MQSPR Modeling in Materials Informatics: A Way to Shorten Design Cycles?

N. Sukumar[1,2], Michael Krein[2], Qiong Luo[2], Curt Breneman[2]

[1] Department of Chemistry, Shiv Nadar University, Chithera, Dadri 203207, UP, INDIA

Email: nagams@rpi.edu; n.sukumar@snu.edu.in

[2] Rensselaer Exploratory Center for Cheminformatics Research and Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180, USA

Abstract:

We demonstrate applications of Quantitative Structure Property Relationship (QSPR) modeling to supplement first-principles computations in materials design. We have here focused on the design of polymers with specific electronic properties. We first show that common materials properties such as the glass transition temperature (Tg) can be effectively modeled by QSPR, to generate highly predictive models that relate polymer repeat unit structure to Tg. Next, QSPR modeling is shown to supplement and guide first-principles DFT computations in the design of polymers with specific dielectric properties, thereby leveraging the power of first-principles computations by providing high-throughput capability. Our approach consists of multiple rounds of validated MQSPR modeling and DFT computations to optimize the polymer skeleton as well as functional group substitutions thereof. Rigorous model validation protocols ensure that the statistical models are able to make valid predictions on molecules outside the training set. Future work with inverse QSPRs have the potential to further reduce the time to optimize materials properties.

Keywords:

polymers, QSPR, informatics, Tg, dielectric, polarizability, statistical modeling

I. Introduction

Discerning and exploiting patterns in chemical data is at the heart of any systematic program for materials design. MQSPR refers to the application of the science of Quantitative Structure-Property Relationship modeling to materials informatics. The existence of quantitative relationships between chemical structure and the properties of materials was first discerned through the study of linear free energy relationships [1-4] early in the last century. These studies quantified the effect of a substituent group on equilibrium or rate constants. In recent years, the tools of statistical learning and pattern recognition have been employed to discover more complex relationships hidden in the wealth of data produced by high-throughput experimentation and robotic assays. Such statistical methods typically use an array of computed structural descriptors and/or process parameters as input to a model that can be trained to predict the value of an experimental quantity. When employed instead to predict a computed rather than an experimental quantity, statistical modeling can also serve to complement and leverage the results from first-principles computations, such as those using *ab initio* quantum chemistry and density functional theory (DFT), thereby enabling quantitative predictions on many more systems than would be possible in the same time span with first-principles computations alone.

This paper deals with applications of MQSPR modeling to supplement first-principles DFT computations in the design of polymers with specific electronic properties, such as high dielectric constant and band gap for capacitors, or a specific range of glass transition temperatures. Our approach consisted of multiple rounds of validated MQSPR modeling and DFT computations to optimize the polymer skeleton as well as functional group substitutions thereof. Model validation ensures that a statistical model is able to make valid predictions on molecules outside the training set, rather simply producing *post facto* correlations on the training data. It is equally important to assess the domain of applicability of a model, *i.e.* to know if a model is capable of predicting materials properties with useful levels of accuracy in a particular part of chemistry space. Methods for model validation and applicability domain assessment are discussed more fully elsewhere [5-8].

II. Background

A. Glass Transition Temperatures T_g

The performance of polymers may be best understood by considering their thermal transitions. Among these, the glass transition temperature (T_g) is one of the most important and widely studied thermal characteristics. When an amorphous polymer undergoes the glass transition, almost all of the properties that relate to its processing, such as heat capacity, coefficient of thermal expansion and viscosity, change dramatically. This temperature represents the ease of long-range motion for polymer chains, which is reflected in several physical properties of polymers. At temperature below T_g, the motion of polymer chains is restricted to local vibration and thus polymers appear hard and glassy. At temperature above T_g, polymers become soft and rubber-like because of the increase of polymer chain mobility.

Glass transition behavior is characteristics of the fundamental dynamics of polymer chains. Though research on glass transition behavior has been an active subject over the past decades, the physics behind this phenomenon is still not fully understood. The glass transition temperatures of polymers may vary widely depending on various factors, such as the molecular weight and chemical structure of the polymer. Based on "free-volume" theory, low molecular weight polymers have lower T_g values, because low molecular weight polymer chains have more ends per unit volume than long chains, and hence more freedom of motion and higher free volume [9,10]. In general, polymers with flexible backbones and small side groups have lower T_g than polymers with rigid backbones, such as polymers containing main-chain aromatic groups. Furthermore, cross-linking, crystallization and co-polymerization all affect the glass transition temperatures of polymers.

Recently, the use of quantitative structure property relationship (QSPR) has emerged as a valuable means of predicting physical properties of polymeric materials [11]. Consequently, it would be advantageous to produce robust QSPR models that could predict T_g values for new polymeric materials. In the process of developing new materials, such models could be used as high throughput screening tools given a list of possible polymer candidates, and to exclude ones that do not fall in the desired T_g range. A substantial savings in time and money can be achieved by focusing on only those materials with appropriate T_g values.

Numerous models have been reported for predicting the T_g of amorphous polymeric materials. Generally these studies involved the use of small data sets, so the range of T_g values is quite limited. Perhaps the most widely referenced model is one generated by Bicerano [12]. This study used a data set including 320 polymer compounds, and the QSPR model utilized connectivity indices of the topology of the repeat unit of a polymer as the principal descriptors. A linear regression procedure was used to build a model with a standard deviation of 24.65 K and a correlation coefficient of 0.975 (R^2 of 0.95). However, no external data set compounds were withheld to validate this model. There is thus a need for a model that allows prediction of T_g for polymers spanning a large variety of structures. The goal of the present study was to produce robust QSPR models that can predict T_g values for a diverse set of polymers.

B. Dielectric Properties

The primary objective of this part of the study is the navigation of the polymer chemical space to screen polymers appropriate for high energy density capacitor dielectrics with an attractive combination of high dielectric constant, fast response and low loss. Thus the first objective is to scan the polymer chemical landscape rapidly in order to identify systems with a large dielectric constant and band gap adequate to provide reasonable insulating properties. We use the normalized polarizability as an indicator of the dielectric constant ε. The polarizability may be normalized either on a per volume basis (*X*) or on a per mass basis (Φ). Larger values of X and Φ lead to larger dielectric constant (ε) on account of the Clausius-Mossotti equation:

$$\frac{\varepsilon-1}{\varepsilon+2}=\frac{4\pi\alpha}{3\,V}= \qquad (1),$$

where ε is the relative permittivity (i.e., the dielectric constant) of the medium made up of polarizable units; α and V are, respectively, the polarizability (in units of volume) and the volume occupied by the polarizable units. The polarizabilities X and Φ are related through the relation:

$$X = \frac{4\pi}{3}\frac{\alpha}{V} = \frac{4\pi}{3}\frac{\alpha A}{M}\rho = \Phi\rho \qquad (2),$$

where A, M and ρ are the Avogadro number, molecular weight and density, respectively. X is a dimensionless quantity, whereas Φ has units of volume/mass.

The polarizability α, being the change in the dipole moment **μ** in response to an applied electric field **E**, is a tensor. The quantity used for modeling is the trace:

$$\mathrm{Trace}(\alpha) = \partial\mu_X/\partial E_X + \partial\mu_Y/\partial E_Y + \partial\mu_Z/\partial E_Z \qquad (3).$$

The total polarizability may be thought of as consisting of two main contributions: an electronic contribution, arising from the polarization of the electron density by the field (with the nuclei staying fixed), and the other an ionic contribution, arising from the reorientation of the molecular scaffold. The ionic contribution is often small in comparison to the electronic component, except in cases where there is significant flexibility of the molecular scaffold and a large permanent dipole moment. These are the most interesting systems from the point of view of polymer design.

One approach to the optimization of polymer dielectric properties is to identify highly polar, highly polarizable and highly rotatable functional groups to manipulate the dielectric response. This line of investigation is described in Section III.B. Another approach, described in Section III.C, is to optimize the polymer backbone to identify elemental substitutions favorable for high dielectric properties. The basic idea in either case is to collect enough information on the relationship between composition and configuration on the one hand, and dielectric properties on the other. The ultimate goal of the study is to use this knowledge to solve the "inverse problem", namely, identification of classes of polymers with an attractive set of dielectric properties.

III. Materials and Methods

A. Modeling the Glass Transition Temperatures of homopolymers with TAE descriptors

Transferable Atom Equivalent (TAE) RECON descriptors [13-15], which are based on Bader's Atoms in Molecules formalism [16] have been used for the prediction of glass transition temperatures of homopolymers [17,18]. The computational bottleneck associated with the generation of molecular descriptors from *ab initio* quantum calculations in circumvented by pre-computing a library of transferable atomic fragment densities and density-derived atomic fragment properties (the TAE library) from *ab initio* wave functions. The RECON algorithm [14,15] exploits the fact that atomic fragments constructed as in Bader's theory of Atoms-In-Molecules possess properties that are approximately additive and transferable between molecules, to enable rapid, high throughput computation of molecular electronic properties from the atomic charge density fragments stored in the TAE library. TAE descriptors encode the distributions of electron density-based molecular properties, which are much more sensitive indicators of the local chemical environment than is the density itself. The electron density-derived descriptors employed in this study are listed in Table 1. Surface extrema, surface integral averages and histogram bins derived from surface distributions for each property were used in this study.

