# General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks

M. Karthikeyan

Information Division, National Chemical Laboratory, Pune - 411 008, India

Robert C. Glen and Andreas Bender*

Unilever Centre for Molecular Informatics, Chemistry Department, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Received January 12, 2005

We report the development of a robust and general model for the prediction of melting points. It is based on a diverse data set of 4173 compounds and employs a large number of 2D and 3D descriptors to capture molecular physicochemical and other graph-based properties. Dimensionality reduction is performed by principal component analysis, while a fully connected feed-forward back-propagation artificial neural network is employed for model generation. The melting point is a fundamental physicochemical property of a molecule that is controlled by both single-molecule properties and intermolecular interactions due to packing in the solid state. Thus, it is difficult to predict, and previously only melting point models for clearly defined and smaller compound sets have been developed. Here we derive the first general model that covers a comparatively large and relevant part of organic chemical space. The final model is based on 2D descriptors, which are found to contain more relevant information than the 3D descriptors calculated. Internal random validation of the model achieves a correlation coefficient of $R^2 = 0.661$ with an average absolute error of 37.6 °C. The model is internally consistent with a correlation coefficient of the test set of $Q^2 = 0.658$ (average absolute error 38.2 °C) and a correlation coefficient of the internal validation set of $Q^2 = 0.645$ (average absolute error 39.8 °C). Additional validation was performed on an external drug data set consisting of 277 compounds. On this external data set a correlation coefficient of $Q^2 = 0.662$ (average absolute error 32.6 °C) was achieved, showing ability of the model to generalize. Compared to an earlier model for the prediction of melting points of druglike compounds our model exhibits slightly improved performance, despite the much larger chemical space covered. The remaining model error is due to molecular properties that are not captured using single-molecule based descriptors, namely both inter- and intramolecular interactions and crystal packing, for which examples of and reasons for outliers are given.

## 1. INTRODUCTION

The melting point of a compound, which denotes its transition from solid to liquid state and vice versa, is one of its fundamental characteristics. It can be used e.g. for rapid determination of purity and identity of a substance in organic synthesis and also for the prediction of other properties such as boiling point or water solubility.[1,2] Since in particular water solubility is important in the pharmaceutical industry (for example as a determinant of intestinal absorption) the reliable prediction of melting points is, apart from the purely scientific point of view, also hugely useful in practice.

We can make an assumption that melting points of different compounds can be estimated based on the "molecular similarity principle" which states that structurally similar molecules tend to have similar properties.[3,4] Nonetheless, mainly due to ignorance of the solid-state properties, predictions of melting points are generally more difficult than the prediction of related properties such as boiling points.[5,6] Polymorphism, the formation of different crystal habits of otherwise identical compounds, as well as phase transitions within the solid state, such as into liquid crystalline states, show that one compound is not necessarily associated with only one, clearly defined melting point. Every anomaly in e.g. heat capacity—or a different property—can give rise to a different state of the system. Transitions into the liquid−crystalline state are indicated if part of the translational symmetry of the crystal is lost and relatively fast rotations around one axis of the aligned molecules occur. An estimated 5% of organic compounds pass through a liquid−crystalline state before assuming a liquid state at rising temperature,[7] increasing the complexity of the problem.

In general the melting point of a substance depends on single-molecule properties and the spatial arrangement of molecules in the solid state (their packing), both of which influence the strength of intermolecular interactions.[1,8] For organic molecules hydrogen-bond donors and acceptors are considered to be the dominating intermolecular forces.[1] The larger the molecules, the more important induced dipole interactions become since electrons in larger molecules are generally easier to polarize.

* Corresponding author phone: +44 (1223) 763 073; fax: +44 (1223) 763 076; e-mail: andreas.bender@cantab.net.

10.1021/ci0500132 CCC: $30.25 
Published on Web 05/03/2005

The influence of different molecular parameters on melting points has already been subject to earlier research.[9,10] For an isotopological data set comprising simple molecules (which were monatomic rare gases, diatomic, trigonal, pyramidal, and tetrahedral structures) it was found that polarizability and dipole moment were capable of explaining variations in melting points and boiling points.[9] This is intuitively plausible since polarizability and dipole moments correspond to induced and static partial spatial charges which in turn influence intramolecular interactions that are important determinants of the melting point.

Bergstrom et al.[10] investigated the influence of molecular properties (descriptors) on melting points for a comparably large (277 compounds) and diverse drug data set. They attempted both to determine the relative importance of descriptors for predicting melting points and to develop a classification as well as a correlation model. A total of 612 (121 after elimination of skewed variables) 2D and 3D descriptors were employed in combination with PLS (partial least squares) for this purpose. It was found that hydrophilic as well as polar surface areas increase melting point, while nonpolar surface area reduces it. This could be explained by more stable crystal structures via intermolecular interactions which is in agreement with the findings by Charton.[9] In addition, ring structures and rigidity were found to increase the melting point, whereas structures dominated by chains and generally more flexible structures tended to have lower melting points.

Earlier work on melting point prediction focused on very restricted classes of compounds, e.g. on oligoribonucleotides[11] in combination with a neural network or alkanes containing between 10 and 20 carbon atoms in combination with topological indices and nonlinear regression.[12] A "Group Vector Space" method for the prediction of melting points was applied to the prediction of boiling points of hydrocarbons.[13] For a series of 42 anilines an equation containing five parameters—hydrogen-bond donor ability, hydrophobic substituent constants, molar refractivity, the Sterimol parameters B2, and an indicator variable for meta-substitution—achieved good correlation.[14] For 1,2,3-diazaborines charge and size/weight were confirmed as determinants of the melting point, along with the sum of different Randic connectivity indices.[15] Using experimentally determined solubility data it was possible to estimate melting points reliably for a small (81 compounds) but relatively diverse data set, comprising mainly (partly atypical) drug compounds such as steroids.[16] Predictive models were also developed for the melting points of different classes of ionic liquids[17,18] and substituted benzenes.[19] Among the more general methods are the ones by Jain and Yalkowsky[5] and Zhao and Yalkowsky.[20] The former[5] predicted, among other properties, melting points for 405 organic compounds but only took into account "rigid, non-hydrogen-bonding aromatic" structures which does not include most of today's organic compounds and drugs. The latter[20] predicted melting points for 1040 aliphatic compounds with a variety of functional groups, but it in turn excluded aromatic entities from the model which are present in many of today's chemicals and drugs.

In this work we developed an artificial neural network (ANN) based melting point model that differs from previous approaches in three ways.

