = 50.3 °C, $r^2 = 0.20$). The optimized model yields an average RMSE as low as 46.2 °C ($r^2 = 0.49$) for data set 1, and an average RMSE of 42.2 °C ($r^2 = 0.42$) for data set 2. It is shown that the *k*NN method inherently introduces a systematic error in melting point prediction. Much of the remaining error can be attributed to the lack of information about interactions in the liquid state, which are not well-captured by molecular descriptors.

INTRODUCTION

One of the characteristic properties of a chemical compound is its melting point. It is the temperature at which the solid phase of a compound is in equilibrium with its liquid phase at atmospheric pressure. As the melting point is a defined property of any pure substance, it is often used routinely for the determination of identity and purity in synthetic laboratories. Other physicochemical properties are closely related to the melting point of a compound, for example, solubility and boiling point.[1−3] The former is crucial in the pharmaceutical industry for the prediction of the bioavailability of drugs,[4−7] whereas the latter is important for environmental purposes.[8] Recently, interest in the prediction of melting points has been fostered by the growing body of work on ionic liquids.[9−11]

For the prediction of melting points, Karthikeyan et al.[12] used an artificial neural network (ANN), whereas more traditional approaches apply different regression techniques to a multitude of molecular descriptors giving quantitative structure−property relationship (QSPR)-type[13−15] models which relate structural features to a physical property. The *k*-nearest neighbor algorithm (*k*NN), however, has not previously been used for the prediction of melting points. Karthikeyan et al. discuss several modeling techniques developed by different groups. We will only give a brief overview of newer methods and papers not discussed there. A summary of the results obtained with different methods and data sets is given in Table 1.

Yalkowsky and co-workers[13,16−18] have produced a combined model that incorporates the enthalpy of melting[16] (ΔH_m) and the total entropy of melting[17,18] (ΔS_m). Because at the equilibrium of the solid and liquid states the change in Gibbs free energy is zero, the melting point can be expressed as $T_m = \Delta H_m/\Delta S_m$. Two sets of parameters were introduced to describe the entropy of melting: the first pair consisting of molecular flexibility number and molecular symmetry number,[18] the second pair being molecular eccentricity and molecular spirality.[17] The molecular flexibility number is derived from the number of twist angles present in the molecule, whereas the molecular symmetry number corresponds to the number of rotations (but not reflections) yielding an equivalent spatial arrangement as an arbitrarily chosen reference.[19] With only these two parameters, they were able to predict the entropies of melting of a set of 376 environmentally and pharmaceutically important molecules with an average error of 21%. When these were combined with enthalpies of fusion obtained by a group contribution

* Corresponding author phone: +44 (0)1223 762 983; fax: +44 (0)1223 763 076; e-mail: jbom1@cam.ac.uk.
[†] University of Cambridge.
[‡] Novartis Institutes for BioMedical Research Inc.
[§] Unilever R&D Vlaardingen.
[||] Vienna University of Technology.

10.1021/ci060149f CCC: $33.50
Published on Web 09/16/2006

Table 1. Overview of Published Models, Including the Numbers of Molecules They Employ, Methodology, and Model Performance Statistics

reference	method	total compounds	training set	test set	RMSE[a] training set	r^2 training set	AAE[b] training set	RMSE test set	q^2 test set	AAE test set
Bergström[14]	PLS[c]	277	185	92	36.6	0.57		49.8	0.54	
Bergström/Karthikeyan[12]	ANN[d]	2366	2089	277	48.0	0.66	37.6	41.4	0.66	32.6
Karthikeyan	ANN	4173	2089	1042	48.0	0.66	37.6	49.3	0.66	38.2
Modaressi[20]	MLR[e]	323	278	45	40.9	0.66		42.0	0.79	
Jain[13]	MLR	1215	1094	119			33.2			36
Trohalaki[11]	MLR	13	13	LOO		0.91/0.93		5/14	0.78/0.87	
Johnson[17]	MLR	106	106			0.90	43			

[a] Root mean square error (in °C). [b] Absolute average error (in °C). [c] Partial least squares. [d] Artificial neural network. [e] Multiple linear regression.

approach, the authors were able to predict a test set of 120 compounds, based on a training set of 1095 molecules, with an absolute average error of about 36 °C [no root-mean-squared error (RMSE) or r^2 data given].[13] The second set of entropic parameters, eccentricity and spirality, account for entropy of expansion and configurational entropy, respectively. The incorporation of these two quantities as descriptor variables for the prediction of melting entropies, combined with another group contribution approach for the enthalpy of fusion, for 106 compounds resulted in the reduction of the average absolute error in the melting point by 52% (from 90 to 43 °C), with increased correlation ($r^2 = 0.90$).[17]

Trohalaki et al.[10,11] constructed models to predict the melting points of 26 new putative energetic ionic liquids based on a cationic triazolium scaffold with either bromide or nitrate as the anion. The proposed models for the two sets of salts are based on three parameters yielding accurate prediction and high correlation. The RMSEs were 5 and 14 °C and the correlations (in terms of q^2) were 0.78 and 0.87, respectively, for bromides and nitrates.

Modarressi et al.[20] used the Bergström[14] data set as a training set and an additional 45 compounds extracted from MDPI[21] as an external validation set. They used molecular descriptors from different sources (CODESSA,[22] Dragon,[23] and TSAR[24]) and multiple linear regression with stepwise regression and genetic algorithm descriptor selection. Their best model is comparable to the model obtained by Bergström et al. in the first place. With a larger training set (277 instead of 185), Modaressi et al. obtain the following results for a smaller test set (45 instead of 92): the RMSE is 40.9 and 42.0 °C with correlation coefficients of $r^2 = 0.66$ and 0.57, respectively, for the training and test sets. A comparison with the original data from Bergström et al. (RMSEs of 36.6 and 49.8 °C with correlation coefficients of $r^2 = 0.57$ and 0.54, respectively, for the training and test sets) shows that both models are of comparable predictive capability.

Given the size of the data sets at hand, the *k*NN method seemed a valuable approach for the prediction of melting points. In contrast to other methods such as neural networks, the *k*NN method needs no training or parametrization. The present method can be readily applied to a (sufficiently large) set of molecules once their molecular descriptors have been calculated. Moreover, because there is no training step involved, the model is easily extendable for the inclusion of additional molecules.

In the Methods section, we will introduce the model that we have used and its optimization, while the subsequent Results and Discussion sections present the results that we obtained for different data sets and a full interpretation of them. We summarize with our conclusions.

METHODS

1. Data Sets. In the present study, we used two different sets of molecules to build predictive models and to assess their performance and accuracy. We used a diverse set of 4119 compounds (subsequently referred to as data set 1) that has been previously reported in a paper published by Karthikeyan et al.[12] Its compounds range from small unsubstituted hydrocarbons to heavily functionalized heterocyclic structures, with melting points in the range from 14 to 392.5 °C. The second data set used is a set of 277 diverse drugs (subsequently referred to as data set 2), originally compiled by Bergström et al.[14] Both data sets span approximately the same range of melting points, though in different regions of chemical space, and have been prepared in the same manner as previously described by Karthikeyan et al. They are available for download from the ACS Web site[25] as well as from http://www.cheminformatics.org/.