Table 1: Transferable Atom Equivalent Electron density-derived properties

<table>
<tr><td colspan="3"><b>Surface electronic properties:</b> surface extrema, surface integral averages and histogram bins derived from surface distributions are available for each property.</td></tr>
<tr><td><i>EP</i></td><td>Electrostatic potential</td><td>$EP(r)=\sum_{\alpha}\frac{Z_{\alpha}}{|r-R_{\alpha}|}-\int\frac{\rho(r')dr'}{|r-r'|}$</td></tr>
<tr><td><i>DRN</i></td><td>Electron density gradient normal to 0.002 e/bohr^3 electron density isosurface</td><td>$\nabla\rho.\mathbf{n}$</td></tr>
<tr><td><i>G</i></td><td>Electronic gradient kinetic energy density</td><td>$G(r)=-(1/2)(\nabla\psi^{*}.\nabla\psi)$</td></tr>
<tr><td><i>K</i></td><td>Electronic Schrödinger kinetic energy density</td><td>$K(r)=-(1/2)(\psi^{*}\nabla^{2}\psi+\psi\nabla^{2}\psi^{*})$</td></tr>
<tr><td><i>DGN</i></td><td>Gradient of the Schrödinger kinetic energy density normal to surface</td><td>$\nabla K.\mathbf{n}$</td></tr>
<tr><td><i>DGN</i></td><td>Gradient of the gradient kinetic energy density normal to surface</td><td>$\nabla G.\mathbf{n}$</td></tr>
<tr><td><i>F</i></td><td>Fukui F$^-$ function scalar value</td><td>$F^{-}(r)=\left[\frac{\partial\rho(r)}{\partial N}\right]_{v}\approx\rho_{HOMO}(r)$</td></tr>
<tr><td><i>L</i></td><td>Laplacian of the electron density</td><td>$L(r)=-\frac{1}{4}\nabla^{2}\rho(r)=K(r)-G(r)$</td></tr>
<tr><td><i>BNP</i></td><td>Bare nuclear potential</td><td>$BNP(r)=\sum_{\alpha}\frac{Z_{\alpha}}{|r-R_{\alpha}|}$</td></tr>
<tr><td><i>PIP</i></td><td>Local average ionization potential</td><td>$PIP(r)=\sum_{i}\frac{\rho_{i}(r)|\varepsilon_{i}|}{\rho(r)}$</td></tr>
<tr><td colspan="2">Integrated atomic properties</td><td>Energy, Integrated electron population, Volume, Surface area</td></tr>
<tr><td colspan="2">Rot (Topological)</td><td>Total number of rotatable single bonds in the repeat unit</td></tr>
<tr><td colspan="2">RRot (Topological)</td><td>Ratio of the number of rotatable single bonds in the side chain to that in the main chain of the repeat unit</td></tr>
</table>

Some of the most commonly employed electron density-derived descriptors are the molecular electrostatic potential and Politzer's local average ionization potential [19-22]. Quantitative correlations of the values of electrostatic potential minima with the carcinogenic activities of molecules[23] have shown the promise of this descriptor for biological and environmental applications, leading to its extensive use in QSAR and drug design [24]. Maxima of the local average ionization potential identify regions in the molecule that do not give up electrons readily, while electron-donor or hydrogen bond acceptor regions correspond to minima of this function.

In this study, the polymer repeat unit was used as representative of the polymeric material. The repeat unit structures end-capped with monomers were used to construct the appropriate atom type environments for descriptor generation in the quantitative structure-property relationships (QSPR) investigation. This study also utilizes an external prediction set which validates models based on their ability to predict properties of polymeric materials that were not used in the training. The size and diversity of the external prediction set is much larger than those considered in most earlier studies.

The 300 polymer compounds used in this study came from Bicerano [12] and are listed in Table S1 of the Supplementary Information. This data set contains non-cross-linked polymers with a large variety of compositions and structural features, T_g values range from 130 K to 685 K. Polymer repeat units for 300 polymers were represented in SMILES format. The RECON algorithm uses the SMILES format for the repeat units as input, determines atom types and environments, assigns the closest match to each atom from a library of atom types (TAE library) and combines the densities and properties of the atomic fragments to compute the TAE QSPR descriptors. Polymer TAE descriptors are constructed by summing the respective atomic descriptors stored in the data files that constitute the TAE library. In addition to TAE descriptors, two new polymer descriptors, Rot and RRot, describing polymer flexibility, were imployed: Rot is the total number of rotatable single bonds in the repeat unit, RRot is the ratio of the number of rotatable single bonds in the side chain to that in the main chain of the repeat unit. A training set consisting of 214 polymer compounds was selected from the 300 polymers; the remaining 86 compounds were used as an external prediction set.

Faced with this large pool of potentially useful polymer descriptors, objective feature selection was performed to remove descriptors that contain identical information or that are highly correlated with other descriptors ("cousin" descriptors). All descriptors with zero variance were removed. Pairwise correlations were examined to remove descriptors that are highly correlated with other descriptors. If two descriptors were highly correlated, one was randomly removed from the descriptor pool. By eliminating redundant and low-information-content descriptors, high-quality data for use in modeling analysis is obtained. The reduction of the descriptor pool is also done to ensure that the ratio of descriptors to training set observations does not exceed 0.6, thereby reducing the risk of chance correlations during model development [25]. Partial Least Square (PLS) regression and Kernel Partial Least Squares (KPLS) regression were used to generate linear or nonlinear models after objective feature selection. The ultimate goal of the QSPR models is to predict T_g values for unknown polymeric materials with a useful level of accuracy. To test the robustness of the final model for extrapolative predictions, the full data set was divided into a training set and a validation set, which were together used to build a learning model, and an external prediction set, which was used only for model evaluation. The most unbiased method of building models is to employ the external prediction set only after the model is trained. In this study, a training set consisting 214 polymer compounds was selected from the 300 polymers; the remaining 86 polymers were used as an external test set to assess the predictive ability of the derived models. 149 polymer descriptors were generated, including 147 TAE descriptors from RECON and two rotatable bond descriptors. After objective feature selection, one constant feature and 65 cousin features were eliminated. QSPR models were generated using the remaining 83 features and different machine learning methods, such as Bootstrap PLS and Bootstrap KPLS.

Bootstrap PLS

The PLS method used in this study transforms the original variables into a few orthogonal "latent" variables, which are linear combinations of the original variables. PLS calculates one latent variable at a time and stops when the added information becomes insignificant, as determined by a bootstrap procedure (Figure 1). In such a procedure, 10% of the training set was randomly selected as a validation set and removed from the training set; the remaining compounds in the training set were used to develop each model. The compounds left out were then predicted from the developed model. Such a process was repeated 100 times, each time a different validation set was selected randomly and then predicted from each model developed. The sum of the squared differences between the predicted property and the experimental property for the compounds left out (predictive residual sum of squares: PRESS) was computed. The number of latent variables giving the smallest computed PRESS value was used. In this study, the optimal number (9) of latent variables was determined using 100 bootstraps. Once the optimal number of latent variables was determined, a PLS model was built on the whole training set including 214 polymer compounds. The external prediction set, which was set aside during the model development stages, was predicted to assess the predictive ability of the model (Figures 2 and 3).

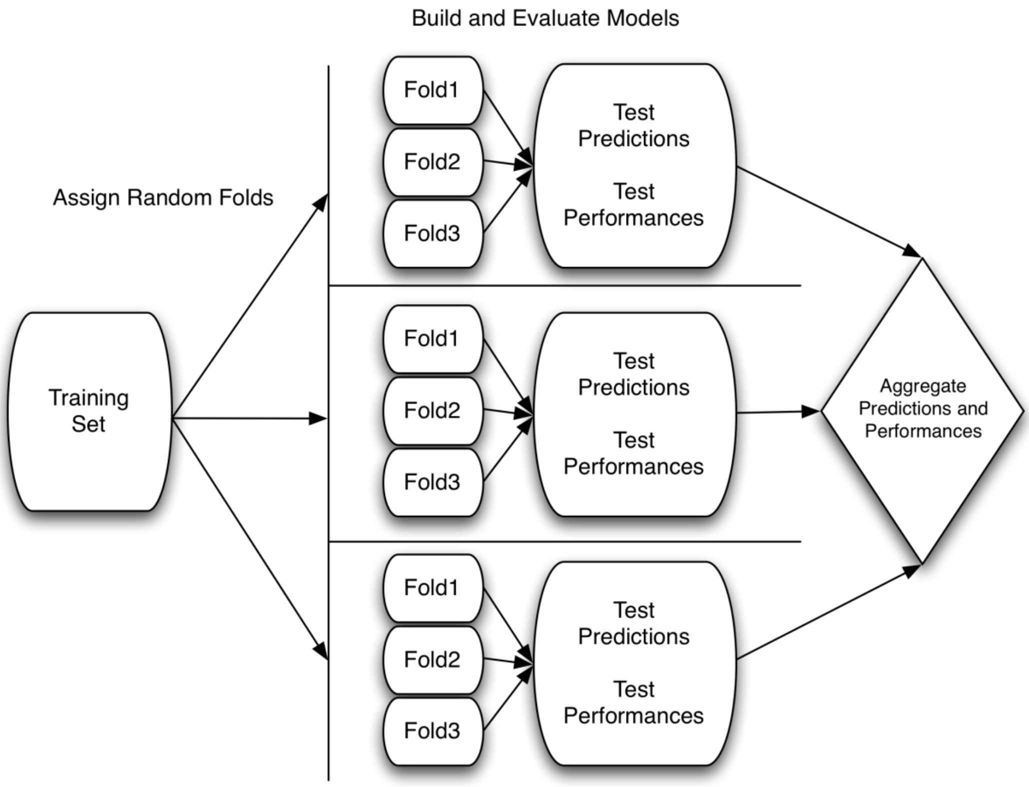

Figure 1: Example 3-Round, 3-Fold Cross-validation Procedure.

