Prediction of Polymer Properties Using Infinite Chain Descriptors (ICD) and Machine Learning: Toward Optimized Dielectric Polymeric Materials

K. Wu,[1] N. Sukumar,[1,2] N. A. Lanzillo,[1] C. Wang,[3] Ramamurthy "Rampi" Ramprasad,[3] R. Ma,[4] A. F. Baldwin,[4] G. Sotzing,[4,5] C. Breneman[1]

[1]Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180
[2]Department of Chemistry and Center for Informatics, School of Natural Sciences, Shiv Nadar University, NH-91, Tehsil Dadri, Gautam Budh Nagar 201314, Uttar Pradesh, India
[3]Department of Materials Science and Engineering, University of Connecticut, Storrs, Connecticut 06269
[4]Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, Connecticut 06269
[5]Deparment of Chemistry, University of Connecticut, Storrs, Connecticut 06269
Correspondence to: C. Breneman (E-mail: brenec@rpi.edu)

Received 7 January 2016; accepted 27 May 2016; published online 2 August 2016
DOI: 10.1002/polb.24117

ABSTRACT: To facilitate the development of new polymeric materials, we report the development of new heuristic models to predict the dielectric constant, band gap, dielectric loss tangent, and glass transition temperatures for organic polymers. A new set of features called infinite chain descriptors (ICDs) was designed and developed especially to characterize organic polymers, utilizing methods with minimal dependence on predefined fragment libraries. Machine learning models were built for the aforementioned properties incorporating best practices in the field such as objective feature selection, cross-validation and external test sets. All models produced in this study showed good performance in prediction. A web tool has been developed and has been made available that supports the input of novel structures. © 2016 Wiley Periodicals, Inc. J. Polym. Sci., Part B: Polym. Phys. **2016**, *54*, 2082–2091

KEYWORDS: dielectric properties; heuristic model; machine learning; MQSPR; polymer

INTRODUCTION The first and most important step in shortening the cycle of new material discovery is creating models that relate controllable parameters in synthesis to a target property or set of properties. To this end, Material Quantitative Structure–Property Relationship (MQSPR) can be applied to create robust models, particularly when coupled with physics-based descriptors and appropriate validation methods.[1,2] MQSPR refers to the application of quantitative structure–property/activity relationship (QSPR/QSAR) modeling within the domain of materials informatics. In its original application as a component of drug discovery workflows to predict small molecule physical and biological properties, modern QSPR/QSAR has been shown to be most effective when used within well-defined domains of applicability. The motivation for introducing MQSPR in polymer dielectric design is to create a time-efficient and low-cost screening method for materials with desired properties by relating encoded structural information (descriptors) to macroscopic properties.

Dielectric polymeric materials have been used in energy storage devices more than other alternatives, such as ceramics, due to their high breakdown strengths, low fabricating temperatures, and structural flexibility.[3] The utility of the design tool described here is to enable the design of low loss and high energy density materials. In this process, a variety of properties must be considered. For example, for capacitor applications, suitable solvents need to be identified that can facilitate film fabrication, and the glass transition temperatures of the resulting polymers will need to be predicted to assure that they will not be within the working temperature range of the capacitor system. In this application, to have high energy density, polymers need to have both a high dielectric constant and high breakdown strength. Dielectric loss must also be considered, and this property is often correlated with high dielectric constants in polar polymers such as PVC and PVF. To find structures that meet all of these design goals, a set of models needs to be built that can be quickly applied as part of a materials design workflow. To be most useful for new material discovery, model universality is highly desirable.

