CHEMICAL REVIEWS

REVIEW

pubs.acs.org/CR

Quantitative Structure−Property Relationship Modeling of Diverse Materials Properties

Tu Le,[†] V. Chandana Epa,[‡] Frank R. Burden,[†] and David A. Winkler*[,†,§]

[†]CSIRO Materials Science and Engineering, Bag 10, Clayton South MDC 3169, Australia
[‡]CSIRO Materials Science and Engineering, 343 Royal Parade, Parkville 3052, Australia
[§]Monash Institute of Pharmaceutical Sciences, 381 Royal Parade, Parkville 3052, Australia

S Supporting Information

CONTENTS

1. Introduction 2889
2. Brief Survey of Methods Used for QSPR Modeling of Materials 2890
2.1. Descriptors 2891
2.2. QSPR Modeling Methods 2891
3. Assessing the Performance of Models 2892
4. Common QSPR Modeling Errors and Pitfalls 2893
4.1. Uninformative Descriptors 2893
4.2. Overfitting and Grossly Underdetermined Systems 2893
4.3. Descriptor Selection and Chance Correlations 2893
4.4. Modeling Complex, Nonlinear Structure−Property Relationships 2894
4.5. Validating QSPR Models 2894
4.6. Domain of Applicability of Models 2894
4.7. Incorrect Handling of Outliers 2895
5. Materials Classes 2895
5.1. Nanomaterials 2895
5.1.1. Fullerenes and Nanotubes 2895
5.1.2. Other Nanomaterials 2898
5.2. Catalysts 2899
5.2.1. Homogeneous Catalysts 2900
5.2.2. Heterogeneous Catalysts 2901
5.2.3. Electrocatalysts 2903
5.3. Biomaterials 2903
5.4. Polymers (Nonbiological Applications) 2904
5.4.1. QSPR Models of Glass Transition Temperature T_g 2905
5.4.2. Models of Other Synthetic Polymer Properties 2909
5.5. Ionic Liquids 2911
5.5.1. Melting Points 2911
5.5.2. Models of Other Ionic Liquid Properties 2912
5.6. Supercritical Carbon Dioxide 2913
5.7. Ceramics 2913
6. Perspective on QSPR Modeling of Materials 2914
6.1. Summary of Material Property Prediction Examples 2914
6.2. Summary of Pitfalls in QSPR Modeling 2915
7. Conclusion 2916
Author Information 2916
Biographies 2916
References 2917

1. INTRODUCTION

The design and synthesis of materials with useful, novel properties is one of the most active areas of contemporary science, generating a veritable explosion of scientific activity in areas such as biomaterials, cell and tissue engineering, organic photovoltaics and light-emitting materials, and nanomaterials for a myriad of medical and nonmedical applications. This new era of materials design and discovery covers many disciplines from chemistry and biology to physics and engineering. The vast majority of this research effort is in experimental science, with theoretical and computational science lagging somewhat behind. Obviously, the ability to predict the properties of novel materials prior to synthesis, and to understand the relationships between the microscopic properties of molecular components and the macroscopic materials properties, would be of substantial benefit to materials designers. In view of the complexity of many new materials, there is a strong need for machine learning methods that can generate robust, predictive models linking these microscopic and macroscopic properties. The application of such methods to model materials properties is described as quantitative structure−property relationship (QSPR) modeling. Although these, and closely related methods such as quantitative structure−activity relationships (QSAR), have proven to be very successful in other areas of molecular design, there is surprisingly little published work on their applications to materials, as can be seen in Figure 1.

This review summarizes the most commonly used predictive, structure−property modeling methods and their recent applications to materials design. There are no published reviews of this type of materials modeling, in spite of the small but increasing number of materials QSPR modeling papers appearing in the literature across a wide range of materials classes—nanomaterials, catalysts, biomaterials, polymers, ionic liquids, supercritical CO_2, and ceramics, as shown in Figure 2. We anticipate that this comprehensive, critical review will be useful to a broad spectrum

Received: March 1, 2011
Published: January 17, 2012

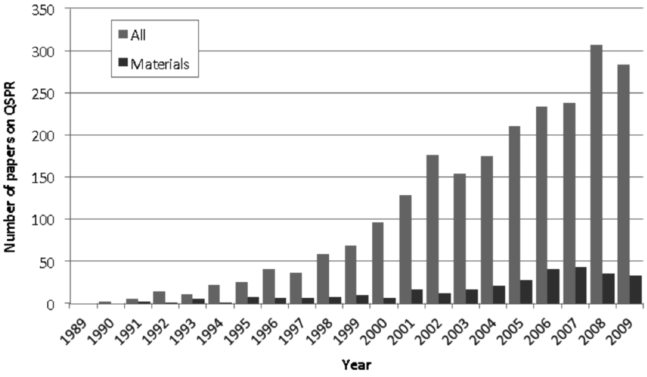


Figure 1. Number of scientific publications on QSPR studies as a function of publication year (ISI Web of Science).

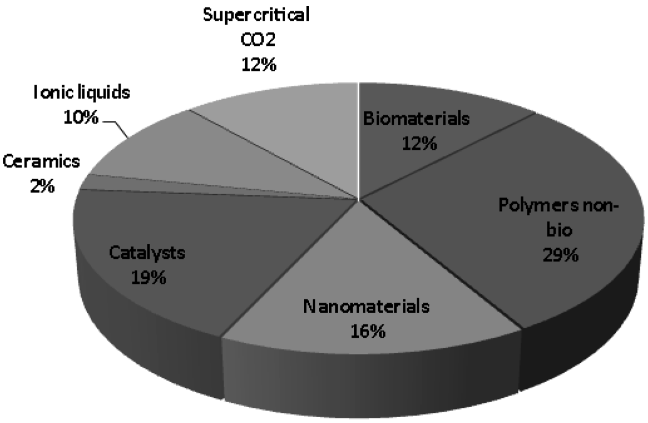


Figure 2. Different classes of materials that have been the subject of QSPR studies.

of materials scientists who are interested in predicting properties of new materials using robust, platform-modeling tools. Given the likely large increase in number and size of materials data sets, particularly as high-throughput methods are more widely adopted, this review will also be of interest to many experimental scientists who wish to extract knowledge and value from their material libraries.

This review is organized as follows. Section 2 gives a brief survey of methodologies used in QSPR modeling of materials in work published to date. Section 3 describes the most commonly used metrics for assessing the quality and predictivity of QSPR models. Section 4 provides a timely summary of the types of problems that can occur when QSPR modeling methods are applied without due care. QSPR models for different materials classes are critically reviewed in section 5. Section 5.1 is devoted to QSPR models of nanomaterials such as fullerenes, carbon nanotubes, and nanoparticles. Published work on modeling of homogeneous, heterogeneous, and electro catalysts is reviewed in section 5.2. Modeling studies of polymers in biomaterials and nonbiological applications are presented in sections 5.3 and 5.4, respectively. These sections are followed by a review of QSPR studies of ionic liquids and supercritical carbon dioxide in sections 5.5 and 5.6, respectively, and of ceramics in section 5.7. Conclusions and recommendations for future work are given in sections 6 and 7. This review aims to provide a concise summary of the QSPR modeling of materials research published to date. We have conducted a *critical* analysis of the published work so that readers will have a good understanding of the methods employed, the quality of the models, and the pitfalls that can occur unless care is taken.

2. BRIEF SURVEY OF METHODS USED FOR QSPR MODELING OF MATERIALS

QSPR methods are based on the hypothesis that changes in molecular structure are reflected in changes in observed macroscopic properties of materials. QSPR modeling is a supervised learning method that extracts the often-complex relationships between the microscopic (usually molecular) structure and properties of materials and their macroscopic properties (e.g., mechanical, thermal, electrical, etc.). Consequently, the key requirement for QSPR modeling, which distinguishes it from other physics-based computational methods like molecular dynamics, is a reliable data set of molecules or materials whose microscopic structures and properties are reasonably well-defined, together with their measured macroscopic properties of interest. This is termed training data. Because all measured data is associated with errors, and these errors will affect the reliability of models, it is important to estimate the reliability of the training data. This is relatively straightforward if all measurements have been made in the same laboratory, but may pose problems if not. The reliability of the experimental property chosen to be modeled is clearly very important, because it is one of the factors that determines the stability and predictivity of models. If the experimental measurements have high uncertainties, or if the chemical diversity or range of measured property values is too small, generation of robust, predictive, and reliable models is not possible. QSPR often assumes a normal distribution of the experimental property values, and this is sometimes not the case. Where the property being modeled spans several orders of magnitude, or the property values deviate greatly from the normal distribution, a transformation such as log(property), log (1/property) is used.[1] The final model cannot be more reliable than the original data. However, machine learning methods used in QSPR models can be quite tolerant of experimental error and missing data.

Materials properties can be modeled in two main ways depending on the type of input variables used to construct the model (although hybrid models in which these are combined are also possible). First, process and synthesis conditions and compositions of starting materials can be used as input data (useful when the exact structure and composition of the final material is uncertain or unknown). Alternatively, molecular descriptors (mathematical objects that capture the microscopic properties of the material or system being modeled) and properties that are related to the nature and connectivity of the microscopic components (often, atoms) of the material can be used as input data. However, some material properties depend not only on the intrinsic material properties but also the history of the material: how it was synthesized, processed, and prepared for testing. Importantly, in some cases QSPR models will need to combine molecular and physicochemical properties descriptors of materials *and* descriptors related to the synthesis, processing, or sample preparation to generate the most predictive and useful material property models.

QSPR modeling usually consist of four main operations: calculating or measuring a pool of descriptors or other input variables; choosing a small subset of these descriptors that are relevant to the macroscopic material properties being modeled (in some cases this step may not be required); generating the often nonlinear relationship between the descriptors and the global material property; and validating the model to assess its reliability, robustness, predictivity, and domain of applicability. A very recent review by Katritzky et al.[1] provides an in-depth and accessible summary of the main concepts involved in QSPR modeling of physical properties of discrete molecules. Therefore, only a summary of the important elements of the QSPR modeling process is provided here.

2.1. Descriptors

Examples are given in section 5 where the synthesis conditions and compositions of starting materials are used as descriptors. Often, simple compositional and process parameters are found to give useful models. However, in the vast majority of cases reviewed, molecular descriptors are used to characterize the microscopic properties of the materials for modeling purposes. Polymers are problematic materials to model, as the chain length and polydispersity are often not well characterized, especially for large polymer libraries. It is also clearly not possible to represent an entire polymer chain in terms of mathematical descriptors. Fortunately, as the examples in section 5 show, it is often possible to model polymer properties effectively by using the monomer or repeating unit to generate molecular descriptors. Many different types of molecular descriptors have been devised, and it is beyond the scope of this review to describe them. Recent books and review papers are available for further information on the myriad of molecular descriptors that have been developed.[2−4] They can be broadly categorized into the following categories:

- constitutional (the relative numbers of various atom types);
- topological (describing properties and connectivity of atoms making up the material);
- physicochemical (relating to the water or lipid solubility, dipole moment, formal charge, etc.);
- structural (describing the three-dimensional size, shape, and surface properties of the molecule; and
- quantum-chemical (e.g., partial charges, polarizabilities, multipole moments, orbital energies, etc. calculated using semiempirical, density functional theory (DFT), and ab initio quantum-chemical programs).

Many molecular descriptors can be calculated easily using software packages such as DRAGON[5,6] and CODESSA.[7] An important class of structural descriptor is three-dimensional property fields, a type of 3D descriptor (describes the way properties of molecules are distributed in three dimensions). These parameters are generated by calculating interaction energies of probe atoms at grid points surrounding a molecule. The property values at grid points around molecules constitute descriptors that capture how the property is distributed in space. The most commonly used molecular field descriptors are calculated using the CoMFA and CoMSIA approaches.[8,9] These 3D descriptor methods require molecules to be aligned in a consistent way in space, and conformational properties of molecules (their flexibility and distribution of 3D shapes) are usually important. Several materials QSPR models described in section 5 have used molecular field property descriptors.

The correct choice of descriptors will clearly have a large impact on the quality of the predictions the model can make. It will also impact on the ability of the model to elucidate the relationships between molecular or other microscopic of physical properties of the material and the useful property being modeled. If effective descriptors can be found, particularly if a sparse subset of them can be identified that model the property of interest well, they can sometimes identify how these microscopic or property features can be changed to improve the property of interest. It is likely that new descriptors will need to be developed for some materials.

2.2. QSPR Modeling Methods

Almost all QSPR modeling methods involve some sort of regression.[10] This can be a simple least-squares, multiple linear regression (MLR) or, where the structure−property relationship is not linear, a polynomial, bilinear, kernel, or neural network method. In some cases, particularly when the property being modeled is a category or class (e.g., strong, moderate, or weak) rather than a continuous variable, other kernel-based methods such as support vector machines can be employed, and decision trees can often provide models that are readily interpretable.

The simplest QSPR modeling method is known as multiple linear regression,[11] and it can be carried out in common spreadsheet programs. It assumes that the property being modeled is a linear function of the descriptors.

$$y_i = w_0 + w_1x_{i1} + w_2x_{i2} + w_3x_{i3} + \ldots \tag{1}$$

where y_i is the property, x_i are the descriptors, and w are fitted coefficients (derived using the least-squares criterion) for the ith material. In matrix notation, this can be written as

$$\mathbf{y} = \mathbf{Xw} \tag{2}$$

This can be extended to polynomial regression.

$$\mathrm{y_i = w_0 + w_1x_{i1} + w_2x_{i1}{}^2 + w_3x_{i1}{}^3 + w_4x_{i2} + w_5x_{i2}{}^2 + w_6x_{i2}{}^3\ldots} \tag{3}$$

where polynomials of any order can be used, and the w are again fitted coefficients derived using the least-squares criterion. Additional cross terms could also be added. However, polynomial regression requires subjective decisions to be made about the largest polynomial order and number of cross-terms to be used.

Because the number of descriptors may be much larger than the size of the data set, MLR models can often be overfitted. Overfitting generates many different equivalent models that use different combinations of descriptors, none of which can reliably predict properties of new data. Linear methods such as partial least-squares (PLS) and principle component regression (PCR) can be used to reduce the size of descriptor space and generate models that are not overfitted. A linear regression model is found by projecting the predicted variables and the observable variables to a new space that is described by latent variables called principal components (PCs). The three regression techniques differ in that MLR aims to achieve the maximum correlation between the X and Y, PCR captures maximum variance in the X, and PLS tries to do both by maximizing covariance between the X and Y. In the context of QSPR modeling, variance is defined as the average of the sum of the squares of the difference between the observed and predicted values. The standard deviation is the square root of the variance.

Linear regression methods can also use a *kernel trick* to convert a nonlinear problem into a linear one. This is achieved by fitting a linear combination of nonlinear kernel functions such as Gaussians to the data. A very useful classification or regression method that exploits the kernel trick is the support vector machine (SVM) first introduced by Cortes and Vapnik.[12] Support vector machines construct a set of hyperplanes in a high-dimensional space, where the data points can be linearly separated.

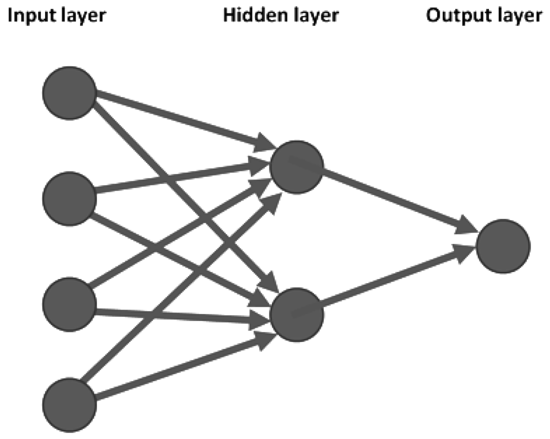


Figure 3. Structure of a simple neural network.

Decision trees (also known as classification trees or regression trees) are classifiers that map observations about an object to conclusions about the object's properties. In these trees, leaves represent classes and branches represent decisions on features that lead to those classifications.[13] Decision trees can be more interpretable than neural networks, and support vector machines and may give insight into the relative importance of the variables.

Artificial neural networks (ANNs)[14,15] (Figure 3) are a more versatile method of modeling nonlinear QSPR relationships than polynomial regression. They are machine learning methods that can model essentially any complex relationship given sufficient data.

The most commonly used neural network is the feed-forward network. It consists of an input layer to receive the input data, one or more hidden layers that perform nonlinear computation, and one or more output layers that generate the value of the response variable or property. The architecture of neural networks is often described by the notation $N_{in}-N_{hidden}-N_{out}$ (e.g., 9−3−1). Nodes in successive layers are connected to each other by weights. The inputs to the hidden and output nodes consist of the sum of the value of each input variable multipled by the weight of all connections leading into the node. This sum is presented to a transfer function (usually linear for input and output nodes and sigmoidal for hidden-layer nodes) to generate the output for that node. When used for QSPR modeling, the descriptors are presented at the input layer and the number of nodes in that layer is the same as the number of descriptors (plus a bias node that essentially plays the role of a constant in the regression). The output layer usually contains a single node corresponding to the property to be modeled, although some examples are given in this review where multiple output nodes have been used to model several materials properties simultaneously. A single hidden layer consisting of several nodes is most commonly employed, although some examples have used multiple hidden layers (but see the section discussing pitfalls). To train a neural network, a data set consisting of a set of descriptors and a measured property to be modeled are partitioned into training, validation, and test sets. When training a neural network, descriptors for each material in the training set are presented to the neural network and a corresponding desired or target response set at the output. The difference between the measured value of the material property and the system output generates an error that is fed backward through the network to adjust the weights and minimize the error in the output. All materials in the training set are presented repeatedly to the neural network until the prediction performance is acceptable. The neural network model is used to predict the properties of the validation set during training, and when the error in this set reaches a minimum, training is stopped.

Standard neural network methods such as fully connected feed-forward networks can still overtrain or overfit data. These problems can be largely overcome when the neural network training includes a regularization step controlled by Bayesian statistics, such as the BRANN[16] network. Regularization can be applied to any statistical modeling method to control the complexity (the degree of nonlinearity) of models and improve their ability to predict the properties of new molecules or materials. Bayesian regularization is able to find the optimum complexity for the neural network model and to define an objective criterion for stopping neural network training so that overtraining does not occur. It is not necessary to keep data aside to form a validation set.

As with the extension of the MLR method by principal components analysis (PCA) to reduce the number of descriptors, so the BRANN method can be extended (BRANNLP,[17]) by incorporating a Laplacian prior that allows some of the neural network weights to be set identically to zero. If all of the weights associated with a given index are set to zero, then the index itself can be eliminated from the model. Some QSAR/QSPR examples that use this approach are described by Burden and Winkler[18] and Tarasova et al.[19]

3. ASSESSING THE PERFORMANCE OF MODELS

The best measure of the usefulness of a model is how well it predicts the properties of materials that were not used to generate it. In practice, this ideal is rarely possible but is approximated by partitioning the available training data into a training set and test set. The training set is used to develop the model, and the test set is used to estimate how well the model predicts unseen data. Statistical criteria are used to assess the quality of QSPR models. Commonly, these are the coefficient of determination and the standard error of estimation or prediction. The coefficient of determination (r^2) is the square of the correlation coefficient (r) between the predicted and measured values of the property being modeled. It describes the proportion of variability in a data set that is accounted for by the statistical model and provides a measure of how well the model can predict new outcomes. The root-mean-square (rms) values, adjusted for degrees of freedom, of the difference between the predicted and measured property values for the training and test sets is called the standard error of estimation (SEE) and standard error of prediction (SEP), respectively. SEE and SEP are more robust estimates of the predictive ability of models because, unlike r^2 and other statistical measures like the F-value, they do not depend on the number of data points in the training set or the number of descriptors in the model. Good QSPR models have r^2 values close to 1.0 and SEE and SEP values that are similar and small. Although the statistics of prediction of an independent external test set provide the best estimate of the performance of a model, cross-validation methods are often also reported. This involves omitting one (leave-one-out, LOO) or more (leave-many-out) data points from the training set, generating a QSPR model using the remaining data points, then predicting the properties of the data point(s) omitted. Each data point or set of data points is omitted in turn, and when all data points have been omitted at least once, the statistics of the predictivity are compiled. It has been shown that LOO methods in particular give an overly optimistic estimate of the predictivity of models.[20] This important point is discussed further in the next section.

4. COMMON QSPR MODELING ERRORS AND PITFALLS

Although QSPR methods are relatively simple to understand and implement, it is also relatively easy for inexperienced modellers to encounter pitfalls that can weaken their models. Some of these are discussed in the recent review by Scior et al.[21] and are summarized below.

4.1. Uninformative Descriptors

To successfully model materials properties, it is essential to understand how to represent the microscopic (usually molecular) properties of the material mathematically as descriptors. Clearly, descriptors must contain information relevant to the property being modeled. In the case of polymeric materials, the selection of an appropriate structural motif for which to generate descriptors is also important and often not intuitive. However, as illustrated below, many polymer properties can be modeled using the structure of the monomer or repeating unit. Once descriptors have been generated, they may be uninformative for two main reasons. First, they may not contain sufficient relevant information, making the construction of a useful model impossible. Usually, uninformative descriptors of this type are not problematic as it is clear when the model is poor. Second, they may contain information relevant to the property being modeled but may be obscure or arcane microscopic properties derived from quantum-chemical calculations or topological properties of the material's components. Although such descriptors can generate successful and useful models, it is hard to understand how the microscopic properties influence the macroscopic (measured) properties in a mechanistic way. This also makes it hard to "reverse engineer" the model to optimize materials directly. However, this can be done very successfully by generating large virtual libraries of materials and using the model to predict properties of, and select suitable materials from, the library. The effect of the domain of applicability (the descriptor and property space of materials used to generate the model) on the accuracy of the predictions must also be considered. Poor descriptor choice may also be more likely when modeling materials properties because relatively few materials-specific descriptors are available. Many of those used in current QSPR models have been adapted from QSAR studies of the biological and physicochemical properties of small organic molecules and may not be completely suitable for materials modeling.