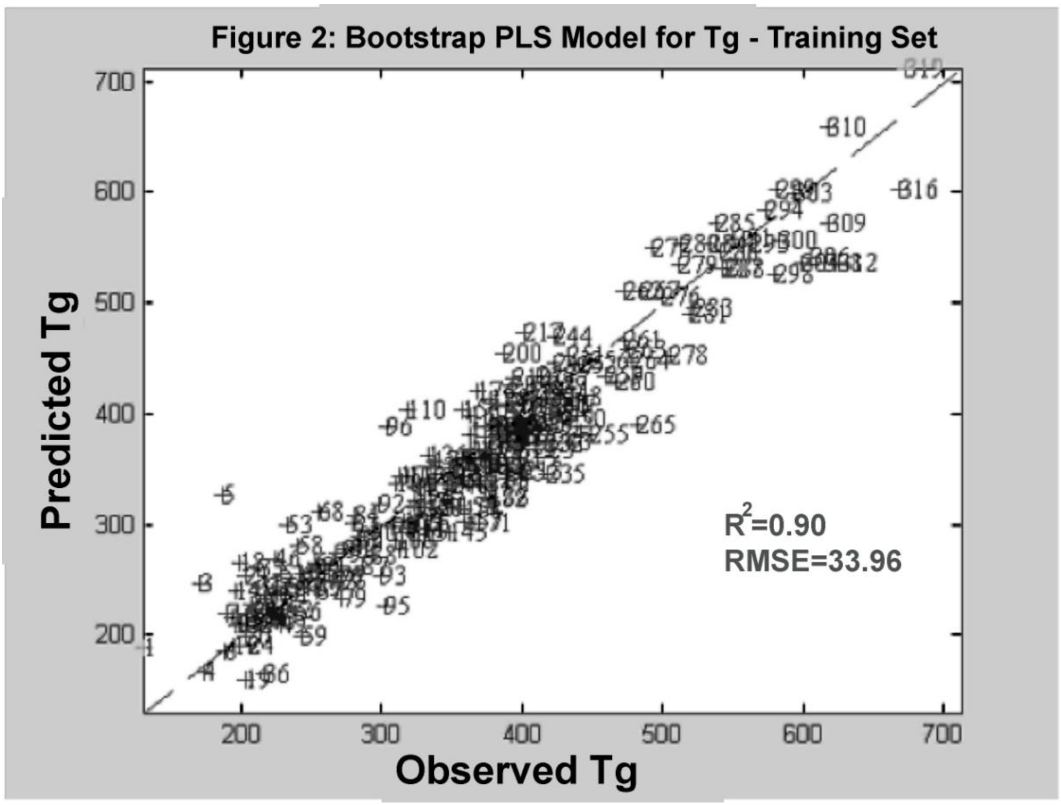

Figure 2: Bootstrap PLS Model for Tg (Training Set)

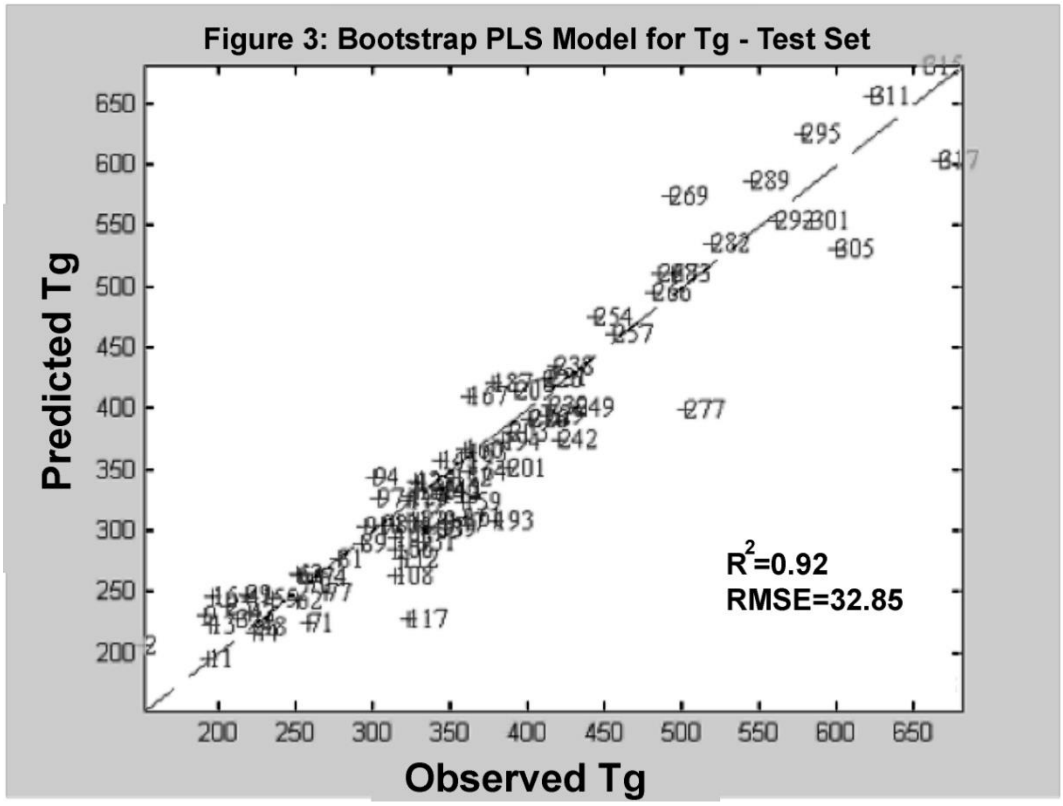

Figure 3: Bootstrap PLS Model for Tg (Test Set)

Bootstrap KPLS

As traditionally implemented, PLS is a linear modeling technique. However, it is possible to construct a nonlinear extension of this method using a nonlinear kernel. The general idea of K-PLS is to apply a kernel matrix in the process of modeling, which can be considered as a nonlinear transformation of the input data. Various K-PLS methods use different choices of the kernel, such as a polynomial kernel or a radial basis function. In this work, a Gaussian kernel, which is a widely used radial basis function, was employed. In a similar manner to PLS, the best exponent sigma for the Gaussian kernel and the number of latent variables were determined using 50 bootstraps. In this study, the optimal number of latent variables was 9 and sigma for the Gaussian kernel was 10. Figures 4 and 5 give the KPLS modeling results on the training set and the external prediction set separately.

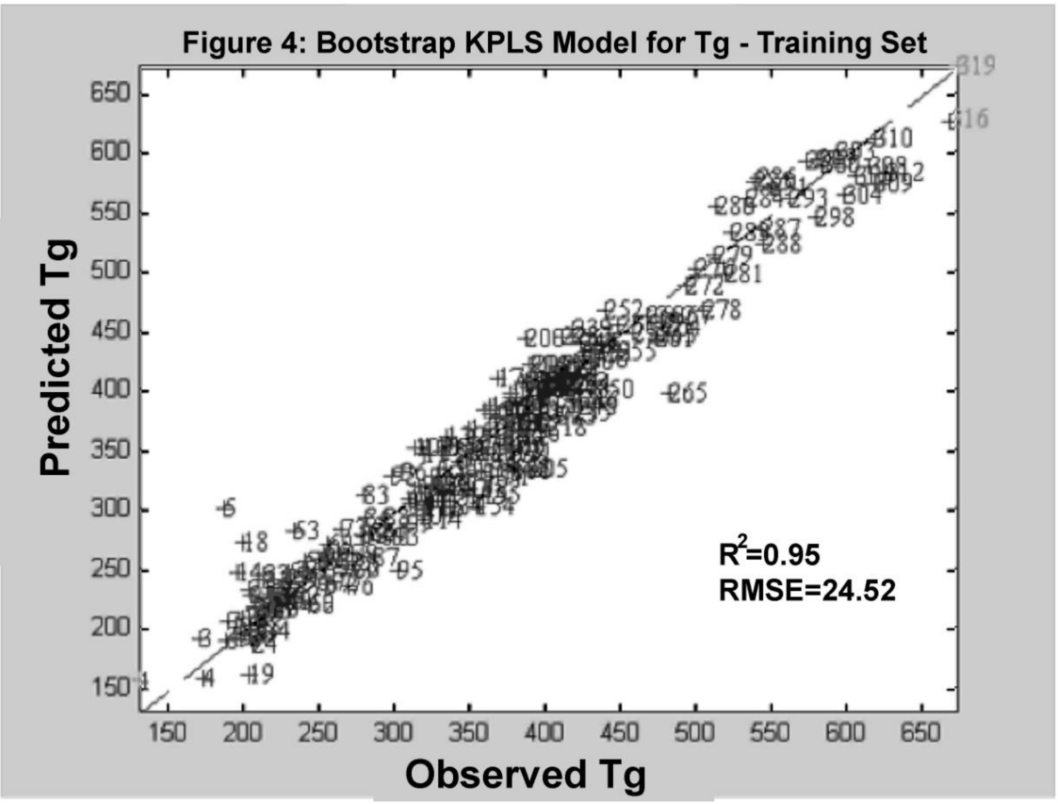

Figure 4: Bootstrap KPLS Model for Tg (Training Set)

