Development of quantitative structure property relationships for predicting the melting point of energetic materials

Jason A. Morrill [a],[*], Edward F.C. Byrd [b]

[a] Department of Chemistry, William Jewell College, 500 College Hill, Liberty, MO 64068, USA
[b] United States Army Research Laboratories, AMSRD-ARL-WM-BD, Aberdeen Proving Ground, MD 21005-5069, USA

ABSTRACT

The accurate prediction of the melting temperature of organic compounds is a significant problem that has eluded researchers for many years. The most common approach used to develop predictive models entails the derivation of quantitative structure-property relationships (QSPRs), which are multivariate linear relationships between calculated quantities that are descriptors of molecular or electronic features and a property of interest. In this report the derivation of QSPRs to predict melting temperatures of energetic materials based on descriptors calculated using the AM1 semiempirical quantum mechanical method are described. In total, the melting points and experimental crystal structures of 148 energetic materials were analyzed. Principal components analysis was performed in order to assess the relative importance and roles of the descriptors in our QSPR models. Also described are the results of k means cluster analysis, performed in order to identify natural groupings within our study set of structures. The QSPR models resulting from these analyses gave training set R^2 values of 0.6085 (RMSE = ± 15.7 °C) and 0.7468 (RMSE = ± 13.2 °C). The test sets for these clusters had R^2 values of 0.9428 (RMSE = ± 7.0 °C) and 0.8974 (RMSE = ± 8.8 °C), respectively. These models are among the best melting point QSPRs yet published for energetic materials.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The range in temperatures over which a substance melts is a physical property of fundamental importance. This is especially true in the energetic material community, where an unacceptably low melting point of a notional material can render it completely unfeasible for usage, regardless of any superior performance capability. When a substance undergoes a change in state from a relatively ordered solid to a disordered liquid the average thermal kinetic energy of the substance is sufficiently high so as to overcome the attractive forces between the constituents of the substance. The amount of thermal energy required to overcome the attractive forces depends on a great many factors including the types of electrostatic attractive forces between the constituents making up the substance, symmetry of the individual constituents, and efficiency with which the constituents are organized in the solid, just to name a few. In addition, so-called bulk effects, such as crystal defects, can affect the melting temperature of a substance. In common use the melting point range of a substance is a measure of its purity

and also an indicator of useful temperature range for a substance. Prior knowledge of the melting point range is therefore an important consideration in designing new materials. Toward designing new materials a great many computationally derived mathematical models have been developed for predicting such physical properties as boiling points, octanol-water partition coefficients [1], glass transition temperatures [2], and melting points [1–9]. While there have been many attempts to develop models for predicting melting points, to date most models have resulted in only moderately accurate predictive capability and very few have been developed specifically for energetic materials. Additionally, several attempts have been made to generate melting point models that are applicable to large numbers of structurally diverse compounds, it is generally known in the QSPR community that a more accurate model is produced when restricting the model to compounds of a common structure type.

Mitchell et al. used an Artificial Neural Network (ANN) approach to explore hypothesized reasons explaining why the QSPR approach sometimes fails to develop good fits to experimental data and/or highly predictive multiple correlation models for various properties, including melting points [10]. Using the molecular operating environment and MMF94× molecular mechanics-based energy minimization the authors computed 567 2D and 3D descrip-

* Corresponding author.
 E-mail address: morrillj@william.jewell.edu (J.A. Morrill).

tors for each molecule under investigation. The best model for each property was determined by using four different multiple correlation equation derivation algorithms: partial least squares, random forest, k nearest neighbors, and support vector machine. These algorithms were used to fit a pre-selected set of descriptors, chosen by the "Ant Colony Optimization" algorithm. The most predictive models were obtained for the octanol-water partition coefficient with an R^2 of 0.87 and the least predictive model was for the melting temperature with an R^2 of 0.46. The authors suggested that having relatively few compounds with melting points at the extreme ends of the range in experimental data may have lead to the poor predictive capability of the model. They also suggested that thermal decomposition of some compounds on melting likely contributed to error in the experimental data that couldn't be accounted for in the regression models.

Karthikeyan et al. employed an ANN approach to derive predictive models for a diverse set of 4173 organic molecules [3]. The descriptor set that they employed was also diverse, having such descriptor types as physical quantities (charges, van der Waals volume, etc.), connectivity and topology indices, pharmacophore features (hydrogen bond donors/acceptors, charged partial surface areas, etc.). The charges and dipole moments were computed using PM3 [11] and AM1 [12] semiempirical quantum mechanical methods, respectively. The authors reduced the dimensionality of their data by performing principle components analysis (PCA). They found that the types of descriptors representing the most variance in the data could generally be categorized as those describing size, polarity, and sign of the surface area partial charge. Further, the authors found that models employing 2D descriptors outperformed both models based solely on 3D and those based on combinations of 2D and 3D descriptors. The 2D (best models) had coefficients of multiple determination of $R^2 = 0.661$ for the training set and an RMSE of 48.0 °C. The corresponding test set gave an RMSE of 49.3 °C.