2. Descriptors. All structures (provided as SD or MOL2 files) have been imported into MOE2004.03[26] molecular databases for further processing. To exclude unwanted salts and solvent molecules present in the structures, all compounds have been treated with the "Wash Molecules" option. At the same time, explicit hydrogens were added, formal charges were set, and acids and bases were (de)protonated according to pH 7. These structures were then subjected to a geometry optimization using the MMFF94[27] force field, preserving existing chirality where applicable. Partial charges were subsequently calculated according to the PM3 method, making use of the MOE options "Optimized Geometry" and "Adjust Hydrogens and Lone Pairs as required". The full range of available MOE descriptors (QuaSAR descriptors[28]) was then calculated, totaling 203 descriptor values for every single molecule, out of which 146 are 2D descriptors and the rest are 3D descriptors. A full list of descriptors along with their description can be found in the QuaSAR reference.[28]

3. Initial Model. Our model is based on the "molecular similarity principle",[29] according to which "similar molecules exhibit similar properties".[30] This basic assumption leads directly to the idea that a nearest neighbor model should be applicable to the problem of prediction of physical properties of molecules. The nearest neighbor technique is a classification and clustering method[31−34] used in a variety of areas that range from image processing[35] to statistical analysis.[36,37] It has been applied in previous work in the area of activity prediction[38,39] (quantitative structure−activity relationship, QSAR) and property prediction[40] (QSPR).

The principle of the method demands two prerequisites. First, a space of arbitrary dimensionality r has to be defined

in which every single molecule is to be identified by its r descriptor values. Second, to determine the environment of a certain point, a distance measure is required. In the present work, we used the Euclidean distance in r dimensions, other possibilities that have not been considered being the Manhattan distance (*1*-norm distance) or higher p-norm distances.[41] Therefore, the distance between two molecules i and j in our r dimensional descriptor space is given by

$$d_{ij} = [\sum_{n=1}^{r}(X_n^i - X_n^j)^2]^{1/2}$$

X_n^i being the value of descriptor n for a given compound i. The distance information for a set of molecules can then easily be used to determine the set of nearest neighbors of a given compound. For a given set of molecules (that may further be partitioned into a training and a test set), the distance matrix is calculated, containing all pairwise distances. The dimensionality of the model space is equal to the number of descriptors included, resulting in a different distance matrix for every combination of descriptors. This distance matrix has to be calculated only once for a given set of molecules and descriptors and is used for the determination of the nearest neighbors.

Once the set of nearest neighbors has been determined, a prediction has to be made on the basis of the property values associated with them, in our case, the melting points of the compounds. To compute a prediction from the melting temperatures of the nearest neighbors, we investigated four different averaging techniques: the first consists of an arithmetic average (eq 1), and the second one is a geometric average (eq 2). A third and fourth method incorporate the distance information that is associated with every neighbor, those being an average weighted by the inverse distance (eq 3) and an exponential weighting scheme (eq 4), respectively. T_n and d_n stand for the melting temperature and distance of neighbor n; k is the number of nearest neighbors used in the prediction.

$$T_{\text{pred}}^{\text{a}} = \frac{1}{k}\sum_{n=1}^{k} T_n \tag{1}$$

$$T_{\text{pred}}^{\text{g}} = \prod_{n=1}^{k} T_n^{1/k} \tag{2}$$

$$T_{\text{pred}}^{\text{i}} = \sum_{n=1}^{k} T_n \frac{1}{d_n} \frac{1}{\sum_{n=1}^{k}\frac{1}{d_n}} \tag{3}$$

$$T_{\text{pred}}^{\text{e}} = \sum_{n=1}^{k} T_n \, \mathrm{e}^{-d_n} \frac{1}{\sum_{n=1}^{k} \mathrm{e}^{-d_n}} \tag{4}$$

As will be shown in the Results section, the incorporation of distance information results in better models and also provides some insight into the applicability of the similarity principle.

To assess model performance, the data set was divided randomly into a training and a test set. For data set 1, the size of the test set was set to 1000 out of 4119 molecules (approximately 25% of the total data) and, for data set 2, to 80 out of 277 molecules (approximately 30% of the total data). In order for all predictions to be based on the training set only, the molecules in the test set were excluded from the search of nearest neighbors. We used 25-fold Monte Carlo cross-validation to assess our models; for each data set, 25 random test sets were predicted on the basis of the remainder of the data.

The nearest neighbor program has been implemented in C++. Preprocessing steps (i.e., scaling and PCA) and statistical analyses including figure plotting have been performed with the freely available statistical package R.[42]

4. Optimization of the Initial Model. In the initial model, the values of every descriptor for all molecules are scaled to zero-mean and unit variance, establishing equal importance for every one of the descriptors. As this does not necessarily correspond to reality (e.g., the number of oxygen atoms may not have the same influence on melting points as does the water-accessible surface area), we optimized the initial model. We constructed a genetic algorithm using the Genetic Algorithm Utility Library (GAUL)[43] around the *k*NN program in order to minimize the global root-mean-squared error of prediction (RMSEP) for a given data set. The genetic algorithm applies transformations to the initial data in order to find the distortion of descriptor space that yields the lowest global RMSEP.

The input to the genetic algorithm is a data matrix whose rows contain all descriptor values for every molecule. The bits of each chromosome are made up of pairs of numbers, every pair corresponding to the mapping of one descriptor from the original space into the new one; these two numbers encode an operation and a parameter. To allow the genetic algorithm to transform the values of a given descriptor, we chose the following mathematical operations: power, logarithm (to base e), and multiplication. The choice of parameters for the power operation is limited to values between 0 and 3 and, for the multiplication, between 0 and 30, whereas the logarithm operation requires no parameter. By its chromosome, one individual therefore specifies the different operations that have to be applied to each descriptor column.

This optimization procedure provided suitable distortions of descriptor space that at the same time decreased the global RMSEP of the model and also augmented the correlation with experimental values. An inherent drawback of this method is the stochastic nature of the resulting solution.[44,45] Each run of a genetic algorithm will result in a different set of optimal transformations after the same number of generations. In the context of building purely predictive models for the characterization of a wide range of different molecules, the use of this method may be justified. The influence of different descriptors on the melting point of chemicals can be investigated by multiple linear regression (MLR) methods[20] or other variable selection techniques (e.g., partial least squares, PLS).[14,15] The importance, however, that is given to specific variables is dependent on the algorithm used for their selection (e.g., variable subset regression), on the underlying data set, and also on the set of available descriptors. The extensive sifting through large numbers of candidate models via stepwise regression means that MLR

Table 2. Results for Data Set 1 Showing RMSE and Correlation for Varying Numbers of Nearest Neighbors, Different Types of Descriptors, and Different Methods of Calculating the Predictions[a]

		method											
		A			G			I			E		
NN	descriptor	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2
1	2D	57.9	0.35	0.20									
1	3D	62.3	0.27	0.07									
1	23D	57.6	0.36	0.21									
5	2D	49.0	0.44	0.43	50.1	0.43	0.40	48.2	0.45	0.45	48.6	0.46	0.44
5	3D	53.1	0.34	0.33	54.1	0.34	0.30	52.2	0.36	0.35	52.0	0.37	0.36
5	23D	48.9	0.44	0.43	50.0	0.43	0.40	48.2	0.45	0.45	48.8	0.45	0.43
10	2D	49.0	0.43	0.43	50.5	0.42	0.39	47.9	0.46	0.45	47.5	0.47	0.46
10	3D	52.1	0.36	0.35	53.4	0.35	0.32	51.2	0.38	0.37	50.8	0.39	0.38
10	23D	48.8	0.44	0.43	50.2	0.43	0.40	47.7	0.46	0.46	47.6	0.47	0.46
15	2D	49.4	0.42	0.42	51.1	0.42	0.38	48.2	0.45	0.45	47.2	0.47	0.47
15	3D	52.3	0.35	0.35	53.6	0.34	0.31	51.4	0.37	0.37	50.7	0.39	0.39
15	23D	49.3	0.43	0.42	50.9	0.42	0.38	48.1	0.46	0.45	47.3	0.48	0.47

[a] Numbers shown are averages over 25 runs where 1000 random molecules were predicted from the other 3119. RMSE is given in °C. A, arithmetic average; G, geometric average; I, inverse distance weighting; E, exponential distance weighting.

is prone to learn the idiosyncrasies of the data and yield overfitted, unstable models.[46−48]

5. Figures of Merit. To assess model performance and provide statistically meaningful data, we performed a 25-fold Monte Carlo cross-validation in which we predicted 25 random test sets on the remainder of the data for both data sets. Unless otherwise stated, all results in the following are averages over 25 such runs. These figures of merit have been calculated in the following way:

RMSEP is the root-mean-squared error of the predictions calculated according to

$$\mathrm{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{MP}_i^{\mathrm{obs}} - \mathrm{MP}_i^{\mathrm{pred}})^2}$$

where n is the size of the test set and $\mathrm{MP}^{\mathrm{obs}}$ and $\mathrm{MP}^{\mathrm{pred}}$ are the observed and predicted melting points, respectively. For every experiment, 25 distinct values of RMSEP are obtained.