4.2. Overfitting and Grossly Underdetermined Systems

QSPR modeling is a supervised statistical method. Models are generated from independent variables that represent the molecular or microscopic properties of the material and the dependent variable, a measured material property. Like any regression technique, QSPR models can suffer from overfitting. This occurs when the number of adjustable parameters in the model (e.g., coefficients in an MLR model, or weights in a neural network) exceeds the number of data points available to be modeled. In statistical terms, these are called grossly underdetermined systems. The result is that the model becomes very good at predicting the training data and very bad at predicting new data not used in training (that is, the ability of the model to predict the properties of new materials falls to zero). Parsimonious models (those with a relatively simple structure and small number of descriptors and fitted variables) generally have higher predictive power than more complex models and are preferred. Overfitting is not difficult to detect. Estimating the number of fitted variables in the model and ensuring they do not exceed roughly half of the number of data points is a useful rule of thumb. The statistics from the models can also provide warning that overfitting is a problem. If the statistical parameters (e.g., r^2 and standard error) for the training set and independent test set are similar, the model is probably valid. If the training set statistics are very good (high r^2 and low standard error) and substantially different to those of the test set, overfitting should be suspected.

4.3. Descriptor Selection and Chance Correlations

It is possible to generate thousands of descriptors for a given molecular structure. To avoid overfitting a QSPR model, the number of descriptors must be reduced. The most common way to do this employs PCA, a means of generating a smaller number of orthogonal latent variables from a larger number of descriptors. This can be very effective but suffers from being more difficult to interpret than individual descriptors, and in not eliminating descriptors that are truly uninformative and therefore contribute noise to the model. A common error made by QSAR and QSPR modellers is incorrectly sampling the large pool of possible descriptors to generate a large number of smaller subsets of descriptors. This is sometimes performed automatically by evolutionary methods like genetic algorithms. Topliss[22,23] and others showed that this could generate spurious models that have deceptively good statistical quality. Generating a large set of random numbers as descriptors and then choosing a large number of smaller subsets for QSPR models can often generate apparently good models. Topliss described how selection of smaller subsets of descriptors can be achieved while minimizing chance correlations. Good statistically sound methods of "feature detection" or descriptor selection have been developed. A review of variable reduction methods is given by Livingstone and Salt.[24] Simple modeling methods such as principal component regression (PCR) and partial least squares (PLS)[11] are often used in these circumstances (see paper by Taylor et al.[25] for an example of the application of PLS to modeling properties of a large combinatorial polymer library). PCR and PLS reduce the dimensionality of the space, often considerably, by generating a new set of orthogonal descriptors that are linear combinations of the original descriptors. However, as with PCA, this can make interpretation of models more difficult. Ideally, a variable-reduction (feature-selection) method should completely remove uninformative descriptors and retain only those that are relevant to the problem under study, to make the interpretation of the model easier.

Sparse feature-detection methods employing novel mathematical techniques have been discovered for handling grossly underdetermined systems.[17,18] A method that removes unnecessary variables by reducing their weights to zero makes use of Bayesian statistics and a Laplacian prior which minimizes the modulus $\sum_{i=1}^{n}|\hat{y}_i - y_i|$ rather than the square $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ of the error in y that is commonly employed in the ordinary least-squares method. This method, known as multiple linear regression with expectation maximization (MLREM), has been successfully applied to QSAR/QSPR modeling.[18,19] The aim of feature selection is to choose the smallest number of the most informative descriptors in a way that is sensitive to the context of the model (i.e., the property being modeled).

4.4. Modeling Complex, Nonlinear Structure−Property Relationships

In many cases, simple linear statistical methods like multiple linear regression can generate good models. This is possible when the relationship between the microscopic or structural properties of the material (as exemplified by descriptors) and the property being modeled is approximately linear and additive. Whether or not the relationship is linear will depend on both the property being modeled and the type and relevance of the descriptors employed. In a substantial number of cases, the structure−property relationship is nonlinear. Polynomial regression methods, nonlinear kernel methods, and neural networks are then the methods of choice for QSPR modeling. Properly applied, neural networks are very useful because they do not rely on subjective decisions on the complexity of the polynomial or type of kernel function to be made, as they are universal approximators. This means they can model any continuous function to any level of accuracy given sufficient training data. However, the most widely used neural network type, the feed-forward back-propagation neural network, can also generate poor models if care is not taken.

As discussed above, neural networks can overfit models if the number of adjustable weights in the network exceeds the number of data points available. This is relatively easy to detect. Most back-propagation neural networks can also be overtrained, where they become better and better at predicting (memorizing) the training data and worse at predicting new data. Traditionally this is avoided by partitioning some of the data into an additional validation set not used in training or testing. While the error in the training set continues to become smaller as training continues, the error in the validation set drops to a minimum then increases when overtraining begins. Training is stopped when the validation set error is minimum. While this is a useful technique, it requires further partitioning of the data (which may be scarce or expensive to generate) from the model. Newer, robust neural networks like the Bayesian regularized neural network[16,17] use Bayesian evidence to stop training and do not require a validation data set. Another issue with neural networks is that they require an architecture to be defined—that is, how many hidden layers and how many nodes per hidden layer. If too many hidden layers and nodes are used, the number of weights in the network also increases, making overfitting more likely and compromising the predictivity of the model. Usually a single hidden layer is used, and the number of nodes in this layer is set by trial and error to find the model with the best ability to predict data not used to generate it. The choice of transfer function incorporated in the nodes is also important. For regression models, linear transfer functions are used for input and output layers, and nonlinear (sigmoidal) functions are used for the hidden layer(s). For classification models, a sigmoidal transfer function is used for the output layer nodes as well. The newer Bayesian neural networks[16,17] also eliminate the problem of optimizing the architecture of the neural network as they are relatively insensitive to the network architecture and automatically optimize the model complexity and predictivity. Applying Bayesian regularization to a back-propagation neural network optimizes the balance between bias (model is too simple to capture the underlying QSPR relationship) and variance (model is fitting the noise as well as the underlying QSPR relationship)

4.5. Validating QSPR Models

If all available data is used in training the model, it not possible to estimate how predictive the model is when applied to new data not used in training. Two main methods are used to estimate model predictivity. Cross-validation methods involve leaving one or more data points out of the training data, building the model, and predicting the property for the omitted data. This is done multiple times until all data in the set has appeared in the training and cross-validation sets at least once. Although widely used, as previously mentioned this method has been shown to give an overly optimistic estimate of the predictive power of the model.[26] The gold standard for estimating model predictive power is to predict properties for new data or materials that were not used in building the model. This situation can be approximated by partitioning the data set into training and test sets. It is not clear that there is a single, best way to partition the data into training and test sets, and this issue is still debated by the QSPR modeling community. If the partitioning is only done once, the training and test sets are almost independent, especially if the portioning is done by random selection. This provides the least optimistic, but arguably most realistic, estimate of the ability of the model to make new predictions. However, other QSAR modellers stress the need to make multiple partitions of training and test sets so that better statistical reliability can be achieved. However, in this case the independence of the training and test sets is reduced, as compounds in the data set appear more than once in both training and test sets. A method that improves statistical reliability (over random selection) but chooses the test set only once is to select the test set by a supervised clustering technique. This reduces the independence of training and test sets, but less so than LOO and multiple test set selection methods. For large data sets, random selection of the test set usually works well. Partitioning of small data sets into training and test sets is problematic as this leaves very few points in the test set, and different random selections can generate widely differing estimates of model predictivity. There is no ideal solution, and it may be argued that very small data sets should not be modeled. However, choosing the test set to be representative of the training set using a clustering method, or multiple random selections of the test set and combining the results of predictions, are workable approximations, albeit also slightly optimistic in their estimates of model predictivity. Kubinyi et al.[27] and Golbraikh and Tropsha[28] have shown that the estimates of model predictivity from leave-one-out (LOO) cross-validation and external test sets do not correlate significantly. Therefore, the use of an independent external test set not used in training is preferred when possible.

4.6. Domain of Applicability of Models

Models are trained using data that have a defined range of property values and a specific range of molecule types that constitute the domain of applicability of the model. Like any form of extrapolation, predicting materials properties outside of the molecular or property space used to develop the model must be done with care. Molecular similarity based on descriptors used to develop the model can be a guide to how far outside the model domain the prediction lies. Some probabilistic modeling methods such as Bayesian regularized neural networks that generate a distribution of weights rather than a single set of weights can also be used to estimate the domain of applicability of the model. As models are likely to be used to predict properties of larger virtual libraries of materials, the extent to which the library encompasses the domain of applicability of the model needs to be carefully considered. Done carefully, such analysis can yield useful information on the likely reliability of a prediction, depending on how far from the domain of applicability a given library member, or new material proposed for synthesis, lies. This important issue has been reviewed for the very similar quantitative structure−activity relationship (QSAR) method.[29−32]

4.7. Incorrect Handling of Outliers

Models can sometimes contain one or more data points that are poorly predicted, although the remainder of the data set is predicted well. Care must be taken when excluding these outliers. Sometimes the data point has been incorrectly measured or transcribed and when the property is measured again it conforms to the prediction. However, the presence of outliers can also indicate that the model is not capturing some important attribute of the material. This may be due to the outlier being the only material in the set with a specific attribute (e.g., a particular type of functional group or atom), or it may be that an important microscopic property of the material has not been accounted for in the model and that the outlier represents an extreme point for this property. Reexamination of the parameters used to train the model, and the structure of the outlier material, can often identify useful information relevant to the property being predicted that has been overlooked in the modeling process. It is not acceptable to exclude outliers simply because they "do not fit".

5. MATERIALS CLASSES

We summarize the QSPR models of properties of materials. The range of properties and material types modeled to data are relatively limited, but the results show that the QSPR methods have great potential to model a wider range of materials and properties. However, some of the studies reported have substantial weaknesses in implementation, and these have been critically evaluated where appropriate. In almost all studies, as is standard practice in QSAR/QSPR modeling, the materials property being modeled is converted into the logarithm of the property. This is largely necessary because many properties span several orders of magnitude in the data set. However, the resulting log−log plots of predicted versus measured properties may give a misleading visual impression of quality of the model. We have included expanded descriptions of some of the studies to illustrate more clearly how different types of QSPR approaches are used.

5.1. Nanomaterials

Nanomaterials are defined as structures with at least one dimension of 100 nm or less. They have attracted considerable interest since they were discovered in the 1990s due to their novel properties that make them valuable as catalysts and semiconductors and in space technology, cosmetics, environmental engineering, medicine, and pharmacy.[33,34] Factors such as the small size, large surface area-to-volume ratio, shape, chemical composition, and surface structure contribute to the unusual physicochemical and biological properties of nanomaterials.

5.1.1. Fullerenes and Nanotubes. Most of the reports on QSPR models for nanomaterials focus on the solubilities of fullerenes, C_{70} and especially C_{60}, because they were the first fullerenes discovered and there is considerable interest in their reactions and properties. In early reports, only a small number of descriptors such as surface area, refractive index, polarity, and polarizability were used to build QSPR models. Later reports employed a large number of descriptors that included constitutional, topological, geometrical, electrostatic, and quantum-chemical because of their ease of calculation using CODESSA[35] or DRAGON[5] and quantum-chemical software. Different approaches such as MLR, heuristic, least-squares support vector machine (LSSVM), and neural networks were applied, but, unfortunately, the predictive ability of most of the models was not reported.

The first semiquantitative fullerene structure−property model was reported by Ruoff et al.[36] They modeled the solubility of C_{60} in 47 solvents using diverse solvent parameters such as polarizability, polarity, molecular size, and Hildebrand solubility. It was found that no single parameter could adequately model the solubility in all solvents, and they suggested that a combination of solvent parameters would be more successful. Hildebrand solubility parameters for 7 alcohols were also used by Heymann[37] to model the solubilities of C_{60} and C_{70} . They estimated the solubilities of the fullerenes in water with an uncertainty of 1 order of magnitude by extrapolation of the alcohol models.

The solubilities of C_{60} were also the subject of a QSPR study by Murray et al.[38] Twenty-two organic solvents were studied in this work. Geometries were optimized using quantum-chemical calculations, and several descriptors related to the electrostatic potential on the solvent surface were calculated. A complex, nonlinear model describing the relationship between solubilities of C_{60} and these solvent descriptors was obtained. The model had an r^2 value of 0.91 and could predict the solubilities of the data set within a factor of 3. The model showed that solubility was increased by large solvent molecules and by moderately attractive interactions that are relatively balanced between positive and negative regions on the solute and solvent molecular surfaces. However, the predictive ability of this model was not reported.

Stepwise linear regression was first used by Marcus[39] to obtain the QSPR models for solubilities of C_{60} and SF_6, both of which are globular, large, and nonpolar and interact with neighboring molecules by dispersion forces. Different sets of solvent property descriptors were found to describe the solubilities of these two substances in organic and inorganic solvents.

The QSPR models for the solubility of C_{60} at 298 and 303 K were

$$\begin{aligned}\log x_2 = & -4.58(\pm 0.59) + 7.72(\pm 0.84) \times 10^{-2}R \\ & + 2.53(\pm 0.30)\pi^* - 8.05(\pm 1.27) \\ & \times 10^{-2}E_{\mathrm{T}}(30) - 0.190(\pm 0.080)\mu \\ n = & \ 55, r^2 = 0.86, F_{4,48} = 79, s = 0.40\end{aligned} \tag{4}$$

$$\begin{aligned}\log x_2 = & \ 3.59(\pm 3.31) + 3.46(\pm 1.08) \times 10^{-2}R \\ & + 5.10(\pm 0.91)\pi^* - 31.7(\pm 10.2) \times 10^{-2}E_{\mathrm{T}}(30) \\ n = & \ 20, r^2 = 0.72, \quad F_{3,17} = 17, \quad s = 0.52\end{aligned} \tag{5}$$

where x_2 is the mole fraction of C_{60} in saturated solution, R is the molar refractivity, π^* is the polarity/polarizability, $E_{\mathrm{T}}(30)$ is the polarity index, and μ is the dipole moment of the solvent. Several compounds that did not fit the model were removed as outliers. For C_{60}, solvent polarizability increases the solubility, while the dipole moment reduces it. The different signs for the constants in the models at 298 and 303 K suggest the models lack robustness, probably due to lack of diversity in the solvents used to obtain the 303 K model. No assessment of the ability of the model to predict solubility of C_{60} in new solvents was made. In a later study, this research group generated a QSPR model of C_{60} solubilities in a larger number of solvents.[40] The linear regression models described the log of the solubility of C_{60} in terms of four parameters—molar volume, molar refractivity, a parameter β that is closely related to Abraham's hydrogen bond basicity parameter, and a complex parameter related to the hydrogen bond donor ability (for protic solvents) and solvent polarity/polarizability.[40] Training set r^2 values of 0.99 and 0.98 were obtained for the four-descriptor models of these data at 298 and 303 K, and the QSPR equations were more similar than in the first study, showing improved robustness. However, 18 of the 113 solvents were excluded from the model as outliers simply because they were not well-predicted by the model (prediction errors larger than 2 standard deviations). The model was used to predict the solubilities of C_{60} in a range of polar solvents. Using selected data from this data set in combination with other data, a QSPR model was built by Makitra et al.[41] that described the solubilities of fullerene C_{60}. Descriptors for three properties were significant in the model: polarizability, polarity, and cohesive energy density. Linear models were obtained with $r^2 = 0.71$ for 89 solvents. When some solvents were excluded, the following three-descriptor model was obtained

$$\begin{aligned} \log X = & -11.1 + 30.1(\pm 1.44)f(n) \\ & -(2.75 \pm 0.78)f(\varepsilon) - 1.94(\pm 0.68) \times 10^{-3}\delta^2 \\ N = & \ 81, r^2 = 0.87, s = 0.467. \end{aligned} \tag{6}$$

where $f(n)$ is the polarizability, $f(\varepsilon)$ is the polarity, and δ^2 is Hildebrand's solubility parameter. Abraham et al.[42] reported an indirect QSPR-based method for predicting the solubilities of C_{60} in 20 solvents using five solute descriptors relating to solvent hydrogen bond acidity and basicity, volume, polarizability, and molar refractivity. The authors did not report model statistics except for standard deviation between the observed and calculated log solubilities, 0.36 log S (i.e., the model predicted solubility to within a factor of 2.3). Apart from the solubilities, other physicochemical properties such as excess molar refraction, dipolarity, air—water partition, blood—brain barrier distribution, and lipophilicity for C_{60} were also estimated.

Linear regression was also used by Sivaraman et al.[43] to model the solubility of C_{60} in 75 organic solvents. Three types of descriptors were used: topological indices such as Randic connectivity indices and Hall and Kier indices, polarizability, and indicator variables denoting the numbers of various types of atoms in the solvent, or position of substitution in aromatic solvents. QSPR models were built for subsets of solvent types as well as for combinations of these sets. Subsets often contained small numbers of solvents (five or more), but they could be modeled well by one or two parameters, usually polarizability and indicator variables. However, model validation was only done for the combination of the alkane, alkyl halide, and alcohol data sets with the training set containing 29 solvents and validation set containing 16 solvents. The training data was modeled more accurately for individual subsets of data than for combined data sets, and in most cases the best models had r^2 values of ≥0.95. Although polarizability and topological indices were the most significant parameters in most QSPR models, indicator variables (set to 1 if an specific attribute exists in the molecule, or zero if not) were essential in several solubility models. The solubility of C_{60} in a test set of 16 solvents was predicted quite reliably.

The first QSPR study of the solubilities of C_{60} using neural networks was reported 10 years ago.[44] Molecular descriptors such as molar volume, polarizability, LUMO (lowest unoccupied molecular orbital) energy, and saturated surface were calculated

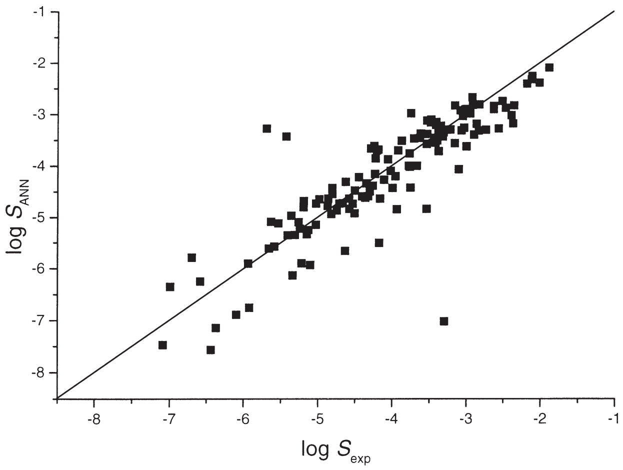


Figure 4. Ability of neural network model to predict the log of the solubility of C_{60} in 134 solvents. Adapted with permission from ref 44. Copyright 2000 American Chemical Society.

for 134 organic solvents using semiempirical quantum-chemical methods. A back-propagation neural network with five input nodes, five hidden nodes, and one output node was used to generate a solubility model having a standard deviation for training set prediction of 0.58 log S (i.e., the model could predict solubility within a factor of 4). Excluding several large outliers from the model reduced the standard deviation to 0.45 log S (the model could predict solubility within a factor of 3). However, this model was not tested on external data sets and the network architecture was relatively complex, meaning overfitting of the data cannot be excluded. The ability of the model to predict the training data is illustrated in Figure 4.

Neural networks were also used by Danauskas and Jurs to model the solubility of C_{60} in 96 solvents.[45] They divided the data into three sets, a training set (76 solvents), a cross-validation set for stopping network training (10 solvents), and an external prediction set (10 solvents). Four types of quantum-chemical and topological descriptors were used, individually and in combination. Three types of models were generated: type I employed multiple linear regression; type II used a three-layer, feed-forward neural network with the best descriptors from type I models; type III used a genetic algorithm to select descriptors and a neural network to build the QSPR model. The simple MLR type I solubility model employed 9 descriptors chosen from a reduced pool of 85 descriptors and exhibited a training set root-mean-square error (RMSE) of 0.42 log S and a test set prediction error of 0.50 log S. The best type II model, with fixed 9—3—1 network architecture, generated RMSE values of 0.30 log S for the training set, 0.45 log S for the cross-validation set, and 0.52 log S for the prediction set. The architecture of the type II ANN was optimized and an improved cost function was employed in the network training, resulting in a new ANN model with a training set RMSE of 0.26 log S, cross-validation RMSE of 0.25 log S, and test (prediction) set RMSE of 0.35 log S. The type III modeling approach offered no advantages over type II.

Liu et al. reported C_{60} solubility models developed using a novel type of support vector machine (LSSVM).[46] Five classes of molecular descriptors consisting of constitutional, topological, geometrical, electrostatic, and quantum-chemical descriptors were calculated for 128 organic solvents. A heuristic method was used to select subsets of descriptors based on variance across the data set and correlation with the log of the solubility and with each other. Linear and nonlinear solubility models were generated for 128 solvents and 122 solvents, respectively. In the nonlinear model, the data set was separated into a training set of 92 compounds and a test set of 30 compounds. Furthermore, a process of leave-one-out cross-validation of the training set was performed. The training set r^2 values were 0.76 and 0.89 and the RMSE values were 0.32 and 0.12 for the linear and nonlinear models, respectively, suggesting the relationship between descriptors and solubility was nonlinear. The LSSVM method was quite complex, with additional steps required to optimize the SVM kernel and parameters for the method.