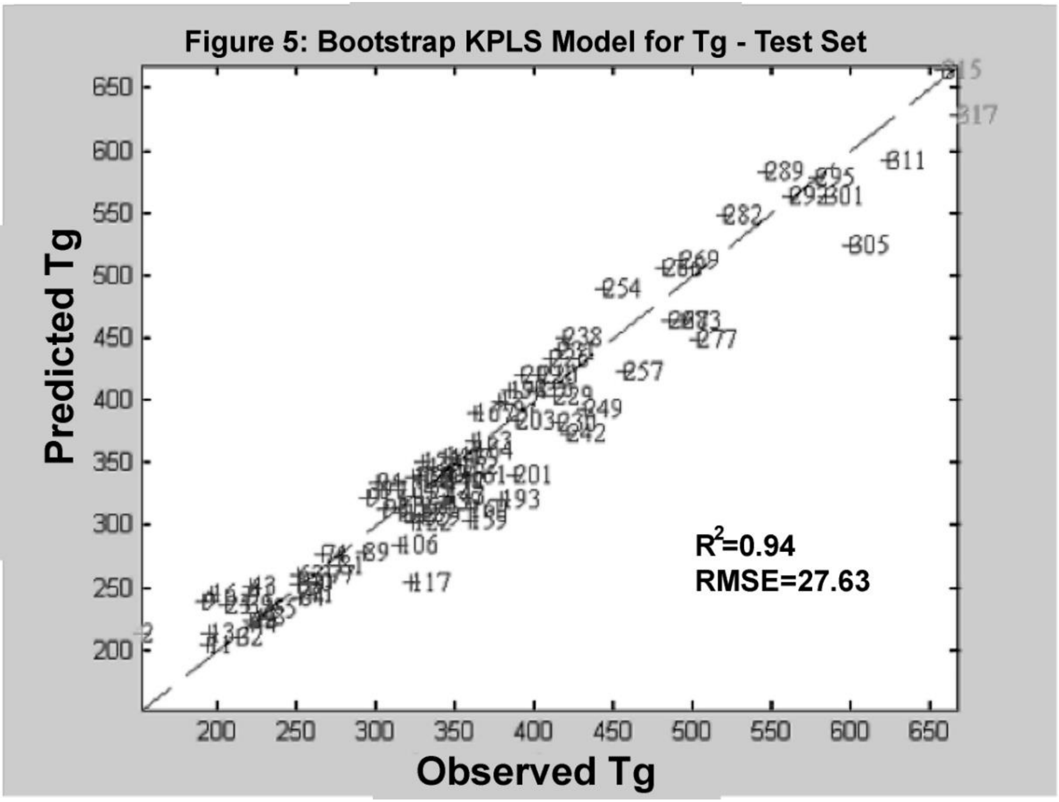

Figure 5: Bootstrap KPLS Model for Tg (Test Set)

In accordance with cheminformatics best practices, we performed cross-validation and used external test sets to determine model performance. In addition, y-scrambling was employed as a test of susceptibility of the method to over-fitting

[26,27]. In this technique, the response vector is shuffled multiple times, and "fake" QSPR models are built using the original, unscrambled chemical descriptors with shuffled responses, which are then compared to the real model. If many or all of the scrambled QSPR models built using a specific set of descriptors and a given machine learning method appear to have relatively high performance, it implies that QSPR models based on the given modeling method and descriptors demonstrate many plausible hypotheses and a lack of differentiation between them. This may be due to model overfitting, and less overall confidence should be placed in these models. Y-scrambling results on the model are shown in Figures 6.

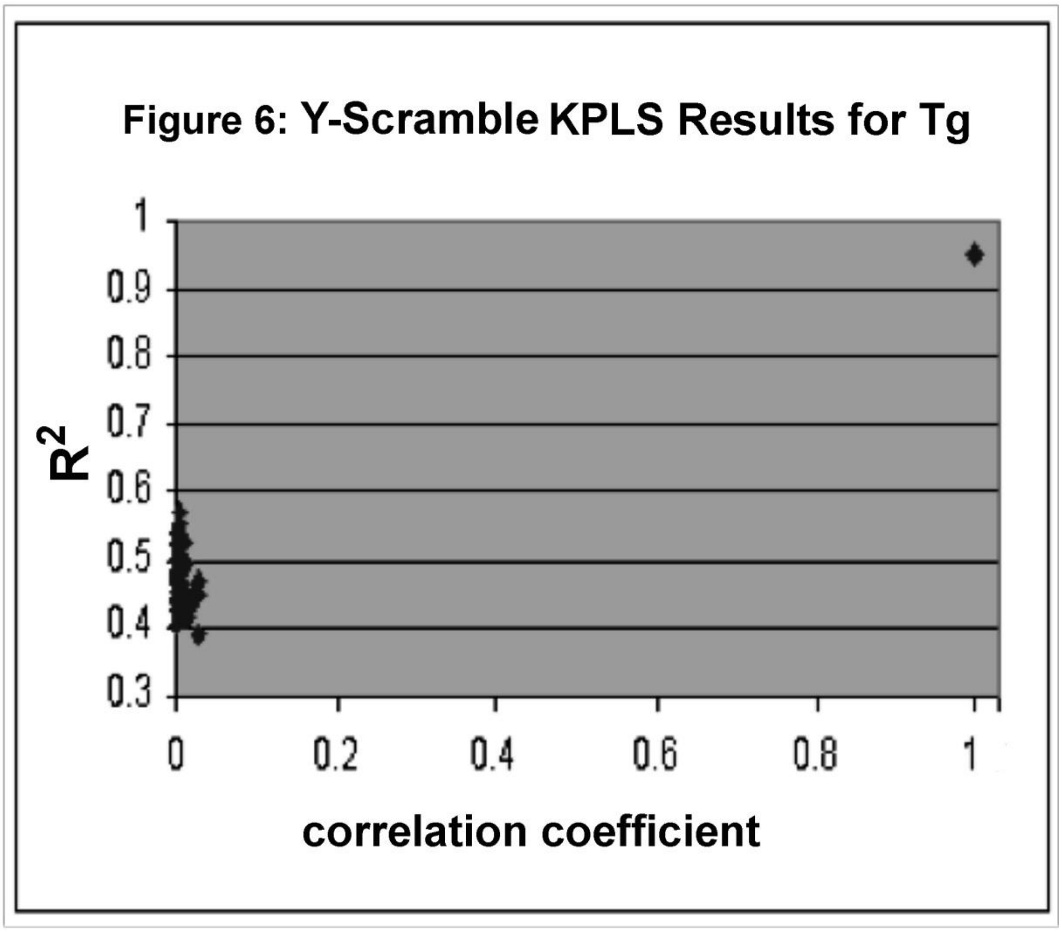

Figure 6: Y-Scrambling KPLS Results for Tg

B. Modeling the Electronic Polarizability contributions of Functional Groups with MOE descriptors

In an effort to design high dielectric constant polymers with moderately high band gaps, we constructed QSPR models for HOMO-LUMO gaps and electronic polarizabilities of polymer fragments with a variety of side-chain functionalization. The models were trained on and validated with electronic polarizabilities from *ab initio* DFT computations and HOMO-LUMO gaps from *ab initio* Hartree-Fock computations. Validated QSPR models [5,28] were constructed for HOMO-LUMO gaps and the trace of the electronic component of the polarizability tensor for polyethylene (PE) functionalized with various groups, employing constitutional and topological descriptors from Molecular Operating Environment [29] and Transferable Atom Equivalent (TAE) [15] descriptors. The functional groups used are shown in Table S2 of the Supplementary Information.

The models were trained on HOMO-LUMO gaps and polarizabilities computed from DFT (using Gaussian '03) [30] for polymer fragments of varying lengths. Model building was wrapped by 10 rounds of 10-fold cross-validation [31-33,5], where multiple data points were successively withheld for evaluation, as shown in Figure 1.

Optimum model parameters were chosen via multiple rounds of Leave-N-Out cross-validation and a grid parameter search, as shown in Figure 7. PLS latent variable selection was performed with 20 rounds of 4-fold cross-validation. All parameters were set via a default grid search so as to optimize cross-validated model R^2. Feature selection [34] was employed based on correlation and sensitivity analysis to select relevant descriptors and generate robust models, together with model validation protocols.

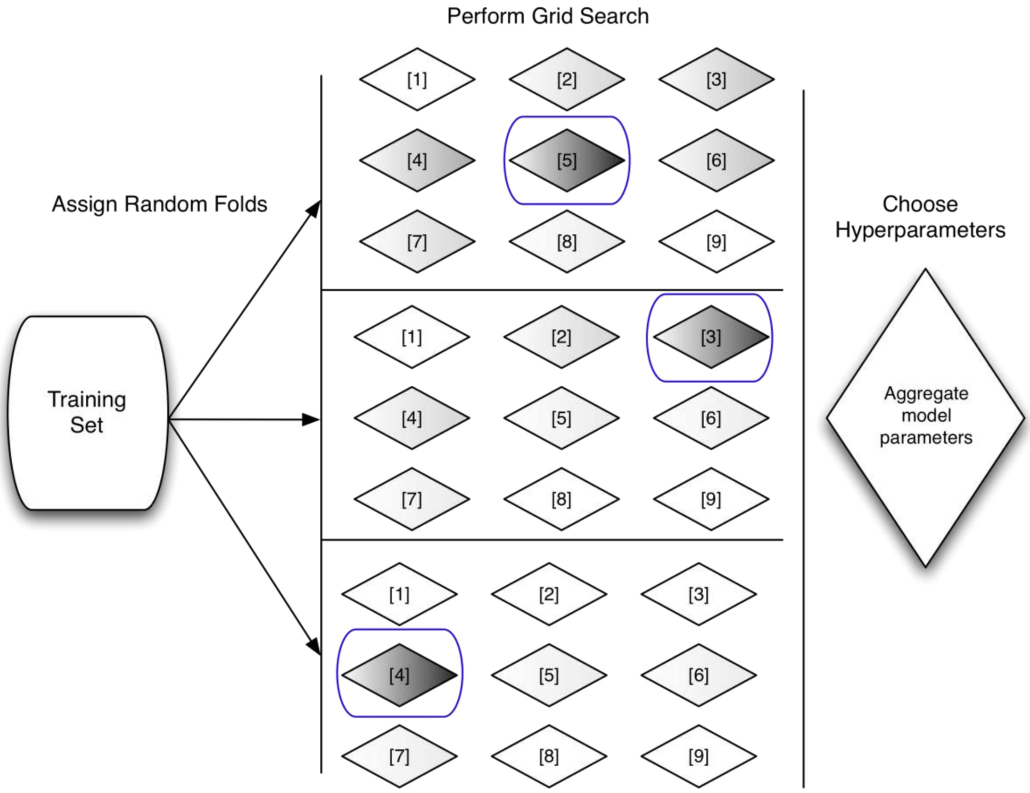

Figure 7: Model parameter selection through cross-validation and grid search. In this figure, a 3-fold cross-validation is performed.