RMSECV is the aggregated root-mean-squared error of the cross-validation. For an M-fold cross-validation, it is defined as

$$\mathrm{RMSECV} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\mathrm{RMSEP}^2}$$

RMSETR is the root-mean-squared error of the training set, calculated as

$$\mathrm{RMSETR} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\mathrm{MP}_i^{\mathrm{obs}} - \mathrm{MP}_i^{\mathrm{pred}})^2}$$

where m is the size of the training set.

Values for q^2 have been calculated as

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(\mathrm{MP}_i^{\mathrm{obs}} - \mathrm{MP}_i^{\mathrm{pred}})^2}{\sum_{i=1}^{n}(\mathrm{MP}_i^{\mathrm{obs}} - \overline{\mathrm{MP}_{\mathrm{train}}})^2}$$

with a different

$$\overline{\mathrm{MP}_{\mathrm{train}}} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{MP}_i^{\mathrm{obs}}$$

for every single one of the 25 runs; the sizes of the test and training sets are n and m, respectively. The final q^2 for an experiment is reported as the arithmetic average over 25 runs and is a leave-multiple-out q^2 for the test set and a leave-one-out (LOO) q^2 for the training set, where summations run over all compounds in the training set. Negative values for q^2 stemming from poor models have not been reported. The coefficients of determination (r^2) are also reported as arithmetic averages over 25 runs. Unless otherwise stated, RMSE is used synonymously with RMSECV hereafter. The full characterization of the training sets for all different models is contained in the Supporting Information.

RESULTS

We examined the influence of different types of descriptors in conjunction with different numbers of nearest neighbors and the four averaging schemes mentioned above. For the number of nearest neighbors, we limited the maximum to 15, because there was no significant improvement in preliminary experiments with more nearest neighbors (data not shown). The results for data sets 1 and 2 are summarized in Tables 2 and 3, respectively.

The influence of different types of descriptors is already apparent when only the single nearest neighbor is used. For data set 1, performance is almost the same for 2D descriptors alone (RMSE = 57.9 °C, $r^2 = 0.35$) and the combination of 2D and 3D descriptors (RMSE = 57.6 °C, $r^2 = 0.36$), whereas 3D descriptors on their own perform worse in comparison (RMSE = 62.3 °C, $r^2 = 0.27$). The same is true for all other experiments (Tables 2 and 3).

Of the four different methods of calculating the prediction, the exponential weighting method is the only one that is able to produce a lower RMSE with more nearest neighbor information. For data set 1, the best overall performance over 25 runs is achieved by using 15 nearest neighbors, 2D and 3D descriptors, and exponential weighting, resulting in a RMSE of 47.3 °C (with standard deviation 1.1 °C; best individual run gives a RMSEP of 45.0 °C) and a mean r^2 of 0.48 (Table 2). See Figure 1a and b for the typical scatterplots

Table 3. Results for Data Set 2 Showing RMSE and Correlation for Varying Numbers of Nearest Neighbors, Different Types of Descriptors, and Different Methods of Calculating the Predictions[a]

		method											
		A			G			I			E		
NN	descriptor	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2
1	2D	71.2	0.08										
1	3D	74.9	0.03										
1	23D	67.8	0.09										
5	2D	53.4	0.16	0.12	53.6	0.15	0.11	53.4	0.16	0.12	54.0	0.16	0.10
5	3D	60.4	0.03		60.5	0.03		60.3	0.03		60.3	0.04	
5	23D	52.1	0.19	0.16	52.3	0.18	0.15	52.0	0.19	0.16	52.6	0.18	0.14
10	2D	52.1	0.16	0.16	52.2	0.15	0.16	52.0	0.16	0.16	52.1	0.17	0.16
10	3D	58.2	0.04		57.9	0.04		58.0	0.04		57.8	0.04	
10	23D	50.7	0.19	0.20	50.7	0.18	0.20	50.6	0.19	0.20	50.9	0.20	0.20
15	2D	51.7	0.16	0.17	51.8	0.15	0.17	51.6	0.16	0.17	51.6	0.17	0.18
15	3D	58.6	0.03		58.1	0.03		58.3	0.03		57.6	0.04	
15	23D	50.4	0.19	0.21	50.4	0.18	0.21	50.3	0.19	0.21	50.3	0.20	0.21

[a] Numbers shown are averages over 25 runs where 277 druglike molecules are predicted on the basis of 3119 random molecules of data set 1. RMSE is given in °C. A, arithmetic average; G, geometric average; I, inverse distance weighting; E, exponential distance weighting.

of the predictions and a histogram of their residuals, respectively.

These results suggest that 2D descriptors contain most of the essential information, whereas 3D descriptors do not add significantly to overall performance. This is partly due to the additional noise introduced by the conformation dependence of the 3D descriptors. A principal components analysis (PCA) has been performed on data set 1 with two aims in mind: first, the reduction of dimensionality which results in a faster computation of the pairwise Euclidean distances and, second, optimal use of all information contained in the data set and the reduction of noise. Of the 203 linear combinations of original descriptors, the first 27 principal components (PCs), explaining 90% of the total variance, had an eigenvalue above 1. The predictive performance has been assessed for different numbers of PCs (Table 4), up to 30 PCs in steps of 5 (5, 10, 15, 20, 25, and 30) and then with larger gaps (50, 100, 150, and 200) due to the small gain in total variance explained (see Figure 2). As the number of PCs increases, the RMSE decreases progressively; this is what is expected because, with more PCs, more information is available for the predictions (Figure 3). The results obtained when 25 principal components were used with exponential weighting over 15 nearest neighbors are a RMSE of 47.7 °C and a mean correlation of $r^2 = 0.46$.

Using the same training set with which the 1000 random molecules of data set 1 were predicted, we then predicted all molecules of data set 2 (Table 3). The method that performed best was a combination of 2D and 3D descriptors with exponential weighting of the melting temperatures of 15 nearest neighbors (RMSE = 50.3 °C, $r^2 = 0.20$). The assessment of a model to predict drugs on the basis of druglike molecules (the training and a separate test set were both taken from data set 2) yielded a completely different result (Table 5). The lowest RMSE was obtained when 2D and 3D descriptors were used in combination with inverse distance weighting for only 10 nearest neighbors (RMSE = 46.3 °C, $r^2 = 0.30$). The prediction error of this model is lower than the one obtained by Bergström et al.[14] for the same set of compounds (Bergström et al. values: RMSE = 51.7 °C, $q^2 = 0.53$). With a slightly worse correlation with the experimental values, we have an improvement of 11% (5.4 °C) in RMSE. See Figure 4 for a plot of one of the 25 predictions.