Toropov and co-workers[47−49] created QSPR models to predict the solubilities of fullerene C_{60} in organic solvents using a Monte Carlo optimization procedure. The model descriptors were correlation weights calculated from SMILES (simplified molecular input line entry system) or InChI (international chemical identifier) text-based descriptions of organic structures.[50] The correlation weights were based on the number of occurrences of each SMILES or InChI character in the training set, and the resulting model had very good predictive performance. The first study modeled a data set of 36 substituted benzene solvents, separated into a training set of 25 compounds and a test set of 11 compounds. The training set model had r^2 values of 0.81 and 0.79 and standard errors of estimation of 3.60 and 4.67 solubility units for the training and test sets, respectively. Unlike other solubility models, and standard practice in QSPR, the authors used the solubility rather than the log of the solubility as the property being modeled. This is likely to have compromised the quality of the model as a second study by the same authors suggests. This study modeled a substantially larger data set of 122 solvents, 92 of which were included in the training set and 30 of which were in the test set. Toporov et al. again used descriptors based on SMILES characters, but this time generated models of log solubility rather than solubility. In spite of larger and more chemically diverse training and test sets, the QSPR model had r^2 values of 0.86 and 0.89 and standard errors of 0.40 and 0.44 for training and test sets, respectively. In comparison with the model built using quantum-chemical descriptors,[46] this model has higher statistical quality. A third study used the same 122 solvent data set but employed a different type of text-based description of the molecular structure, the InChI system. Descriptors based on InChI appeared to be more information-rich, as the resulting linear QSPR model achieved higher statistical quality, with r^2 values of 0.94 and 0.94 and standard errors of 0.25 and 0.35 for the training and test sets, respectively. The errors in these models correspond to uncertainties in solubility of a factor of 2.

In contrast to the above reports where the solubility of C_{60} in various solvents was modeled, the study of Martin et al.[51] considered the solubilities of C_{60} and other polyaromatic compounds in two solvents. Three hundred twenty-eight constitutional, topological and geometrical, quantum-chemical descriptors, polarizabilities, and dipole moments were calculated. The QSPR models were derived using a heuristic forward selection approach to generate the best multiple linear regression models using CODESSA package.[7] Leave-one-out and leave-50%-out validations were performed for all proposed models. For the three-descriptor QSPR model of solubility in *n*-heptane, the training, leave-one-out, and leave-50%-out validation correlation coefficients were 0.90, 0.84, and 0.82, respectively. For the three-descriptor model of solubility in 1-octanol, the training, leave-one-out, and leave-50%-out validation correlation coefficients were 0.97, 0.93, and 0.96, respectively. Significantly, the solubilities of C_{60} in *n*-heptane and 1-octanol were predicted correctly by QSPR models, even when these solvents were not present in the training set. However, the applicability of these models for predicting solubility of other fullerenes has been questioned by Puzyn et al.[33] due to the large structural difference between the spherical fullerene and the near-planar hydrocarbons (Figure 5), as well as the large difference in observed solubilities of fullerene compared to the polycyclic aromatic compounds.

The biological properties of fullerene and its derivatives were the subject of two 3D QSAR modeling studies by Durdagi and co-workers.[52,53] Fullerenes may be considered a crossover point between single molecules and materials. In the first study, the HIV-1 protease inhibitor (HIV-1 PR) activity of 49 fullerene derivatives was modeled. The data set was partitioned into a training set of 43 fullerenes and a test set of 6. The binding interactions, binding energy, and binding affinity with HIV-1 PR residues were analyzed. The optimized fullerene structures were docked into the binding site of the protease using the FlexX program, and molecular dynamics simulations were performed for ligand-free and inhibitor-bound HIV-1 PR systems to provide proper input structure of HIV-1 PR in docking simulations. The 3D QSAR molecular field-based method, comparative molecular similarity indices analysis (CoMSIA),[9] was then employed to model the protease structure−activity relationships and to predict novel compounds with improved inhibition effect. The best CoMSIA model had a cross-validated r^2 value of 0.74 and a noncross-validated r^2 value of 0.99. The CoMSIA models were able to predict the activity of the 6 fullerenes in the test set to within 1 order of magnitude. In a second study, a very similar approach was employed to model the HIV-PR activities of 20 fullerene derivatives, most of which were in the training set of the first study. These were partitioned into a training set of 17 compounds and a test set of 3 compounds. Both 3D QSAR/comparative molecular field analysis (CoMFA) and CoMSIA methods were used to derive 3D structure−activity models for the training set. The two models generated LOO cross-validated r^2 values of 0.55 and 0.56 and noncross-validated r^2 values of 0.99 and 1.00 for CoMFA and CoMSIA, respectively. Although the LOO validation statistics were modest, the CoMSIA model could adequately predict the activity of the 3 fullerenes in the test set. Toropov et al.[54] also analyzed this data set using the same SMILES-derived optimum correlation weights descriptors successfully employed in the fullerene solubility modeling described previously. The study compared the classical statistical modeling paradigm, where data is split into training and test sets, with a balance of correlations paradigm where subtraining, calibration, and test sets were used. The rationale for the latter paradigm is not clear, as the classical paradigm gave the most robust models with r^2 values of 0.75 and 0.67 and standard errors of 0.5 and 1.6 log EC_{50} for training and test sets, respectively.

Carbon nanotubes are another new class of nanoscale substances that were discovered in the early 1990s and have also drawn intensive research interest because of their fascinating structural features and properties as well as potential technological applications.[55] However, there has been only one report on a QSPR study of carbon nanotubes,[56] which unfortunately was derived from computational models of water solubility and octanol−water partition coefficients, so the QSPR model was effectively a “model of a model”, not a model of experimental data.

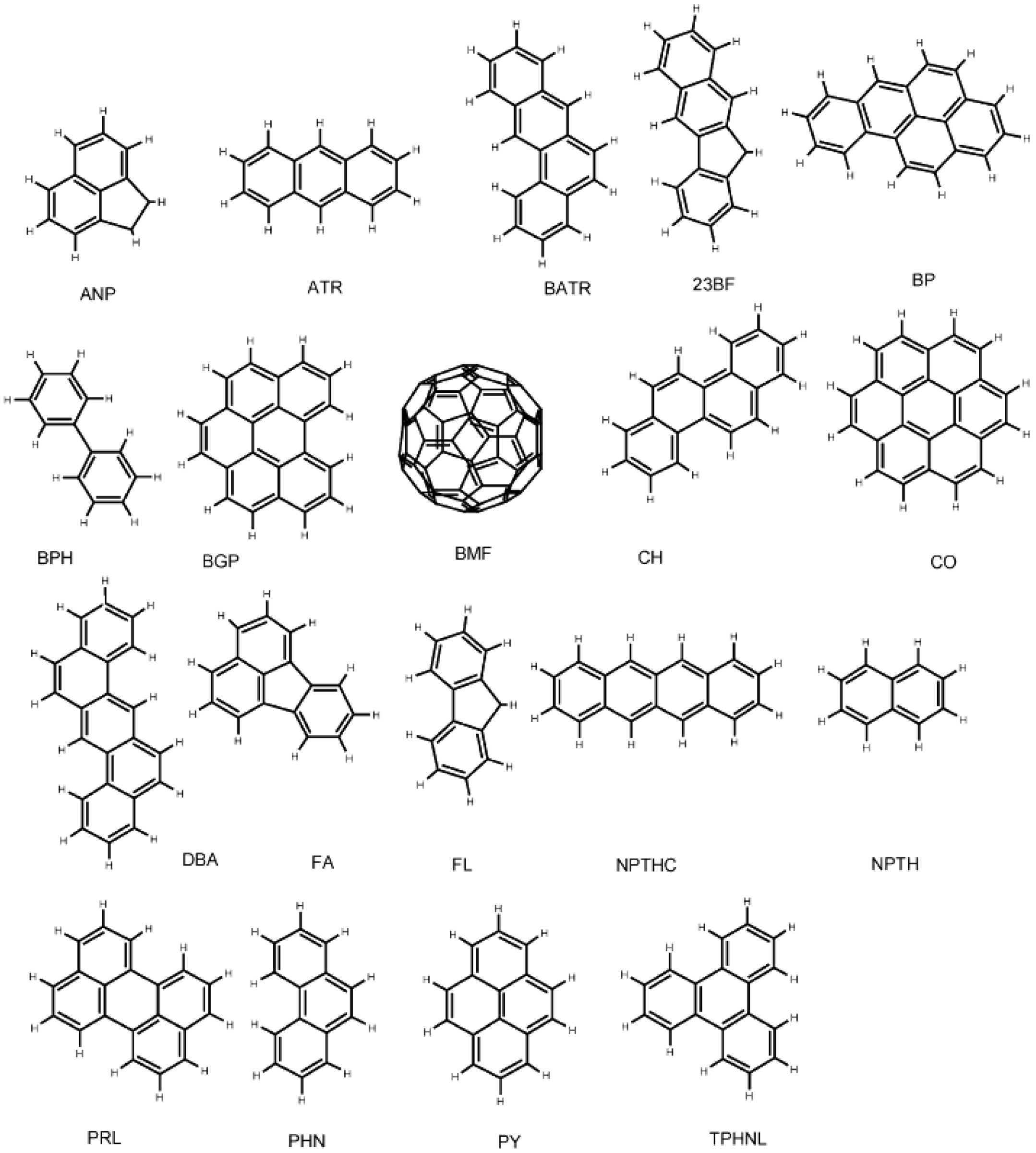


Figure 5. Molecular structures of polyaromatic hydrocarbons and fullerene studied by Martin et al.[51] Reprinted with permission from ref 51. Copyright 2007 American Chemical Society.

5.1.2. Other Nanomaterials. Toropov's group[57,58] also generated QSPR models for Young's modulus and thermal conductivity of nanomaterial data sets consisting of metal oxides, carbides, nitrides, and silicides. They again used correlation weights calculated from SMILES-like nomenclature and parameters relating to conditions of syntheses as descriptors. The QSPR models were generated using Monte Carlo and least-squares methods. The Young's modulus QSPR model was based on a data set of 29 nanomaterials, which was randomly separated into a training set of 21 and a test set of 8 nanomaterials. The model had r^2 values of 0.98 and 0.90 and standard errors of 18 GPa and 35 GPa for the training and test sets, respectively. Young's modulus of nanomaterials could be predicted with an error of ~10%. Such models are useful for predicting the properties of materials not yet synthesized. However, because of the complexity of the descriptors used, it is not clear how the model can be used to optimize properties of new nanomaterials and how the domain of applicability of the model could be estimated.

In a second example, a QSPR model of the thermal conductivity of nanomaterials was built using a data set consisting of 58 nanomaterials, 43 of which were included in the training set and 15 of which were in the test set. As with the Young's modulus model, a SMILES-like representation of the materials was used, essentially capturing the identities of the atoms making up the material, the bulk properties, and the temperature of synthesis. The statistical significance of the model was characterized by r^2 values of 0.87 and 0.86 and standard errors of 5.1 and 4.9 W/m/K for the training and test sets, respectively. Both of the above models demonstrated a linear relationship between thermal conductivity or Young's modulus and the descriptor correlation weights.

QSPR modeling has also been applied to assist the synthesis of magnetite nanoparticles in the presence of amino acids[59] to

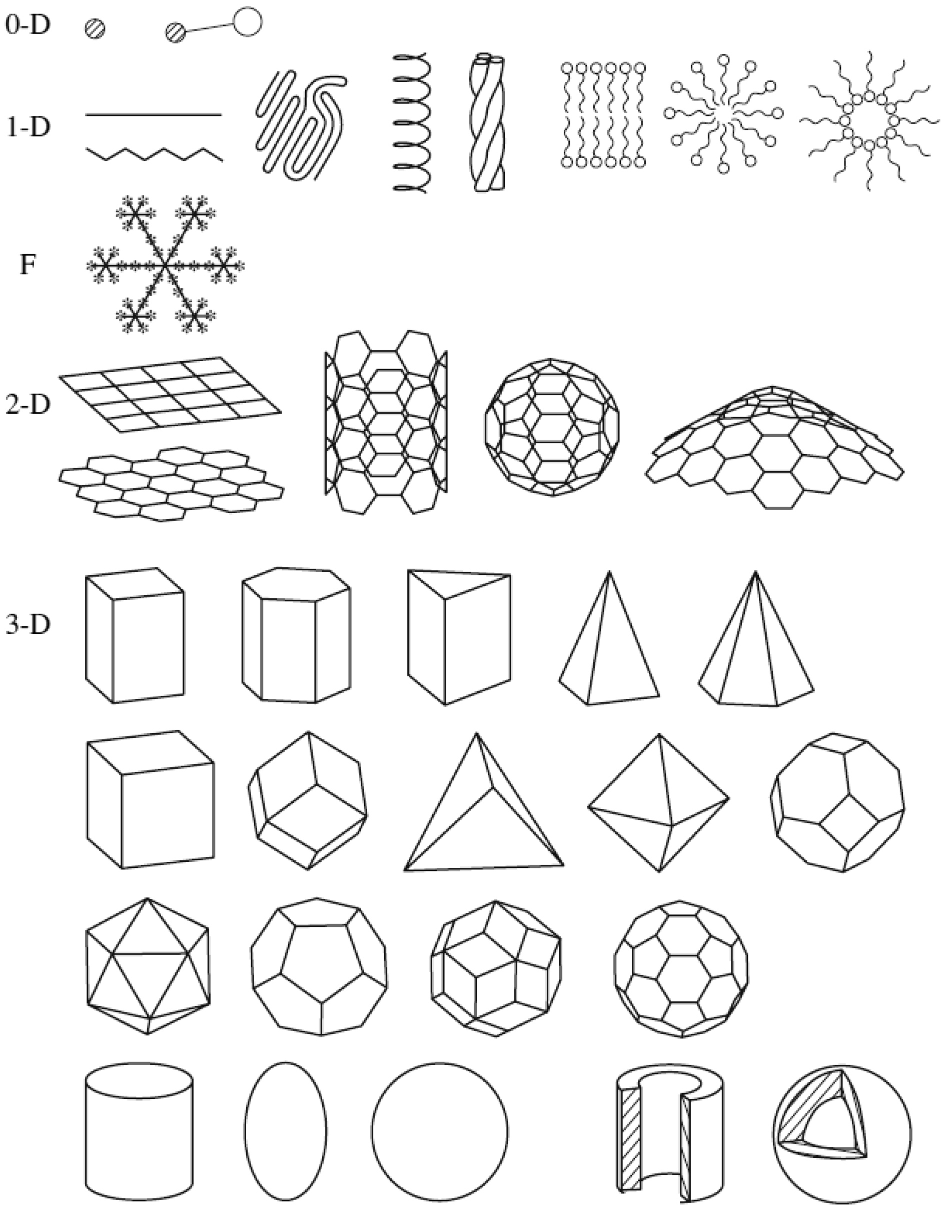


Figure 6. Structural diversity of the nanoworld: zero-dimensional (point), one-dimensional (linear), fractal, two-dimensional, and three-dimensional nanoparticle fragments. Reprinted with permission from ref 65. Copyright 2003 Maik Nauka Interperiodica.

elucidate factors that contribute to the self-assembly of amino acid/magnetite nanoparticles. Linear regression methods were used to identify parameters that correlated with the number of unit cells of magnetite contained in the nanoparticle core. The saturation magnetization was found to be influenced by molecular descriptors for electronegativity computed using the atomic electric charges from quantum-chemical ab initio calculations. The predictive ability of the model was not determined. Analysis not only showed that carbon atoms play an important role in the formation and self-assembling of the nanoparticles of amino acid/magnetite but also confirmed the existence of the chemical bonds between the oxygen atoms from the amino acid molecules in the layer and the iron ions situated at the margin of the magnetite core. However, care must be taken to not over-interpret this model, as it is almost certainly a correlative rather than causative relationship.

Although there have been a number of studies on the toxic potential of materials at the nanoscale,[34,60−63] only one QSAR/QSPR model predicting the toxicity or biological effects of nanomaterials has been reported.[64] QSAR/QSPR studies of nanomaterials generally lack a description of the domain of applicability of the model,[33] and calculation of descriptors is difficult for some classes of materials. The challenge facing development of the descriptors for these materials is partly due to their structural diversity, as illustrated in Figure 6.

5.2. Catalysts

QSAR/QSPR methods have been employed in catalyst design and modeling. These models can be used to find optimum reaction conditions, examine the effects of different factors on the catalytic reactions, create virtual catalyst libraries, design new catalysts with better performance, or extract general principles

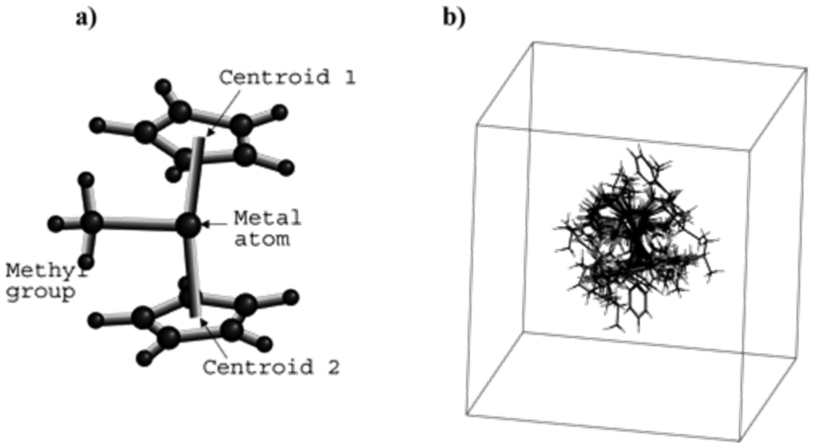


Figure 7. Structure of metallocene catalysts and alignment used to generate 3D QSPR model for polymerization activity. Reprinted with permission from ref 68. Copyright 2005 American Chemical Society.

from high-dimensional data resulting from catalysis high-throughput experimentation. QSPR methods have been applied successfully to the modeling of chemical reactions primarily involving small molecules. To keep this review focused on materials, we describe only QSPR studies that relate to complex catalyst materials or that employ simple or complex catalysts to generate polymers or other materials, rather than discrete, small organic molecules.

5.2.1. Homogeneous Catalysts. The properties of relatively diverse homogeneous catalysts have been modeled by QSPR methods. Some studies selected and designed catalysts using data obtained by high-throughput synthesis methods.

There have been a small number of reports of QSPR studies of olefin polymerization using metallocene catalysts. These catalysts possess homogeneous active sites that lead to uniformity of polymer microstructure and narrow molecular weight distribution.[66] One of the earliest attempts to generate QSPR models for metallocene catalysis was published in 1999.[66] As metallocene catalysts do not show any activity without cocatalysts, and their activity depends on the content and species of cocatalysts, five cocatalysts were also studied in this work. Molecular mechanics and molecular dynamics calculations were employed to calculate parameters such as atomic distances and charges for the most stable structures. The authors interpreted the QSPR models in terms of structural requirements for catalysis. However, the model used three parameters plus a constant, and there were only five measured catalyst activity data points (three of the catalysts being inactive), making overfitting a significant issue. Cruz and co-workers [67−69] also studied the catalytic activity of metallocene catalysts. Data sets of 7,[69] 25,[68] and 22[67] metallocene catalysts were modeled. The geometry of the active catalyst species was optimized using ab initio and density functional theory quantum-chemical calculations. The 3D molecular field QSAR method, CoMFA, was used to build a model predicting the catalytic activity of metallocene catalysts in ethylene polymerization, as well as the molecular weight of resulting polyethylenes. Models employing one principal component had the best predictive power. The steric (size and shape) properties of the catalysts contributed >90% to both of the CoMFA models for the activity and the polymer molecular weight. The cross-validated q^2 and the final noncross-validated r^2 values of all models spanned the range of 0.40−0.71 and 0.78−1.00 for polymerization activity and polymer molecular weight, respectively. The details of the 3D QSPR modeling of the 25 metallocene catalysts data set are illustrated below. Figure 7 shows the structures of the catalysts and the alignment of the data set that was used to generate the 3D QSPR models. Four types of molecular fields were used: steric, electrostatic, LUMO, and local softness.[68] The LUMO field model was the most predictive with a LOO q^2 value of 0.53 for five PCs, SEP of 11.4×10^3 kg PE (mol metallocene × h × [ethylene])$^{-1}$, and a noncross-validated r^2 value of 0.97, SEE of 5.5×10^3 kg PE (mol metallocene × h × [ethylene])$^{-1}$.

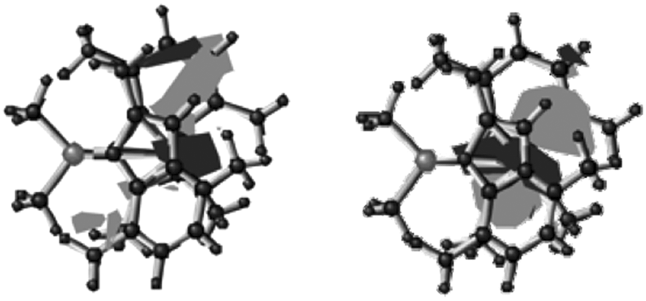


Figure 8. (Left) CoMFA maps for the LUMO field. (Right) 3D shape of the LUMO field. Reprinted with permission from ref 68. Copyright 2005 American Chemical Society.

The model showed how the spatial distribution of these molecular properties enhanced or decreased catalysis by these compounds. The light and dark gray areas in Figure 8a are the LUMO field map, showing where the LUMO enhanced polymerization activity. Those areas corresponded to the location of the LUMO orbital, illustrated in Figure 8b.[68] The models were used to successfully predict the activity of three additional catalysts within experimental error.

In a separate study, QSPR models predicting the ethene/1-hexene copolymer melting temperature were built using a data set of 11 metallocene catalysts.[70] Five structural descriptors were calculated using quantum-chemical techniques, and partial least-squares regression models were generated. The best QSPR model for 8 catalysts used two PCs and had r^2 values of 0.87 and 0.78 for training and leave-one-out cross-validation, respectively. The predicted copolymer melting temperatures for the 3 test set catalysts were in error by up to 7 K. When all catalysts were used in the model, a single PC yielded r^2 values of 0.78 and LOO q^2 values of 0.66. The properties of an additional 11 catalysts were also predicted using the model. A larger training set would be required if more accurate quantitative prediction was necessary.