Firstly, a large and diverse data set of 4173 compounds was used which includes compounds ranging from small unsubstituted hydrocarbons to heavily functionalized heterocyclic structures, examples of which are given in Figure 1. To our knowledge, this is the most diverse data set employed yet for melting point prediction, which covers a large region of (relevant) chemical space.

Secondly, feature selection was performed via principal component analysis and selection of only those components with the largest eigenvalues. This step reduces dimensionality of the problem while retaining the information content of the descriptors. Interpretability of the model was not feasible anyway due to high correlations between many of the descriptors employed, as it is often the case if a very high number of initial descriptors are used.

Thirdly, an artificial neural network was used for model building. A neural network was favored over linear techniques to allow for the modeling of nonlinear relationships between descriptor space and output variable. Again, linear regression using a large number of (correlated) variables can be misleading if interpreted; in addition poorly understood theoretical models of melting do not provide sufficient support for the assumption of linear descriptor-property relationships.

Selection of training, test, and validation sets are crucial in the construction of an internally consistent model that is in addition able to generalize to previously unseen compounds. Leave-one-out (LOO) cross validation is generally considered not to be sufficient for model validation; in addition to leave-multiple-out protocols the employment of an external validation set is recommended.[21] Thus we employ a two-step validation protocol. The model is first validated internally on the data set comprising 4173 compounds (the "principal data set") by using separate training ($n = 2087$), test ($n = 1043$), and validation sets ($n = 1043$). In addition, an external drug data set of 277 compounds is used to gauge generalizability on an important subclass of compounds (drugs), which only represents part of the principal data set. The compounds of the external validation set were not used at any stage of the model building process.

In the following section 2, we describe the data set, the calculation, and selection of descriptors as well as the computational methods employed. In section 3 we present our results which are discussed fully in section 4. The final section 5 concludes our work.

## 2. MATERIALS AND METHODS

**(a) Data Sets.** 4173 structures with melting points were extracted from the Molecular Diversity Preservation International (MDPI) database[22] possessing melting points in the range between 14 °C and 392.5 °C. Apart from the exclusion of purely inorganic substances it was attempted to collect a data set as diverse as possible in order to obtain a broadly applicable model. Structures with melting point ranges of larger than 5 °C were excluded from the MDPI database as well as structures containing heavy metals (Sn, Se) and those sublimating or decomposing.

Global molecular properties of compounds in the data set vary within large ranges: molecular weight of the compounds used was between $MW_{min} = 84.1$ g/mol and $MW_{max} = 815.6$ g/mol, the number of heavy atoms between 6 and 59, and

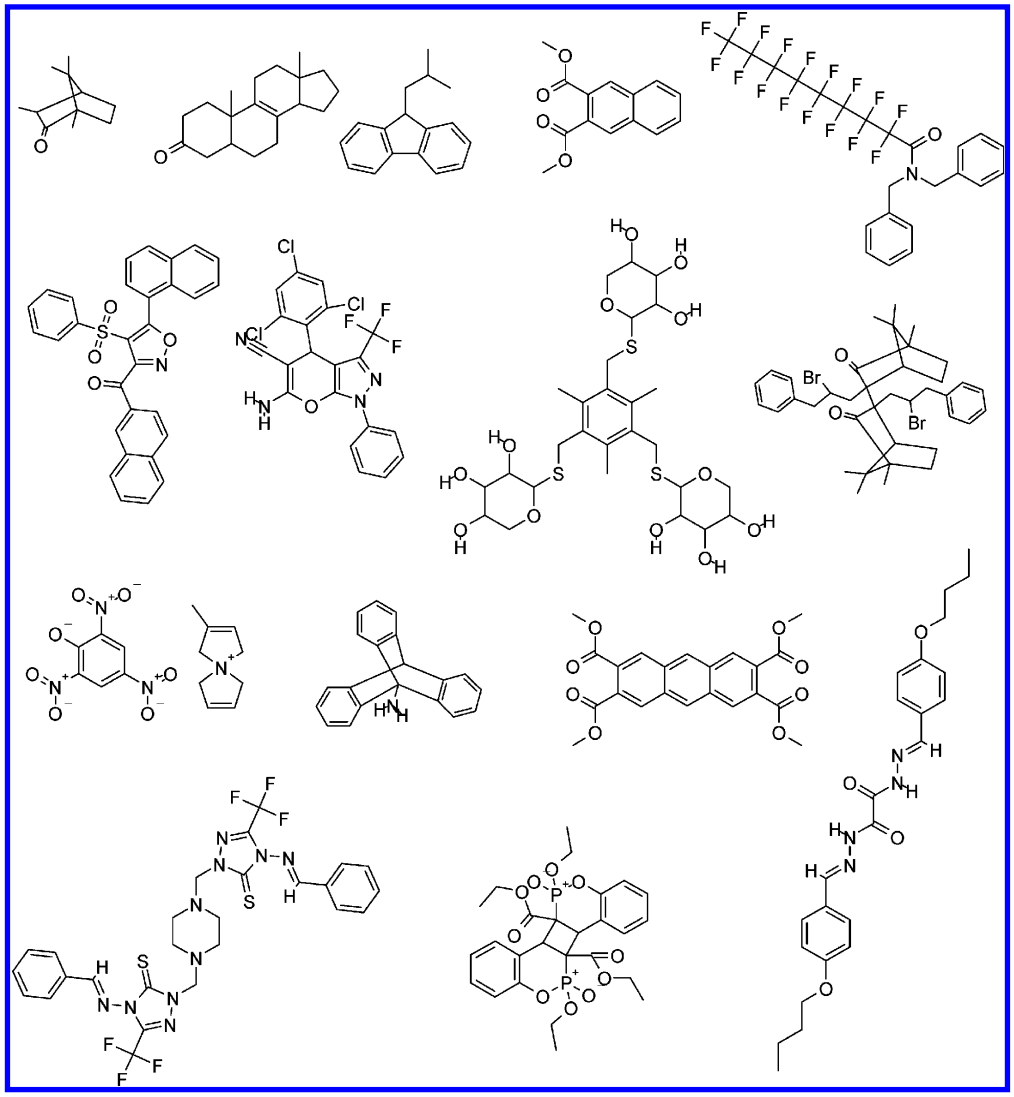


**Figure 1.** Fifteen sample training set compounds, reflecting diversity of the data set used with respect to size and chemical features present in the structures.