To make simpler models that are more interpretable, we used atom counts as descriptors.[49] For every molecule, the occurrences of different atom types are counted; these data are then used as input for the *k*NN program. We used the Sybyl[50] atom types because they contain implicit information about the hybridization state and environment of a given atom. Given the simplicity of the input, these models performed better than expected (Table 6). For data set 1, the best result is obtained when using exponential weighting of 15 neighbors (RMSE = 51.8 °C, $r^2 = 0.36$); the same is true for data set 2 (RMSE = 49.2 °C, $r^2 = 0.20$).

A genetic algorithm was used to optimize the models using 2D descriptors only, and for Sybyl atom counts; in both cases, the exponential weighting scheme has been used with 10 nearest neighbors (Table 7). For data set 1 and 2D descriptors, the improvement is only marginal (RMSE = 46.2 °C, $r^2 = 0.49$), whereas for data set 2, the performance is increased for both RMSE and r^2 (RMSE = 42.2 °C, $r^2 = 0.42$). The optimized model for data set 1 when using Sybyl atom counts yields a RMSE of 49.1 °C and a r^2 of 0.40; for data set 2, the RMSE is 46.7 °C with a r^2 of 0.29.

DISCUSSION

The performance of the nearest neighbor model depends on the descriptor space that distances between molecules are evaluated in. From our results, it is evident that, compared to two-dimensional descriptors, the information contained in their three-dimensional counterparts is less useful in the current context, resulting in different model performances. This result on the information content of molecular descriptors is in agreement with those of other authors.[51,52] For data set 1, the difference in RMSE between 2D and 3D descriptors is 3.5 °C (average over 25 runs, 15 nearest neighbors, exponential weighting, predictions based on data set 1), the best individual runs being 44.9 and 48.7 °C, respectively. For data set 2, the difference in RMSE between 2D and 3D descriptors is 1.9 °C (average over 25 runs, 10 nearest neighbors, inverse distance weighting, predictions based on data set 2), the best individual runs being 35.6 and 40.7 °C, respectively. Even though 3D descriptors perform worse on

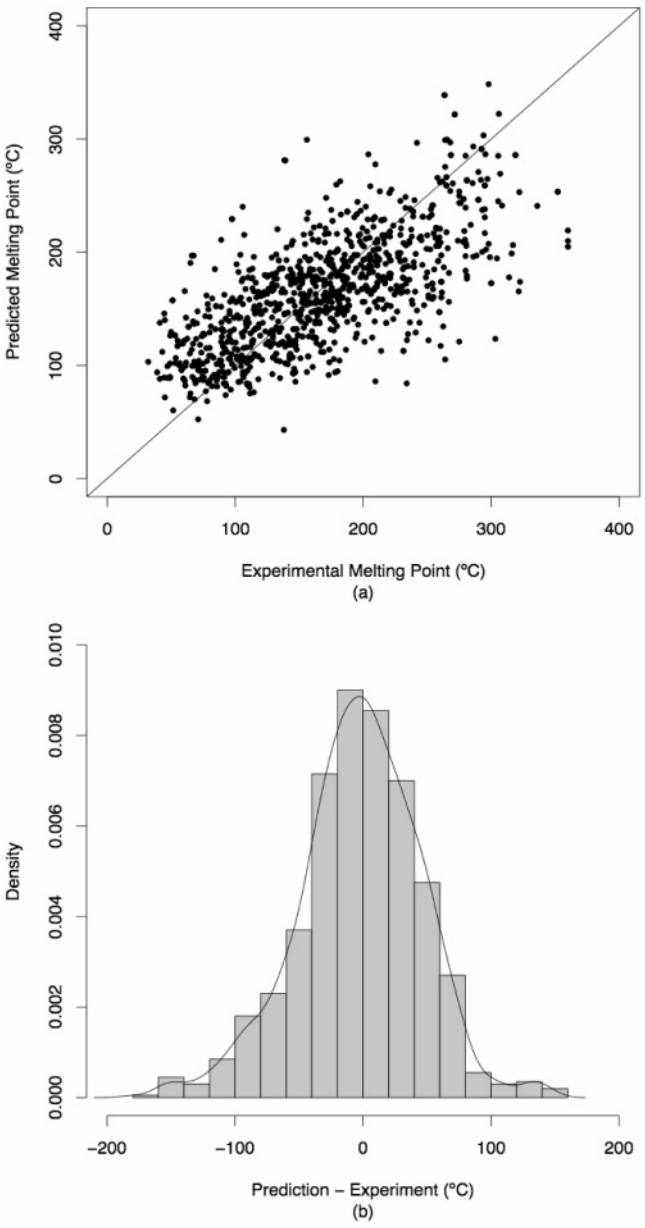

Figure 1. (a) Predictions for 1000 random molecules from data set 1. The predictions are based on the remainder of data set 1 (3119 molecules). (b) Residuals for the predictions of Figure 1a with an overlaid kernel density estimate.

Table 4. Results for Predictions Based on Principal Components[a]

Number of PCs	RMSECV	r^2	q^2	Number of PCs	RMSECV	r^2	q^2
5	54.5	0.30	0.29	30	47.5	0.47	0.46
10	49.1	0.43	0.43	50	47.0	0.50	0.49
15	48.2	0.45	0.45	100	46.8	0.51	0.50
20	48.0	0.46	0.45	150	46.8	0.51	0.50
25	47.7	0.46	0.46	200	46.8	0.51	0.50

[a] Only the results using the exponential weighting scheme for 15 nearest neighbors are given. RMSE is given in °C.

their own, their inclusion in the nearest neighbor search is beneficial to the overall performance.

To make use of all of the information contained in both 2D and 3D descriptors, a reduced model in principal component space confirms that 3D descriptors do indeed contain valuable information and do not merely add noise.

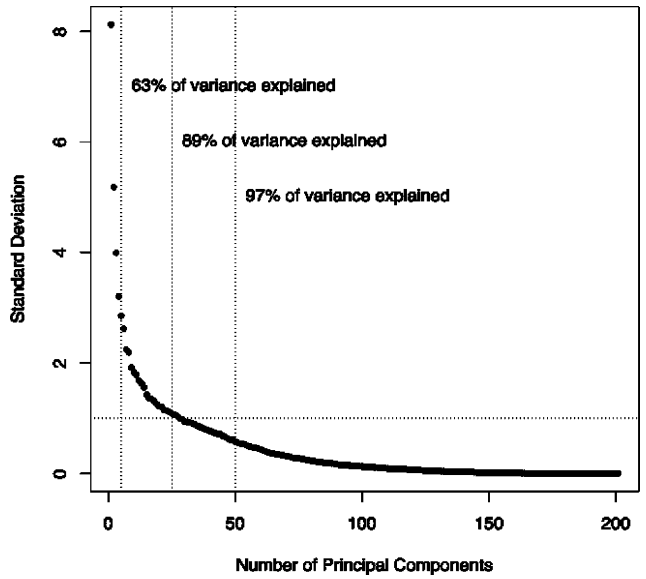

Figure 2. Standard deviations of the principal components analysis of data set 1. The vertical lines are drawn at the following numbers of principal components with the percentage of total variance explained in parentheses: 5 (63%), 25 (89%), and 50 (97%). A total of 27 principal components have a standard deviation above 1.