Cross-coupling reactions have also been the subject of several QSPR studies. Burello et al.[71] analyzed 412 Heck cross-coupling reactions to generate a QSPR model predicting catalyst performance (turnover number and turnover frequency). The models employed descriptors, such as dipole moment, HOMO (highest occupied molecular orbital) and LUMO energies, atom charges, and structural parameters that were relevant to the Heck reaction. Descriptor correlations and PCA were used to eliminate descriptors and reduce the dimensionality of the problem to increase robustness and predictivity of the models. Linear regression, artificial neural networks, and classification tree methods were used to model the structure–property relationships. The best models had prediction confidence levels as high as 93%, but unfortunately, no statistics were provided for the models. Palladium loading was the most relevant descriptor for both turnover number and turnover frequency. The models were used to predict the performance of 60 000 combinations of virtual catalysts and reaction conditions in silico, although no consideration was given to the domain of applicability of the model in predicting properties of these combinations. an der Heiden et al. modeled rate constants for almost 500 Sonogashira cross-coupling reactions from a high-throughput study of reaction kinetics in homogeneous catalysis.[72] Density functional theory was used to compute steric and electronic parameters, which were used to build a statistical model using a goal-seeking function. The model had a training set r^2 value of 0.98. an der Heiden et al. demonstrated that parallel multisubstrate screening was useful to understand the kinetics of coupling reactions. No tests of the model predictivity were reported.

In the final organometallic catalyst example, both 2D and 3D QSPR models were built for a data set of 23 Ti−N═P organometallic ethylene polymerization catalysts.[73] The activity being modeled was defined as the amount of polymer in g/mmol/h/atm produced by each catalyst. The 3D QSPR model used the SOMFA (self-organizing molecular field analysis) molecular field-based modeling method[74] employing only steric fields. The 2D QSPR model employed diverse molecular descriptors calculated using the DRAGON software. A small set of descriptors was selected from the large pool calculated by DRAGON using an unsupervised feature selection method and a genetic algorithm. The 3D model had relatively poor statistical quality, with an r^2 value of 0.67 and a LOO cross-validated q^2 value of only 0.22, which may be due to the SOMFA method not being statistically rigorous.[75] When the catalytic performance of a test set of 5 compounds was predicted using a reduced training set, the quality of the prediction was substantially better, with the test set r^2 value exceeding 0.6 for some test set selections, worse for others. Allowing for stereoisomerism in the cyclopentadiene group improved the 3D model. The 2D QSPR model also used the same training and test sets. As the number of descriptors in the original pool (1 600) was much larger than the number of reactions in the training set (∼70), the selection of a relevant subset of descriptors for the QSPR model had to be done with care to avoid chance correlations. The best 2D catalyst activity model had r^2 values of 0.97 and 0.92 for the training and test sets, respectively. A limitation of this study was that titanium was not parametrized in the DRAGON software, so it was replaced by carbon. Performance of the model was very dependent on which compounds were chosen for the test set, and the model appeared to have low robustness. This variability was due to the small size of the data set.

5.2.2. Heterogeneous Catalysts.

QSAR/QSPR analyses for heterogeneous catalysts have also drawn considerable interest. Diverse approaches have been used to model the catalytic performance of a relatively wide range of materials. In most cases, process variables such as synthesis conditions and catalyst composition were used as descriptors in the QSPR models.

Applying the same method that was used to build the QSPR model for homogeneous metallocene catalysts,[66] Yao and Tanaka[76] studied a set of 10 heterogeneous Ziegler−Natta catalyst−external donor systems. The catalytic activity and molecular weight distribution of polymers synthesized using the catalysts were modeled. Descriptors such as the interaction energy between catalyst and external donor, distances between Ti and Si atoms, principal moment of inertia, radius of gyration, and molar refractivity were used in the model. A catalyst activity model with four parameters (a relatively large number for the size of the data set) had an r^2 value of 0.93, and a two-parameter model had an r^2 value of 0.88. The molecular weight distribution model had an r^2 value of 0.88 for a four-parameter model and 0.75 for the more parsimonious model with a single parameter. The main factors influencing the catalyst activity were the interaction between the active site and the external donor and the polarity of the external donor. The main factor influencing the molecular weight distribution was the principal moment of inertia of the external donor. The small size of the data set and limited dynamic range in the properties being modeled were limitations.

Neural networks have been used widely for designing heterogeneous catalysts. In one of the earliest studies,[77] the acid strengths of mixed oxides, the activity of a series of lanthanide oxides in catalyzing the oxidation of butane, and the selectivities toward various products in the oxidative dehydrogenation of ethylbenzene on promoted SnO_2 catalysts were modeled. Although little statistical information was provided in the paper and the details of the neural network architecture used were not provided, the results suggested that neural networks were able to make useful predictions of these properties.

Neural networks were also used to model the Cu/ZSM-5 zeolite-catalyzed NO decomposition reaction. Reaction conditions such as temperature, Cu loading, O_2 concentration, NO concentration, and SiO_2/Al_2O_3 mole ratio were used as descriptors.[78] No statistical results were presented, and the models appeared to make good predictions of the training set data. However, the complexity of the neural network used (in particular, the number of weights) relative to the number of experiments and the long neural net training time suggests the models may have been overtrained and overfitted. However, the conversion was predicted for one set of conditions not used in training with good accuracy.

Hou et al.[79] applied the same technique to design VSbWSn (P, K, Cr, Mo)/SIAL catalysts for acrylonitrile synthesis from propane using a training set of 19 and a test set of 4 catalysts. The conversion of propane and the selectivity of acrylonitrile were modeled using a neural network. Catalyst components were represented as atom number fractions of P, K, Cr, Mo, and V and the weight ratio of $A1_20_3/SiO_2$. In spite of a very long neural network training time (which can result in overtraining) and the large number of weights in the neural network relative to the size of the training set (which can result in overfitting), the performance of the model on the training and tests sets was good. The model was used to optimize catalyst performance. However, the dynamic range of all variables in the model was quite small, suggesting that it may not make reliable predictions of catalysts outside the property space (domain of applicability) of the training set.

Huang et al.[80] modeled catalytic oxidative coupling of methane using back-propagation neural networks. The aim was to predict the C_2 selectivity and the conversion of methane based on the catalyst components. The training set comprised 25 catalysts, and the test set comprised 8 catalysts. The descriptors used in the model were mol % of elements in the catalyst. A number of relatively complex four-layer neural networks were used to develop the QSPR models. The ability of all network architectures to predict the training set was excellent despite the number of network weights exceeding the number of data points in the training set. The dynamic range of experimental data placed a limit on the generalization ability of the model. In a later related study,[81] the same method was used to predict oxidative dehydrogenation of ethane by 50 catalysts. Forty catalysts were used in the training set, and 10 were used in the test set. The molar composition of 13 elements present in the catalysts was again used as descriptors for QSPR models that predicted the mole fractions of 6 reaction products. As with previous neural network studies, the architecture of the network was overly complex, and the ability of the models to accurately predict the test set was modest. The models were clearly overtrained because the number of weights was much larger than the number of data points; the training set was quite well-predicted, but the test set was less so. Consequently, simpler neural network architectures are likely to make substantially better predictions of this data set.

Moliner et al.[82] reported a high-throughput synthesis study of zeolites using factorial design. Larger data sets, such as those generated by high-throughput experiments, are ideally suited to QSPR modeling. The data for this study were generated by a $3^2 \times 4^2$ factorial design, so they consisted of 144 points. These data were partitioned into a training set (100 samples), neural net validation set (used as a stopping criterion for the neural net training, 20 samples), and a test set (24 samples). The inputs to the neural network were the concentrations of reagents used in the synthesis, and the properties being modeled were the crystallinity and relative amounts of two different zeolite phases. They investigated a range of neural network topologies, ranging from sparse, three-layer networks to more complex four-layer networks. The sparsest models, with the least complex neural network architecture, were able to predict the crystallinity of the two phases at least as accurately as the more complex network architectures. The best QSPR models could predict the training set percent crystallinity to within 5% and the test set crystallinity to within 10%. These studies showed that neural networks show considerable promise for accelerating development and optimization of new catalysts.

Neural networks have been also shown to give similar or better prediction than classification/decision trees and support vector machines for some data sets. A highly diverse data set of 467 catalysts used to oxidize propene with oxygen was used to compare the modeling ability of neural networks and classification trees.[83,84] Three thousand one hundred seventy-nine attributes were measured or calculated for the catalytic oxidation experiments. These included the concentrations of elements, methods of synthesis, enthalpies of formation, coordination numbers, ionization energies, electronegativities, etc. Feature selection based on chemical intuition rather than algorithms was used to reduce this large number of attributes to 75. The large data set was partitioned into training, validation, and test sets (50%, 25%, 25%) for the neural network models and training and test sets (67%, 33%) for the classification trees. A neural network-based clustering method further reduced the number of attributes to 45, and the classification trees algorithm reduced the number of attributes to 23. The neural network consistently, and sometimes markedly, outperformed both the classification tree and a standard statistical method in spite of the very high complexity of the neural network. Contingency tables were the only statistical properties reported for the models.

Neural networks and decision trees were also compared for their abilities to model zeolite crystallinity for 144 samples, using synthesis conditions and X-ray diffraction data as descriptors.[85] Principal components analysis and *k*-means clustering techniques were employed to analyze data and for dimensional reduction. The models built using neural networks with two hidden layers and a back-propagation training algorithm yielded good predictive models, with an r^2 value of 0.92 for both ITQ-21 and ITQ-30 catalysts. However, although data were held aside in a test set, the statistics for prediction of these data were not reported. Given the complexity of the network used relative to the data set size, the models may have been overfitted. It has also been shown that the predictive ability of the neural network model depends on the choice of descriptors.[86]

Support vector machine classification methods were applied to heterogeneous catalysis by Baumes et al.[87] Two data sets were considered. The first one used data for 26 olefin epoxidation catalysts to build QSPR models predicting the yield of cyclohexene epoxide. The second data set of 24 isomerization catalysts was used to model isomerization yield for light paraffins. Numeric data on epoxidation and isomerization were converted into two categories using a threshold. The inputs to these models were the molar concentrations of several components of the starting gel. The support vector machine models were shown to be superior to those generated using classification trees, with recognition rates for the two reactions as high as 90%. The paper also provided a very useful summary of causes of overfitting in QSPR models and how they can be overcome.

Recently, Wang et al. reported useful comparison of the ability of neural networks, support vector machines, and classification trees to model four different data catalyst sets.[13] This work suggested that, although multilayer neural networks performed strongly, none of these algorithms gave the best performance on all data sets.

Linear QSPR modeling methods were used to model Michael addition to different substrates using $ZrOCl_2$, silica gel, and sodium dodecyl sulfate catalysts.[88] The data set containing 46 reactions and 16 quantum-chemical molecular descriptors was used to generate QSPR models of the logarithm of the yield/time ratio. The models were able to predict the training set well, with r^2 values of 0.82−0.93 and test set r^2 values of 0.76−0.85. Such models would be useful for making quantitative predictions of the reactivity of related reactions not used to train the model.

Bis(imino)pyridine and bis(arylimino)pyridine iron catalysts of ethylamine polymerization were modeled in two separate QSPR studies.[89,90] Nineteen catalysts were used in the first study to generate 3D QSPR models. Molecules were aligned using a rule, and the QSAR method CoMFA and partial least-squares analysis were used to generate the model. Although attempts were made to model polymerization activity and the molecular weight of the resultant polymer, only the molecular weight model was statistically significant. The cross-validated q^2 value of this model was 0.63 whereas the r^2 value for the training set was 0.94.

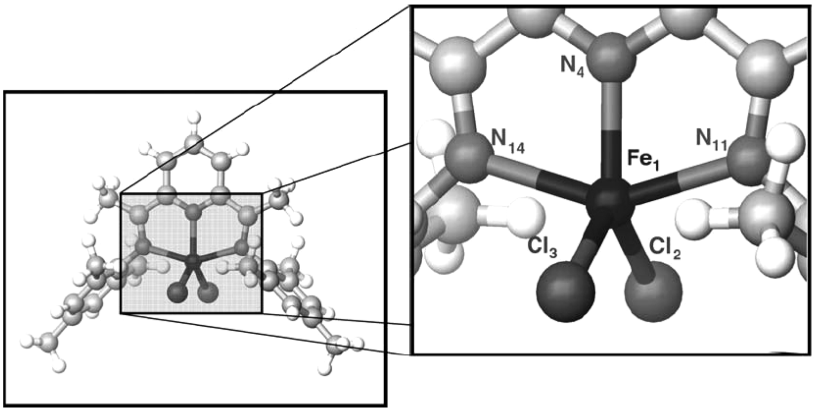


Figure 9. Atom numbering around the metal center of the bis(arylimino)pyridyl catalyst core. Reprinted with permission from ref 90. Copyright 2009 Elsevier.

The model could make good predictions of the polymer molecular weight for 5 of 6 new catalysts not used in training the model.

In the second study, simple linear correlation and multivariate analyses were used to model catalytic activity of 22 bis(arylimino)pyridyl precatalysts (Figure 9).

Molecular descriptors derived from density functional theory calculations were used to generate the models using 10 catalysts as a training set.

$$\text{Activity}(10^6\text{g/mmol h bar}) = 18.51d_{1-2} - 18.49\theta_{4-1-14} - 11.28\phi + 15.61 \qquad (7)$$

where d_{1-2} is the distance Fe1−Cl2, θ_{4-1-14} is the angle N_4−Fe_1−N_{14}, and φ is the dihedral angle (C_{im}−N_{im}−C_{aryl}−C_{aryl}(Me)) (Figure 9). This simple three-term linear model exhibited good predictivity, with the training set r^2 value being 0.98 and the LOO cross-validated q^2 value being 0.92.[90] The model made accurate predictions of catalytic activities of four catalysts not used in constructing the model. However, it was not possible to explain the relationship between these structural descriptors and the mechanism of catalysis.

In some studies, the relationship between structural descriptors and the desired property cannot be modeled and alternative methods of optimization need to be employed. Beckers et al.[91] analyzed 61 doped ceria catalysts in an attempt to predict the performance in selective hydrogen combustion expressed by a fitness value. Descriptors chosen included the dopant electronegativity, ionic radius, and dopant concentration. The results showed that there was only a low correlation between the predicted and real fitness of the catalysts. However, genetic algorithms were able to screen doped ceria compositions for their performance and could increase average fitness of the catalysts over three optimization cycles. Tognetti et al.[92] very recently reported a QSPR study predicting the butane selectivity in mixed (P,N)-nickel(II) catalyzed ethylene dimerization reactions using a data set of 29 active species. The catalysts were characterized by different substitution patterns on the nitrogen and phosphorus atoms. Nineteen quantum-chemical descriptors, such as geometrical parameters, atomic charges, isodesmic energies, polarizabilities, etc., from DFT calculations were used to develop a linear QSPR model of selectivity. The model had very modest predictive ability, thought to be due to the descriptors not capturing the electronic properties of the catalysts adequately.

5.2.3. Electrocatalysts.

Only one report of QSPR modeling of electrocatalysts[93] has been published. The electrochemical performance of six samples of nonplatinum porphyrin-based catalysts of oxygen reduction was predicted based on 24 XPS spectral variables and electrochemical measurements. The combination of genetic algorithm and multiple linear regression generated a model that had excellent predictivity for the training set and good cross-validation performance. However, the imbalance between the small data set size and number of descriptors risks overfitting the QSPR model.

In conclusion, various QSPR algorithms have been developed and applied successfully to different areas of catalysis. However, as this is a relatively new area of modeling, there is a risk of generating flawed QSPR models because of poor choice of descriptors, overfitting, or neural network overtraining issues. This should not detract from the general high utility of the QSPR method, because expertise and experience will grow as they become more widely adopted and as larger data sets become available. With the rapid growth of the high-throughput synthesis techniques, QSPR models are powerful tools to design experiments, screen very large catalyst libraries, and increase the scope of the search space as well as the chance of discovering better, cheaper, or more eco-friendly catalysts or processes.[94]

5.3. Biomaterials

Biomaterials, i.e., materials used in biological (usually medical) applications, include metals, ceramics, and most commonly polymers. This research area is undergoing a very exciting expansionary phase and is starting to embrace high-throughput methods.[95] Common medical applications of biomaterials are intraocular lenses in cataract surgery, prostheses in hip and knee replacement, pacemakers, heart valves and stents in cardiovascular disease treatment, skin grafts for burns victims, and artificial vasculature.[96,97] Another highly active biomaterials research area is targeted drug delivery.[98] Clearly, it would be very useful to be able to predict suitable properties of biomaterials for a particular application before synthesis. For the types of applications mentioned above, properties such as protein adsorption, cell attachment, and cellular proliferation on the biomaterial surface are very important. The Kohn biomaterials research group at Rutgers University has pioneered the application of QSPR modeling to biomaterials.

Most protein adsorption modeling on polymer surfaces has involved the important protein fibrinogen.[99] For example, Tang and Eaton[100] showed that an acute inflammatory response to implanted biomaterials appeared to be initiated by adsorbed fibrinogen. Fibrinogen is also the key protein that initiates blood clotting. Smith et al.[101] used neural networks to model the fibrinogen adsorption on 45 polyarylate surfaces. These polymers had been synthesized using 14 tyrosine-derived diphenol and 8 diacid components. For each polyarylate (represented as the monomer units), 104 2D largely topographical molecular descriptors were computed.[102] As they could not generate QSPR models of high quality using computed molecular descriptors alone, Smith et al. also employed two experimentally derived quantities, glass transition temperature (T_g) and air—water contact angle (CA), a measure of the hydrophobicity of the surface. In an earlier study of fibrinogen adsorption on the same polyarylate library, Weber at al.[99] found that adsorption decreases as CA increases. Decision tree and Monte Carlo methods were used to select the most relevant descriptors from a derived larger pool. They obtained a modest correlation coefficient of 0.78 (explaining 61% of the variance in the data) for the training set and 0.54 (explaining 29% of data variance) for the test set that constituted half of the data set. The three most significant descriptors in modeling fibrinogen adsorption were T_g, number of hydrogen atoms, and the octanol—water partition coefficient, log P. The same group revisited this data set subsequently[103] using molecular dynamics simulations to compute 3D molecular conformations that were input to the DRAGON program to compute the molecular descriptors. The rationale for introducing this protocol was that the strength of the interactions between the protein and the polymer is strongly dependent on the three-dimensional conformation of the polymer. Using this approach, they required only computed molecular descriptors, removing the need for experimentally measured properties like CA and T_g. These models exhibited a modest improvement in the ability to predict the test set. The Rutgers group reported another QSPR model using the same data and PCA to select a subset of descriptors from a pool of 109. They employed a neural network to model fibrinogen adsorption, contact angle, and glass transition temperature.[104] These models showed good ability to predict the training and test set properties, consistent with the relatively large experimental error in measurement of fibrinogen adsorption.

Freely accessible protein adsorption data sets facilitate the development of novel QSPR methodologies to model and predict this important property. The Biomolecular Adsorption database (BAD) (http://dbweb.liv.ac.uk/bad) established by the Nicolau group (University of Liverpool) provides a valuable public resource. This database[105] was compiled from published literature and contains data from >700 protein adsorption experiments. Twenty-one proteins, including albumin, fibrinogen, lysozyme, immunoglobulin G, α-lactalbumin, myoglobulin, fibronectin, and ribonuclease, are represented giving adsorption data (protein concentration in solution and on surface) with their dependence on temperature, buffer, pH, ionic strength, etc. In addition to the curated experimental data, the Web site also provides neural network applets to predict the adsorbed protein concentration, with user-provided input of protein, solution concentration, pH, ionic strength, and contact angle. Neural network models of the data[105] suggested that the predictive accuracy of models could be improved if the BAD data are divided into classes for hydrophobic and hydrophilic surfaces. Finally, it should be mentioned that protein adsorption is important not only for biomaterials in medical applications but also for protein microarrays in proteomics experiments, as well as for diagnostics in "lab-on-a-chip" microfluidic devices.

In biomaterial applications, the adsorbed proteins can mediate interactions between the material and the cells in vivo. Tissue engineering has a pressing need to be able to correlate quantitative biological responses such as cellular proliferation to the molecular structure of the biomaterial. With the complexity of the interactions precluding an atomistic treatment, statistical QSPR techniques are again an obvious method for tackling this problem. Kholodovych et al.[106] modeled the cellular response (CR) of fetal rat lung fibroblasts (FRLF) to the same 112-member polymer library described previously using PLS regression and PCA. The CR was defined as the metabolic activity of the cells measured using a colorimetric assay normalized with respect to a control (normalized metabolic activity (NMA)). As the entire polymer structure could not be used to generate descriptors, a linear chain consisting of three monomer units was used to compute 15 molecular descriptors for each polymer. A satisfactory r^2 value of 0.62 was obtained for the NMA of a training set of 62 polyarylates. The Rutgers group also modeled the NMA of fetal rat lung fibroblasts for the same polymer library using neural networks[107] and MOE[102] and DRAGON descriptors. Again, they used T_g and contact angle as additional descriptors and a decision tree algorithm with pseudo-Monte Carlo experiments to rank and select final sets of descriptors. Using between 3 and 17 descriptors, they obtained QSPR models with an average correlation coefficient of 0.75—0.79 and rms error of 16—20% for the training set. Interestingly, they found that very short R- (i.e., pendant) groups combined with short diacid monomer components were required for high normalized metabolic activity in the FRLFs. Kholodovych et al.[108] modeled cell attachment and cell growth of NIH3T3 cells from mouse embryos and fibrinogen adsorption for a library of 79 polymethacrylates. Twenty to 22 polymers provided measurable biological responses. They employed 2D molecular descriptors calculated with MOE for the polymer units and used a novel type of neural network, a polynomial neural network.[109] The resulting models had r^2 values of 0.78, 0.95, and 0.78 for cell attachment, cell growth, and fibrinogen adsorption data sets, respectively (Figure 10). The models could make good predictions of the properties of 4 or 5 test set compounds. The models were used to predict the biological properties of 40 000 polymers in a virtual library. However, the domain of applicability of these models must be well-understood before an extrapolation from a few tens to 40 000 polymers can be made with confidence.