Our studies indicated that the sum of the atomic polarizabilities (apol) from the CRC Handbook of Chemistry and Physics [35], from the MOE 2-D set [29], provides a good approximation to determine functional group contributions to electronic polarizabilities. Almost all the deviation arises from groups with Li, Na and K atoms. When these systems are removed, one obtains a much better regression, with a correlation coefficient of 0.98.

Functional group contributions with polyethylene (PE), polyacetylene (PA) and polysilene backbones exhibit similar trends, showing that apol may be used to determine additive contributions to the electronic polarizability in each case. The only major deviations were the groups with alkali atoms, indicating poor parameterization for these atoms. A 5-descriptor model, including apol and bpol (the sum of the absolute values of the difference between atomic polarizabilities of all bonded atoms in the molecule taken from the CRC Handbook of Chemistry and Physics [29]), produces a good cross-validated regression for the electronic component of the polarizability tensor, even for systems containing alkali groups.

Since the apol descriptor (sum of the atomic polarizabilities) was found to provide a good approximation to determine functional group contributions to electronic polarizabilities in most cases, and semi-empirical AM1 computations gave a reasonable estimate of Hartree-Fock HOMO-LUMO gaps, these descriptors were then used to screen a large fragment library of 13,300 functional groups from the MOE [Labute, 2000] fragment library, and 150 of the most promising functional groups with high apol per van der Waals volume and high HOMO-LUMO gap (shown as red circles in Figure 8) were identified for further analysis. The functional groups were neutralized (deprotonated) and DFT (PBE/6-31G* [36,37] computations, with full geometry optimization, were performed on polymer fragments with these 150 groups. Since the computational bottleneck is geometry optimization, which grows with length of the backbone chain, a very short two-carbon PE chain was employed to represent the polymer backbone in the *ab initio* computations.

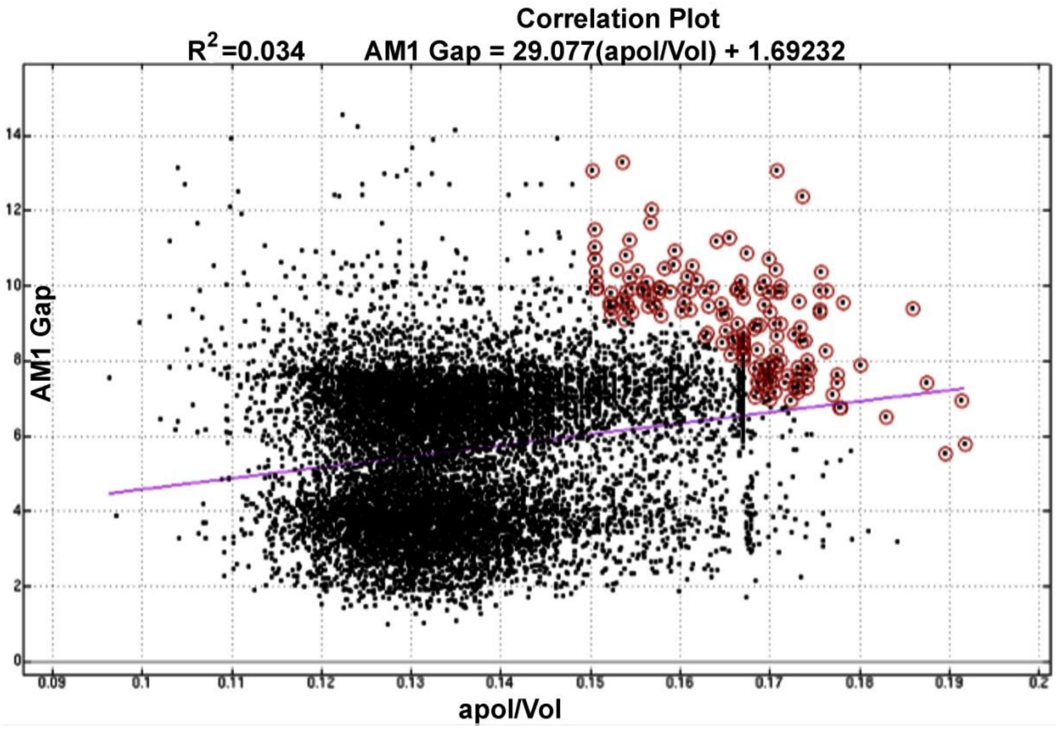

Figure 8: Library view of 13,300 functional groups. Shown highlighted are the 150 most promising compounds, in terms of high apol per van der Waals volume and high HOMO-LUMO gap.

Dipole moments and total polarizabilities were computed using DFT with the PBE functional [36,37] and 6-31G* basis in Gaussian'03 [30]. Comparing the total versus electronic polarizabilities identifies the functional groups with high ionic polarizabilities (shown as red circles in figure 9, with structures in figure 10). The functional groups were then ranked by their ionic polarizabilities, as determined from *ab initio* DFT computations, and by other criteria such as the total dipole moment and the number of rotatable bonds.

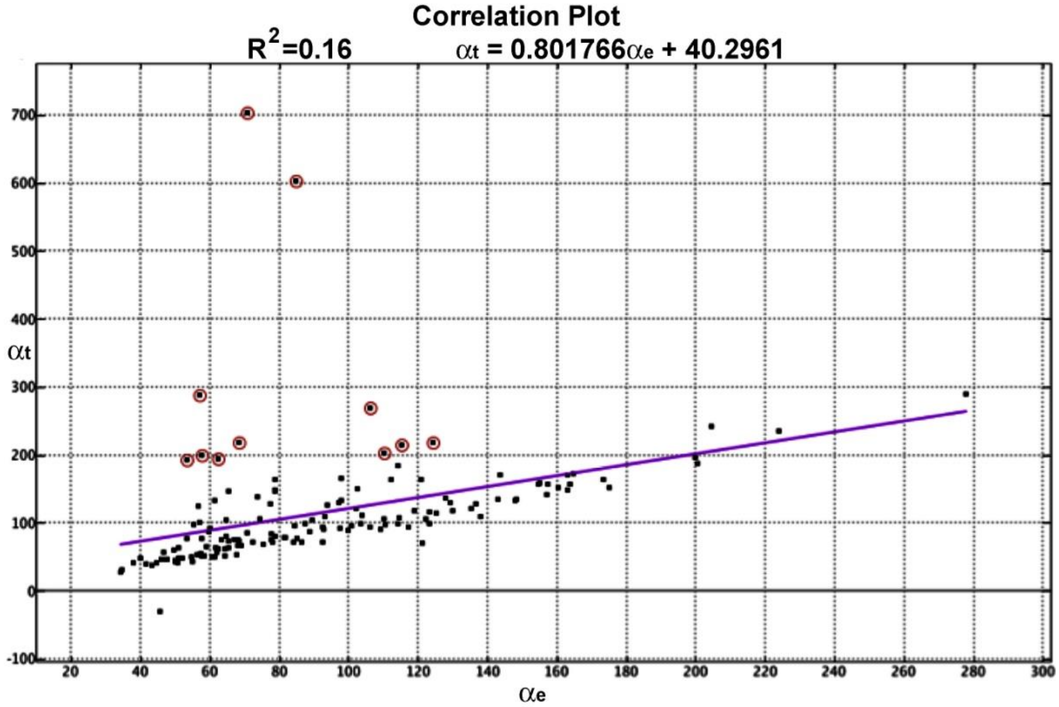

Figure 9: Total polarizability vs. electronic polarizability for the most promising functional groups.