**Table 1.** Characterization of the 4173 Compound Principal Data Set, Comprising Structures of Very Different Properties Which Capture a Large Area of Relevant Chemical Space

| | melting point/°C | molecular weight/g/mol | number of heavy atoms | SlogP | molar refractivity/cm$^3$ | dipole moment (AM1)/Debye |
|---|---|---|---|---|---|---|
| minimum | 14 | 84.1 | 6 | −6.0 | 2.0 | 0.005 |
| maximum | 392.5 | 815.6 | 59 | 12.8 | 19.4 | 250.8 |
| mean | 166.3 | 318.7 | 22.3 | 3.4 | 8.5 | 4.8 |
| standard deviation | 64.6 | 105.0 | 7.2 | 1.9 | 2.7 | 9.4 |
| 1. quartile | 118 | 243.3 | 17 | 2.2 | 6.6 | 2.4 |
| 2. quartile | 163 | 309.8 | 22 | 3.3 | 8.3 | 3.8 |
| 3. quartile | 210.5 | 376.5 | 27 | 4.6 | 10.2 | 5.5 |

the (calculated) molar refractivity between 2.0 cm$^3$ and 19.4 cm$^3$. More detailed information about the data set is given in Table 1, also including information on SlogP[23] (which uses an atom contribution method for the calculation of logP), and the molecular dipole moment using the AM1 Hamiltonian.[24] This data set was used for internal validation of the model, and it is publicly available including structures, melting points, calculated properties, and principal component analysis factor scores (see Supporting Information).

For external validation, 277 drugs with melting points extracted from the Merck index compiled by Bergstrom et al. were used[10] that were not included in the principal data set. For this external data set, the influence of molecular parameters on the melting points has been established[10] based on PLS analysis[25] of 121 (left after exclusion of variables too skewed from a set of 612 variables) descriptors calculated by Molconn-Z[26] and the AstraZeneca in-house-software Selma, employing either 2D, 3D, or both 2D and 3D descriptors. This data set is used as an external validation set here. Correlation coefficients and 7-fold cross-validated correlation coefficients obtained by Bergstrom et al.[10] as well as root-mean-square errors (RMSE) are given in Table 2.

**(b) Descriptors.** Structures were obtained in SD format from the MDPI database.[22] 3D-structures were generated using Concord.[27,28] All subsequent steps were performed in MOE2004.03.[29] Salts were removed using the "Wash Mol-

**Table 2.** Correlation Coefficients and Root Mean Square Errors of the External Validation Set, as Obtained from 612 (121 after Exclusion of Skewed Variables) 2D and 3D Descriptors and Partial Least Squares Analysis by Bergstrom et al.[10]

| model | $R^2$ | $Q^2$ | RMSE train/°C | RMSE test/°C |
|---|---|---|---|---|
| 2D | 0.56 | 0.53 | 36.9 | 51.7 |
| 3D | 0.31 | 0.30 | 46.1 | 50.3 |
| 2D/3D | 0.57 | 0.54 | 36.6 | 49.8 |

ecules" option, checking the options "Add Explicit Hydrogen Atoms", "Set Atom Ionization to Formal Charge", and "Deprotonate Acids and Protonate Bases". Structures were minimized using the MMFF94 force field,[30] preserving existing chirality. Partial charges were calculated for all structures using the PM3 Hamiltonian, checking the options "Optimized Geometry" and "Adjust Hydrogens and Lone Pairs as required". In a subsequent step, all 203 2D and alignment-independent 3D descriptors from MOE were calculated from the optimized structures. The 2D descriptors include physical properties (such as charge, van der Waals volume, and molecular refractivity), subdivided surface areas (atomic contributions to logP and molecular refractivity), counts of elemental atom types and of bond types, Kier/Hall connectivity and kappa shape indices, topological indices (Wiener index and Balaban index), pharmacophore feature counts (number of acidic and basic groups and hydrogen bond donors and acceptors), and partial charge descriptors. The conformation-independent 3D descriptors include potential energy terms (such as total potential energy and contributions of angle bend, electrostatic, out-of-plane, solvation, etc. terms) and surface area, volume, and shape descriptors (among them water accessible surface area, mass density, and principal moments of inertia).

**(c) Model Construction.** Models were constructed separately for 2D descriptors only, for 3D descriptors only, and for the whole descriptor space spanned by 2D and 3D descriptors. To reduce dimensionality of the feature space, principal component analysis was performed using the statistical software R.[31] All structures were projected onto principal components space, and the first 30 principal components were used for further analysis. Feed-forward back-propagation artificial neural networks were built using the Neural Network toolbox of Statistica 6[32] following the general procedure below. (The final optimal model was different in every descriptor space, and it is given in the Results section.)

First, model building employing the data set comprising 4173 compounds was performed using separate training ($n = 2087$), test ($n = 1043$) and validation sets ($n = 1043$) with randomly chosen compounds. The classification error was defined by entropy with a linear regression output function. Nodes were initialized by a Gaussian distribution with a mean of zero and a standard deviation of one. The learning rate was fixed at 0.01 and the momentum at 0.3. Presentation order of cases was shuffled each epoch with no weight decay regularization. Input nodes with fan-out weights below 0.05 were subject to pruning. Back-propagation[33] training was performed for 100 iterations, followed by conjugate gradient optimization of weights until no improvement prediction on the test data set occurred. The final network architecture and training parameters differed between the 2D, 3D, and combined 2D/3D data sets (see Results section for details). Network performance was measured on training set, test set, and internal validation set.

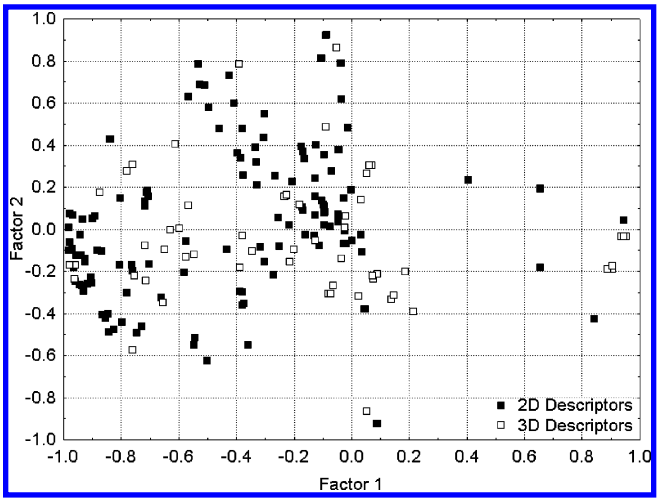



**Figure 2.** Factor scores of the variables (descriptors), showing high correlation of many of the descriptors used in this study (only partly visible in two dimensions). High correlations led us to choose PCA to decrease dimensionality of descriptor space.

Second, the network constructed in the first step was used to predict melting points of the external drug validation set of 277 compounds. Correlation coefficient as well as average absolute errors and RMSE values were determined for each descriptor space used.