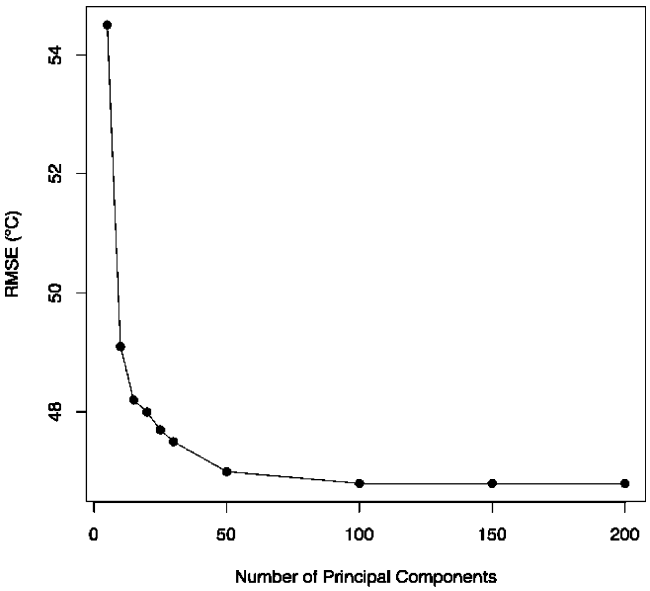

Figure 3. RMSE for a test set of 1000 random molecules (average over 25 runs). Inclusion of the 160 principal components ranked from 21st to 180th in importance improves the average RMSE by only 1.2 °C (2.5%) relative to the model formed by the best 20 principal components.

A model using only the first 25 PCs (accounting for 89% of the total variance) yields an average RMSE of 47.7 °C and performs therefore almost as well as the best model using only 2D descriptors. The essential information present in all descriptors is confined to the reduced principal component space that accounts for the largest part of the variance in the data set in much lower dimensionality, whereas most of the noise is contained in the remaining principal components (see Figures 2 and 3).

A comparison with other models reveals and underlines the reliability of the nearest neighbor approach. Given the size and diversity of the data sets we employed, we will compare our model to the most recent ones dealing with

Table 5. Results for Drugs Predicted by Drugs Showing RMSE and Correlation for Varying Numbers of Nearest Neighbors, Different Types of Descriptors, and Different Methods of Calculating the Predictions[a]

		method											
		A			G			I			E		
NN	descriptor	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2
1	2D	57.7	0.18										
1	3D	59.8	0.16										
1	23D	57.1	0.19										
5	2D	47.7	0.27	0.25	48.9	0.26	0.21	47.1	0.28	0.27	47.8	0.28	0.25
5	3D	51.0	0.19	0.14	51.8	0.19	0.12	50.2	0.21	0.17	49.8	0.23	0.18
5	23D	47.7	0.27	0.25	48.8	0.26	0.22	47.1	0.29	0.27	48.5	0.28	0.22
10	2D	47.1	0.28	0.27	48.3	0.29	0.23	46.5	0.30	0.29	46.7	0.30	0.28
10	3D	49.0	0.22	0.21	50.0	0.22	0.18	48.4	0.24	0.23	47.8	0.26	0.24
10	23D	46.9	0.29	0.28	47.8	0.30	0.25	46.3	0.30	0.29	47.4	0.29	0.26
15	2D	47.8	0.26	0.25	49.1	0.27	0.21	47.1	0.29	0.27	46.5	0.30	0.29
15	3D	49.2	0.21	0.20	50.0	0.22	0.18	48.5	0.24	0.23	47.4	0.27	0.26
15	23D	48.0	0.25	0.24	49.0	0.26	0.21	47.2	0.28	0.27	47.2	0.29	0.27

[a] Numbers shown are averages over 25 runs where 80 random molecules out of 277 are predicted on the basis of the remaining 197 molecules. RMSE is given in °C. A, arithmetic average; G, geometric average; I, inverse distance weighting; E, exponential distance weighting.

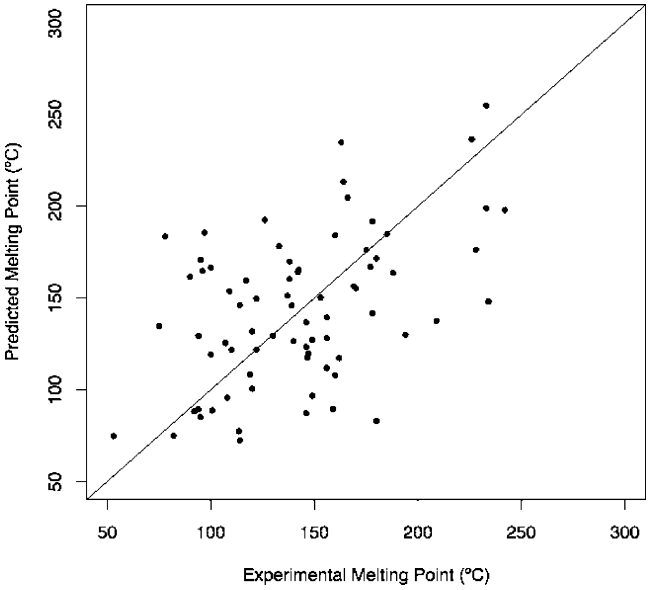

Figure 4. Predictions for 80 random molecules of data set 2. The predictions are based on the remainder of data set 2 (197 molecules).

similar complexity. Bergström et al.[14] used a diverse set of 277 drugs in combination with a PLS method. Their best model achieved a RMSE of 44.6 °C for the test set (no correlation coefficient reported); a model using both 2D and 3D descriptors yields a RMSE of 49.8 °C with a r^2 of 0.54. For the same data set, we achieve with the optimized model an average RMSE as low as 42.2 °C ($r^2 = 0.42$); the best individual run has a RMSE of 35.2 °C. In terms of correlation with the experimental melting point, our model performs slightly worse. It has to be pointed out, however, that in the context of making predictions for unknown compounds our model performs better, which is reflected by a statistically lower value for the average prediction error expressed in terms of the RMSE. Our simplest model employing only Sybyl[50] atom counts predicts a test set of equal size out of the same molecules with a lower RMSE than the PLS model by Bergström et al. (Table 7).

Employing an artificial neural network, Karthikeyan et al.[12] report for the same set of 277 drugs a RMSE as low as 41.4 °C ($r^2 = 0.66$), which is their best result when only using 2D descriptors as input variables. Their model is comparable in RMSE and exceeds the nearest neighbor approach in terms of correlation with the experimental data. They, however, only report values for the prediction of one fixed test set based on a similarly fixed training set, whereas we assessed our model more thoroughly. The process of 25-fold validation using predictions for a random test set based on the remaining molecules gives higher relevance to our reported averages for RMSE and correlation values.

In a direct comparison of the ANN and *k*NN approaches, even the unoptimized version of the latter presents an improvement of almost 4% in RMSE (47.3 vs 49.3 °C), the genetically optimized version yielding an improvement of 6% (46.2 vs 49.3 °C) for the prediction of data set 1. When we compare our best single run to the only reported single run provided by Karthikeyan et al.,[12] we have an improvement of 9% in RMSE (44.9 vs 49.3 °C) with a slightly lower overall correlation. However, whereas the ANN method is able to interpolate within chemical space to predict drugs, the *k*NN method is unable to reliably predict drugs only on the basis of knowledge of nondrugs. The increase in accuracy when the nearest neighbors of druglike molecules are chosen from a set of other druglike molecules is as high as 8% for the averages over 25 runs and 30% for the best individual run. ANNs are nonlinear and global; the totality of available training data is used (via the preceding adjustment of weights in the neural net during training) to make predictions. The *k*NN method, on the other hand, is nonlinear as well, but local instead, simply because only a limited number of nearest neighbors is used. The larger this number, *k*, the less local the model becomes.

The number of parameters adjusted by the genetic algorithm is small in comparison to that used in ANNs. In the approach used by Karthikeyan et al.,[12] the trained fully connected network[53] with the 26−12−1 architecture has 337 parameters.[54] The optimized version of the *k*NN model performs better in terms of RMSE with only 146 parameters. Moreover, the unoptimized *k*NN model already performs as well as the ANN, suggesting that the ANN employed by Karthikeyan et al. may have been overfitted.