Finally, it should be mentioned that Linati et al.[110] have begun applying QSPR techniques to bioactive glasses used in bone defect repair applications. They derived simple linear relationships between structural descriptors from molecular dynamics simulations and various experimentally measured properties. In spite of very small data sets, they obtained good single-variable models with r^2 values of 0.99 (density), 0.82 (crystallization temperature), 0.94 (glass transition temperature), and between 0.74 and 0.93 for leaching of various elements from the glass.

5.4. Polymers (Nonbiological Applications)

Although natural polymers such as silk, cellulose, and rubber have existed for a very long time, synthetic polymers have only existed in industrial and consumer applications since the 19th century. Given the diversity of applications of synthetic polymers, a range of polymer properties have been the subject of QSPR studies. These include thermophysical properties such as glass transition temperature; thermal decomposition temperature; Flory—Huggins parameters; electrical and optical

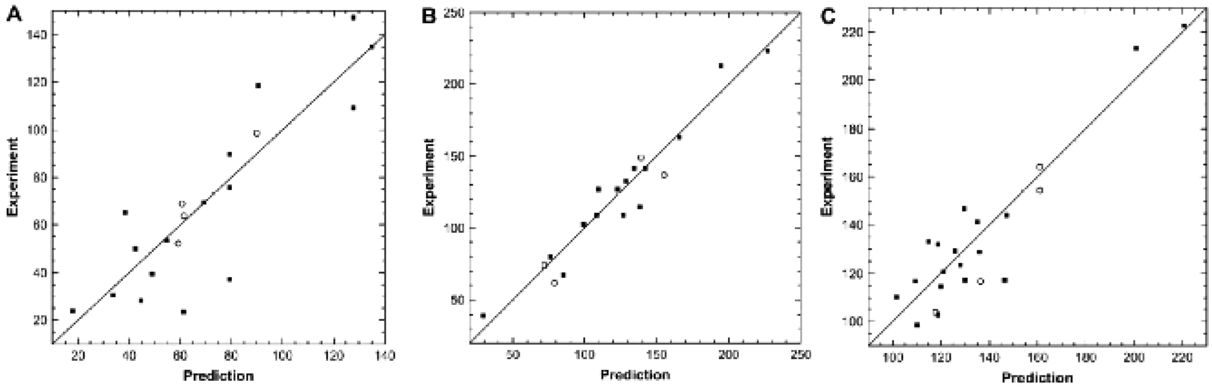


Figure 10. Performance of QSPR models for (a) cell attachment, (b) cell growth, and (c) fibrinogen absorption for a data set of 22 polymethacrylates. Reprinted with permission from 108. Copyright 2008 Elsevier.

properties such as dielectric constant, electrical conductivity, and refractive index; transport properties such as gas and aqueous diffusion and intrinsic viscosity; mechanical properties such as impact resistance; etc.

5.4.1. QSPR Models of Glass Transition Temperature T_g.

The polymer property that has been most widely modeled by QSPR is glass transition temperature, T_g. This is the point at which a reversible transition between a hard and relatively brittle state to a molten or rubber-like state occurs. By determining T_g, the temperature range in which the polymers have useful properties can be found. T_g can be difficult to determine experimentally because the phase transition may occur over a relatively wide temperature range and depends on measurement technique, duration, and pressure. Hence, since the 1960s, scientists have attempted to predict the glass transition temperatures of polymers using theoretical and computational methods.[111] In the interest of a balanced review, we have summarized many of the QSPR models of T_g in Table 1 rather than describe them fully. We discuss a subset of these that are of particular interest, because of the modeling technique used or because they illustrate some of the pitfalls in QSPR modeling summarized previously.

Early T_g modeling work was empirical and used employed group additive property (GAP) methods. These commonly use the Van Krevelen paradigm[148] (a weighted sum of scalar quantities associated with functional groups commonly occurring in polymers) or Bicerano's method[149] (solubility parameters and topological considerations independent of specific functional groups). The disadvantage of the GAP approach is that it is limited to polymers that contain previously investigated structural groups. Computer-aided molecular modeling in combination with GAP method can partially overcome the GAP theory limitation, as was first shown by Hopfinger et al.[111] Two molecular properties, conformational entropy and mass moment, for 30 structurally diverse polymers were modeled using conformational energy calculations. The QSPR model used MLR and had an r^2 value of 0.86; the entropy terms accounted for >70% of the variance in the T_g values. These early models did not include a test set or a cross-validation estimate of the predictivity of the models. This modeling method was modified[150] by using charges calculated using the empirical Gasteiger−Marsili[151] method. This improved the model predictability for polyolefins, polyacrylates, and polymethacrylates. It was also shown that the prediction error could be reduced if T_g values were predicted for the same class of polymers after hierarchical cluster analysis. Subsequently, partial atomic charge descriptors derived from semiempirical quantum-chemical calculations were shown to generate QSPR models with superior predictivity compared to those using the Gasteiger−Hückel method.[151,113] The r^2 values for models of T_g for 62 polymers, built using quantum-chemical descriptors were >0.98, compared with 0.96 for Gasteiger−Hückel charges. Standard errors of prediction were almost halved when semiempirical charges were used. Schut et al. used the well-known dependence of T_g on molecular weight and chain flexibility of polymers to model this property for a library of 132 L-tyrosine derived homo-, co-, and terpolymers.[115] In this work, the ratio of mass-per-effective flexible bond was employed as a descriptor. The QSPR models had high prediction accuracy (4−6 K) and were applicable to polymers with any number of comonomers.

MLR and simple descriptors were used to model T_g of styrene copolymers and poly(acrylonitrile-*co*-methyl acrylate) (ANMA) copolymers using a training set of 32 and test set of 16 polymers. Three thermodynamic and intermolecular force descriptors were used in the model.[117]

$$\begin{aligned} T_g &= 316.4 + 5.58\alpha + 735.5q^{+} - 19.5C_v \\ r^2 &= 0.98; s = 6.9; F = 527; n = 32 \end{aligned} \quad (8)$$

where q^+ is the most positive net atomic charge on hydrogen atoms in a molecule, α is the average polarizability of the molecule, and C_v is the heat capacity. The QSPR model had an r^2 value of 0.98 for the training set and 0.98 for the test set. The model was of particular interest because it was trained using only data for the styrene polymers and yet could predict the properties of the ANMA polymers very well.

Computationally inexpensive topological and geometrical descriptors have also been successful in modeling and predicting T_g for a representative subset of 17 polymers from a library of 112. QSPR models of T_g and CA were generated and used to predict the properties of the entire library.[121] Simple QSPR models with a small number of descriptors and r^2 values of 0.94 for T_g and 0.95 for CA were obtained. The most relevant descriptors were molecular flexibility (number of rotatable bonds) for the T_g model and the log P for the CA model. Indeed, models using only these single descriptors could predict the training set with an r^2 value of 0.82. The T_g and CA of the remaining 95 polymers were then used to validate the predictivity of the models, and test set r^2 values of 0.89 for T_g and 0.92 for CA

Table 1. Summary of QSPR Models of Polymer Glass Transition Temperature

					training		validation or test	
author	polymer	data size	descriptors	modeling method	r^2	SEE	q^2/r^2	SEP
Hopfinger et al.[111]	diverse	30	conformational entropy, mass moments	MLR	0.86			
Koehler and Hopfinger[112]	diverse	35	entropy, intramolecular energy	MLR	0.92			
Tan and Rode[113]	diverse	62	quantum-chemical charges	MLR	0.98			
Hamerton et al.[114]	poly(arylene ether)s	10	quantum-chemical	MLR	0.98			
Schut et al.[115]	L-tyrosine derived homo, co-, and terpolymers	132	mass per flexible bond	MLR				
Yu et al.[116]	polystyrenes	107	side-chain rigidity, main chain stiffness, density of hydrogen bonds, polarizability	MLR	0.92	15 K	0.91_{cv}, 0.89_{test}	
Yu et al.[117]	random styrenic copolymers and poly(acrylonitrile-*co*-methyl acrylate) (ANMA) copolymers	48	quantum-chemical	MLR	0.98	7 K	0.98_{cv} 0.98_{test}	
Liu et al.[118]	polyacrylamides	20	quantum-chemical	MLR	0.92	22 K	0.88_{cv}	
Yu et al.[119]	polyvinyls, polyethylenes, and polymethacrylates	60	quantum-chemical	MLR	0.91	27 K	0.86_{cv} 0.90_{test}	
Yu et al.[120]	polyacrylates	60	quantum-chemical	MLR	0.96	14 K	0.94_{cv} 0.87_{test}	
Reynolds[121]	diverse	112	topological	MLR	0.94		0.89_{test}	
Garcia-Domenech and de Julian-Ortiz[122]	diverse linear addition	88	chemical graph/topological	MLR	0.89		0.84_{cv}	
Cao and Lin[123]	diverse linear addition	88	chain stiffness and intermolecular forces	MLR	0.91	21 K		
Xu and Chen[124]	OLED materials	80	topological	MLR	0.93	10 K	0.92_{cv}	
Katritzky et al.[125]	linear	22	CODESSA	MLR	0.93		0.89_{cv}	
Katritzky et al.[126]	uncross-linked homopolymers	88	CODESSA	MLR	0.95		0.94_{cv}	
Morrill et al.[127]	amine-epoxy copolymers	13	quantum-chemical	MLR	1.00		1.00_{cv}	
Cypcar et al.[128]	polyacrylates and polymethacrylates	47	energy, volume, mass-related	MLR	0.96	17 K		
Camelio et al.[129]	aliphatic acrylate and methacrylate homopolymers	30	energy, volume, mass-related	MLR	0.96	12 K		
Carro et al.[130]	aliphatic acrylate and methacrylate polymers	39	energy, volume, mass-related	MLR				
Rauzy et al.[131]	poly(methyl methacrylate)-naphthopyran system	9	energy, volume, mass-related	MLR	0.97			
Sumpter and Noid[132]	diverse	357	topological	neural network		3%		

Table 1. Continued

author	polymer	data size	descriptors	modeling method	training		validation or test	
					r^2	SEE	q^2/r^2	SEP
Ulmer et al.[133]	diverse	320	vector representations of molecular structure and atomic composition	neural network		10 K		
Joyce et al.[134]	diverse	449	SMILES strings	neural network		20 K		35 K
Sun et al.[135]	homopolymers	271	functional group	neural network	0.69−0.98	5−26 K	0.10−0.50	7−28
KMattioni and Jurs[136]	diverse	165	topological, electronic and geometric	neural network	0.98	10 K	0.92_{test}	22 K
Mattioni and Jurs[136]	diverse	251	topological, electronic, and geometric	neural network	0.96	21 K	0.96_{test}	22 K
Gao et al.[137]	polyamides	87	quantum-chemical	neural network	0.86	15 K	0.81	16 K
Liu et al.[138]	polymethacrylates	35	quantum-chemical	neural network	0.92−0.98	16 K	0.88−0.94	16 K_{cv}
Liu and Cao[139]	polyacrylates and polystyrenes	113	quantum-chemical	neural network	0.95	11 K	0.91_{test}	17 K
Liu[140]	aromatic heterocyclic polyimides	54	topological and quantum-chemical	neural network	0.87	12 K	0.88_{test}	16 K
Afantitis et al.[141]	diverse	88	polarity, dipole of side groups, no free rotation part of side chains, and bond count of free rotation part of side chains	radial basis function neural network	1.00		0.93_{test} 0.93_{cv}	
Duce et al.[142]	acyclic polymers	154	labeled directed positional acyclic graphs (DPAGs)	recursive neural network	0.90−1.00	4−21 K	0.80−0.90	19−27 K
Duce et al.[143]	(meth)acrylate polymers	95	labeled structures	recursive neural network	0.98	10 K	0.92	13 K
Bertinetto et al.[144]	(meth)acrylic polymers containing phenyl groups	277	DPAGs, SMILES, and InChI	recursive neural network	0.97	11 K	0.88	21 K
Duce et al.[145]	diverse	~170	DPAGs	recursive neural network	0.90−1.00	4−21 K	0.80−0.90	19−28 K
Bertinetto et al.[146]	(meth)acrylic random copolymers	615	DPAGs	recursive neural network	0.92−0.98	6−12 K	0.80−0.98	6−24 K
Sun et al.[147]	diverse	256	fuzzy set membership function	fuzzy set theory		20 K		30 K_{test}

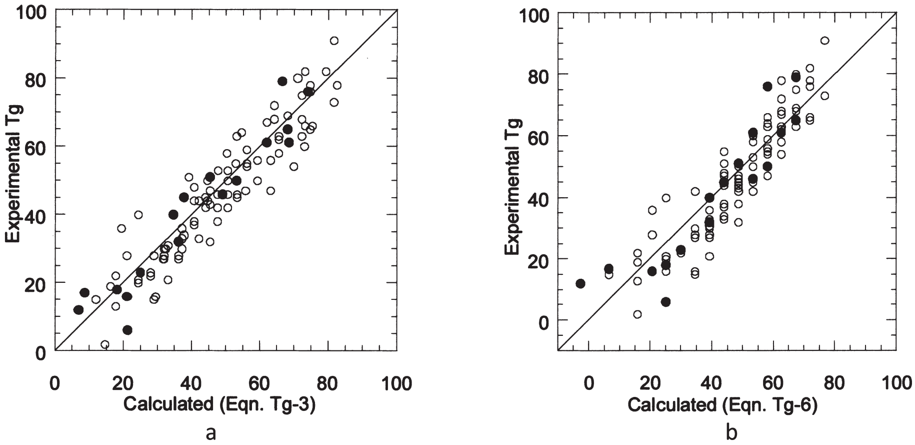


Figure 11. Experimental versus predicted T_g values (°C) for a polymer library for (a) two-variable QSPR model performance and (b) simple model using number of rotatable bonds only. Black marker for training set, white marker for test set. Reprinted with permission from ref 121. Copyright 1999 American Chemical Society.

were obtained. The performance of the best T_g model (model Tg-3 in Reynolds's paper) and a simple model with a single descriptor (model Tg-6) in Reynolds's paper) (number of rotatable bonds) is illustrated in Figure 11. The QSPR models were used successfully to build focused libraries with specific values of T_g and contact angle.

It is possible to predict properties of polymers using only the chemical graph (the connectivity of each atom to the others in a molecule) and the topological properties derived from this. Garcia-Domenech and de Julian-Ortiz modeled the T_g and refractive index of 88 structurally heterogeneous linear addition polymers using this approach.[122] The QSPR model for T_g that used 10 descriptors had an r^2 value of 0.89 for training set and 0.84 for LOO cross-validation, respectively. Xu and Chen reported another QSPR study that used topological descriptors to model T_g for a set of 80 organic light-emitting diode (OLED) materials.[124] A five-parameter QSPR model generated an r^2 value of 0.93 using stepwise MLR and leave-one-out cross-validation.

Morrill et al. also used the CODESSA descriptor program to generate QSPR models of T_g for 13 amine−epoxy copolymers with high accuracy (leave-one-out cross-validation r^2 of 1.00) using quantum-chemical descriptors.[127] Different combinations of descriptors generated equally valid models. It is not clear how large a pool of possible descriptors was sampled to generate the final models, so chance correlations could not be ruled out. The small data set meant that an independent test set could not be used to assess predictivity of the models.

Quite complex molecular mechanics and dynamics methods have been used in some QSPR studies to predict the polymer T_g values. Cypcar et al. modeled a set of 47 multicyclic and bulky substituted acrylate and methacrylate polymers.[128] Polymer geometries and conformations were simulated with three different force fields, and energy minimization and molecular dynamics calculations were performed to generate energy, volume, and mass descriptors. The best models had high statistical significance, with r^2 values of 0.96 and standard error as low as 17 K. This method was also applied successfully to a set of 20 linear and branched alkyl acrylate and methacrylate polymers[129] and a set of 39 aliphatic acrylate and methacrylate polymers.[130] In another study the T_g of polymer−naphthopyran systems was studied,[131] generating a model with an r^2 value of 0.97. This work also investigated interactions between naphthopyran photochromic pigments and the polymer matrix, showing that nonbonded van der Waals interactions were important for the photochromic behavior of the optical material and of commercial available glasses. Given that T_g can be modeled and predicted using much simpler descriptors, the use of complex, computationally intensive molecular dynamics methods appears to not be justified.

Neural networks were first applied to model T_g values of polymers by Sumpter and Noid.[132] Multilayer feed-forward neural networks and 18 topological indices for the repeat units in the polymer were used to generate the models. The neural networks employed 3 nodes in the hidden layer, and 9 polymer properties were modeled. These were the molar volume, heat capacity, change in heat capacity at the glass transition temperature, cohesive energy, solubility, glass transition temperature, refractive index, thermal conductivity, and dielectric constant. Unlike most of the other QSPR models using neural networks, this study trained the network to predict several properties simultaneously by using multiple output nodes. Three hundred fifty-seven different polymers were used to build the QSPR models, which had an average prediction error less than 3%. Neural network models were derived for each physical property separately, resulting in even higher accuracy of prediction.

The neural network approach was also applied by the same authors to model a data set of 320 polymers consisting of 23 different classes of polymers.[133] T_g models using a variety of descriptor families could predict the training set with high fidelity, with r^2 values of 0.98. They were also successful in building models for degradation temperature, tensile strength, dielectric constant, Rockwell hardness, and several other useful properties. Joyce et al.[134] employed monomer structures represented as SMILES strings as the inputs for their neural network models of T_g. Three hundred sixty polymers, represented as their monomer structures, were used to train the network, and an independent set of 89 monomer structures was used for testing model performance. The neural networks were very complex, with upward of 1 000 input descriptors, between 1 and 3 hidden layers, each of which contained between 40 and 240 nodes. The number of weights in such networks vastly exceeds the number of polymers in the training set, making overfitting very likely. These models gave RMSE values of <50 K for the training set and maximum errors of 150−200 K for polymers in the test set. This large discrepancy between training and test set errors, and the poor prediction of the test set, indicate relatively poor generalization consistent with overfitting. A more sparse neural network architecture would improve the predictive power of the models.

Sun et al.[135] also used neural networks to model 271 homopolymers, 251 of which were included in the training set and 20 in the test set. Six types of input vectors consisting of various types of weighted functional group frequencies were used as the inputs to a neural network. The r^2 values for the models ranged from 0.69 to 0.98 for the training set and from 0.10 to 0.50 for the test set. Although the best model could predict the T_g of 20 polymers in the training set, it is not clear how many hidden layer nodes were used in the neural networks and, therefore, whether the model could have overfitted the data. The relatively poor predictivity of most models on the test set, and large discrepancies between training and test set performance, is indicative of overfitting.

Quantum-chemical descriptors combined with neural networks were used by several authors to create QSPR models for T_g. Mattioni and Jurs[136] studied two different sets of polymers using topological, electronic, and geometric descriptors. The monomer structure was used to model T_g for the first set of 165 polymers, 17 of which were used in a test set and the remainder in training the model. The repeating unit structures of a second superset of 251 polymers were used to model T_g, with 86 additional polymers added to the polymers in set 1. More than 200 descriptors were calculated for the first set, and 128 topological descriptors were computed for the second set. Both linear and nonlinear feature selection methods were employed to reduce the number of descriptors to 62 and 35 members for sets 1 and 2, respectively. The best model for set 1, built from monomer structure descriptors, had r^2 values of 0.98 and 0.92 and standard errors of 10 and 22 K for the training and test sets, respectively. The optimum model for set 2, employing repeat unit structure descriptors, had r^2 values of 0.96 and 0.96 and standard errors of 21 and 22 K for training and test sets, respectively. Randomization experiments destroyed the models, showing the original models were valid and not the result of chance correlations.

Gao et al.[137] generated back-propagation neural network models for T_g, density, and refractive index of 87 polyamides. Descriptors were calculated for repeating units end-capped by hydrogen. Density functional theory quantum-chemical methods were used to generate descriptors for the repeat units. The neural network model for T_g gave an r^2 value of 0.81 and standard error of 16 K for the test set, whereas the MLR model had an r^2 value of 0.79 and standard error of 23 K. The neural network model for density gave an r^2 value of 0.81 and standard error of 0.042 for the test set, compared to the MLR model that had an r^2 value of 0.80 and standard error of 0.051. The neural network model for refractive index gave an r^2 value of 0.81 and standard error of 0.027 for the test set, whereas the MLR model had an r^2 value of 0.82 and standard error of 0.018. Although these models had lower statistical significance compared to others reported in the literature, the similarity between the neural network and multiple linear regression models suggests that all three polymer properties are relatively linear functions of the chosen descriptors.

Liu and co-workers[138,139] created QSPR models of molar volume, refractive index, and glass transition temperature for a data set of 35 polymethacrylates using stepwise regression and neural network methods. MLR and neural network models had very high statistical significance, with LOO cross-validated rms errors of 16 K. No test set was used in this study. The T_g values of 113 polyacrylates and polystyrenes were also modeled using neural networks of various degrees of complexity and quantum-chemical descriptors, generating rms errors of 11 and 17 K for training and test sets, respectively. The more complex neural network architectures (4−8−4−1) were overfitted and had lower predictivity.

Radial basis function (RBF) neural networks were employed by Afantitis et al.[141] to create QSPR models for T_g using the data set of 88 polymers compiled by Katritzky et al.[126] The model used a set of four descriptors calculated by Cao and Lin.[123] The training set contained 44 polymers, and the test set contained 40 polymers (4 polymers were rejected as outliers). The RBF neural network model had a test set r^2 value of 0.93, a considerable improvement over MLR models with r^2 values of 0.82.