## 3. RESULTS

**(a) Analysis of Chemical Space.** Virtually all correlations between descriptor variables are significant at $p < 0.05$ which made us believe that principal component analysis is indicated in order to decorrelate variables. Variable factor loadings obtained by PCA are shown in Figure 2, distinguished between 2D and 3D descriptors. In two dimensions 2D and 3D descriptors still seem to convey slightly different information, and associated highest scoring contributing are variables given in Table 3.

Component 1 of the full descriptor set PCA explains 32.61% of the variance and it is defined by the size of the molecule, positive partial charge, and polarizability, on one hand, and the total molecular and electronic energy of the system on the other (descriptors belonging to these properties possess factor loadings > 0.70). Total molecular and electronic energy are closely related to the size of the molecule but point in opposite directions due to a negative sign. Component 2 explains 12.81% of the variance, and it is defined by polar and negative surface areas in the one direction and fractional hydrophobic water accessible surface area in the other. Component 3 explains 7.61% of the variance, and it distinguishes between positively and negatively charged surface areas. Thus, the first three principal components are strongly correlated to properties we might call size, polarity, and sign of surface area partial charge.

Factor scores of cases (representations of structures in principle component space) using 2D descriptors only are given in Figure 3. We restricted this figure to the PCA of 2D properties since for the final model they were found to outperform both 3D descriptors and the full set of 2D and 3D descriptors. The definition of the principal axes still

**Table 3.** MOE Descriptors Spanning the First Three Principal Axes of the PCA Coordinate System, Using 2D and 3D Descriptors[a]

| principal component (PC) | explained variance (%) | positive factor loadings > 0.7 | negative factor loadings < −0.7 |
|---|---|---|---|
| PC1 | 32.61 | VAdjEq, PEOE_PC-, PC-, Q_PC-, *AM1_E, AM1_Eele, MNDO_E, MNDO_Eele, PM3_E, PM3_Eele* | diameter, VdistEq, VdistMa, weinerPath, weinerPol, a_count, a_IC, b_1rotN, b_count, b_rotN, b_single, chi0v, chi0v_C, chi1v, chi1v_C, Weight, a_heavy, a_nC, b_heavy, chi0, chi0_C, chi1, chi1_C, VadjMa, zagreb, PEOE_PC+, PEOE_VSA_HYD, PEOE_VSA_NEG, PEOE_VSA_POS, PC+, Q_PC+, Q_VSA_HYD, Q_VSA_NEG, Q_VSA_POS, Kier1, Kier2, Kier3, KierA1, KierA2, KierA3, KierFlex, apol, bpol, mr, a_hyd, vsa_hyd, SMR, vdw_area, vdw_vol, *E_vdw, rgyr, ASA, ASA+, ASA_H, CASA+, CASA-, FCASA+, VSA, vol* |
| PC2 | 12.81 | PEOE_VSA_FPNEG, PEOE_VSA_FPOL, PEOE_VSA_FPPOS, PEOE_VSA_POL, PEOE_VSA_PPOS, Q_VSA_FPNEG, Q_VSA_FPOL, Q_VSA_FPPOS, Q_VSA_POL, Q_VSA_PPOS, *ASA_P, FASA_P* | PEOE_VSA_FHYD, Q_VSA_FHYD, *FASA_H* |
| PC3 | 7.61 | PEOE_VSA+0, *FASA+* | *ASA-*, *FASA-* |

[a] 3D descriptors in italics. All descriptors possessing absolute factor loadings > 0.7 are given.

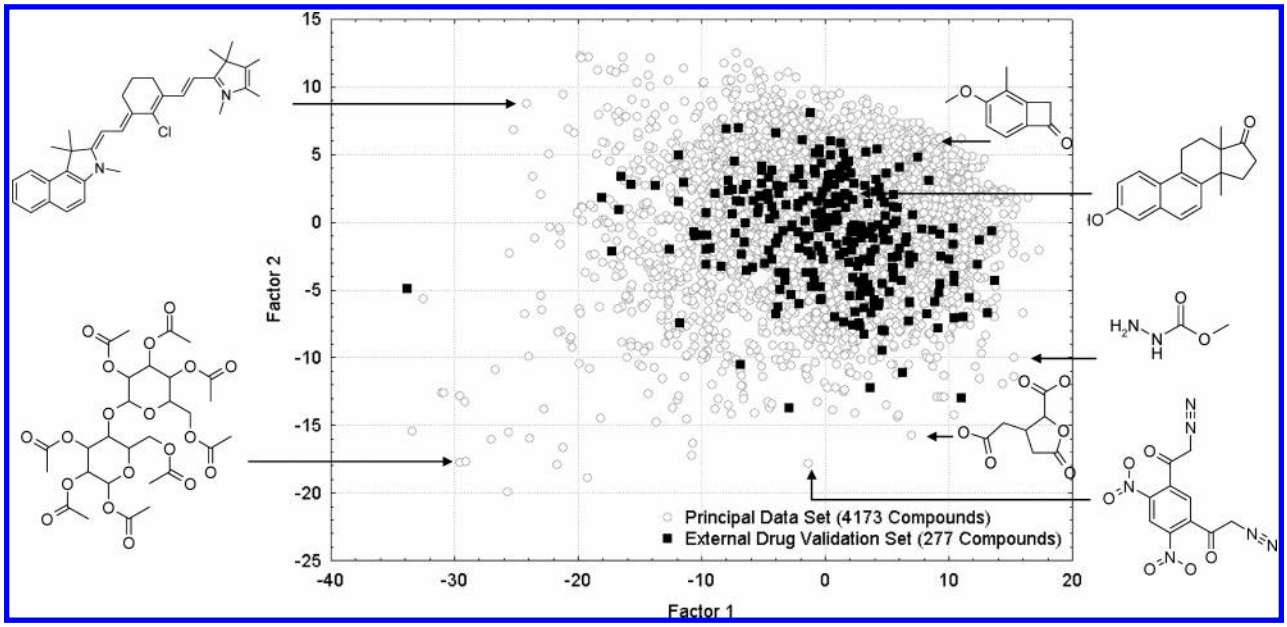



**Figure 3.** First two principal components of the principal MDPI data set (gray), compared to the external validation drug data set (black) in 2D descriptor space. The first two principal axes explain 32.61% and 12.81% of the variance, respectively. Owing to the physicochemical diversity of the principal data set (gray) it comprises a much larger area in chemical space than the, comparably, more homogeneous drug data set.

broadly agrees with the definitions of overall property space (results not shown). Along with the scatterplot of all compounds in the first two dimensions of PCA space, examples of molecules found in different regions of chemical space are given in Figure 3.