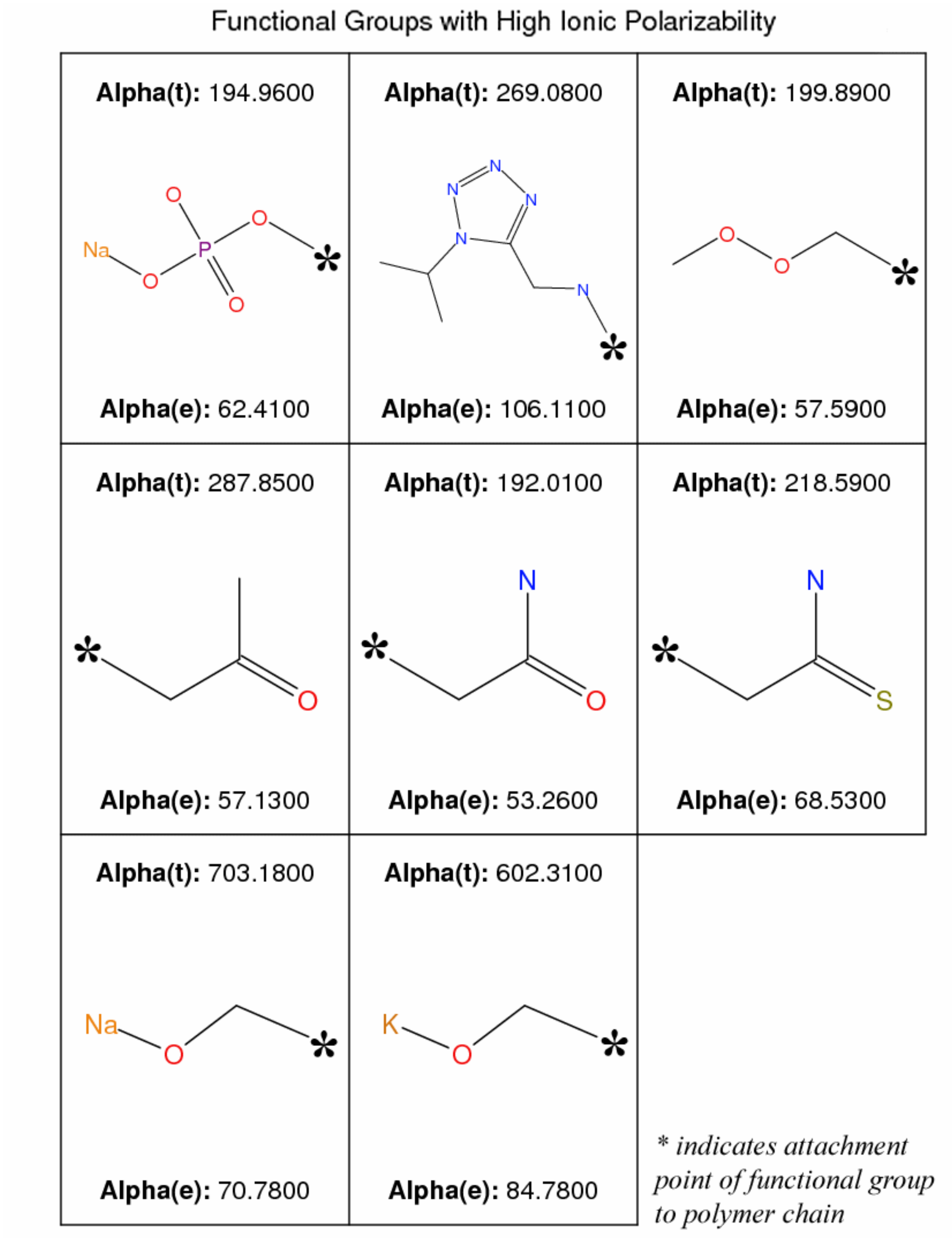

Figure 10: Most promising functional groups selected in terms of high ionic polarizability.

C. Modeling the HOMO-LUMO Gaps and Electronic Polarizabilities of block copolymers

Modification of the polymer backbone is another potential approach to optimize polymer dielectric properties. Thus we constructed validated QSPR models for the HOMO-LUMO gaps and electronic polarizabilities of representative carbon-, silicon- and germanium-containing block copolymers. The polymer fragments employed are shown in Table S3 of the Supplementary Information. The HOMO-LUMO gaps and electronic polarizabilities were computed [38] using DFT with van der Waals-augmented functionals. Partial least squares models for their HOMO-LUMO gaps are shown in Figures 11(a) and (b).

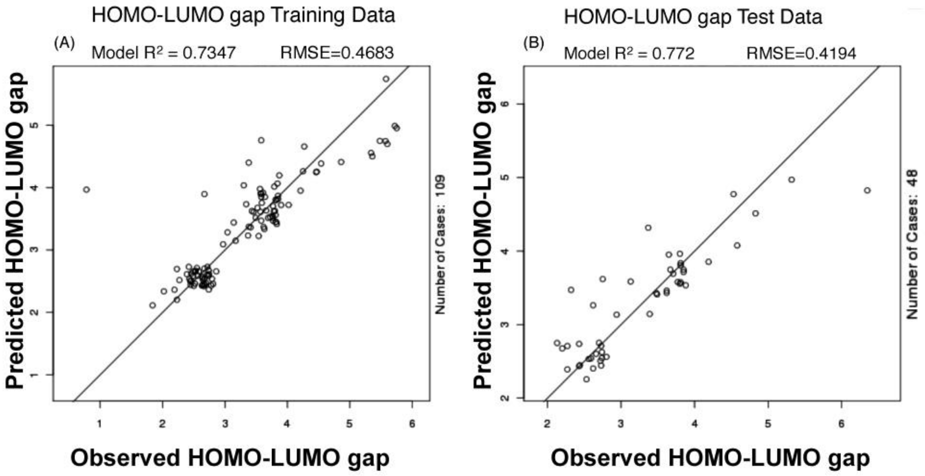

Figure 11: HOMO-LUMO gap models. Shown in (A) on the left is the performance of the model training data. Shown in (B) is the performance of the model applied to previously withheld test data.

By performing sensitivity analysis (shown in Figure S1 in the Supporting Information) on the model, we find the most important descriptors, in decreasing order of sensitivity, to be: the relative negative partial charge, the bond stretch potential energy, the Carbon valence connectivity index of order 0, the first shape moment with respect to the closest atom to the molecular centroid, the fractional positive van der Waals surface area, the second bin of GCUT descriptors calculated from the eigenvalues of a modified graph distance adjacency matrix, and the sum of the accessible van der Waals surface areas for atoms with partial charges falling in the first bin. The results of y-scrambling validation are shown in Figures 12 (a) – (d).

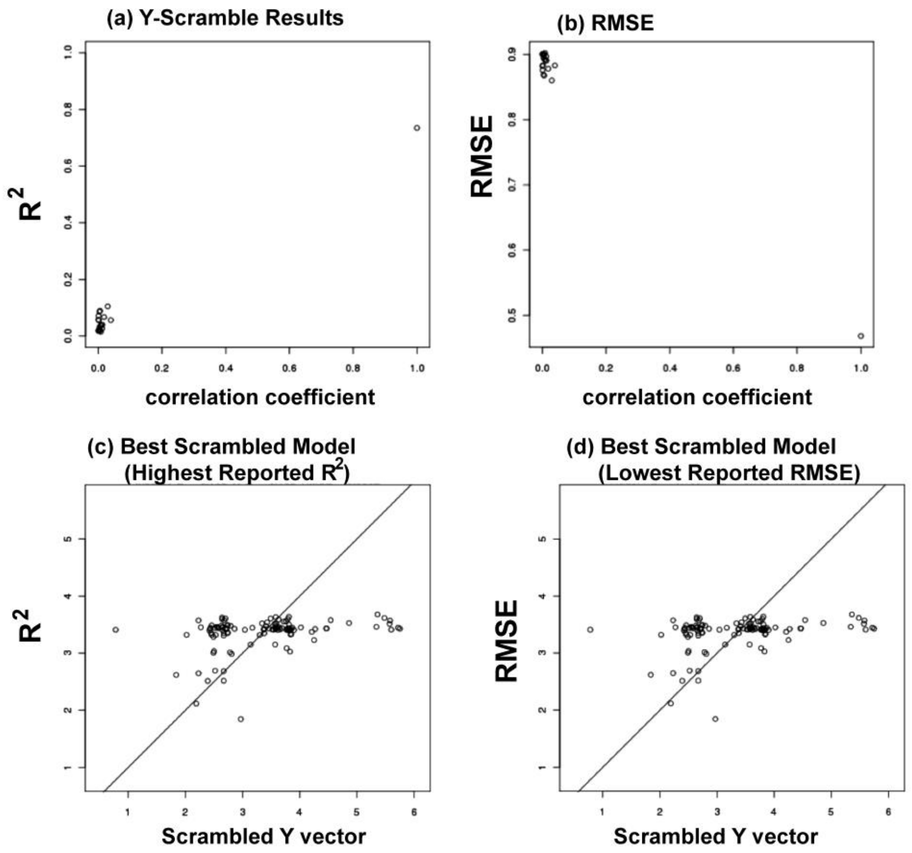

Figure 12: Y-scrambling validations of PLS models for the HOMO-LUMO gap. (a) R^2, (b) root mean squared error - RMSE. (c) Scrambled Model with the highest R^2, (d) Scrambled Model with the lowest RMSE.

Partial least squares models for the electronic polarizability of block copolymers are shown in Figures 13 (a) and (b).

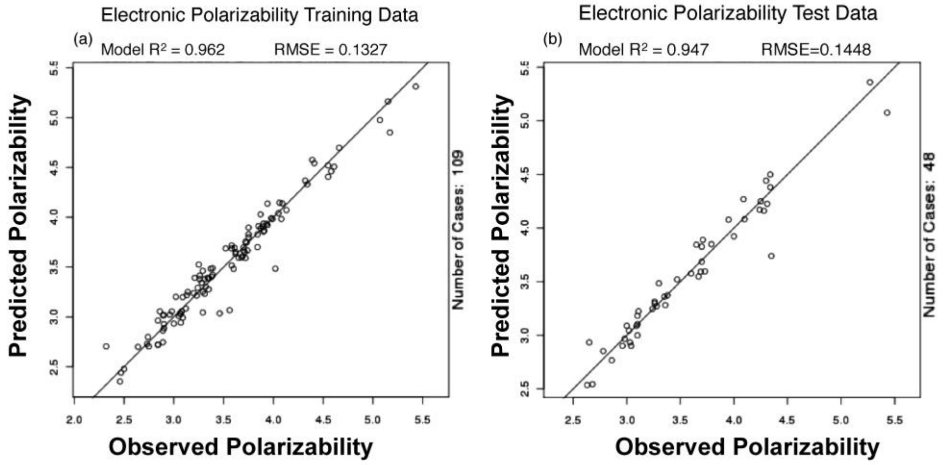

Figure 13: PLS models for the electronic polarizability of block copolymers (a) Training data, (b) Test set.