Lucic and Trinajstic[55] showed that MLR models outperformed neural networks in nonlinear problems like, for example, QSAR/QSPR[56,57] and the prediction of chemical

Table 6. Results Obtained with Sybyl Atom Counts as Descriptors[a]

		method											
		A			G			I			E		
data set	NN	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2	RMSECV	r^2	q^2
1	1	65.1	0.24										
1	5	53.6	0.33	0.31	54.8	0.33	0.27	52.9	0.35	0.32	52.9	0.35	0.32
1	10	53.3	0.33	0.32	55.0	0.33	0.27	52.0	0.36	0.35	51.9	0.36	0.35
1	15	53.7	0.32	0.30	55.6	0.31	0.25	52.2	0.36	0.34	51.8	0.36	0.35
2	1	62.9	0.10										
2	5	51.4	0.15	0.08	52.8	0.14	0.03	51.1	0.16	0.09	51.6	0.16	0.07
2	10	50.0	0.17	0.13	51.7	0.17	0.08	49.3	0.19	0.16	49.7	0.18	0.14
2	15	50.0	0.17	0.13	52.1	0.17	0.06	49.1	0.20	0.16	49.2	0.20	0.16

[a] Results shown are averages over 25 runs. Training and test sets are taken from the same data set (3119:1000 for data set 1 and 197:80 for data set 2).

Table 7. Results after Optimization by a Genetic Algorithm[a]

					% improvement		
data set	descriptors	RMSECV	r^2	q^2	RMSECV	r^2	q^2
1	2D	46.2	0.49	0.49	2	4	4
2	2D	42.2	0.42	0.40	9	29	28
1	SAC[b]	49.1	0.40	0.39	5	10	10
2	SAC	46.7	0.29	0.27	6	38	48

[a] The numbers shown are averages over 25 runs; for data set 1, 1000 random molecules and, for data set 2, 80 random molecules were predicted on the remaining molecules in the respective data sets. RMSE is given in °C. All predictions use 10 nearest neighbors. [b] Sybyl atom counts.

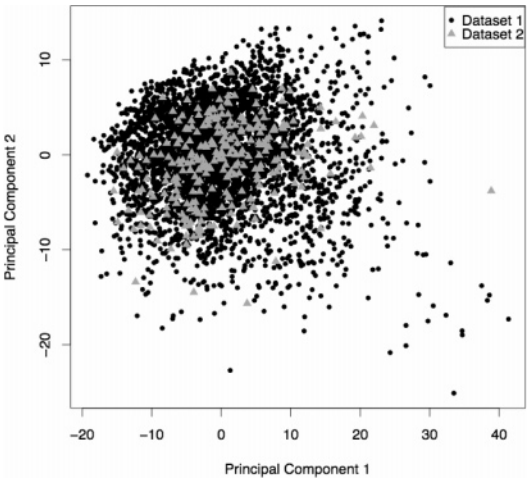

Figure 5. Plot of the first two principal components of the union of data set 1 and data set 2. Circles correspond to data set 1, triangles to data set 2. The more diverse and larger data set 1 covers more area in chemical space than data set 2.

shifts.[56] Whereas neural networks deal with nonlinearity via their hidden layer(s), with MLR techniques a different approach has to be taken. To this end, Lucic and Trinajstic augmented the pool of initial descriptors by their 2- or 3-fold products. Their formally linear models, which however include such products in their descriptor sets, outperform a set of neural networks in several prediction problems, yet have considerably fewer parameters. Our optimization procedure is comparable in the sense that we also apply a mathematical device to transform the underlying data; in our case, this is a distortion of descriptor space.

Visual inspection of a plot of the first two principal components (Figure 5) might suggest that the druglike molecules are well-embedded in the chemical space of nondrugs. This visual approach is, however, misleading, as the problem is much more complex than the two-dimensional projection onto the first two principal components. The results of the *k*NN model also confirm that the distribution of points in chemical space is different for the two data sets, resulting in poor performance when predicting one from the other.

We have shown that the way in which the nearest neighbor information is combined into a prediction for an unknown molecule has an influence. Of the four averaging methods (arithmetic, geometric, weighted by inverse distance, and weights exponentially decreasing with distance), the exponential weighting method provides the best results. In the process of screening virtual libraries, the aim is to find structures similar to a given reference structure. The present prediction problem, where first similar structures are determined (the nearest neighbors) and then a prediction has to be calculated on the basis of this set of similar compounds, is somewhat different in nature. Because the *k*-nearest neighbor method is inherently sensitive to the change in similarity with distance, it should be able to identify to a certain extent how similarity depends on distance.

One of the main problems in the area of the prediction of melting points is the systematic failure[12] of all models to accurately estimate compounds that have either low (below approximately 100 °C) or high (above approximately 250 °C) melting points. For illustration, we calculated for a test set of 1000 random molecules from data set 1 the signed RMSE (see the next paragraph for the definition) for intervals of 30 °C over the whole range of experimental melting temperatures (Figure 6). Low-melting compounds are generally predicted to have too-high a melting point, whereas the opposite is the case for high-melting compounds.

It is clear that the nearest neighbors of molecules located near the extremes of the temperature range (14−395 °C) have melting points shifted toward more moderate values. This is partly because there is a lower density of data points near both ends of the scale than in the middle, so that the majority of a given high-melting molecule's neighbors are likely to have lower melting points, closer to the middle of the overall range. We defined a signed version of the RMSE (eqs 5−7), where e_i is the difference between experimental value x_i and its prediction y_i and sgn is the signum function (i.e., +1 for positive numbers, −1 for negative ones, and 0 otherwise).

$$e_i = x_i - y_i \tag{5}$$

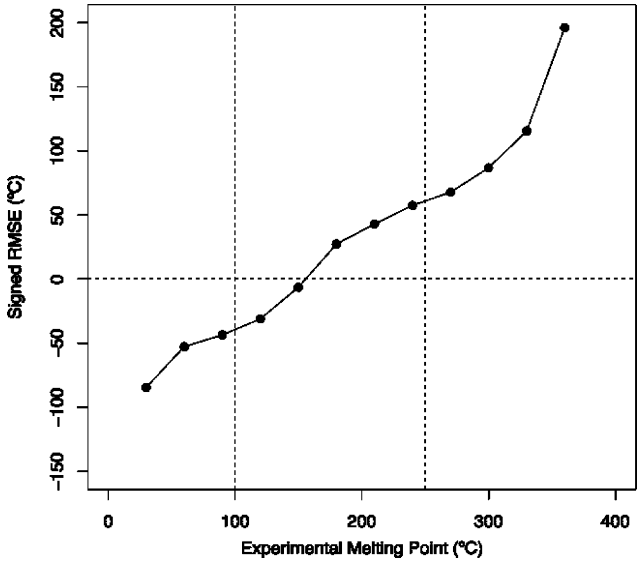

Figure 6. Signed RMSE (experiment minus prediction) for 1000 test molecules of data set 1 in a 30 °C interval around the plotted points. The vertical lines are at 100 and 250 °C. For compounds below ca. 100 °C and above ca. 250 °C, the (absolute) RMSE increases significantly.

The signed error for high-melting compounds based on

$$A = \sum e_i^2 \,\mathrm{sgn}(e_i) \tag{6}$$

$$\mathrm{SRMSE} = \mathrm{sgn}(A)\sqrt{|\frac{A}{N}|} \tag{7}$$

15 neighbors (Figure 6) is typically positive, implying underprediction of the melting point. Similarly, a given low-melting molecule will have more neighbors with melting points higher than its own, resulting in the negative signed error illustrated in Figure 6. Thus, there is a "reversion to the mean" inherent in the *k*NN approach.