In the past 30 years, numerous attempts have been made to predict polymer properties, where the glass transition temperature

Additional Supporting Information may be found in the online version of this article.

may be the most studied one.[1,4–17] Studies on the dielectric constant and refractive index have also been reported.[16–21] Le et al.[22] did a comprehensive review on the QSPR modeling for material properties in 2012. Basically, these models can be divided into two categories: group contribution techniques and descriptor-based methods. Group contribution methods, like those built by van Krevelen,[17] are fast and easy to interpret. The well-known models built by Bicerano[16] can be considered a mixture of these two categories since molecular connectivity index descriptors[23] and modifications based on substructure types are both used. Other models utilize more abstract descriptors that may not be as easily interpreted—for example, topological descriptors or quantum-based descriptors. These two model types are most widely used because both methods have different and complementary limitations. There are always tradeoffs between interpretability and prediction power for models. While very effective within its defined applicability domain, a group contribution model is unable to account for the effect of new substructures not contained in the substructure/fragment library, leading to instant failure of the method for truly novel materials. It becomes even more severe when the curated training data set is small, which is often the case in materials studies. An additional issue is that the process of partitioning group contributions between sets of substructures is empirical and not deterministic, especially when the size of the training set is small. Consequently, group contribution-based methods may not be the optimal choice for new material discovery, but can be useful tools for informing designers of the likely outcome of including specific functional groups in a system. On the other hand, descriptor-based methods use a more generalized numerical description of structure and thus can cover a wider domain of applicability if a sufficient level of physics is embedded in the descriptors. Descriptor-based modeling practitioners often face the problem that polymer chain features, especially those derived from quantum computation, cannot be easily calculated on a whole realistic polymer chain. Most of the published studies use features of a minimal repeat unit with hydrogen atoms as terminating atoms. In this paper and other informatics studies, the minimal repeat units are usually called monomers. It should be noted that the usage of this term here is not the same as in polymer synthesis, where monomers refer to the structure of small-molecule building blocks prior to polymerization. Similarly, we use dimer and oligomer to describe the multiplication of the repeat units.

The first question to be answered when modeling polymers is to define how they should be represented on a molecular level. Direct computation on polymer chains of realistic lengths is infeasible, not only because polymer chains are usually much larger than classical small molecules, but also due to polydispersity in polymer systems that preclude the use of a single representative molecule. A natural and common idea is to use the polymer repeat unit to characterize unbranched polymers, which essentially uses a small, capped molecule to represent the chemical features of polymer chains. Representations of oligomers that include statistical treatments have also been used.[9] However, none of these methods have adequately addressed the limitations of "local" molecular representations. Even though studies such as the one described in Katrizky's 1996 paper[6] considered the extension of descriptor values for long chains using a numerical treatment, most of the models simply consider a repeat unit with different end-cap atoms as a complete molecular representation. Monomers with carbons as end-cap atoms[14], dimers,[19] and ring-like dimers[21] have been used as ways of incorporating features of the polymer chemical environment. With these models, two theoretical issues arise: If a descriptor value is an "extensive" property that varies with the scale of the local molecular representation (e.g., the monomer and dimer give different descriptor values), the information contained in this descriptor is contaminated by the choice of the molecular representation. This issue makes creating a universal model trained on data that includes large variation in repeat unit size problematic. In addition, since the dimer representation can be seen as the "copolymer" of the two identical monomers, if the model predicts dimers and monomers differently, a natural consequence is the difficulty in dealing with alternating copolymers. Another issue, also mentioned by Mattioni and Jurs,[24] is that if oxygen is the terminating atom, then using hydrogen as end-cap atoms will introduce properties of a nonexistent hydroxyl group. Such problems are usually seen in the treatments of polycarbonate and nylon.

The premise described here is that descriptors suitable for predicting polymer properties that converge at higher chain lengths must have three properties: (i) the descriptors should represent intensive properties instead of extensive ones; (ii) the descriptors should provide correct chemical environments for all atoms; (iii) the descriptors should provide information about the functionalities of side groups and backbones separately, as well as overall information. The first property is quite easy to approximate by using normalized descriptors (normalized by the number of atoms or the volume of the repeat unit in use). Our contribution here is that we developed descriptors that are not only free from the influence by the end-capped atoms, but also provide information relevant to both backbone and side-chains (see more details in Methods Sections and Supporting Information). The second feature mentioned above has also been recognized by other groups, but can only be partially realized using different end-cap atoms, as mentioned previously. In our approach, we maintain this feature in the descriptor computation level: The converged descriptor values are computed directly from the fragment of interest, rather than from an end-capped approximation of the repeat unit. Figure 1 shows a comparison between the partial charge (using Partial Equalization of Orbital Electronegativities, PEOE[34]) computed for hydrogen-capped monomers and for the corresponding polymers. It is clear from this example that the classical local representation provides inaccurate information, but ICD can directly produce correct values. The third property is less commonly discussed in QSPR studies, but is much more in polymer physics. The side groups and backbones often contribute differently to a certain property;