**(b) Model Building and Performance.** Results for the model building step are given in Table 4, distinguishing between dimensionality of descriptors used and showing performance separately for training set, test set, and internal validation set as well as for the external validation set.

For 2D descriptors, pruning and optimization of the number of nodes in the hidden layer resulted in a 26-12-1 ANN which was trained using back-propagation for 100 iterations, followed by 94 iterations of conjugated gradient optimization. Correlations between experimental and predicted melting points vary around 0.65 (0.645 and 0.662). Absolute mean errors are between 37.6 °C for the training set and 39.8 °C for the internal validation set. For the external validation set an absolute mean error of 32.6 °C was obtained.

For 3D descriptors, pruning and optimization of the number of nodes in the hidden layer resulted in a slightly larger 30-17-1 ANN which was trained using back-propagation for 100 iterations, followed by 20 iterations of conjugated gradient optimization. Correlations between experimental and predicted melting points vary around 0.5 (0.547 and 0.566) for all data sets except the external validation set for which much worse a correlation of 0.327 was obtained. Absolute mean errors are between 45.2 °C for the training set and 45.8 °C for the internal validation set. For the external validation set an absolute mean error of 50.5 °C was obtained.

**Table 4.** Correlation Coefficients, Mean Absolute Errors, and RMSE Values Obtained from Neural Network Training[a]

| | | | training set ($n = 2089$) | | | test set ($n = 1042$) | | | internalvalidation set ($n = 1042$) | | | external validation set ($n = 277$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| descriptors | network architecture | training | $R^2$ | absolute mean error/ °C | RMSE/ °C | $Q^2$ | absolute mean error/ °C | RMSE/ °C | $Q^2$ | absolute mean error/ °C | RMSE/ °C | $Q^2$ | absolute mean error/ °C | RMSE/ °C |
| 2D | 26-12-1 MLP | BP 100, CG 94 | 0.661 | 37.6 | 48.0 | 0.658 | 38.2 | 49.3 | 0.645 | 39.8 | 50.4 | 0.662 | 32.6 | 41.4 |
| 3D | 30-17-1 MLP | BP 100, CG 20 | 0.566 | 45.2 | 54.9 | 0.547 | 45.6 | 55.5 | 0.564 | 45.8 | 56.0 | 0.327 | 50.5 | 60.6 |
| 2D/3D | 17-10-1 MLP | BP 100, CG 47 | 0.630 | 38.5 | 48.4 | 0.628 | 41.2 | 51.4 | 0.648 | 41.3 | 52.0 | 0.544 | 43.1 | 43.5 |

[a] Results are given for the training set, the test set, and the internal and external validation sets. (MLP = multilayer perceptron, BP = back propagation, CG = conjugate gradient.).

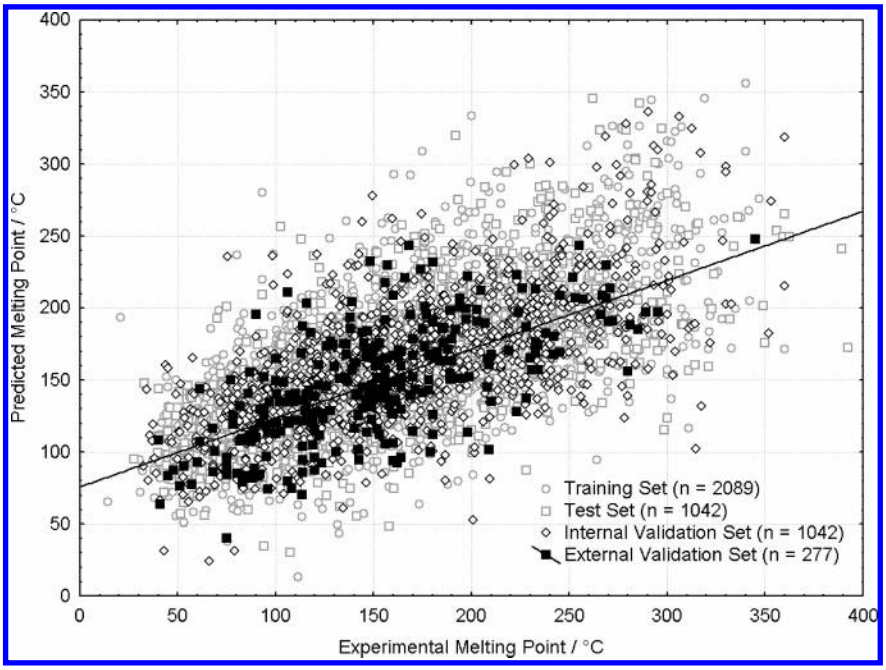



**Figure 4.** Plot obtained for a training set (unfilled gray circles), a test set (gray squares), and an internal (black rhombi) and an external validation set (black squares) between experimental and predicted melting points for the 2D descriptor set. For correlation coefficients see Table 4.

For the combination of 2D and 3D descriptors, pruning and optimization of the number of nodes in the hidden layer resulted in a 17-10-1 ANN which was trained using back-propagation for 100 iterations, followed by 47 iterations of conjugated gradient optimization. Correlations between experimental and predicted melting points vary around 0.63 (0.628 and 0.648) for all data sets except the external validation set for which a correlation of 0.544 was obtained. Absolute mean errors are between 38.5 °C for the training set and 41.3 °C for the internal validation set. For the external validation set an absolute mean error of 43.1 °C was obtained.

Predicted vs observed melting points are shown in Figure 4 for the best performing model, employing 2D descriptors only, for training set, test set, internal validation set, and external validation set.

## 4. DISCUSSION

**(a) Analysis of Chemical Space.** The expectation that 2D and 3D descriptors due to different descriptions of chemical space also convey complementary information (Figure 2 and Table 3) was not fulfilled. Although both 2D and 3D descriptors seem to sample the combined 2D/3D property chemical space well (given the smaller number of 3D descriptors) employing only 2D descriptors gave consistently superior results. The 3D descriptors do not seem to capture as much information relevant to melting point prediction as 2D descriptors do (or alternatively, they introduce additional noise).