We now consider possible variation in the distance of the nearest neighbors in the chosen descriptor space from one side of the temperature scale to the other. A calculation of the mean distance of the nearest neighbors of all molecules of data set 1 showed an increase with experimental melting temperature (Figure 7). The mean distance of the 15 nearest neighbors of low-melting compounds (6.7 au, arbitrary units) is significantly lower (ca. 25%) than that for high-melting compounds (9.2 au); for midrange compounds, the average distance is 7.7 au. This explains to a certain extent why high-melting compounds are associated with a larger error value. For the low-melting compounds, it implies that the typical distance to the neighbors is very similar to that for midrange ones; we believe that it is the uneven distribution of the neighbors' melting points (most being higher, i.e., closer to average) that gives rise to the larger overall error.

Furthermore, nearest neighbors in descriptor space may not be near in terms of melting temperature. An analysis of the number of nearest neighbors further away than a specified temperature cutoff is shown in Figure 8. A nearest neighbor (NN) for which the difference (the melting point of a test molecule minus the melting point of a nearest neighbor) is positive is counted as positive; in the case of a negative difference, it is counted as negative. The (absolute) numbers

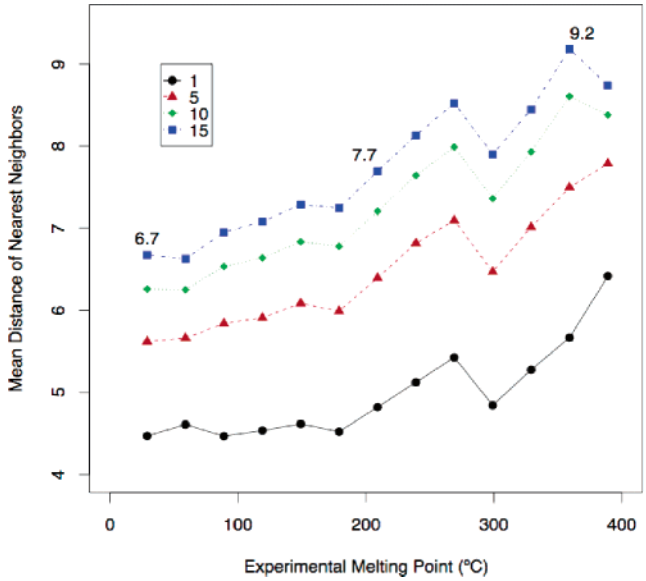

Figure 7. Mean distance (in arbitrary units) of the nearest neighbors, analyzed in 30 °C intervals of the total temperature range covered by data set 1. Analysis is for all 4119 molecules of data set 1.

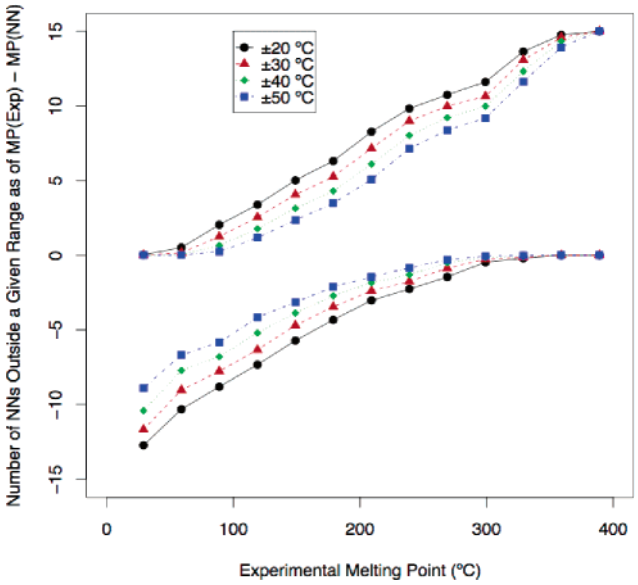

Figure 8. Number of nearest neighbors with an associated melting temperature at least ±20, ±30, ±40, or ±50 °C different from the experimental melting point, analyzed in 30 °C intervals of the total temperature range covered by data set 1. Analysis is for all 4119 molecules of data set 1. The sign of MP(Exp) minus MP(NN) determines if the nearest neighbor is counted as positive or negative.

of positive and negative NNs show complementary behavior; as one increases, the other decreases across the range of experimental melting points. With more negative neighbors, the prediction is more likely to be too high, whereas with more positive neighbors, the prediction is likely to be too low.

It can be seen from Figure 8 that nearest neighbors very distant in temperature are more common on the extremities of the analyzed temperature range, notably more so for high-melting compounds. In the middle of the range, there is almost an equal spread of positive and negative NNs. This helps to explain the larger RMSE on both sides of the spectrum, and also the magnitude of their absolute values

(twice as large for high-melting compounds than for low-melting ones, Figure 6).

A second category of errors stem from the complexity of the liquid phase which results from the melting process. Ouvrard and Mitchell[49] have shown that with very simple models it is possible to predict sublimation energies of molecular crystals from the molecular structure, suggesting that simple QSPR methods can accurately describe properties depending only on the solid and gaseous states. The melting point, notoriously harder to predict,[58] inherently involves not just the solid but also the liquid phase, $T_m = (H_{liquid} - H_{solid})/(S_{liquid} - S_{solid})$. It seems very likely that a significant error results from our insufficient understanding and less satisfactory description of the interactions in the liquid phase.

CONCLUSIONS

We used two chemically distinct groups of molecules (4119 structurally diverse organic molecules and 277 druglike molecules) to build a conceptually simple and reliable method for the prediction of melting points using a nearest neighbor approach. Different methods of combining the information from nearest neighbors have been employed, resulting in valuable insights into the applicability of the "molecular similarity principle" that provides the basis for the *k*NN method in the field of property prediction. The method that performs best in calculating a prediction on the basis of the melting temperatures of the nearest neighbors is a weighted average using an exponential weighting scheme. Both classes of models, the initial one and the genetically optimized one, perform better in terms of RMSE when compared to most other published models. With respect to artificial neural networks, an optimized version of the *k*NN model has equal or better predictive capability with many fewer parameters. The ability of neural networks to extrapolate from nondrugs to drugs may be due to their larger number of parameters. An analysis of the distribution and properties of the nearest neighbors showed that the *k*NN method is prone to a systematic error at the extremes of the range of melting temperatures. This error means that low melting points are typically overestimated and high melting points underestimated. The comparison of models based solely on atom counts with more sophisticated models published earlier shows the need for other descriptors that are able to describe more accurately the interactions in the solid and, especially, in the liquid state.

ACKNOWLEDGMENT

F.N. and J.B.O.M. are grateful to Unilever for their funding. A.B. thanks the Gates Cambridge Trust for financial support.

Supporting Information Available: Excel spreadsheets characterizing all of the different training and test sets for all models, including results for partitions obtained by the Kennard-Stone algorithm.