Given the documented difficulty in generating highly accurate melting point QSPRs, Preiss et al. posed the question of whether it was even possible to generate a chemically reasonable, universal, and simple melting point predictive model. Their melting point model was based on the 1:1 salts of 520 organic compounds [8]. Giving much consideration of the theoretical understanding of the phase change process of melting the authors ultimately derived a nine descriptor model. The descriptors were various types including computed entropy and enthalpy of solvation terms. The enthalpy of solvation was computed from geometries treated with the BP86 [13–16] density functional theory (DFT) optimized geometries using the TZVP basis set [17] and employing the COSMO solvation model [18]. Their model also included torsional flexibility, electrostatic, and molecular volume type descriptors. Their model provided an R^2 value of 0.537 and an average error of 33.5 °C or 9.3%.

Salahinejad, et al., attempted to generate a generally useful model for the prediction of the enthalpy of sublimation (ΔH_{sub}) of small organic molecules [19]. The authors fit a simple multiple linear regression model consisting of four preselected descriptors based on hydrophilicity and charged partial surface area descriptors (CPSA) using an artificial neural network (ANN) approach. The molecular basis set consisted of a very large and structurally diverse set of 8241 small organic molecules, which was divided into training and test sets of 80% and 20% of the total number of molecules, respectively. Using this approach they derived various QSPR models for ΔH_{sub} having R^2 of 0.954 for the test set and $R^2 = 0.950$ for the test set. Subsequently, they treated the predicted ΔH_{sub} values as a descriptor and incorporated additional descriptors by an artificial neural network approach to derive a QSPR model for melting points. The authors found it necessary to add 97 descriptors in multiple ANN-derived models to derive predictive QSPRs having R^2

values that varied from 0.77 to 0.79 for the training sets and from 0.75 to 0.79 for the test sets.

Lazzus developed a QSPR that utilized descriptors from the results of calculations employing the PM3 semiempirical quantum mechanical method [20]. His models were derived, first using CODESSA [21] (Comprehensive Descriptors for Structural and Statistical Analysis) to compute and pre-analyze the descriptors to eliminate descriptors which were significantly intercorrelated or not significantly correlated with the experimental data. Following the filtering of descriptors Lazzus derived his predictive models using a back propagation ANN (BP ANN) method. For an eleven descriptor model fit to a training set of 260 compounds he found an average absolute deviation of 5.2% and for a 73 molecule test set the deviation was 4.8%.

It is well known that QSPR models based on congeneric sets of structures generally give better results. By focusing their work more narrowly on compounds of a similar type, a number of investigators attempted to generate melting point QSPR models that would have to account for less variance in the experimental data. Guan et al. derived QSPR models on the basis of structures that were optimized using the AM1 semiempirical quantum mechanical method [22]. The compounds under study were the bromide and chloride salts of various imidazoliums. Three sets of compounds were used and within each set there were training and test sets. The training sets consisted of 24, 16, and 40 compounds and the test sets consisted of a number of compounds that was 25% of the number in the corresponding training set. The authors carried out a comparison of multiple linear regression (MLR) models obtained by Genetic Algorithms to those obtained by BP ANN methods. In general they found that the BP ANN models gave better results than the MLR models. The mean absolute errors (MAEs) for each of the respective MLR model test sets were 20.52 K, 13.59 K, and 21.95 K. The MAEs for the BP ANN derived model test sets were 8.77 K, 4.98 K, and 9.31 K, respectively.

Brauner and Shacham examined the melting point of structurally homologous series of compounds, which varied only in the number of carbon atoms present in the molecule [23]. Their multiple linear regression model was based only on functions of constitutional (number of carbon atoms) type descriptors and were regressed on melting points for varying series of long-chain hydrocarbon containing compounds (n-alkanes, n-alkanoic acids, n-alkenes, n-mercaptans, and n-alkyl acetates). The average absolute relative errors reported for the different sets were between 1% and 9%.

Holder and Liu derived a melting point QSPR for organosilicon compounds [24]. Through a forward stepping multiple linear regression approach they developed a more comprehensive organosilicon melting point model, which was later refined for silanes only. All structures in Holder's work were modeled using the SAM1 semiempirical quantum mechanical method [25]. The descriptors were obtained directly from these calculations or were calculated from quantities obtained in these calculations using CODESSA. For the general organosilicon model they found a six descriptor model for 97 training compounds having $R^2 = 0.789$ and an RMSE of 50.45 °C. For the general organosilicon model an external test set of 32 compounds was found to have $R^2 = 0.710$ and an RMSE of 42.44 °C. A better fit was found for a more focused set of 62 structurally similar silanes. For this set Holder and Liu derived a 4 descriptor model having $R^2 = 0.889$ and an RMSE of 34.09 °C for the training set and an external test set ($N = 19$) was found to have $R^2 = 0.858$ and an RMSE = 38.44 °C. Holder's model was also refined by an iterative robust regression analysis. The descriptors in the obtained models were generally constitutional, electrostatic, and thermodynamic in nature and were found to be chemically sensible on the basis of theoretical considerations of melting behavior.