Recursive neural networks that allow variable-size label structures to be used as descriptors have been applied intensively by Duce et al.[142−146,152] The model input consisted of a hierarchical set of labeled vertexes connected by edges that belong to subclasses of chemical graphs (graphs describing how various atoms are connected in a molecule). Labeled structures are highly abstract graphical tools that can convey both the occurrences of specific atoms/groups in the compound and the topological relationships expressed by the structure. This type of descriptor generated QSPR models for different polymer classes using the 2D graph of the repeating units. The could be extended to both homopolymers and copolymers. These authors initially studied a data set of 95 diverse polymers, 80 of which were included in the training set and 15 in the test set. They built T_g models using recursive neural networks[143] that could predict the test set with a mean average error of 10 K. They subsequently modeled a set of 154 acyclic polymers[142] using a training set of 127 compounds and a test set of 27 compounds. Both models could predict T_g values with good accuracy (errors of ∼20 K in test set). An extended data set including both acyclic and cyclic structures were also analyzed using the same method.[144] The T_g values for 277 polymethacrylates were divided into training, validation, and test sets containing 217, 54, and 6 polymers, respectively. T_g for the test set was predicted with mean absolute error of 15 K and standard deviation of 20 K. The largest data set studied by these authors consisted of 615 polymers (340 homopolymers and 275 copolymers).[146] This was split into a training set of 494 and a test set of 121 compounds. The best model gave good prediction with standard deviation of 11 K for the training set and 18 K for the test set.

Fuzzy set theory was employed to study the relationship between T_g and the structure of polymers by Sun et al.[147] Two hundred forty-one polymers were included in the training set used to build QSPR models. T_g values for these polymers were predicted with a standard deviation of 20 K. The model was tested on 15 additional polymers, and the standard deviation for the predictions was 30 K. These results suggested that fuzzy set theory has potential for analyzing and predicting properties of polymers.

Given the difficulties in measuring some polymer properties such as T_g, models able to explain such a high amount of variance in the data are essentially extracting all the information from the data. The wide variety of successful and relatively simple QSPR models for T_g suggests that this important polymer property is relatively easy to predict using this modeling method.

5.4.2. Models of Other Synthetic Polymer Properties.

Another characteristic temperature of polymers that is very important, particularly in high-temperature applications, is the temperature of half-decomposition, $T_{d,1/2}$, a measure of thermal stability. This is defined[153] as the temperature at which polymer loss of weight has reached half its final value during pyrolysis at a constant rate of temperature rise. Yu et al.[154] predicted the molar thermal decomposition function, $Y_{d,1/2}$ (approximately $T_{d,1/2}$ multiplied by the repeat unit molecular weight[149]) for a set of 72 vinyl polymers using a stepwise MLR model. The final model consisted of just 2 molecular descriptors, the quadrupole moment and the total energy of the monomer (computed using DFT). The $Y_{d,1/2}$ model had an r^2 value of 0.98 for the training set and 0.97 for a 17-polymer test set, with standard error of prediction of 5.2 (K kg)/mol.

The solubility of polymers in solvents is important for processes such as removal and recovery of synthesis byproducts, unreacted substrates, and process solvents as well as for environmental considerations. Several related properties or metrics of polymers such as Flory−Huggins interaction parameter χ_{12}, cohesive energy E_{coh}, solubility parameter δ, and lower critical solution temperature θ have been modeled by QSPR methods. According to Flory−Huggins solution theory, the enthalpy of mixing between components 1 and 2 is proportional to χ_{12}.[149] Xu et al.[155] used a genetic algorithm (GA) to generate a predictive model of χ_{12} for a data set consisting of 7 polymers and 15 solvents (a total of 104 data points). DRAGON was used to compute descriptors for each polymer and each solvent. Eliminating highly correlated descriptors and use of a genetic algorithm reduced the large number of descriptors to a small subset. The final cubic polynomial model predicted χ_{12} in the test set with an r^2 of 0.96. Xu et al.[156] used MLR and neural network models to predict the lower critical solution temperature (LCST, θ) for a data set of 12 polymers and 67 solvents (a total of 169 data points). Below the lower critical solution temperature θ, the polymer is considered miscible in the solvent. DRAGON molecular descriptors for the polymer repeating units were used in the MLR model. The 9 best descriptors from the MLR model were used to develop a nonlinear neural network QSPR model. Better results were obtained with the neural network model, which had a standard error of 13 K for the training set, compared to 26 K for the MLR model. The neural network model showed a similar improvement for the test set containing 57 data points.

The solubility parameter δ for 51 polymers of structure $(-C^1H_2-C^2R^3R^4-)$ was modeled by Yu et al.[157] using stepwise MLR. This resulted in a six-descriptor final model that gave a standard error of 0.75 $(J/cc)^{1/2}$ for the training set of 51 polymers and 1.01 $(J/cc)^{1/2}$ for the test set of 46 polymers. Note that the solubility parameter is related to the cohesive energy by $\delta = (E_{coh}/V)^{1/2}$, where V is the molar volume, and quantities such as δ and E_{coh} can be measured only indirectly by experiment.

QSPR methods also have been used to model the optical and electrical properties of polymers. Optical properties of polymers are important in coating applications (such as optical lenses) and in the packaging industry. The refractive index, n, is the most commonly modeled optical property of polymers. Katritzky et al.[158] modeled this property for a set of 95 amorphous homopolymers using the CODESSA program. They obtained an r^2 value of 0.94 and a standard error of 0.018 for the best five-parameter model using descriptors for the repeating unit derived from quantum-chemical calculations. Similarly, in a comparative study of oligomers of the repeating unit of the polymers, Holder et al.[159] found that dimers gave the most accurate QSPR model for the refractive index of 70 polymers, including those used as the resin component of dental restorative materials. Descriptors from quantum-mechanical calculations were employed, i.e., the HOMO−LUMO energy gap and a polarizability parameter expressed as the difference between the maximum and minimum partial charges in the molecule ($Q_{max} - Q_{min}$):

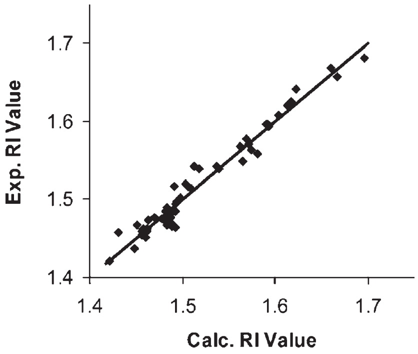


Figure 12. Calculated vs experimental polymer refractive index values for the training set using the dimer QSAR. Reprinted with permission from 159. Copyright 2006 Wiley VCH.

$$\mathrm{RI} = 1.997 - 0.0338(\mathrm{HOMO} - \mathrm{LUMO}) - 0.429(Q_{max} - Q_{min})$$
$$n = 60, r^2 = 0.96, \mathrm{LOO}\ q^2 = 0.96, F = 740, s = 0.014 \qquad (9)$$

The performance of the refractive index (RI) model is illustrated in Figure 12.

The HOMO−LUMO energy gap was the descriptor that made the largest contribution to the refractive index. The model generated accurate predictions for the RI of 10 additional polymers in a test set not used to generate the model.

The most important electrical applications of polymers include use as insulation in cabling and to encapsulate electronic devices. The dielectric constant ε and the dielectric dissipation factor or power factor (loss tangent, tan δ) of polymers have been modeled with stepwise MLR[118] and neural networks,[160] respectively. Liu et al.[118] computed quantum-mechanical descriptors for the repeating units of 22 polyalkenes. The best resultant three-variable MLR model predicted ε with an r^2 value of 0.91. The molecular descriptors in this model were E_{LUMO}, q^- (the most negative net atomic charge on the molecule), and S (entropy), and the authors rationalize the positive correlation of these properties with ε in their QSPR model. Yu et al.'s neural network model was able to predict the loss tangent for 92 diverse polymers with standard errors of 0.011 for the training set and 0.025 for the test set, substantially better than for the MLR model.

Gas, aqueous, and other small-molecule diffusion, solubility, and permeability in polymers are transport properties that are important in practical applications ranging from food packaging to pharmaceutical controlled release. Permeability is defined as the product of solubility and diffusivity.[149] Patel et al.[161] used MLR and genetic algorithms to construct a model for CO_2, O_2, and N_2 diffusion in 16 polymers. They computed 9 descriptors such as conformational entropy, bulk modulus, cohesive energy density, etc. for the polymers using the group additivity method. Relatively parsimonious models containing 2 and 3 descriptors had very good statistical quality and predictive performance. They obtained r^2 values of 0.85−0.87 for the gas diffusion coefficient models, if 1 polymer outlier was excluded. Overwhelmingly, the bulk modulus was found to be the most significant descriptor for gas diffusion in polymers. This same

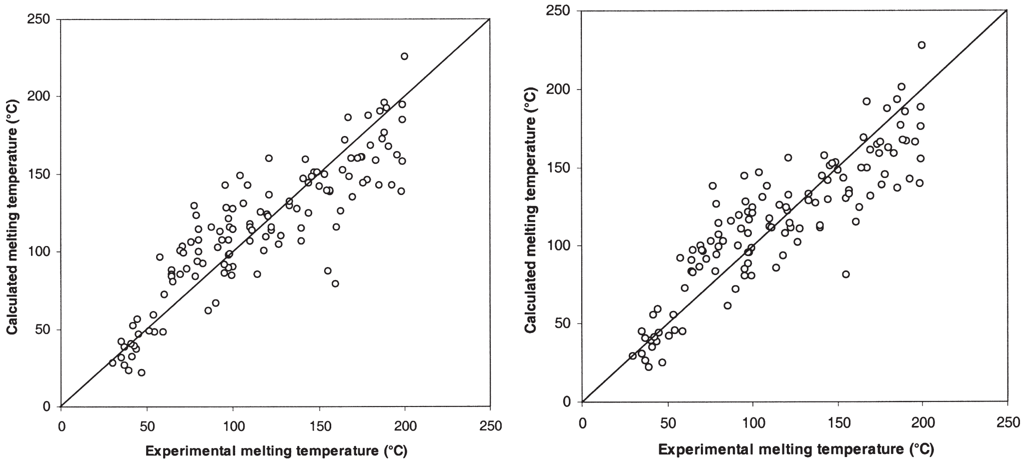


Figure 13. Graphs of calculated versus measured ionic liquid melting points for (a) all data and (b) leave-1/3-out cross-validation sets. Reprinted with permission from ref 174. Copyright 2002 American Chemical Society.

conclusion was reached by another study from the same group reported by Tokarski et al.[162] They used molecular dynamics calculations to characterize the polymers and diffusants and found that O_2 permeation was most highly correlated with the cohesive energy of the polymer matrix.

Other important polymer properties that have been modeled successfully with QSPR include intrinsic viscosity of polymer solutions,[163,164] impact resistance of polymers,[165] plasticization efficiency or the low-temperature flex point of poly(vinyl chloride) (PVC) plasticizers,[166] flexural modulus of dental polymer materials,[167] and imprinting factors of molecular imprinting polymer compounds.[168]

As mentioned earlier, freely accessible databases of materials structural and experimentally determined property data greatly facilitate the development of QSPR modeling. Takaeda and Yagi[169] have compiled a database of polymer structural and property data, PoLyInfo,[170] from more than 12 000 literature articles. This currently contains >220 000 data points, ranging from properties such as T_g and thermal decomposition temperature to dielectric constant and tensile modulus, for monomers, homopolymers, copolymers, blends, and composites.

5.5. Ionic Liquids

Ionic liquids (ILs) can be defined simply as ionic salts that have low melting points. For example, ethylammonium nitrate (melting point 285.7 K) is liquid at room temperature. This definition can be broadened to include what are called “molten salts”, with melting points higher than 100 °C. ILs can be further classified as protic or aprotic depending on whether the cation is protonated or not. A comprehensive review of protic ionic liquids has been published recently by Greaves and Drummond.[171] Common cations in ILs include imidazolium, pyridinium, alkyl ammonium, etc., while common anions are halides, carboxylates, and nitrates. Many ILs are able to dissolve a diversity of chemical compounds. They also are quite thermostable and exhibit low vapor pressures. Such properties make them ideal as solvents for many chemical and industrial processes particularly “green” chemistry processes involving separation, solvent extraction, and catalysis. Their intrinsic conductivity can also be utilized in electrochemical applications, recently reviewed by Armand et al.[172]

With the recent rapid growth in research and application of ionic liquids, it is only natural that there is also increasing interest in modeling and predicting their properties using both molecular simulations and statistical structure−property relationship analyses. The current status of molecular simulations of ILs has been reviewed recently by Maginn[173] and will not be discussed here.

5.5.1. Melting Points.

As one of the dominant materials properties of ILs is their melting point T_m, it is not surprising that this was one of the first properties to be examined by QSPR analyses. One of the issues encountered frequently with ILs is that the melting points can vary depending on the experimental technique, and even within a given technique because of the existence of polymorphs. In some cases, the melting points can be uncertain by >50 °C. Katritzky et al.[174] extracted data for 126 structurally diverse pyridinium bromides from the Beilstein database[175] and modeled their melting points using MLR. Experimental uncertainty was minimized by using only the most recent data in the model. CODESSA was used to compute a large pool of molecular descriptors. A heuristic approach was used to eliminate descriptors from the pool. The best melting point QSPR model that employed six descriptors had an r^2 value of 0.79 and standard error of 23.0 K. Six data points lay outside the range of $\pm 2\sigma$ (95% confidence limit) from the predicted value. A leave-1/3-out cross-validation study showed that the model had good predictivity within the chemical space of the training data. The quality of the predictions of the entire data set and the cross-validation sets is illustrated in Figure 13.

Although several topological descriptors and indices made major contributions to the QSPR model, they were not easily interpreted in terms of chemical structure, i.e., they would not be very helpful to a synthetic organic chemist in deciding which compound to synthesize to achieve a particular value of melting point. The issue of the most appropriate descriptors and the appropriate level of theory to compute them was considered in a QSPR study of the melting points of 13 high-energy ILs based on 1-substituted 4-amino 1,2,4-triazolium bromide, nitrate, and nitrocyanide salts.[176] The MLR models for each anion employed descriptors calculated in the gas phase using ab initio quantum-chemical methods. Although the models were quite parsimonious and statistically valid (r^2 values of 0.87−0.89 and standard errors of 21 and 23 K for the bromide class), they were less accurate than those obtained in another study by the same authors[177] using descriptors computed at a lower level of theory.

The following QSPR model was derived for T_m of the nitrate salts:

$$T_m = -284 - 214\text{HDCA1}_Z - 3.94 \times 10^4\text{NRI}_{\text{min, C}} + 3.16 \times 10^3\text{FHDCA}$$
$$N = 13, r^2 = 0.933, F = 41.5, s = 14, q^2 = 0.872 \quad (10)$$

HDCA1_Z and FHDCA are measures of the hydrogen bond-donating ability of the cation, and $\text{NRI}_{\text{min,C}}$ is the minimum nucleophilic reactivity index for a carbon atom. No attempt was made to generate a global model containing data from all three anion classes. IL densities could also be predicted very accurately using QSPR models. The authors presented the important conclusion that QSPR models would be improved by the design of ionic liquid-specific descriptors rather than resorting to higher levels of theory to compute the descriptors. This is a conclusion that could be generalized to QSPR of all classes of materials. It should also be noted that for a few of the ILs in this study the glass transition temperature was used instead of the melting point, as the glass formed much more readily than the crystal from the liquid state. The ILs in the study were also a homologous series, so they were probably easier to model than the more diverse set reported by Katritzky et al.

The largest IL QSPR study to date has been that of Varnek et al.[178] who studied the melting points of 717 ILs consisting of 126 pyridinium, 384 imidazolium, and 207 quaternary ammonium bromides. They employed a range of modeling methods including MLR, PLS, support vector machines, and neural networks. The nonlinear SVM and neural network models were only slightly better than the linear models. Very large pools of descriptors, ranging in size from 2 000 to 13 500, were generated for some of the models. Some of the models reported used relatively large subsets of these descriptors in the model, ranging between 50 and 1 500, so variable reduction methods such as PLS were employed. Most QSPR modeling methods of the combined data set generated melting point models of similar predictive power as assessed by 5-fold cross-validation. Test set r^2 values ranged between 0.52 and 0.62, except for the MLR model, and rms errors of prediction were 40 ± 2 K. When models were generated for the pyridinium bromide, imidazolinium bromide, and ammonium bromide data subsets individually, the quality of prediction was similar to the combined set, except for the small pyridinium set that was predicted with lower rms error (26−35 K). The authors concluded that neural networks generated the most predictive models, and MLR and clustering methods, such as *k*-nearest neighbors, generated the least. The melting point QSPR model rms errors of 38−46 °C for the full set of data are reasonable given the difficulties in accurately measuring the melting points because of polymorphs and glass formation. The study was important because it showed that it was possible to generate predictive models of properties of ionic liquids from large, moderately diverse data sets. Carrera et al.[179] published a QSPR study of the melting points of 101 guanidinium salts using counterpropagation neural networks. The model that included all four counterions exhibited r^2 values of 0.87 and 0.82 and rms errors of 30 and 24 K for training and test sets, respectively. This was one of the few studies where the QSPR model was used to design new ILs with low melting points that were subsequently synthesized and found to have properties in reasonable agreement with the model predictions.

5.5.2. Models of Other Ionic Liquid Properties.

There have been a number of studies where QSPR methods were employed to model and predict the activity coefficient at infinite dilution, γ_i^∞,[180,181] or related quantities such as the Ostwald solubility coefficient log L or partition coefficient log P[182] of organic compounds in ILs. Both Eike et al.[181] and Tamm and Burk[180] modeled the same set of 38 solutes in the same 3 ILs, but at different temperatures, whereas Katritzky et al.[182] modeled data for 92 organic solutes in 7 imidazolium ILs and 1 pyridinium IL. All groups reported good models employing a small number descriptors and having r^2 values >0.90. However, this level of performance is achieved by modeling each IL as a separate chemical class. At infinite dilution, solute−solvent interactions are paramount, and accordingly, molecular descriptors such as number of hydrogen donor sites and charge-related descriptors were found to make the most significant contributions. Xi et al.[183] developed a QSPR model with an inversely linear dependence on the temperature for a set of 39 solutes in the IL trihexyl(tetradecyl)phosphonium bis(trifluoromethylsulfonyl)imide. It is interesting to speculate on whether an accurate single model across spanning diverse classes of ILs could be obtained using descriptors specific to ILs.

Viscosity and conductivity are two properties that are important in ionic liquid applications such as catalysis, electrodeposition, and electrolyte solutions in batteries. In general, there is an inverse relationship between viscosity and conductivity in ILs. Bini et al.[184] reported a QSPR study of 33 imidazolium, pyridium, piperidinium, and morpholinium ILs and summarized the difficulties in using data collected from various literature sources, as well as in measuring viscosity accurately for ILs. To eliminate this problem, Bini et al. used data collected from ILs synthesized and measured at their own laboratory. They also attempted to use descriptors that were more easily interpretable in terms of structure. Four-descriptor models for conductivity and viscosity having r^2 values of 0.94 and 0.95, respectively, were obtained. A lower-quality model for viscosity was obtained for viscosity at a lower temperature, which was ascribed to non-Newtonian behavior of the ILs. Tochigi and co-workers,[185,186] differently from most of the other IL QSPR studies, used a genetic algorithm (GA) to select the best coefficients in polynomial regression QSPR models of conductivity and viscosity of ILs. Their regression method was found to give better models than did MLR. This group[186] also showed how QSPR models of IL properties can be used to "reverse engineer" ILs. They used their models to calculate property values for a large virtual library of ILs generated by varying the anion species and side chain, until combinations of cations and anions that were predicted to have the desired values of conductivity or viscosity were obtained. This method is applicable to all QSPR models of materials properties, even when relatively obscure or arcane descriptors are used, provided that the domains of applicability of the models are understood. Finally, Billard et al.[187] constructed a neural network model using the viscosity data for 99 ILs at 25 °C. After using 5-fold external cross-validation to assess the predictive performance of the model (R^2 = 0.73, RMSE = 67.5 cP), they predicted the viscosity of 23 new ILs and obtained a prediction error of 73 cP. Finally, they synthesized three new ILs and compared the measured value of the viscosity with that predicted by the model: the agreement was only qualitative. They ascribed the relatively modest quantitative precision of the models to the error in the experimental data collected from different sources. However, it must also be said that 5-fold cross-validation can give overly optimistic assessments of the predictivity of the model.

Although many ILs exhibit physical properties that make them attractive as green chemistry solvents, they may be toxic to cells and organisms. Cytotoxicity of ILs was the subject of QSPR studies by Couling et al.[188] (on the aquatic organisms *V. fischeri* and *D. magna*), Garcia-Lorenzo et al.[189] (on human Caco-2 cells), and Torrecilla et al.[190] (on a leukemia rat cell line, and on acetylcholinesterase). In the largest of these studies, Torrecilla et al.[190] modeled toxicity of 153 ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium, and quinolinium ILs. They used simple constitutional descriptors (number of rings, number of C, H, N, O atoms, and molecular weight), PCA, MLR, and neural network methods to model IPC-81 cell line toxicity (rat leukemia) and acetylcholinesterase inhibition. The neural network models were the most predictive, with r^2 values of 0.97−0.98 for both biological properties for the training set. Couling et al.[188] found that toxicity of their ILs toward the two aquatic organisms increases with the number of nitrogen atoms in a cationic aromatic ring and also increases with the alkyl chain length. Garcia-Lorenzo et al.[189] observed a similar trend in the relationship between chain length and toxicity of imidazolium-based ILs toward human Caco-2 cells.

A very useful resource for the study of ionic liquids, including the QSPR modeling of their properties, is the NIST ILThermo database[191] that has free web access.[192] As of mid-2010, this database contained more than 94 600 data points, amounting to 339 ionic liquids. The compiled thermodynamic data (including properties such as electrical conductivity, heat capacity, melting temperature, refractive index, thermal diffusivity, viscosity, molar volume, etc.) are available not only for pure ionic liquids but also for binary and ternary mixtures.

5.6. Supercritical Carbon Dioxide

Like ionic liquids, supercritical fluids such as supercritical carbon dioxide ($scCO_2$) have become increasingly attractive as solvents in chemical and industrial processes from a green chemistry or environmental protection point of view. Carbon dioxide becomes a supercritical fluid, at and above its critical temperature of 31.1 °C and critical pressure of 7.38 MPa. $scCO_2$ has properties combining those of gases and liquids; for example, it diffuses through solids as well as solvates many organic compounds. Such behavior, combined with its low cost, lack of toxicity and flammability, low viscosity, and ease of removal and recycling, lends itself to a wide range of applications. Examples include decaffeinating coffee beans, use as a reaction medium for organometallic compound synthesis, use as a solvent for dyes in the textile industry, and manufacture of photovoltaic devices. This section of the review will summarize only QSPR studies of $scCO_2$ as a solvent of organic dyes.