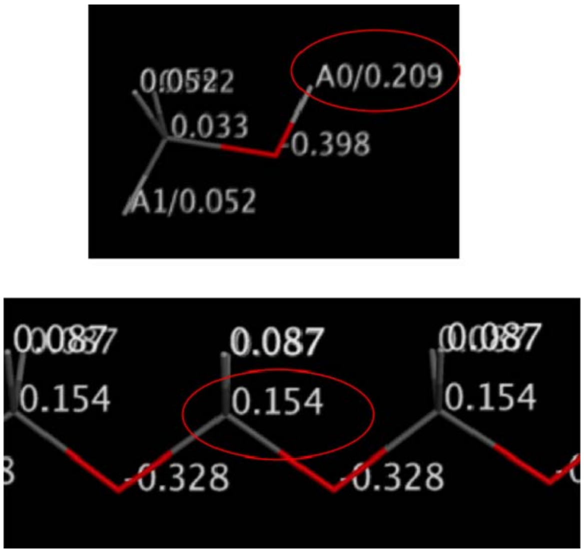


FIGURE 1 Partial charges computed for a hydrogen-capped monomer (upper) and the polymer (lower) for a 20-mer polymethylene oxide oligomer. The maximum positively charged atoms are circled. In this example, the charges computed for the end-capped monomer are quite different from those computed for a representative polymer. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

consequently, using descriptors that can be computed separately for each of these components should be a better choice.

In this study, we define a set of infinite chain descriptors (ICDs) which were developed partly based on modifications of some traditional descriptors used in QSPR/QSAR. These descriptors directly describe the properties of infinite chains, which are free from all the problems described above. By utilizing machine-learning approaches, especially those with ability to deal with nonlinearity, models for glass transition temperature, dielectric constant, dielectric loss, and band gap for linear polymers were built with minimal dependence on specific fragment/atom type library entries. A web tool with all the built models implemented in this study was also developed and is ready for use at reccr.chem.rpi.edu/polymerdesign.

METHODS

Dataset

Dielectric constants and band gaps for 155 polymers were computed using DFT methods.[25] This dataset included 118 backbone structures containing four components (CH_2, NH, $C=O$, benzene ring and thiophene ring) and 37 common polymers (like PMMA and PEO). A set of 48 polymers with dielectric loss tangent data at 100 Hz and 44 polymers at 1kHz were obtained from Ku and Liepins[26] and Sotzing et al.[27–29] Properties of 262 polymers including glass transition temperature values and 58 polymers with dielectric constant data were taken from the set reported by Bicerano.[16] All structures (in SMILES format[30]) and associated data are listed in the Supporting Information.

Infinite Chain Descriptors

Descriptors are numerical representations of a variety of molecular characteristics of chemical entities that encode specific structural and/or electronic information. When chosen appropriately, descriptor sets can provide meaningful correlations between molecular structures and target properties. The descriptors used in this study were created by modifying traditional descriptors that have been previously used in (M)QSPR modeling for polymers and small molecules, while retaining the chemical meaning of these descriptors. Since properties for polymers usually converge as the chain grows, a reasonable procedure to employ is to use descriptors designed to represent infinite chains instead of those designed for a local structural representation, which either lack information on the connection environment or are time-consuming to compute for oligomers. It is also unacceptable for modeling if monomers and dimers give different predictions, since they represent the same polymer. Consequently, we seek descriptors that are scale-insensitive, but represent pertinent features of the repeat units. The problem of lack of information concerning 3D conformations introduces a stochastic issue, so primarily 2D descriptors were considered. The descriptors considered can be divided into three categories:

- Topological descriptors: Those containing information about shape and structural flexibility, based on classic topological descriptors like Kier's shape descriptors[31] and Balaban's BalabanJ descriptor.[32] Such descriptors are usually derived from graph theory. A polymer can be viewed as (nearly) an infinite graph, rather than a finite graph usually used for small molecules. Therefore, the descriptors are generally redesigned to describe intensive properties, for example, the polymer version of Zagreb descriptor[33] is defined as the number density of the squares of the vertex degrees. These descriptors provide information related to structural flexibility. For example, main chain flexibility is related to the joint effects of rotatable bond density, side group size, backbone connectivity, and related factors.
- Partial charge descriptors: Charges are based on the Partial Equalization of Orbital Electronegativities (PEOE) 2D partial charge algorithm,[34] providing information on partial charge distributions and electrostatic interactions. These descriptors can provide information on interchain and intrachain electrostatic forces and charge distributions, which are related to chain mobility and dipole distributions.
- Electronic transferable atom equivalents (TAE) descriptors[35,36]: derived from atom-based transferable electron density distributions, also containing information such as local average ionization potentials and other electronic properties. These descriptors provide information similar to that of traditional quantum mechanical descriptors, but are based on atomic level property distributions. Changes in TAE electronic descriptor values have been correlated with properties such as polarizabilities and band gaps.[35,36]
- Three types of modifications have been applied to convert these descriptors into the corresponding infinite chain versions, using the following algorithm:

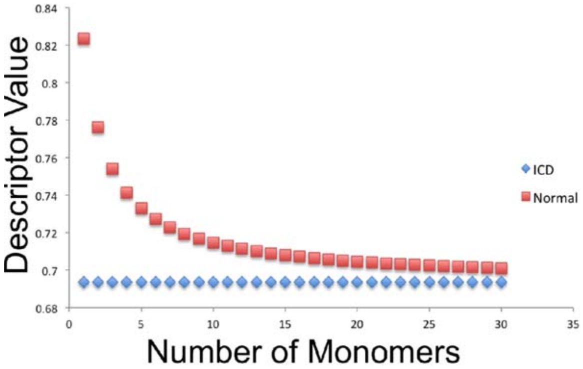


FIGURE 2 Comparison of descriptor values between the ICD and normal versions of PEOE_VSA_FHYD (fractional hydrophobic Van der Waals surface) descriptors for polyvinyl fluoride. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

- Descriptors are calculated independently on backbones and side chain groups. This procedure is based on the observation that side chain and backbone structures often contribute differently to polymer properties and have distinct structural effects. Descriptors for side chain groups are identified with the prefix "sg," while those for backbones are denoted as "BB" in their names.
- Correct connection conditions are set for connecting atoms (head and tail of the repeat unit) within the descriptor algorithm. All TAE and PEOE descriptors are calculated with the correct chemical environment for all atoms. TAE descriptors arise from a connectivity-based atom type search that combines atomic electronic properties into molecular ones.[37] A precomputed atomic fragment library is used to "reconstruct" new molecules. For each atom in a new molecule, the chemical environment information is used to define/encode its type and the corresponding entry with that type is retrieved from the fragment library. In the case of ICDs, the searching algorithm is modified so that all the atoms have the correct encodings as in a polymer chain. The PEOE algorithm within MOE is modified in a similar way, such that the terminal atoms are defined as being connected to each other. The algorithm provides partial charge values for a long chain polymer based on those of the repeat unit.
- For 2D topological descriptors, the connection condition is redefined, and descriptor values are determined in two ways: (1) normalized by the total number of bonds or atoms (with prefix "polyb") within a repeat unit; (2) normalized by the topological displacement from head to tail (designated with the prefix "polyq"). The latter can be viewed as a measure of the extension of the polymer chain in a direction perpendicular to that of the backbone. The label "q" was also described in Balaban's study[38] related to the infinite value for the BalabanJ descriptor.