Overall extracted principal components can be interpreted in a chemically meaningful way, which corresponds to important determinants of melting points as determined in earlier work.[6,10,34] These are, broadly speaking, size, polarity, and sign of surface area partial charge. Since the extraction of principal components simply represents compounds in a different coordinate system, without regard to the property one attempts to predict, this means that representation in reduced-dimensional space decreases the number of weights in the neural network (or any other learning algorithm) while at the same time retaining the relevant information of the data. The principal component analysis of chemical space spanned by the compounds (Figure 2, Table 3) shows the following properties. Component 1 is defined by size of the molecule, positive partial charge, and polarizability in one direction and the total molecular and electronic energy of the system in the opposite direction. This partly resembles

Pearson's concept of "hard and soft acids and bases" where large, polarizable compounds are classified as "soft" acids and bases, and small less polarizable molecules are classified as "hard" acids and bases.[35] Size (or molecular weight) of a molecule has been associated with its melting point from early on,[34] so its prominent representation in the first principal component is consistent with earlier attempts at prediction of melting point. The total electronic energy is also a measure, closely negatively correlated, with the number of atoms and thus electrons in the molecule. Component 2 is defined by polar and negative surface areas on one hand and fractional hydrophobic water accessible surface area on the other. This corresponds to the popular notion of more hydrophilic and more lipophilic compounds, so this axis should also be crucial for the prediction of melting points. Component 3 reflects both different molecular overall charges and partial surface charges, which depends on acidic and basic groups in the molecule as well as on the number and type of heteroatoms.

Representations of structures in principle component space (Figure 3) using 2D descriptors only are given in Figure 3. It can be seen that the principal data set well encapsulates the more physicochemically similar drug-only (external validation) set. The *X*-axis is mainly dominated by the size of the molecule (larger molecules are placed to the left and smaller molecules to the right). The *Y*-axis in turn is mainly dominated by polarity of the molecule: less polar molecules occur at the top of this representation of chemical space and more polar and charged molecules at the bottom of the plot. (Note that the two "outliers" to the left of the plot are structurally very different with nonetheless similar melting points: While one of the structures contains three ether-linked benzene rings around a central benzene entity and two methyl ester groups per auxiliary benzene, the second structure contains nonaromatic hexose units and amide and alcohol groups, which are not present at all in the first structure. Melting points are 188 °C and 135 °C, respectively.)

**(b) Model Building and Performance.** Results for the model building step are given in Table 4, distinguishing between dimensionality of descriptors used and showing performance individually for the training set, the test set, and the internal as well as for the external validation set.

Employing 2D descriptors only gave the best results on all data sets with respect to correlation coefficients, absolute mean error, and root-mean-square error. The network employing 3D descriptors performed worst, with the network using both 2D and 3D descriptors showing intermediate performance. For the external validation set, an absolute mean error of 32.6 °C was obtained. On first sight it might be surprising that the error on the external validation set is better than that on the training set (37.6 °C). This can be explained by the fact that only part of the training set compounds fall into the druglike area of the external validation set, which is probably better modeled by the neural network than more atypical areas of chemical space.

3D descriptors perform considerably worse than 2D descriptors, and even the combination of 2D and 3D descriptors gives worse results, in particular on the external validation set. This corroborates the finding that the information content (signal-to-noise ratio) of 2D descriptors is often equivalent or even superior to that of 3D descriptors,[36] which in addition have to deal with conformational problems.

A plot of predicted vs observed melting points shown in Figure 4 exhibits good correlation for the majority of compounds of all data sets. Despite the temptation to remove "outliers" and improve the model (in a statistical sense, that is) we decided against this approach to retain a model as universally applicable as possible.

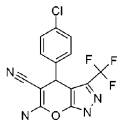
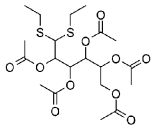
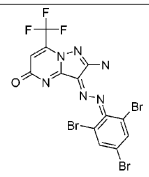
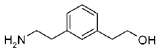
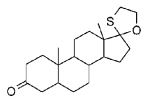
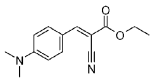
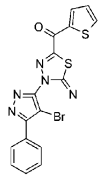
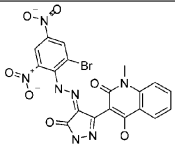
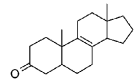
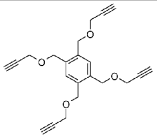
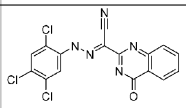
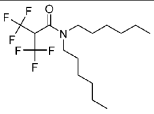
We will now compare our model to earlier melting point predictions. Zhao and Yalkowsky[20] obtained a RMSE of around 34 °C in a 10-fold cross validation using only aliphatic compounds. While our RMSE value is about 7 °C larger, this can still be seen as acceptable, given the diversity of our data set. Jain and Yalkowsky[5] report an average absolute error of 23 °C for 338 "rigid, non-hydrogen-bonding aromatic" (training set) structures. Again, we also include flexible, hydrogen-bonding, and nonaromatic structures which gives an average absolute error about 10 °C higher. The data set used by Bergstrom[10] is the most diverse data set used for melting point prediction so far. Their values compare to those obtained by us as follows. In our model we obtain a slightly higher correlation coefficient on the training set ($R^2 = 0.661$ for our model vs $R^2 = 0.31-0.57$ for the model of Bergstrom et al.) accompanied by similar RMSE values (41.4 °C vs 36.6−46.1 °C). For the internal validation set, we obtain larger correlation coefficients (0.645 vs 0.30−0.53) with comparable RMSE values (50.4 °C vs 49.8−51.7 °C) as well. Both the correlation coefficient and RMSE of our model excel on the external validation set, with a correlation of 0.662 and a RMSE of 41.4 °C vs a correlation of between 0.30 and 0.54 and an RMSE value of between 49.8 and 51.7. Thus our model seems to be able to generalize to the prediction of melting points of previously unseen compounds.

Despite overall sensible predictivity there remain compounds (rather classes of compounds) whose melting points are consistently over- or underpredicted. Examples of both classes of compounds are given in Table 5, along with the data set they belong to and the experimental as well as the predicted melting point. Here we will only comment on some of the examples.

The compounds showing overestimated melting points (left-hand side of Table 5) are generally either heterocycles containing multiple nitrogen atoms or nonaromatic steroids. Melting points of steroids are probably predicted to be very high due to their rigidity (low number of rotatable bonds). While the trend that more rigid compounds possess higher melting points is true in general, in case of steroids this leads to overestimations of melting points. The reason why mainly multiple-nitrogen heterocycles belong to the class of compounds with overestimated melting points is yet unknown.

Compounds from Table 5 whose melting points are underestimated (right-hand side of Table 5) mainly belong to one of three classes: They are either conformationally less flexible than might be guessed from the formal number of rotatable bonds (compound 1), they are small molecules which self-organize in the solid state and form stronger interactions than typical for molecules of this size (compounds 2, 4, and 5), or they are larger molecules whose physicochemically similar neighbors are just in an area of chemical space that does not allow for formation of the interactions they are capable of forming (compound 6).