Most early attempts to model correlate solubility of dyes in $scCO_2$ have used empirical, semiempirical, or theoretical equations of state specifying the relationship between solubility and quantities such as temperature, pressure, or density in some form or other. The most well-known and widely used of these methods is the Bartle equation.[193] The largest study of this type was published by Ferri et al.,[194] who fitted a set of >400 data points describing solubility of 16 azo and anthraquinone dyes in $scCO_2$ to five literature equations and one novel semiempirical equation. The best model gave predicted solubilities that differed from the measured values by an average absolute percentage deviation of 9.0%. However, these types of models require at least one empirical parameter to be derived from experimental solubility data.

More recently, QSPR studies of solubility of dyes in $scCO_2$ have used MLR and neural network methods to model the relationship between measured solubility and selected molecular descriptors over a wide range of temperatures and pressures. One of the largest of such studies was by Hemmateenejad et al.,[195] who applied these methods to 1 190 data points describing the solubility of 29 anthraquinone, anthrone, and xanthone derivatives at different temperatures and pressures. One hundred eight molecular quantum-chemical, physicochemical, constitutional, and topological descriptors were computed for each molecule. The set of best descriptors selected for the MLR models was then used in the ANN models. The rms error of prediction of pS (− logarithm of solubility) from the MLR model was 0.28, whereas that from the neural network model was significantly better at 0.10.

Studies by Burden and Winkler[17,18] and Tabaraki et al.[196,197] confirmed the superiority of nonlinear neural networks over linear methods for modeling solubility in $scCO_2$. Tarasova et al. reported the most diverse set to date.[19] They modeled $scCO_2$ solubility of 67 dye compounds across a temperature range of 286−423 K and pressure range of 60−1400 bar, amounting to a total of 685 data points. They generated linear and nonlinear models using MLR with expectation maximization (MLREM) and Bayesian regularized artificial neural network with Laplacian prior (BRANNLP) methods. The former method was used to select a sparse subset of context-relevant descriptors for the MLR model,[18] whereas BRANNLP uses a sparse Laplacian prior to select descriptors in a nonlinear way and to prune weights in the network.[17] For the training set of 584 points, they obtain a standard error of estimation of 0.34 and r^2 value of 0.90 for the best BRANNLP model, whereas the respective values with the linear MLREM method were only 0.56 and 0.77, respectively. Predictive performance on an independent test set was similar to that of the training set, showing the model was quite robust (Figure 14).

Temperature and pressure, hydrophobicity, total area, hydrogen-bond donor, number of rotatable bonds, positively charged polar surface area, and molar refractivity were found to be the most relevant descriptors in the final sparse BRANNLP model. Khayamian and Esteki[198] applied a wavelet neural network (WNN) method to analyze the solubility of polycyclic aromatic hydrocarbons,[198] anthraquinone dyes,[197] and azo dyes[196] in $scCO_2$. The WNN method uses wavelets as basis functions to construct a feed-forward neural network and has been found to be effective in solving convergence problems.[198] Finally, we anticipate that QSPR modeling of solubility in $scCO_2$ may be inherently easier than modeling aqueous solubility or log *P* (octanol/water partition coefficient). This is because of the much more complex nature of the hydrogen-bond networks and molecular interactions in water. For an excellent recent review of log *P* modeling, see paper by Mannhold et al.[199]

5.7. Ceramics

QSAR/QSPR methods have only been applied to a small number of studies of ceramics. The main purpose was to understand the effect of ceramic components or structure on measured properties, or to formulate ceramics that exhibit a desired property.

Guo et al.[200,201] investigated a data set of 21 $BaTiO_3$-based dielectric ceramic samples. The compositions of the ceramics

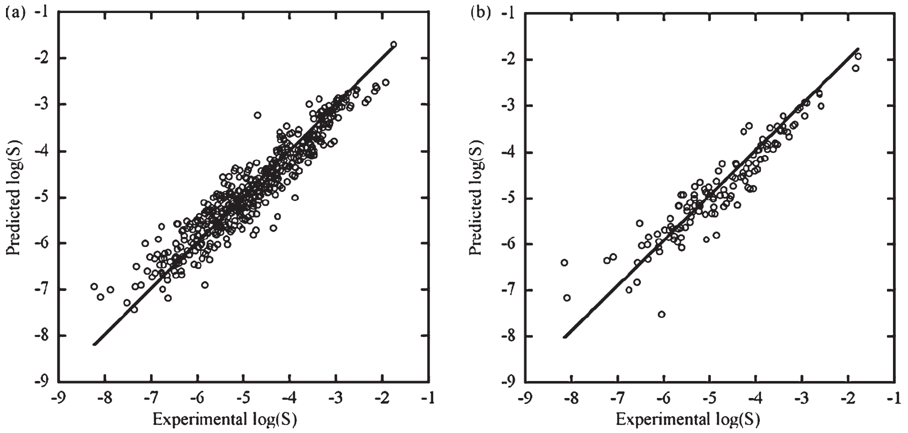


Figure 14. Graphical representation of the performance of the BRANNLP model showing the BRANNLP predicted versus experimental solubility ($\log(S)$) values for the (a) training and (b) test sets. Reprinted with permission from ref 19. Copyright 2010 Elsevier.

were used as descriptors to build QSPR models predicting different properties such as room-temperature dielectric loss and permittivity and maximum and minimum temperature coefficients of capacitance. A three-layer feed-forward neural network was employed, with 4 output nodes that allowed 4 measured properties to be modeled simultaneously. Unfortunately, the number of weights in the neural network with 5 input nodes, 8 hidden layer nodes, and 4 output nodes was much greater than the number of points in the data set, making overfitting likely. This was also suggested by the very high predictivity of the training set (r^2 values of 1.00). The dynamic ranges of some of the measured properties were also quite small (1 log or less in some cases), also making the development of a good model problematic. Lack of a test set made it hard to detect anomalies in the models. The neural network model was used to predict the properties of new compositions with modest success. The data were also modeled by MLR model, using the ceramic compositions and cross-terms (products of two compositional variables) as independent variables. A 10-variable model was reported but the large values for the descriptor coefficients suggest the model was not robust.

Compositional descriptors and three-layer back-propagation ANNs were also employed to model the electrical properties, piezoelectric coefficient, and planar coupling coefficient for 21 samples of donor-doped piezoelectric ceramics.[202] These authors stated that these methods were effective and useful tools for modeling the performance and composition of multicomponent ceramic materials. As with the ceramic modeling studies reported by Guo, this study by Cai et al. used a neural network architecture that contained many more adjustable weights than the number of samples in the training set, so this model risked overfitting as well. However, the properties of 6 ceramics not used in training were predicted with good accuracy, with test set r^2 values between 0.88 and 0.94. No statistics were quoted for the QSPR model derived from the training set.

The largest ceramic data set that has been modeled consisted of 700 samples of dielectric materials and 1 100 samples of ion-diffusion materials.[203] The two data sets contained materials comprised of 53 and 32 elements, respectively. The proportion of each of these elements in the ceramic material formed the input to the network. For ion-diffusion materials, the temperature at which the coefficients were measured was also used as a network input. PCA was used to reduce the dimensionality of the data sets while neural networks with 10-fold cross-validation were employed to build the models predicting the relative permittivity and oxygen diffusion coefficient of the materials. The networks were relatively complex, employing 15 hidden layer nodes and between 16 and 21 PCs in the input layer. A test set was used to assess the accuracy and predictivity of the models. The dielectric models had relatively low statistical significance, with r^2 values close to 0.42 (i.e., 42% of the variance in the data explained). The ion-diffusion data set provided better quality QSPR models, with r^2 values of 0.77 and rms errors of 2.1.

In summary, the ANNs provide a fast and robust tool to model properties and support the formulation design of ceramic materials, provided they are used correctly. The current limited number of applications of QSPR methods to ceramics is probably due to the difficulties in gathering high-quality, large data sets that can be used to generate reliable and robust models, as well as the challenges of developing novel descriptors for this type of material.

6. PERSPECTIVE ON QSPR MODELING OF MATERIALS

In compiling the information for this review, two points emerge quite clearly: almost all material properties investigated to date can be modeled by QSPR quite successfully and accurately; some of the published studies are flawed in their execution.

6.1. Summary of Material Property Prediction Examples

As stated in section 5, there are very few examples where new materials have been synthesized on the basis of model prediction and the properties of these new materials have been tested. Those discussed in detail in the above section are summarized in Table 2. The promise of QSPR models is that they have the capacity to make good predictions, at least near their domains of applicability, so they merit being used more frequently to design materials with improved properties.

It is very encouraging that quite complex properties of complicated materials can be modeled and predicted for new materials with a surprising degree of fidelity, albeit largely assessed through the use of independent test sets rather than by accurate predictions of materials subsequently synthesized. In all cases we have discussed, valid statistical models, some with relatively few descriptors, have been generated. Properties like T_g

Table 2. Summary of New Materials Synthesized and Tested on the Basis of QSPR Modeling

study	outcome
solubility of C_{60} in various solvents[51]	solubility of C_{60} in *n*-heptane and 1-octanol was predicted correctly by QSPR models, even when these solvents were not present in the training set; see, however, comments by Puzyn et al.[33]
catalytic activity of metallocene catalysts in ethylene polymerization[68]	predicted normalized catalytic activity of three new catalysts measured in another laboratory to within experimental error
T_g of styrene copolymers and poly(acrylonitrile-*co*-methyl acrylate) (ANMA) copolymers[117]	model trained using only data for styrene polymers and could predict the properties of the ANMA polymers very well

Table 3. Common QSPR modelling pitfalls and methods of avoiding them

pitfall	recommendation to minimize or avoid
use of uninformative descriptors	use descriptors that are related to the molecular structure where possible, use virtual screening methods when complex descriptors are necessary, and develop new materials descriptors
overfitting, and grossly underdetermined systems	reduce size of descriptor pool before building models,[24] monitor number of fitted parameters (descriptor weights or neural network weights) to ensure they are substantially less than the number of experiments, and check that training and test set statistics are similar
descriptor selection and chance correlations	use Topliss criteria[22,23] to estimate probability of chance correlations and descriptor scrambling; avoid methods where repeated sampling of a larger pool of descriptors is done to obtain the optimum subset of descriptors; and use sparse, context-dependent feature-selection methods[17,18]
modeling complex, nonlinear structure−property relationships	avoid overly complex nonlinear models, compare nonlinear model statistics with linear models, and use regularizing methods that attempt to optimize model complexity[16,17]
validating QSPR models	synthesize new materials that models predict to be superior and test if feasible, use independent test sets to assess model predictivity otherwise, and employ cross-validation methods with caution[27,28]
domain of applicability of models	calculate the range of all descriptors used to develop the model,[29−32] avoid extrapolations using descriptor space distant from that used in model, and use probabilistic modeling methods (e.g., Bayesian regularization[16]) that allow estimation of likely prediction error
incorrect handling of outliers	avoid removing outliers wherever possible, check whether outlier lies well within domain before removing it, remove outliers sparingly and describe why they were omitted, and retest properties for outliers to eliminate measurement of transcription errors

appear to be easy to model using relatively simple and fast QSPR methods. This augurs well for the future, when materials will be generated by high-throughput synthesis and characterization methods like those that transformed pharmaceutical and gene research. QSPR methods will provide a fast and effective method for extracting information and knowledge from very large data sets, allowing materials property prediction and optimization to be achieved relatively quickly. The key research component of developing QSPR methods for materials modeling will largely be the discovery of novel, improved ways of describing materials (i.e., more efficient descriptors for materials than those that currently exist). Another observation from the review is that relatively few QSPR models have been used to design or predict properties of new materials that were subsequently synthesized. This is the true test of a predictive model, and the studies reviewed above suggest that real predictions are achievable.

6.2. Summary of Pitfalls in QSPR Modeling

Section 4 described in detail the types of pitfalls new researchers employing QSPR can encounter. These pitfalls and methods for avoiding them are summarized in Table 3.

It is clear that the QSPR community, historically based in the pharmaceutical and environmental research arenas, will need to ensure that the mistakes and pitfalls that inexperienced QSPR practitioners commonly make are better understood. Referees of materials-related journals will need to recognize poorly executed QSPR models and return them for correction. The incidence of flawed QSPR studies should not distract from the great utility of the method when properly executed.

Two additional important needs that must be addressed by the research community are development of new types of mathematical descriptors to describe diverse, complex materials and development of methods for incorporating sample history into QSPR models. The mathematical descriptors used in almost all published studies of QSPR modeling of materials to date have been those developed for QSAR modeling of small, discrete organic molecules to a large extent. Although these can work well in many cases, the correct way to encode the microscopic properties of different types of polymers (e.g., homo, block, cross-linked, doped, polymer blends, etc.), nanoparticles (e.g., metal oxides, fullerenes, nanotubes, etc.), catalysts, etc. is an important and challenging research topic. Although synthesis and process variables can often be easily combined with other types of materials intrinsic descriptors to generate models (where sample history is important), recording and mathematical encoding of sample history is in its infancy. The development of rapid and automated high-throughput methods of materials synthesis should greatly improve this situation.

7. CONCLUSION

Diverse QSPR techniques and algorithms have been developed and applied successfully for a wide range of material properties from physical, chemical, and biological to mechanical, electronic, and optical properties. They have become essential technologies in a broad variety of research fields because of their computational efficiency, scalability, robustness, and predictability. There has been an enormous growth in high-performance computing power that will benefit more compute-intensive and complementary materials modeling methods like molecular dynamics and quantum mechanics. However, QSPR has the advantage of being capable of developing robust predictive models of complex materials properties without requiring access to specialist computing facilities. QSPR modeling methods are ideally suited to study large libraries of materials, from nanoscale synthetic structures to complex biological materials. They have also been used to fill large data gaps, significantly reducing time and cost of the experimental work. However, useful QSPR models can only be built based on reliable data sets that are obtained from well-designed experiments. As such, they should ideally be combined with experimental design. Furthermore, the current QSPR paradigm is facing challenges in identifying novel descriptors that are relevant for modeling properties of new materials such as nanoparticles. There is also a need to develop improved rational prediction algorithms that can be applied efficiently to model large data sets generated by high-throughput experiments. A close collaboration between QSPR modellers and experimentalists therefore plays an important role in helping elucidate the relationship between the microscopic properties of material, their macroscopic properties, and synthesis.

ASSOCIATED CONTENT

Supporting Information

Table providing a glossary of common terms in QSPR. This information is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

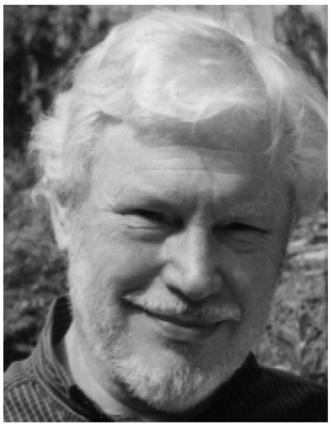
Corresponding Author

*E-mail: dave.winkler@csiro.au.

BIOGRAPHIES

Tu C. Le is a postdoctoral fellow at the Division of Materials Science and Engineering, CSIRO, Australia. She is working with Prof. David A. Winkler on studying different materials properties using quantitative structure−property relationship modeling techniques. She is also involved in simulating a number of molecular systems using molecular dynamics methods. She completed her Ph.D. on hyperbranched polymer melt simulations at Swinburne University of Technology, Australia, in 2010 under the supervision of Prof. Billy D. Todd (Swinburne University of Technology), Prof. Peter Daivis (RMIT University), and Dr. Alfred Uhlherr (CSIRO).

V. Chandana Epa is a Principal Research Scientist in the Division of Materials Science and Engineering of CSIRO. He received his Ph.D. in chemistry from the University of Alberta, Edmonton, Canada, and worked at the Herzberg Institute of Astrophysics, National Research Council of Canada, Ottawa, and the Biomolecular Research Institute, Melbourne, Australia, before joining CSIRO in 2001. His current research interests include materials simulations and modeling, in silico screening of chemical databases, structure-based ligand design, quantum mechanical studies of reaction mechanisms, application of molecular simulation methodologies to study protein dynamics and structure, and structural bioinformatics.



Frank R. Burden is an adjunct associate professor of pharmacy at Monash University, Melbourne, Australia. He has worked in the fields of quantum chemistry, microwave spectroscopy, stratospheric ozone kinetics, and various environmental problems. His current research focuses on QSAR modelling of advanced materials at the nanometer scale and investigating their constitutional-, size-, and shape-dependent properties. He has developed some unique molecular descriptors, BCUT or Burden indices, which are widely used in QSAR studies, and has also developed Bayesian likelihood linear and nonlinear regression methods.

He has published 130+ papers and spoken at numerous international conferences. He is the proprietor of SciMetrics (Victoria) and acts as a consultant to CSIRO.

David A.Winkler is a Senior Principal Research Scientist with CSIRO and an Adjunct Professor at Monash University. His research interests have mainly involved molecular design and complex systems. Recently, he has moved into regenerative medicine where he collaborates with international stem cell biologists and tissue engineers. He employs computational methods developed for small molecule research and uses them in these complex biological systems. He was awarded traveling fellowships to Kyoto and Oxford and a Newton Turner Fellowship in 2009. He is a past Board Chairman of the Royal Australian Chemical Institute and past President of Asian Federation for Medicinal Chemistry. He also represents Australia on the committee organizing of Pacifichem. He has published over 150 scientific papers and patents.

REFERENCES

(1) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. *Chem. Rev.* **2010**, *110*, 5714.

(2) Helguera, A. M.; Combes, R. D.; Gonzalez, M. P.; Cordeiro, M. N. D. S. *Curr. Top. Med. Chem.* **2008**, *8*, 1628.

(3) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, Germany, Chichester, U.K., 2000.

(4) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.

(5) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. *MATCH-Commun. Math. Co.* **2006**, *56*, 237.

(6) Dragon software, TALETE srl, Via V. Pisani, 13-20124 Milano, Italy.

(7) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551.

(8) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.

(9) Klebe, G. *Perspect. Drug Discovery* **1998**, *12*, 87.

(10) Dudek, A. Z.; Arodz, T.; Galvez, J. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213.

(11) Brereton, R. G. *Applied Chemometrics for Scientists*; John Wiley & Sons, Ltd.: Chichester, U.K., 2007.

(12) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273.

(13) Wang, X. Z.; Perston, B.; Yang, Y.; Lin, T.; Darr, J. A. *Chem. Eng. Res. Des.* **2009**, *87*, 1420.

(14) Livingstone, D. J. *Aritificial neural networks: Methods and Applications*; Springer-Verlag: New York, 2008.

(15) Winkler, D. A.; Burden, F. R. *Methods Mol. Biol.* **2002**, *201*, 325.

(16) Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.

(17) Burden, F. R.; Winkler, D. A. *QSAR Comb. Sci.* **2009**, *28*, 1092.

(18) Burden, F. R.; Winkler, D. A. *QSAR Comb. Sci.* **2009**, *28*, 645.

(19) Tarasova, A.; Burden, F.; Gasteiger, J.; Winkler, D. A. *J. Mol. Graph. Model.* **2010**, *28*, 593.

(20) Baumann, K. *TrAC, Trends Anal. Chem.* **2003**, *22*, 395.

(21) Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martinez-Mayorga, K.; Rojas, J. A. Y.; Bernard, P. *Curr. Med. Chem.* **2009**, *16*, 4297.

(22) Topliss, J. G.; Costello, R. J. *J. Med. Chem.* **1972**, *15*, 1066.

(23) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238.

(24) Livingstone, D. J.; Salt, D. W. *Rev. Comp. Chem.* **2005**, *21*, 287.

(25) Taylor, M.; Urquhart, A. J.; Anderson, D. G.; Langer, R.; Davies, M. C.; Alexander, M. R. *Surf. Interface Anal.* **2009**, *41*, 127.

(26) Konovalov, D. A.; Llewellyn, L. E.; Heyden, Y. V.; Coomans, D. *J. Chem. Inf. Model.* **2008**, *48*, 2081.

(27) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. *J. Med. Chem.* **1998**, *41*, 2553.

(28) Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.

(29) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. *J. Chem. Inf. Model.* **2008**, *48*, 1733.

(30) Weaver, S.; Gleeson, N. P. *J. Mol. Graph. Model.* **2008**, *26*, 1315.

(31) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Mueller, K. R. *J. Comput. Aid. Mol. Des.* **2007**, *21*, 485.

(32) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 445.

(33) Puzyn, T.; Leszczynska, D.; Leszczynski, J. *Small* **2009**, *5*, 2494.

(34) Nel, A.; Xia, T.; Madler, L.; Li, N. *Science* **2006**, *311*, 622.

(35) Karelson, M.; Maran, U.; Wang, Y. L.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551.

(36) Ruoff, R. S.; Tse, D. S.; Malhotra, R.; Lorents, D. C. *J. Phys. Chem.* **1993**, *97*, 3379.

(37) Heymann, D. *Fullerene Sci. Tech.* **1996**, *4*, 509.

(38) Murray, J. S.; Gagarin, S. G.; Politzer, P. *J. Phys. Chem.* **1995**, *99*, 12081.

(39) Marcus, Y. *J. Phys. Chem. B* **1997**, *101*, 8617.

(40) Marcus, Y.; Smith, A. L.; Korobov, M. V.; Mirakyan, A. L.; Avramenko, N. V.; Stukalin, E. B. *J. Phys. Chem. B* **2001**, *105*, 2499.

(41) Makitra, R. G.; Pristanskii, R. E.; Flyunt, R. I. *Russ. J. Gen. Chem.* **2003**, *73*, 1227.

(42) Abraham, M. H.; Green, C. E.; Acree, W. E. *J. Chem. Soc., Perkin Trans. 2* **2000**, 281.

(43) Sivaraman, N.; Srinivasan, T. G.; Rao, P. R. V.; Natarajan, R. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1067.