Using these rules together with existing descriptors, a total of 277 descriptors were implemented in MOE,[39] including 143 2D descriptors and 134 TAE/RECON descriptors. Figure 2 shows the fundamental concept of infinite chain descriptors.

An additional descriptor is used to represent the band gap of the relevant polymer systems. This is based on a reciprocal extrapolation of the HOMO-LUMO gap determined using a series of three oligomers (GAP_inf3), and designed to represent the observation that HOMO-LUMO gaps of linear polymers generally decrease as approximately 1/n, where n is the number of repeat units. In conjugated systems, the Kuhn equation[40] and Zade's study[41] have shown such a relation. We also find for several non-conjugated polymer systems (Supporting Information), the HOMO-LUMO gap calculated from the semiempirical PM3 method can also be approximately described by the following rule, where A is a constant:

$$GAP_n=\frac{A}{n}+GAP_{\text{inf}} \tag{1}$$

Based on this equation, the infinite chain HOMO-LUMO gap can be estimated by comparing the HOMO-LUMO gaps of several oligomers. To avoid uncertainty from a single calculation (arising mostly from the uncertainty of the optimized conformation and the influence of the end-cap hydrogens) and to decrease the computation time, the original repeat unit, the HOMO-LUMO gaps of the dimer and the trimer were computed in this way. The following equation was then used to calculate the GAP descriptor (t, d, and m refer to trimer, dimer, and monomer, respectively).

$$GAP_{\text{inf}}3=\frac{6GAP_t-2GAP_d-GAP_m}{3} \tag{2}$$

In practice, we found that the GAP_inf3 descriptor value varies slightly with conformation, and with the size of the initial repeat unit (such as CC or CCCC for polyethylene). In these studies, the variance between different representations is within 5% and the difference compared to the actual computed value for a 30-mer is within 10%. Therefore, the GAP_inf3 descriptor is designed to describe the effect of polymerization upon the polarizabilities of polymers.

All infinite chain descriptors were implemented using MOE[39] and its associated SVL language. In this study, the GAP_inf3 is calculated using MOE 2010 and MOPAC2012.[42] Details of all descriptor computations can be found in the Supporting Information.

Model Development Using SVM Regression and Classification

Support vector machine regression (SVR)[43] is a machine learning technique that utilizes a "kernel trick" in order to find linear relationships in a high dimensional space that correspond to non-linear relationships in descriptor space. In contrast to traditional multiple linear regression (MLR), SVR shows good generalizability and robustness to outliers in the training set. K-fold cross-validation employs an iterative approach, whereby a subset of training data is withheld, models are built on the remaining data, and the withheld data are used to assess the generalizability of the model. In this study, each example within the training data was randomly assigned to one of 10 cross-validation groups. The randomization was performed 10 times. As part of the model optimization process, feature selection was used to reduce the number of descriptors to avoid overfitting. Objective feature selection (removal of collinear features and high variance features that span more than 6 standard deviations) and recursive feature selection (remove the least important features iteratively) were performed using partial least square (PLS), support vector regression (SVR), or random forest (RF) methods. Linearly correlated descriptors with correlation coefficients larger than a specified cutoff value (0.85 by default) were removed at random until only one of them was left in the descriptor set. Descriptors with large deviations are removed because they are more likely to be spurious values, or be the product of outliers. The importance of descriptors is computed based on sensitivity analysis: measured as the prediction change due to a perturbation to the descriptor value. Models with the best cross-validation performance were selected. More discussions of the modeling procedure can be found in Refs. 1 and [44]. As a blind test validation set, 20% of data was retained separately and used to assess the predictive power of the resulting models. The external test set was chosen randomly from the binned (5-bin) target values, in order to make sure the training set and test set have similar ranges and distributions.