**Table 5.** Examples of Structures Whose Melting Points Where Grossly Over- and Underestimated[a]

| Nr. | *Overestimated* Melting Points | Data Set | MP Exp | MP Pred | *Underestimated* Melting Points | Data Set | MP Exp | MP Pred |
|---|---|---|---|---|---|---|---|---|
| 1 |  | Train | 21 | 193 |  | Train | 112 | 13 |
| 2 |  | Train | 93 | 280 |  | Train | 264 | 95 |
| 3 |  | Train | 200 | 334 |  | Train | 311 | 117 |
| 4 |  | Test | 103 | 256 |  | Test | 108 | 30 |
| 5 |  | Test | 192 | 320 |  | Test | 228 | 87 |
| 6 |  | Valid. | 44 | 161 |  | Valid. | 66 | 24 |
| 7 |  | Valid. | 150 | 278 |  | Valid. | 201 | 53 |

[a] Reasons for wrong predictions are mainly the lack of information about both intermolecular and intramolecular interactions (frozen degrees of rotational freedom, see text), which can only be known from the solid-state structure. On the other hand, deviations from predicted values can be used to gauge the extent of intramolecular interactions.

Compound 1 from the right-hand side column of Table 5 formally possesses a large number of rotatable C−C bonds and should thus melt at rather low temperatures. Due to putative O−S interactions[37,38] and electrostatic repulsion between partially negatively charged oxygens, those rotations are not allowed in practice. This shows that not only intermolecular but also intramolecular interactions are not sufficiently captured by the single-molecule based descriptors employed here.

Compounds 2, 4, and 5 show high self-organization in the solid state. This is most pronounced for 3,5-dimethylpyrazole, which effectively exists as a trimer,[39,40] hugely influencing the melting point. Compound 2, possessing both aliphatic alcohol and primary amine function, can be assumed to show strong charge interactions between those functionalities since they possess very similar p$K_a$'s, leading to partial deprotonation of the alcohol and partial protonation of the amine function and thus Coulombic interactions.

Terminal ethyne groups, as in compound 6, may become involved in C≡C−H/$\pi$(arene) and C≡C−H/$\pi$(C≡C) intermolecular interactions such as in ethnylbenzenes.[41] While ethnylbenzene possesses a melting point of −45 °C, lower than benzene itself, the polysubstituted benzenes show much higher melting points: 1,4-diethnylbenzene melts at 97 °C and 1,3,5-triethnylbenzene at 104 °C, respectively. While in ethnylbenzene C≡C−H/$\pi$(arene) interactions dominate, the di- and trisubstituted analogues are dominated by C≡C−H/$\pi$(C≡C) interactions, leading to zigzag arrays and helical chains, respectively.[41] One or both of those interaction types probably also occurs in the case of compound 6.

Finally, the work by Bergstrom et al.[10] shall briefly be commented on and put in context with the work performed here. Both approaches have in common that only single-molecule properties are considered, neglecting information about solid-state packing. As shown above, in addition certain intramolecular interactions are not captured by the descriptors employed. While Bergstrom et al. deleted 491 out of 612 descriptors due to skewedness,[10] we chose to use artificial neural networks instead, which are often able to capture nonlinear properties, thus helping to reduce the influence of skewed variables. In a later step, Bergstrom et al.[10] investigated the influence of each descriptor on the melting point. While the question of which properties influence the melting point is of course of huge relevance, it seems problematic if, firstly, the majority of descriptors is deleted, and, secondly, the remaining parameters are highly correlated. The question turns up then whether the PLS components with the largest weights are really those which—in a physical sense—influence melting points, or whether they are just properties that are highly correlated to them. Since highly correlated descriptors are difficult to interpret anyway, we chose instead to perform PCA to remove correlation of descriptors and, as mentioned above, a nonlinear modeling technique not to run into problems with "strange" behavior of descriptor variables.

So what does the current work teach us, apart from producing a model that works in a practical sense? Single-molecule physicochemical space should be sufficiently captured by the more than 100 descriptors employed here in combination with decorrelation techniques and a nonlinear modeling approach. Thus, we assume that the modeling errors that remain are due to intra- and intermolecular interactions and that here we encounter the theoretical performance barrier of predicting melting points using single-molecule descriptors. The solid-state packing is influenced by factors that can be outside of the simple description of the structure of the molecule itself: e.g. the solvent from which a material is crystallized, which can lead to polymorphs with substantially different melting points. It is thus apparent that different polymorphic forms pose a theoretical upper limit on the prediction of melting points one can hope to achieve. Thinking the other way around, using the model developed here for the estimation of melting points while the experimental value is already known opens up a new possibility: To gauge intra- and intermolecular interactions in the solid state, compared to the "average" intra-/intermolecular interactions of a molecule of a given size. Overestimated melting points, as shown in some examples,

correspond to smaller interactions than expected from the training set, while underestimated melting points indicate formation of oligomers or other forms of higher self-organization.

## 5. CONCLUSIONS

We present a more generally applicable method for the prediction of melting points for diverse compound data sets, ranging from small, unsubstituted hydrocarbons to large, heavily substituted macrocycles. Using a data set comprising 4173 structures extracted from the MDPI database with melting points between 14 °C and 392.5 °C, PCA in combination with an artificial neural network is able to achieve internal consistency and external predictivity of the model.

Compared to an earlier model for the prediction of melting points of druglike compounds, which was based on several hundred descriptors in combination with partial-least squares regression, our model exhibits superior performance, despite the much more diverse training set. The root-mean square error on the external validation set is reduced from 51.7 °C to 41.4 °C, indicating that the model is generally applicable for melting point prediction of a wide variety of organic compounds.

There exist two main reasons for large outliers which are inter- as well as intramolecular interactions. Intermolecular interactions depend on the arrangement of compounds in the solid state which is not taken into account by single-molecule descriptors. Sometimes small structural changes cause new interactions and thus a different solid-state structure to occur, which is not captured by our method. Small heterocycles are a typical class of compounds for which very low melting points are predicted mainly due to their lightness, but where experimental melting points are much higher due to self-organization processes (for example the formation of oligomers). Similar observations can be made with arene/alkyne systems, where terminal ethyne groups may or may not lead to C≡C−H/$\pi$(arene) and C≡C−H/$\pi$(C≡C) interactions of varying strength, depending on the particular compound.

On the other hand, intramolecular interactions may give rise to underestimations of melting points. A molecule with many rotatable bonds is assumed by the model to possess a rather low melting point. Still, if some of the rotational degrees of freedom are frozen due to intramolecular interactions (sulfur−oxygen interactions or electrostatic repulsion in the example presented here) the actual melting point may be much higher than the predicted one.