REFERENCES AND NOTES

(1) Abramowitz, R.; Yalkowsky, S. H. Melting Point, Boiling Point, and Symmetry. *Pharm. Res.* **1990**, *7*, 942−947.
(2) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.
(3) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.
(4) Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME Properties in Silico: Methods and Models. *Drug Discovery Today* **2002**, *7*, S83−S88.
(5) Davis, A. M.; Riley, R. J. Predictive ADMET Studies, the Challenges and the Opportunities. *Curr. Opin. Chem. Biol.* **2004**, *8*, 378−386.
(6) Horter, D.; Dressman, J. B. Influence of Physicochemical Properties on Dissolution of Drugs in the Gastrointestinal Tract. *Adv. Drug Delivery Rev.* **2001**, *46*, 75−87.
(7) Kimura, T.; Higaki, K. Gastrointestinal Transit and Drug Absorption. *Biol. Pharm. Bull.* **2002**, *25*, 149−164.
(8) Dearden, J. C. Quantitative Structure−Property Relationships for Prediction of Boiling Point, Vapor Pressure, and Melting Point. *Environ. Toxicol. Chem.* **2003**, *22*, 1696−1709.
(9) Renner, R. Ionic Liquids: An Industrial Cleanup Solution. *Environ. Sci. Technol.* **2001**, *35*, 410A−413A.
(10) Trohalaki, S.; Pachter, R. Prediction of Melting Points for Ionic Liquids. *QSAR Comb. Sci.* **2005**, *24*, 485−490.
(11) Trohalaki, S.; Pachter, R.; Drake, G. W.; Hawkins, T. Quantitative Structure−Property Relationships for Melting Points and Densities of Ionic Liquids. *Energy Fuels* **2005**, *19*, 279−284.
(12) Karthikeyan, M.; Glen, R. C.; Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *J. Chem. Inf. Model.* **2005**, *45*, 581−590.
(13) Jain, A.; Yang, G.; Yalkowsky, S. H. Estimation of Melting Points of Organic Compounds. *Ind. Eng. Chem. Res.* **2004**, *43*, 7618−7621.
(14) Bergström, 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.
(15) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. Perspective on the Relationship between Melting Points and Chemical Structure. *Cryst. Growth Des.* **2001**, *1*, 261−265.
(16) Zhao, L. W.; Yalkowsky, S. H. A Combined Group Contribution and Molecular Geometry Approach for Predicting Melting Points of Aliphatic Compounds. *Ind. Eng. Chem. Res.* **1999**, *38*, 3581−3584.
(17) Johnson, J. L. H.; Yalkowsky, S. H. Two New Parameters for Predicting the Entropy of Melting: Eccentricity (epsilon) and Spirality (mu). *Ind. Eng. Chem. Res.* **2005**, *44*, 7559−7566.
(18) Dannenfelser, R. M.; Yalkowsky, S. H. Predicting the Total Entropy of Melting: Application to Pharmaceuticals and Environmentally Relevant Compounds. *J. Pharm. Sci.* **1999**, *88*, 722−724.
(19) *IUPAC Compendium of Chemical Terminology*, 2nd ed.; International Union of Pure and Applied Chemistry: Research Triangle Park, NC, 1997.
(20) Modarressi, H.; Dearden, J. C.; Modarress, I. QSPR Correlation of Melting Point for Drug Compounds Based on Different Sources of Molecular Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 930−936.
(21) Molecular Diversity Preservation International (MDPI). http://www.mdpi.org/ (accessed Sep 7, 2006).
(22) *CODESSA*; Semichem, Inc.: Kansas City, MO. http://www.semichem.com/ (accessed Sep 7, 2006).
(23) *Dragon*; Talete srl: Milano, Italy. http://www.talete.mi.it/ (accessed Sep 7, 2006).
(24) *TSAR*; Accelrys Ltd.: Cambridge, United Kingdom. http://www.accelrys.com/ (accessed Sep 7, 2006).
(25) ACS Publications, American Chemical Society. http://pubs.acs.org/ (accessed Sep 7, 2006).
(26) *MOE (Molecular Operating Environment)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada.
(27) Halgren, T. A. Merck Molecular Force Field. 1. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(28) MOE (Molecular Operating Environment), QuaSAR-Descriptor. http://www.chemcomp.com/journal/descr.htm. (accessed Sep 7, 2006).
(29) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(30) Kubinyi, H. Chemical Similarity and Biological Activities. *J. Braz. Chem. Soc.* **2002**, *13*, 717−726.
(31) Asikainen, A.; Kolehmainen, M.; Ruuskanen, J.; Tuppurainen, K. Structure-Based Classification of Active and Inactive Estrogenic Compounds by Decision Tree, LVQ and kNN Methods. *Chemosphere* **2006**, *62*, 658−673.
(32) Pei, T.; Zhu, A. X.; Zhou, C. H.; Li, B. L.; Qin, C. Z. A New Approach to the Nearest-Neighbour Method to Discover Cluster Features in Overlaid Spatial Point Processes. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 153−168.
(33) Chi, M. M.; Bruzzone, L. An Ensemble-Driven k-NN Approach to Ill-Posed Classification Problems. *Pattern Recog. Lett.* **2006**, *27*, 301−307.

(34) Xia, C. Y.; Hsu, W.; Lee, M. L.; Ooi, B. C. BORDER: Efficient Computation of Boundary Points. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 289−303.
(35) Osareh, A.; Shadgar, B.; Markham, R. Comparative Pixel-Level Exudate Recognition in Colour Retinal Images. *Lect. Notes Comput. Sci.* **2005**, *3656*, 894−902.
(36) Sikorska, E.; Gorecki, T.; Khmelinskii, I. V.; Sikorski, M.; De Keukeleire, D. Monitoring Beer during Storage by Fluorescence Spectroscopy. *Food Chem.* **2006**, *96*, 632−639.
(37) Lemm, R.; Vogel, M.; Felber, A.; Thees, O. Applicability of the k-Nearest Neighbours (kNN-) Method to Predict the Productivity of Harvesting − Basic Considerations and First Experiences. *Allg. Forst Jagdztg.* **2005**, *176*, 189−200.
(38) Itskowitz, P.; Tropsha, A. k Nearest Neighbors QSAR Modeling as a Variational Problem: Theory and Applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.
(39) Ajmani, S.; Jadhav, K.; Kulkarni, S. A. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* **2006**, *46*, 24−31.
(40) Neumann, D.; Kohlbacher, O.; Merkwirth, C.; Lengauer, T. Fully Computational Model for Predicting Percutaneous Drug Absorption. *J. Chem. Inf. Model.* **2006**, *46*, 424−429.
(41) Dieudonné; J. A. *Foundations of Modern Analysis*; Academic Press: New York, 1969; pp xviii, 387.
(42) *R: A Language and Environment for Statistical Computing*; R Development Core Team 2005, R Foundation for Statistical Computing: Vienna, Austria. http://www.r-project.org/ (accessed Sep 7, 2006).
(43) GAUL, Genetic Algorithm Utility Library. http://gaul.sourceforge.net/ (accessed Sep 7, 2006).
(44) Elbeltagi, E.; Hegazy, T.; Grierson, D. Comparison among Five Evolutionary-Based Optimization Algorithms. *Adv. Eng. Inf.* **2005**, *19*, 43−53.
(45) Javadi, A. A.; Farmani, R.; Tan, T. P. A Hybrid Intelligent Genetic Algorithm. *Adv. Eng. Inf.* **2005**, *19*, 255−262.
(46) Baumann, K. Cross-Validation as the Objective Function for Variable-Selection Techniques. *Trends Anal. Chem.* **2003**, *22*, 395−406.
(47) Baumann, K.; Stiefl, N. Validation Tools for Variable Subset Regression. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 549−562.
(48) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(49) Ouvrard, C.; Mitchell, J. B. O. Can We Predict Lattice Energy from Molecular Structure? *Acta Crystallogr., Sect. B* **2003**, *59*, 676−685.
(50) *Sybyl 6.9*; Tripos Inc.: St. Louis, Missouri.
(51) 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.
(52) Bender, A.; Glen, R. C. Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.
(53) MacKay; D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, U. K., 2003; pp xii, 628.
(54) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(55) Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.
(56) Lucic, B.; Amic, D.; Trinajstic, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks on Three QSPR Data Sets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403−413.
(57) Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Toward Generating Simpler QSAR Models: Nonlinear Multivariate Regression versus Several Neural Network Ensembles and Some Related Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.
(58) Delaney, J. S. Predicting Aqueous Solubility from Structure. *Drug Discovery Today* **2005**, *10*, 289−295.

CI060149F