Classification models based on support vector machine methods[45] were built to predict the dielectric loss tangent (a measure of energy dissipation) of different polymer systems. From a practical perspective, the demand for prediction accuracy for certain properties like dielectric loss is not as high as for dielectric constant and glass transition temperature. Furthermore, much of the available experimental data on dielectric loss is rather qualitative. For these reasons, a regression model may not provide useful information for guiding synthesis. Classification models, on the other hand, can tolerate errors in experimental data and predict appropriate numerical ranges of the target property. In the models developed to predict dielectric loss tangents, the data were categorized into three classes: low (<0.001), medium (0.001–0.01), and high (>0.01). Then 20% of data from each class was retained as the test set. A different feature selection method specific to classification modeling was applied as follows: The mean-centered and variance-normalized descriptor matrix was first filtered to remove descriptors with extreme values larger than 6 standard deviations from the mean. The correlation matrix was then calculated and presented as a connectivity graph. Descriptors with correlation coefficients higher than 0.85 were considered connected by an edge. Connected descriptors were removed in the order of degree of connectivity and then the number of distinguished values was used to break ties for single pairs until no collinearity above the threshold value remained. The method we used here is a simple approximation, but retains more descriptors than randomly removing one from a pair of correlated descriptors at a time. Principal components were then calculated with whitening,[46] which makes the variance 1.0 for each component. A portion of the total number of principal components was then used as descriptors for modeling after selection using variance analysis (ANOVA). A support vector machine with a radial basis function kernel (RBF)[47] was used as a classifier, and 5-fold cross-validation was used for model parameter selection, including the number of principal components used in the model (the search range being one to ten PCs).

RESULT AND DISCUSSION

Dielectric Constant

Macroscopically, the dielectric constant measures the ability of a material to alter the capacitance of a device, which consists of two parallel plates separated by a dielectric medium. Microscopically, the magnitude of the dielectric constant depends upon the availability of molecular mechanisms that allow for charge polarization to occur. These mechanisms may involve polar group rotation, ion pair separation, and/or electron density deformation. Usually, polar group rotation contributes significantly to the dielectric constant of polymers with small polar side groups such as PVF. The importance of this effect decreases with the frequency of the applied electric field and contributes to dielectric loss. The propensity of electron distribution deformation is related to molecular polarizability. Electronic polarization is independent of frequency in the normal working range and is often denoted as ε_∞. The total dielectric constant at 0Hz (DC) is usually called the static dielectric constant, while at infinite frequency it is called the optical dielectric constant. Theoretical models have been developed to relate the two microscopic properties—dipole moment and molecular polarizability—to the overall dielectric constant. The Claussius-Mossoti equation describes the dielectric behavior of nonpolar gases at low temperature, but can also be applied to polymers.[48]

$$\frac{\varepsilon-1}{\varepsilon+2}=\frac{4\pi N_A \rho \alpha}{3M} \tag{3}$$

where ε is the dielectric constant, N_A is the Avogadro constant, ρ is the density, M is the molecular weight and α is the molecular polarizability. Due to the multiplicity of mechanisms involved, it is well known that dielectric constants for polymers cannot be simply modeled by an ideal single relaxation time model. Alternatively, to describe the dielectric spectrum, the Havriliak–Negami equation, which includes two empirical parameters, is generally used. As two empirical parameters need to be obtained from fitting and are specific for certain polymers, such a model cannot be used as a prediction tool. While computationally intensive, it is possible to use *ab intio* methods to predict dielectric behavior[25]. The advantage of *ab initio* methods lies in their being based on microscopic physics. Unfortunately, good quality results require large computational resources as well as some parameter tuning (e.g., pseudo-potentials, basis set size, reciprocal space sampling), requiring experience, and sometimes case-specific knowledge.

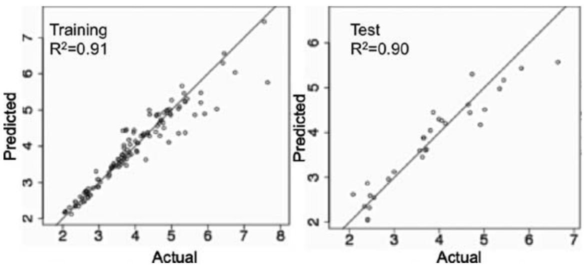


FIGURE 3 Model for the electronic component of the dielectric constant, left: Training set (125 data points), right: Test set data (30 data points).