The most important descriptors used in the model, in decreasing order of sensitivity (shown in Figure S2 in the Supporting Information), are: the molecular mass density, the total atom information content calculated from the entropy of the element distribution in the molecule, the number of violations of Oprea's lead-like test, the molecular weight including implicit hydrogens, the energy of the Highest Occupied Molecular Orbital calculated using the AM1 Hamiltonian, BCUT descriptors using atomic contribution to molar refractivity, and the sum of the accessible van der Waals surface areas for atoms with partial charges falling in the first bin. The results of y-scrambling validation are shown in Figures 14 (a) – (d).

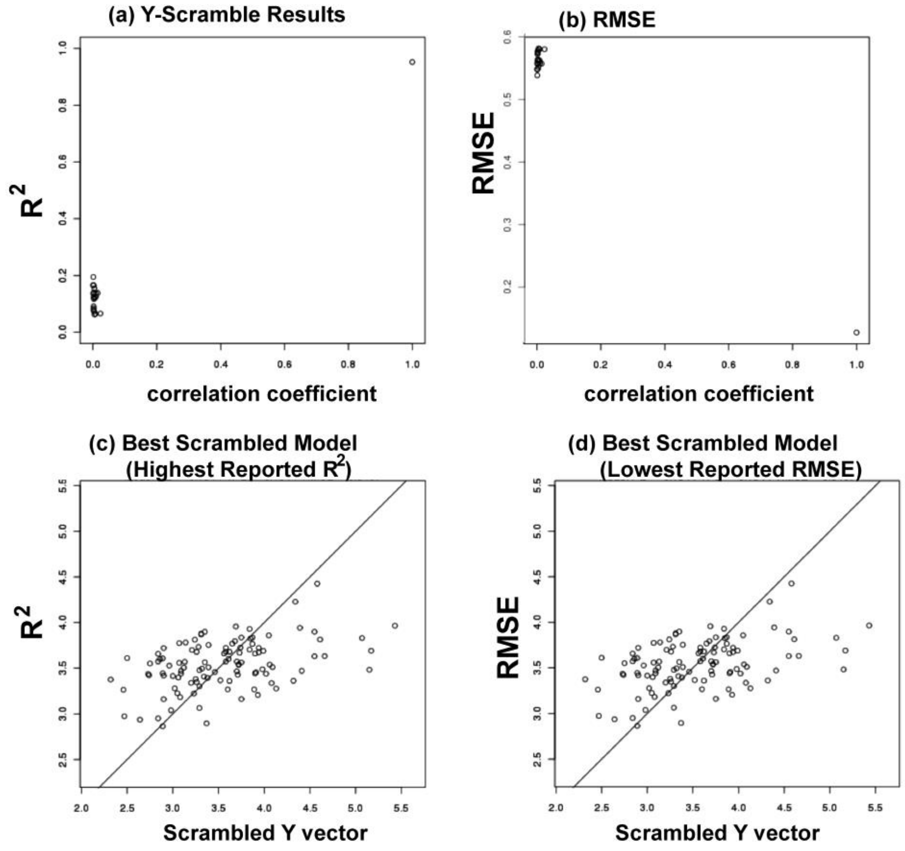

Figure 14: Y-scrambling validations of PLS models for the electronic polarizability of block copolymers. (a) R^2, (b) RMSE. (c) Scrambled Model with the highest R^2, (d) Scrambled Model with the lowest RMSE.

These PLS models were then employed to make blind predictions on new polymers.

IV. Discussion and Summary

From the results of the cross-validated models and external test set data presented above, molecular descriptors derived from the repeat unit structure produced models that quantitatively predicted T_g.

For all models generated, a wide separation was noted between scrambled and real models (Figures 6, 12 and 14), and even the highest reported models generated with fake data showed little ability to predict the scrambled responses. This information, in concert with the excellent performance of the models on test set data indicate that robust QSPR models have been generated. Going forward, these models will be used to shape synthetic decisions in the quest for high-energy materials. MQSPR modeling can thus leverage the power of first-principles DFT computations in the design of polymers with specific electronic properties, by providing high-throughput capability.

While the atom-based apol descriptor was adequate in modeling the electronic polarizability contributions of organic functional groups, this single descriptor was less useful in cases of systems containing atoms that are less well-parameterized, such as the alkali metals. Nevertheless it was possible to develop well-validated models for the electronic polarizability with a broader domain of applicability using a small pool of descriptors, both for functional group substitutions and for homo- and block co-polymers with heteroatom backbones. Topological connectivity descriptors as well as partial charge based surface area descriptors were found to be important in modeling the HOMO-LUMO gaps of polymers.

An existing general criticism about QSPR is that even with model interpretation through multiple means, there is limited-to-no guidance provided towards future studies. While it is important to characterize and understand existing chemical or material spaces, a far more potent position would be to leverage that knowledge to make informed decisions on future experiments. In other words, given the existing data including responses, what chemical structures would have corresponding targeted, or optimized responses? Solving this inverse QSPR problem has long been sought after [39], but as the problem space is large and typically nonlinear and the chemical descriptors used in modeling are often un-interpretable, many non-optimal candidates may be identified [40,41]. Potential solutions should exist on a Pareto front [42] (Fig.18) consisting of those solutions that satisfy the desired outcome, but that cannot be differentiated from each other except on the basis of additional criteria. However, if one restricts attention to this much-reduced space, the process of deciding how to proceed is greatly simplified.

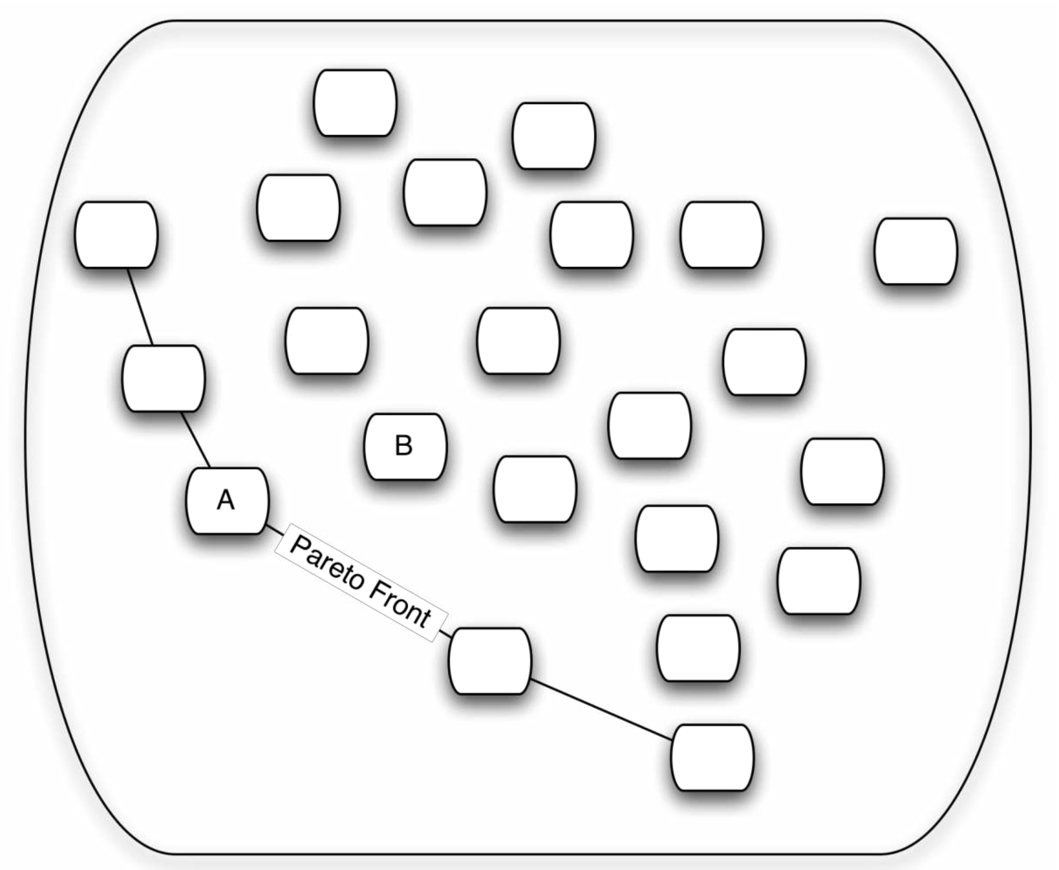

Fig.15: Potential solutions to an optimization problem may be thought to exist on a Pareto front, consisting of all solutions that satisfy the desired outcome. A solution on the Pareto front (A) is more optimal than another (B) that is not on the front.