The interpretation of PLS weights as physical factors influencing melting points (but also other properties) appears problematic if a large number of variables are first excluded from analysis and the remaining variables are highly correlated. Here we follow the route of decorrelation via PCA in combination with a nonlinear modeling technique to alleviate those problems. Due to the large number of descriptors used we can hope that single-molecule physicochemical space to be covered sufficiently well. The remaining model error seems to be due to inter- and intramolecular interactions, so in combination with nonlinear techniques we seem to encounter the theoretical performance limit of single-molecule descriptor based melting point predictions. Inversely, employing a melting-point model, the prediction can be used to gauge solid-state interactions in cases where the experimental melting point is already known: Molecules whose melting points are overpredicted show weaker intermolecular interaction in the solid state than similarly sized molecules on average. Molecules whose melting points are underpredicted show stronger intermolecular interactions in the solid state which may be due to oligomerization processes or other self-organization processes. Examples for both classes of compounds are given.

Finally, the model we present is able to give a very fast estimate of the melting point of a substance. Whether the deviation of our model from the value observed in experiment is acceptable in practice always depends on the particular setting it is applied in.

## ACKNOWLEDGMENT

Christel Bergstrom is thanked for providing us with the external drug validation set. A.B. and R.C.G. thank the Gates Cambridge Foundation and Unilever for funding. M.K. thanks the Department of Science and Technology, New Delhi, India for the award of a BOYSCAST fellowship, and guidance by Prof. Alex Tropsha from School of Pharmacy, University of North Carolina at Chapel Hill is gratefully acknowledged.

**Supporting Information Available:** Complete data set containing optimized molecular structures, associated melting points, and molecular descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) 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.
(2) Abramowitz, R.; Yalkowsky, S. H. Melting point, boiling point, and symmetry. *Pharm. Res.* **1990**, *7* (9), 942−7.
(3) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(4) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204−3218.
(5) Jain, N.; Yalkowsky, S. H. UPPER III: unified physical property estimation relationships. Application to non-hydrogen bonding aromatic compounds. *J. Pharm. Sci.* **1999**, *88* (9), 852−60.
(6) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR − The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.
(7) Steinstrasser, R.; Pohl, L. Chemistry and Applications of Liquid Crystals. *Angew. Chem., Int. Ed. Engl.* **1973**, *12*, 617−630.
(8) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods; Environmental Behavior of Organic Compounds*, 2nd ed.; American Chemical Society: Washington, DC, 1990.
(9) Charton, M. The nature of topological parameters. I. Are topological parameters 'fundamental properties'? *J. Comput.-Aided Mol. Des.* **2003**, *17* (2−4), 197−209.
(10) Bergstrom, C. A.; 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* (4), 1177−85.
(11) Ma, L.; Cheng, C. Predicting melting temperature (Tm) of oligoribonucleotide duplex by neural network. *J. Chemom.* **2002**, *16*, 75−80.
(12) Burch, K. J.; Whitehead, E. G. Melting-point models of alkanes. *J. Chem. Eng. Data* **2004**, *49*, 858−863.
(13) Wen, X.; Qiang, Y. Group Vector Space (GVS) Method for Estimating Boiling and Melting Points of Hydrocarbons. *J. Chem. Eng. Data* **2002**, *47*, 286−288.
(14) Dearden, J. C. The QSAR prediction of melting point, a property of environmental relevance. *Sci. Total Environ.* **1991**, *109−110*, 59−68.
(15) Johnson-Restrepo, B.; Pacheco-Londono, L.; Olivero-Verbel, J. Molecular parameters responsible for the melting point of 1,2,3-diazaborine compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1513−9.

(16) Chickos, J. S.; Nichols, G.; Ruelle, P. The estimation of melting points and fusion enthalpies using experimental solubilities, estimated total phase change entropies, and mobile order and disorder theory. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 368−74.
(17) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. Correlation of the melting points of potential ionic liquids (imidazolium bromides and benzimidazolium bromides) using the CODESSA program. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (2), 225−31.
(18) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 71−4.
(19) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913−919.
(20) Zhao, L.; 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.
(21) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20* (4), 269−76.
(22) Molecular Diversity Preservation International (MDPI), http://www.mdpi.org.
(23) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(24) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(25) Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(26) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 346−56.
(27) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (25), 15869−72.
(28) Pearlman, R. S. CONCORD: Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Autom. News* **1987**, *2*, 5−7.
(29) Balakin, K. V.; Lang, S. A.; Skorenko, A. V.; Tkachenko, S. E.; Ivashchenko, A. A.; Savchuk, N. P. Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1553−62.
(30) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(31) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of 2-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (6), 644−649.
(32) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (4), 651−655.
(33) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*; Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, 1986; Vol. 1, pp 319−362.
(34) Austin, J. B. A relation between the molecular weights and melting points of organic compounds. *J. Am. Chem. Soc.* **1930**, *52*, 1049−1053.
(35) Pearson, R. G. Hard and Soft Acids and Bases. *J. Am. Chem. Soc.* **1963**, *85*, 3533−3543.
(36) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 1−9.
(37) Kalman, A.; Parkanyi, L. Structure of 2-benzoylimino-3-methyl-1,3-thiazolidine: a comparison of intramolecular X−S···O=Y interactions. *Acta Crystallogr. Sect. B* **1980**, *36* (10), 2372−2381.
(38) Kucsman, A.; Kapovits, I.; Czugler, M.; Parkanyi, L.; Kalman, A. Intramolecular Sulfur Oxygen Interaction in Organosulfur Compounds with Different Sulfur Valence State − an X-ray Study of Methyl-2-Nitrobenzene-Sulphenate, Methyl-2-Nitrobenzene- Sulphinate, Methyl-2-Nitrobenzene-Sulphonate and 2- Nitrobenzenesulfenyl Chloride. *J. Mol. Struct.* **1989**, *198*, 339−353.
(39) Stride, J. A.; Jayasooriya, U. A.; Mbogo, N.; White, R. P.; Nicolai, A.; Kearley, G. J. Hydrogen-bonding in the self-organising system 3,5- dimethylpyrazole. *New J. Chem.* **2001**, *25* (8), 1069−1072.
(40) Goddard, R.; Claramunt, R. M.; Escolastico, C.; Elguero, J. Structures of NH-pyrazoles bearing only C-methyl substituents: 4-methylpyrazole is a hydrogen-bonded trimer in the solid (100 K). *New J. Chem.* **1999**, *2*, 237−240.
(41) Nishio, M. CH/pi hydrogen bonds in crystals. *Cryst. Eng. Comm.* **2004**, *6*, 130−158.

CI0500132