(44) Kiss, I. Z.; Mandi, G.; Beck, M. T. *J. Phys. Chem. A* **2000**, *104*, 8081.

(45) Danauskas, S. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 419.

(46) Liu, H.; Yao, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *J. Phys. Chem. B* **2005**, *109*, 20565.

(47) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. *J. Math. Chem.* **2009**, *46*, 1232.

(48) Toropov, A. A.; Rasulev, B. F.; Leszczynski, D.; Leszczynski, J. *Chem. Phys. Lett.* **2007**, *444*, 209.

(49) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. *Chem. Phys. Lett.* **2007**, *441*, 119.

(50) Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97.

(51) Martin, D.; Maran, U.; Sild, S.; Karelson, M. *J. Phys. Chem. B* **2007**, *111*, 9853.

(52) Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M. G. *Bioorg. Med. Chem.* **2008**, *16*, 9957.

(53) Durdagi, S.; Mavromoustakos, T.; Papadopoulos, M. G. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 6283.

(54) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. *J. Comput. Chem.* **2010**, *31*, 381.

(55) Rao, C. N. R.; Satishkumar, B. C.; Govindaraj, A.; Nath, M. *ChemPhysChem* **2001**, *2*, 78.

(56) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. *Comput. Biol. Chem.* **2007**, *31*, 127.

(57) Toropov, A. A.; Leszczynski, J. *Chem. Phys. Lett.* **2006**, *433*, 125.

(58) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. *Mater. Lett.* **2007**, *61*, 4777.

(59) Marinescu, G.; Patron, L.; Culita, D. C.; Neagoe, C.; Lepadatu, C. I.; Balint, I.; Bessais, L.; Cizmas, C. B. *J. Nanopart. Res.* **2006**, *8*, 1045.

(60) Stone, V.; Donaldson, K. *Nat. Nanotechnol.* **2006**, *1*, 23.

(61) Borm, P. J.; Robbins, D.; Haubold, S.; Kuhlbusch, T.; Fissan, H.; Donaldson, K.; Schins, R.; Stone, V.; Kreyling, W.; Lademann, J.; Krutmann, J.; Warheit, D.; Oberdorster, E. *Part. Fibre Toxicol.* **2006**, *3*, 11.

(62) Oberdorster, G.; Stone, V.; Donaldson, K. *Nanotoxicology* **2007**, *1*, 2.

(63) Oberdorster, G.; Maynard, A.; Donaldson, K.; Castranova, V.; Fitzpatrick, J.; Ausman, K.; Carter, J.; Karn, B.; Kreyling, W.; Lai, D.; Olin, S.; Monteiro-Riviere, N.; Warheit, D.; Yang, H. *Part. Fibre Toxicol.* **2005**, *2*, 8.

(64) Fourches, D.; Pu, D. Q. Y.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. *ACS Nano* **2010**, *4*, 5703.

(65) Shevchenko, V. Y.; Madison, A. E.; Shudegov, V. E. *Glass Phys. Chem.* **2003**, *29*, 577.

(66) Yao, S.; Shoji, T.; Iwamoto, Y.; Kamei, E. *Comp. Theor. Polym. Sci.* **1999**, *9*, 41.

(67) Cruz, V. L.; Martinez, S.; Martinez-Salazar, J.; Polo-Ceron, D.; Gomez-Ruiz, S.; Fajardo, M.; Prashar, S. *Polymer* **2007**, *48*, 4663.

(68) Cruz, V. L.; Ramos, J.; Martinez, S.; Munoz-Escalona, A.; Martinez-Salazar, J. *Organometallics* **2005**, *24*, 5095.

(69) Cruz, V.; Ramos, J.; Munoz-Escalona, A.; Lafuente, P.; Pena, B.; Martinez-Salazar, J. *Polymer* **2004**, *45*, 2061.

(70) Wigum, H.; Solli, K. A.; Stovneng, J. A.; Rytter, E. *J. Polym. Sci., Part A* **2003**, *41*, 1622.

(71) Burello, E.; Farrusseng, D.; Rothenberg, G. *Adv. Synth. Catal.* **2004**, *346*, 1844.

(72) an der Heiden, M. R.; Plenio, H.; Immel, S.; Burello, E.; Rothenberg, G.; Hoefsloot, H. C. J. *Chem.—Eur. J.* **2008**, *14*, 2857.

(73) Drummond, M. L.; Sumpter, B. G. *Inorg. Chem.* **2007**, *46*, 8613.

(74) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. *J. Med. Chem.* **1999**, *42*, 573.

(75) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. *Methods Mol. Biol.* **2004**, *275*, 131.

(76) Yao, S. G.; Tanaka, Y. *Macromol. Theor. Simul.* **2001**, *10*, 850.

(77) Hattori, T.; Kito, S. *Catal. Today* **1995**, *23*, 347.

(78) Sasaki, M.; Hamada, H.; Kintaichi, Y.; Ito, T. *Appl. Catal., A* **1995**, *132*, 261.

(79) Hou, Z. Y.; Dai, Q. L.; Wu, X. Q.; Chen, G. T. *Appl. Catal., A* **1997**, *161*, 183.

(80) Huang, K.; Chen, F. Q.; Lu, D. W. *Appl. Catal., A* **2001**, *219*, 61.

(81) Corma, A.; Serra, J. M.; Argente, E.; Botti, V.; Valero, S. *ChemPhysChem* **2002**, *3*, 939.

(82) Moliner, M.; Serra, J. M.; Corma, A.; Argente, E.; Valero, S.; Botti, V. *Microporous Mesoporous Mater.* **2005**, *78*, 73.

(83) Klanner, C.; Farrusseng, D.; Baumes, L.; Lengliz, M.; Mirodatos, C.; Schuth, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 5347.

(84) Farrusseng, D.; Klanner, C.; Baumes, L.; Lengliz, M.; Mirodatos, C.; Schuth, F. *QSAR Comb. Sci.* **2005**, *24*, 78.

(85) Corma, A.; Moliner, M.; Serra, J. M.; Serna, P.; Diaz-Cabanas, M. J.; Baumes, L. A. *Chem. Mater.* **2006**, *18*, 3287.

(86) Corma, A.; Serra, J. M.; Serna, P.; Moliner, M. *J. Catal.* **2005**, *232*, 335.

(87) Baumes, L. A.; Serra, J. M.; Serna, P.; Corma, A. *J. Comb. Chem.* **2006**, *8*, 583.

(88) Hemmateenejad, B.; Sanchooli, M.; Mehdipour, A. *J. Phys. Org. Chem.* **2009**, *22*, 613.

(89) Cruz, V. L.; Martinez, J.; Martinez-Salazar, J.; Ramos, J.; Reyes, M. L.; Toro-Labbe, A.; Gutierrez-Giliva, S. *Polymer* **2007**, *48*, 7672.

(90) Fayet, G.; Raybaud, P.; Toulhoat, H.; de Bruin, T. *J. Mol. Struct., Theochem* **2009**, *903*, 100.

(91) Beckers, J.; Clerc, F.; Blank, J. H.; Rothenberg, G. *Adv. Synth. Catal.* **2008**, *350*, 2237.

(92) Tognetti, V.; Fayet, G.; Adamo, C. *Int. J. Quantum Chem.* **2010**, *110*, 540.

(93) Artyushkova, K.; Pylypenko, S.; Olson, T. S.; Fulghum, J. E.; Atanassov, P. *Langmuir* **2008**, *24*, 9082.

(94) Maldonado, A. G.; Rothenberg, G. *Chem. Eng. Prog.* **2009**, *105*, 26.

(95) Hook, A. L.; Anderson, D. G.; Langer, R.; Williams, P.; Davies, M. C.; Alexander, M. R. *Biomaterials* **2010**, *31*, 187.

(96) Jandt, K. D. *Adv. Eng. Mater.* **2007**, *9*, 1035.

(97) Vert, M. *Prog. Polym. Sci.* **2007**, *32*, 755.

(98) Doshi, N.; Mitragotri, S. *Adv. Funct. Mater.* **2009**, *19*, 3843.

(99) Weber, N.; Bolikal, D.; Bourke, S. L.; Kohn, J. *J. Biomed. Mater. Res., Part A* **2004**, *68A*, 496.

(100) Tang, L. P.; Eaton, J. W. *J. Exp. Med.* **1993**, *178*, 2147.

(101) Smith, J. R.; Knight, D.; Kohn, J.; Rasheed, K.; Weber, N.; Kholodovych, V.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1088.

(102) MOE Molecular Operating Environment. Chemical Computing Group Inc., Montreal, Quebec, Canada, 2010; http://www.chemcomp.com/.

(103) Gubskaya, A. V.; Kholodovych, V.; Knight, D.; Kohn, J.; Welsh, W. J. *Polymer* **2007**, *48*, 5788.

(104) Smith, J. R.; Kholodovych, V.; Knight, D.; Kohn, J.; Welsh, W. J. *Polymer* **2005**, *46*, 4296.

(105) Vasina, E. N.; Paszek, E.; Nicolau, D. V.; Nicolau, D. V. *Lab Chip* **2009**, *9*, 891.

(106) Kholodovych, V.; Smith, J. R.; Knight, D.; Abramson, S.; Kohn, J.; Welsh, W. J. *Polymer* **2004**, *45*, 7367.

(107) Smith, J. R.; Kholodovych, V.; Knight, D.; Welsh, W. J.; Kohn, J. *QSAR Comb. Sci.* **2005**, *24*, 99.

(108) Kholodovych, V.; Gubskaya, A. V.; Bohrer, M.; Harris, N.; Knight, D.; Kohn, J.; Welsh, W. J. *Polymer* **2008**, *49*, 2435.

(109) Aksyonova, T. I.; Volkovich, V. V.; Tetko, I. V. *SAMS* **2003**, *43*, 1331.

(110) Linati, L.; Lusvardi, G.; Malavasi, G.; Menabue, L.; Menziani, M. C.; Mustarelli, P.; Segre, U. *J. Phys. Chem. B* **2005**, *109*, 4989.

(111) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. *J. Polym. Sci., Part B* **1988**, *26*, 2007.

(112) Koehler, M. G.; Hopfinger, A. J. *Polymer* **1989**, *30*, 116.

(113) Tan, T. T. M.; Rode, B. M. *Macromol. Theor. Simul.* **1996**, *5*, 467.

(114) Hamerton, I.; Howlin, B. J.; Larwood, V. *J. Mol. Graph.* **1995**, *13*, 14.

(115) Schut, J.; Bolikal, D.; Khan, I. J.; Pesnell, A.; Rege, A.; Rojas, R.; Sheihet, L.; Murthy, N. S.; Kohn, J. *Polymer* **2007**, *48*, 6115.

(116) Yu, X.; Wang, X.; Li, X.; Gao, J.; Wang, H. *Macromol. Theor. Simul.* **2006**, *15*, 94.

(117) Yu, X.; Wang, X.; Wang, H.; Liu, A.; Zhang, C. *J. Mol. Struct., Theochem* **2006**, *766*, 113.

(118) Liu, A.; Wang, X.; Wang, L.; Wang, H.; Wang, H. *Eur. Polym. J.* **2007**, *43*, 989.

(119) Yu, X.; Yi, B.; Wang, X.; Xie, Z. *Chem. Phys.* **2007**, *332*, 115.

(120) Yu, X. L.; Yu, W. H.; Wang, X. Y. *J. Struct. Chem.* **2009**, *50*, 821.

(121) Reynolds, C. H. *J. Comb. Chem.* **1999**, *1*, 297.

(122) Garcia-Domenech, R.; de Julian-Ortiz, J. V. *J. Phys. Chem. B* **2002**, *106*, 1501.

(123) Cao, C.; Lin, Y. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 643.

(124) Xu, J.; Chen, B. *J. Mol. Model.* **2005**, *12*, 24.

(125) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879.

(126) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300.

(127) Morrill, J. A.; Jensen, R. E.; Madison, P. H.; Chabalowski, C. F. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 912.

(128) Cypcar, C. C.; Camelio, P.; Lazzeri, V.; Mathias, L. J.; Waegell, B. *Macromolecules* **1996**, *29*, 8954.

(129) Camelio, P.; Cypcar, C. C.; Lazzeri, V.; Waegell, B. *J. Polym. Sci., Part A: Polym. Chem.* **1997**, *35*, 2579.

(130) Carro, A. M.; Campisi, B.; Camelio, P.; Phan-Tan-Luu, R. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 79.

(131) Rauzy, E.; Berro, C.; Morel, S.; Herbette, G.; Lazzeri, W.; Guglielmetti, R. *Polym. Int.* **2004**, *53*, 455.

(132) Sumpter, B. G.; Noid, D. W. *Macromol. Theor. Simul.* **1994**, *3*, 363.

(133) Ulmer, C. W.; Smith, D. A.; Sumpter, B. G.; Noid, D. I. *Comp. Theor. Polym. Sci.* **1998**, *8*, 311.

(134) Joyce, S. J.; Osguthorpe, D. J.; Padgett, J. A.; Price, G. J. *J. Chem. Soc., Farad. Trans.* **1995**, *91*, 2491.

(135) Sun, H.; Tang, Y.; Wu, G. *Macromol. Res.* **2002**, *10*, 13.

(136) Mattioni, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232.

(137) Gao, J.; Wang, X.; Li, X.; Yu, X.; Wang, H. *J. Mol. Model.* **2006**, *12*, 513.

(138) Liu, W.; Yi, P.; Tang, Z. *QSAR Comb. Sci.* **2006**, *25*, 936.

(139) Liu, W.; Cao, C. *Colloid Polym. Sci.* **2009**, *287*, 811.

(140) Liu, W. Q. *Polym. Eng. Sci.* **2010**, *50*, 1547.

(141) Afantitis, A.; Melagraki, G.; Makridima, K.; Alexandridis, A.; Sarimveis, H.; Iglessi-Markopoulou, O. *J. Mol. Struct., Theochem* **2005**, *716*, 193.

(142) Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *Macromol. Symp.* **2006**, *234*, 13.

(143) Duce, C.; Micheli, A.; Starita, A.; Tine, M. R.; Solaro, R. *Macromol. Rapid Commun.* **2006**, *27*, 711.

(144) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *Polymer* **2007**, *48*, 7121.

(145) Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *J. Math. Chem.* **2009**, *46*, 729.

(146) Bertinetto, C. G.; Duce, C.; Micheli, A.; Solaro, R.; Tine, M. R. *Mol. Inf.* **2010**, *29*, 635.

(147) Sun, H.; Tang, Y. W.; Wu, G. S.; Zhang, F. S. *J. Polym. Sci., Part B: Polym. Phys.* **2002**, *40*, 454.

(148) Van Krevelen, D. W. *Properties of polymers*; Elsevier: Amsterdam, The Netherlands, 1976.

(149) Bicerano, J. *Prediction of Polymer Property*; Marcel Dekker, Inc: New York, 2002.

(150) Vaz, R. J. *Makromol. Chem., Macromol. Symp.* **1993**, *65*, 261.

(151) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

(152) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tine, M. R. *J. Mol. Graph. Model.* **2009**, *27*, 797.

(153) Van Krevelen, D. W. *Properties of polymers*; Elsevier: Amsterdam, The Netherlands, 1990.

(154) Yu, X.; Xie, Z.; Yi, B.; Wang, X.; Liu, F. *Eur. Polym. J.* **2007**, *43*, 818.

(155) Xu, J.; Liu, H.; Li, W.; Zou, H.; Xu, W. *Macromol. Theor. Simul.* **2008**, *17*, 470.

(156) Xu, J.; Chen, B.; Liang, H. *Macromol. Theor. Simul.* **2008**, *17*, 109.

(157) Yu, X.; Wang, X.; Wang, H.; Li, X.; Gao, J. *QSAR Comb. Sci.* **2006**, *25*, 156.

(158) Katritzky, A. R.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171.

(159) Holder, A.; Ye, L.; Eick, J.; Chappelow, C. *QSAR Comb. Sci.* **2006**, *25*, 905.

(160) Yu, X.; Yi, B.; Liu, F.; Wang, X. *React. Funct. Polym.* **2008**, *68*, 1557.

(161) Patel, H. C.; Tokarski, J. S.; Hopfinger, A. J. *Pharm. Res.* **1997**, *14*, 1349.

(162) Tokarski, J. S.; Hopfinger, A. J.; Hobbs, J. D.; Ford, D. M.; Faulon, J.-L. M. *Comput. Theor. Polym. Sci.* **1997**, *7*, 199.

(163) Gharagheizi, F. *Comput. Mater. Sci.* **2007**, *40*, 159.

(164) Mallakpour, S.; Hatami, M.; Golmohammadi, H. *Polymer* **2010**, *51*, 3568.

(165) Roy, N. K.; Potter, W. D.; Landau, D. P. *IEEE Trans. Neur. Net.* **2006**, *17*, 1001.

(166) Chandola, M.; Marathe, S. *J. Mol. Graph. Model.* **2008**, *26*, 824.

(167) Holder, A. J.; Liu, Y. *Dent. Mater.* **2010**, *26*, 840.

(168) Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. *Biosens. Bioelect.* **2007**, *22*, 3309.

(169) Takaeda, Y.; Yagi, K. *Polym. News* **2003**, *28*, 352.

(170) Japanese National Institute for Materials Science, 2009.

(171) Greaves, T. L.; Drummond, C. J. *Chem. Rev.* **2008**, *108*, 206.

(172) Armand, M.; Endres, F.; MacFarlane, D. R.; Ohno, H.; Scrosati, B. *Nature Mat.* **2009**, *8*, 621.

(173) Maginn, E. J. *J. Phys.: Condens. Matter* **2009**, *21*.

(174) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71.

(175) Crossfire Database Suite; Elsevier B.V., 2010; http://info.crossfiredatabases.com/.

(176) Trohalaki, S.; Pachter, R. *QSAR Comb. Sci.* **2005**, *24*, 485.

(177) Trohalaki, S.; Pachter, R.; Drake, G. W.; Hawkins, T. *Energy Fuels* **2005**, *19*, 279.

(178) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. *J. Chem. Inf. Model.* **2007**, *47*, 1111.

(179) Carrera, G. V. S. M.; Branco, L. C.; Aires-De-Sousa, J.; Afonso, C. A. M. *Tetrahedron* **2008**, *64*, 2216.

(180) Tamm, K.; Burk, P. *J. Mol. Model.* **2006**, *12*, 417.

(181) Eike, D. M.; Brennecke, J. F.; Maginn, E. J. *Ind. Eng. Chem. Res.* **2004**, *43*, 1039.

(182) Katritzky, A. R.; Kuanar, M.; Stoyanova-Slavova, I. B.; Slavov, S. H.; Dobchev, D. A.; Karelson, M.; Acree, W. E. *J. Chem. Eng. Data* **2008**, *53*, 1085.

(183) Xi, L.; Sun, H.; Li, J.; Liu, H.; Yao, X.; Gramatica, P. *Chem. Eng. J.* **2010**, *163*, 195.

(184) Bini, R.; Malvaldi, M.; Pitner, W. R.; Chiappe, C. *J. Phys. Org. Chem.* **2008**, *21*, 622.

(185) Tochigi, K.; Yamamoto, H. *J. Phys. Chem. C* **2007**, *111*, 15989.

(186) Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K. *Fluid Phase Equilib.* **2007**, *261*, 434.

(187) Billard, I.; Marcou, G.; A., O.; Varnek, A. *J. Phys. Chem. B* **2011**, *115*, 93.

(188) Couling, D. J.; Bernot, R. J.; Docherty, K. M.; Dixon, J. K.; Maginn, E. J. *Green Chem.* **2006**, *8*, 82.

(189) Garcia-Lorenzo, A.; Tojo, E.; Tojo, J.; Teijeira, M.; Rodriguez-Berrocal, F. J.; Gonzalez, M. P.; Martinez-Zorzano, V. S. *Green Chem.* **2008**, *10*, 508.

(190) Torrecilla, J. S.; Garcia, J.; Rojo, E.; Rodriguez, F. *J. Hazart. Mater.* **2009**, *164*, 182.

(191) National Institute of Standards and Technology: Boulder, CO, 2006.

(192) Dong, Q.; Muzny, C. D.; Kazakov, A.; Diky, V.; Magee, J. W.; Widegren, J. A.; Chirico, R. D.; Marsh, K. N.; Frenkel, M. *J. Chem. Eng. Data* **2007**, *52*, 1151.

(193) Bartle, K. D.; Clifford, A. A.; Jafar, S. A.; Shilstone, G. F. *J. Phys. Chem. Ref. Data* **1991**, *20*, 713.

(194) Ferri, A.; Banchero, A.; Manna, L.; Sicardi, S. *J. Supercrit. Fluids* **2004**, *32*, 27.

(195) Hemmateenejad, B.; Shamsipur, M.; Miri, R.; Elyasi, M.; Foroghinia, F.; Sharghi, H. *Anal. Chim. Acta* **2008**, *610*, 25.

(196) Tabaraki, R.; Khayamian, T.; Ensafi, A. A. *Dyes Pigm.* **2007**, *73*, 230.

(197) Tabaraki, R.; Khayamian, T.; Ensafi, A. A. *J. Mol. Graph. Model.* **2006**, *25*, 46.

(198) Khayamian, T.; Esteki, M. *J. Supercrit. Fluids* **2004**, *32*, 73.

(199) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. *J. Pharm. Sci.* **2009**, *98*, 861.

(200) Guo, D.; Wang, Y.; Nan, C.; Li, L.; Xia, J. *Sens. Actuators, A* **2002**, *102*, 93.

(201) Guo, D.; Wang, Y.; Xia, J.; Nan, C.; Li, L. *J. Eur. Ceram. Soc.* **2002**, *22*, 1867.

(202) Cai, K.; Xia, J.; Li, L.; Gui, Z. *Comput. Mater. Sci.* **2005**, *34*, 166.

(203) Scott, D. J.; Coveney, P. V.; Kilner, J. A.; Rossiny, J. C. H.; Alford, N. M. N. *J. Eur. Ceram. Soc.* **2007**, *27*, 4425.