Starting from the existing data points, it is possible to systematically alter data points so as to create a design of experiments (DOE) around each data point. To be effective, the DOE must take into consideration the distribution of the original data[43-45] as well as the model confidences in the original data[5]. The objective would be to create a very large virtual library of experiments, to evaluate predictions and their confidences in the context of model domain of applicability[46], so as to identify cases near the Pareto front. Classification or ranking of predictions could occur by estimated prediction reliability, as well as other criteria: total material cost, physical considerations, *etc*. Additionally, this inverse QSPR problem is vastly simplified when considering manufacturing process parameters as materials informatics descriptors, where changes in process parameters have a directly observable impact on material performance, and are thus interpretable. Future work will be focused on closing the loop between MQSPRs and process manufacturing, with the goal of intelligent and semi-automatic process optimization.

V. Conclusions

We have demonstrated the use of QSPR in materials design across several domains by first demonstrating that common materials properties such as the glass transition temperature can be effectively modeled using statistical learning models and appropriate descriptors. In this study, highly predictive QSPR models were developed which relate polymer repeat unit structure to the property of interest

(T_g). It was also demonstrated that the repeat unit structure, end-capped with monomers, can be used to represent the polymeric material effectively. Polymer TAE descriptors and rotatable bond descriptors calculated on the repeat unit structure can be fruitfully employed with PLS and KPLS to develop QSPR models of high predictive ability for the polymer glass transition temperature Tg. This work employed a data set displaying a wide structural diversity, leading to a general method for predicting glass transition temperatures Tg of non-cross-linked polymer materials.

Next, MQSPR modeling was shown to be capable of supplementing and guiding first-principles DFT computations in the design of polymers with specific electronic properties. Model validation ensured that statistical models were able to make valid predictions on molecules outside the training set, thereby enabling quick prioritization and quantitative predictions of many more systems than would be possible in the same time span with first-principles computations alone. Future work with inverse QSPRs may further reduce the time to optimize materials properties.

Acknowledgements

We thank the Office of Naval Research for financial support under a Multi-University Research Initiative, and Prof. R. Ramprasad for communicating his DFT results prior to publication.

References

1. Hammett LP (1937) The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. J Am Chem Soc 59 (1):96-103

2. Hansch C, Muir RM, Fujita T, Maloney PP, Geiger F, Streich M (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc 85:2817-2824

3. Hansch C, Fujita T (eds) (1995) Classical and Three-Dimensional QSAR in Agrochemistry. American Chemical Society Symposium Series,

4. Taft RW (1952) Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. J Am Chem Soc 74:3120-3128

5. Tropsha A (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. Mol Inf 29 (6-7):476-488. doi:10.1002/minf.201000061

6. Breneman CM Model Applicability Domains: When Can I Use my Model? In: American Chemical Society National Meeting, New Orleans, LA, 2008.

7. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity - a review. QSAR Comb Sci 22:1006-1026

8. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: A review. Altern Lab Anim 33 (5):445-459

9. Vrentas JS, Duda JL (1976). Macromolecules 9 (5):785-790

10. Vrentas JS, Duda JL (1977). J Polym Sci 15:403-416

11. Bicerano J (1992) Computational Modeling of Polymers, vol 25. Plastics Engineering Series. Marcel Dekker, New York

12. Bicerano J (1996) Prediction of Polymer Properties. Marcel Dekker, Inc., New York

13. Breneman CM, Thompson T (1993) Modeling the Hydrogen Bond with Transferable Atom Equivalents. In: Smith D (ed) Modeling the Hydrogen Bond. ACS Symposium Series, Washington, D.C., pp 152-174

14. Breneman CM, Thompson TR, Rhem M, Dung M (1995) Electron Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. Comput Chem 19 (3):161

15. Whitehead CE, Breneman CM, Sukumar N, Ryan MD (2003) Transferable Atom Equivalent Multi-Centered Multipole Expansion Method. J Comp Chem **24**:512-529

16. Bader RFW (1990) Atoms in Molecules: A Quantum Theory. Oxford Press, Oxford

17. Nantasenamat C, Naenna T, Ayudhya CIN, Prachasittikul V (2005) Quantitative prediction of imprinting factor of molecularly imprinted polymers by artificial neural network. J Comput-Aided Mol Des 19:509-524

18. Sukumar N, Breneman CM (2007) QTAIM in Drug Discovery and Protein Modeling. In: Matta CF, Boyd RJ (eds) The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design. Wiley-VCH, Weinheim, pp 471-498

19. Murray JS, Politzer P, Famini GR (1998) Theoretical Alternatives to Linear Solvation Energy Relationships. Theochem-J Mol Struct 454 ((2-3)):299-306

20. Sjoberg P, Murray JS, Brinck T, Politzer P (1990) Average local ionization energies on the molecular surfaces of aromatic systems as guides to chemical reactivity. Can J Chem 68 (8):1440-1443

21. Politzer P, Murray JS, Grice ME, Brinck T, Ranganathan S (1991) Radial behavior of the average local ionization energies of atoms. J Chem Phys 95 (9):6699-6704

22. Murray JS, Abu-Awwad F, Politzer P (2000) Characterization of aromatic hydrocarbons by means of average local ionization energies on their molecular surfaces. Journal of Molecular Structure: THEOCHEM 501-502:241-250

23. Politzer P, Daiker K (1981). In: Deb BM (ed) The Force Concept in Chemistry.

24. Politzer P, Truhlar DG (1981) Chemical Applications of Atomic and Molecular Electrostatic Potential. Plenum Press, New York

25. Topliss JG, Edwards RP (1979) Chance Factors in Studies of Quantitative-Structure Property Relationships. J Med Chem 22:1238-1244

26. Rücker C, Rücker G, Meringer M (2007) y-Randomization and Its Variants in QSPR/QSAR. J Chem Inf Model 47 (6):2345-2357. doi:10.1021/ci700157b

27. Clark R, Fox P (2004) Statistical variation in progressive scrambling. J Comput-Aided Mol Des 18 (7):563-576. doi:10.1007/s10822-004-4077-z

28. Breneman CM, Sukumar N, Embrechts MJ, Bennett KP, Sundling CM, Krein M, Hepburn T (2007) Realizing prospective QSAR through data fusion and modern descriptors. 234th National Meeting American Chemical Society

29. Molecular Operating Environment (2008). 2008.10 edn. Chemical Computing Group, Inc., Montreal, Canada

30. Frisch MJ, Trucks GW, Schlegel HB, Gill PMW, Johnson BG, Robb MA, Cheeseman JR, Keith T, Petersson GA, Montgomery JA, Raghavachari K, Al-Laham MA, Zakrzewski VG, Ortiz JV, Foresman JB, Cioslowski J, Stefanov BB, Nanayakkara A, Challacombe M, Peng CY, Ayala PY, Chen W, Wong MW, Andres JL, Replogle ES, Gomperts R, Martin RL, Fox DJ, Binkley JS, Defrees DJ, Baker J, Stewart JP, Head-Gordon M, Gonzalez C, Pople JA (2003) Gaussian 03. Gaussian, Inc.,

31. Tropsha A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR Comb Sci 22 (1):69-77. doi:10.1002/qsar.200390007

32. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26 (5):694-701. doi:10.1002/qsar.200610151

33. Golbraikh A, Tropsha A (2002) Beware of q2. J Molec Graph Model 20 (4):269-276

34. Embrechts M, Bress R, Kewley R (2006) Feature Selection via Sensitivity Analysis with Direct Kernel PLS

Feature Extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh L (eds), vol 207. Studies in Fuzziness and Soft Computing. Springer Berlin / Heidelberg, pp 447-462. doi:10.1007/978-3-540-35488-8_22

35. CRC Handbook of Chemistry and Physics (1994). 75 edn. CRC Press, Boca Raton

36. Zupan A, Burke K, Ernzerhof M, Perdew JP (1997) Distributions and averages of electron density parameters: Explaining the effects of gradient corrections. The Journal of Chemical Physics 106 (24):10184-10193

37. Perdew JP, Ernzerhof M, Zupan A, Burke K (1998) Nonlocality of the density functional for exchange and correlation: Physical origins and chemical consequences. The Journal of Chemical Physics 108 (4):1522-1531

38. Liu C-S, Pilania G, Wang C, Ramprasad R (2011) Unpublished work.

39. De Julian-Ortiz JV (2001) Virtual Darwinian Drug Design QSAR Inverse Problem, Virtual Combinatorial Chemistry, and Computational Screening. Comb Chem High Throughput Screening 4 (3):295-310

40. Visco DP, Pophale RS, Rintoul MD, Faulon J-L (2002) Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. J Mol Graph Model 20 (6):429-438

41. Adams N (2009) Polymer Informatics. In: Meier MAR, Webster DC (eds) Advances in Polymer Science, vol 225. Advances in Polymer Science. Springer Berlin / Heidelberg, pp 107-149. doi:10.1007/12_2009_18

42. Greenwald BC, Stiglitz JE (1986) Externalities in Economies with Imperfect Information and Incomplete Markets. Q J Econ 101 (2):229-264. doi:10.2307/1891114

43. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput-Aided Mol Des 16 (5):357-369. doi:10.1023/a:1020869118689

44. Baroni M, Clementi S, Cruciani G, Kettaneh-Wold N, Wold S (1993) D-Optimal Designs in QSAR. Quant Struct-Act Relat 12 (3):225-231. doi:10.1002/qsar.19930120302

45. Brandmaier S, Sahlin U, Tetko IV, Öberg T (2012) PLS-Optimal: A Stepwise D-Optimal Design Based on Latent Variables. J Chem Inf Model 52 (4):975-983. doi:10.1021/ci3000198

46. Weaver S, Gleeson MP (2008) The importance of the domain of applicability in QSAR modeling. J Mol Graph Model 26 (8):1315-1326