Here, MQSPR modeling has the advantage of allowing a significant increase in prediction speed. In the current work, ICD descriptors were used with machine learning to build models that show good performance and internal consistency, requiring less than 3 min per polymer on average even while using the more computationally intensive GAP_inf3 descriptor.

Two datasets of polymer dielectric constants were used in this study. Experimental data were compiled from the work of Bicerano[16] and data from quantum computation was from DFT calculations performed by Ramprasad et al.[25] In the latter dataset, the electronic and ionic components were calculated separately using density functional perturbation theory (DFPT). Separating the two components of dielectric response can help to better understand the relationship between polymer structures and dielectric constants. It was observed that polymers with large polar components were usually subject to increased dielectric loss. Dielectric loss measures the energy dissipation of the material and for capacitor applications should be as low as possible. The modeling of dielectric loss tangents will be discussed in the next section.

In the model developed for the electronic component of the dielectric constant ε_e, only using ICD (243 descriptors) gave the $R^2 = 0.85$ and $\text{RMSE} = 0.44$ for the training set, and 0.82 and 0.64, respectively, for the test set. When the GAP_inf3_inv (the inverse of GAP_inf3) was added (Fig. 3), the corresponding values were 0.91 and 0.35 for the training and 0.90 and 0.37 for the test set. In the latter case, the cross-validation R^2 was found to be 0.83, demonstrating the ability of GAP_inf3_inv to describe the electronic flexibility of the structures.

The use of semiempirical methods was found to largely broaden the domain of applicability for this model, and by extrapolation (eq 2), the value of GAP_inf3_inv was found to be insensitive to the size of the repeat unit. For repeat units of medium size, such as polycarbonate, the computation time was around 3 min per polymer. A colinearity cutoff of 0.85 was used in descriptor selection for modeling and eight iterations of RF recursive feature selection were used with a 75% retention factor. Nine descriptors were ultimately found to be important, of which the top three were polyb_B_zagreb, GAP_inf3_inv, and RECON_FDel.Rho.NA7. The poly_B_zagreb descriptor was of the same type as the polyb_zagreb but used for the backbone, which basically shows the content of conjugated rings in the structure. RECON_FDel.Rho.NA7 is a TAE/RECON descriptor that is related to the electron density gradient and thus also correlated with the polarizability of the electron distribution.

The model for the ionic component ε_i also shows good performance after log transformation and mean centering (See Supporting information for details)

For feature selection with this dataset, the colinearity cutoff was set to 0.7, and five iterations of PLS recursive feature selection were performed retaining 75% of descriptors in each iteration. A set of 18 descriptors remained after feature selection, resulting in $R^2 = 0.88$ and $\text{RMSE} = 0.11$ for the training set and 0.73 and 0.17, respectively, for the test set, as shown in Figure 4. Important descriptors included poly_PEOE_VSA + 2, b_doubleR, RECON_FDel.G.NA5, poly_PEOE_maxPC-, and poly_PEOE_VSA + 3. Most of these descriptors are related to features of the charge distribution, indicating large differences in partial charges between atoms of varying electronegativity, leading to large ionic components. The descriptor poly_PEOE_maxPC- reflects the maximum negative partial charge value, which showed clear negative correlation with the response, that is, the larger the negative charge a structure has, the larger will be the ionic component of its polarizability. This finding agrees well with physical intuition since more extreme partial charges should lead to more polar groups. In the study of the ionic component of the dielectric constant, we found that descriptors commonly used in QSPR rarely include bond strengths or rotational barrier information, which should be directly related to the deformations responsible for the ionic polarization mechanism. Such deficits may not be easily resolved, especially in the case of polymers, given uncertainties in their conformations. Information derived from IR spectra calculated by semiempirical methods may prove to be useful; this idea is similar to that of EVA descriptors.[49] However, we found that the model developed using the IR information calculated by PM3 on monomers does not show better performance, which may be due to the inconsistencies of the conformations and chain lengths.

Even though the model statistics for predictions of the ionic component are not as good as for the other models, in most cases the ionic component only accounts for a small part of

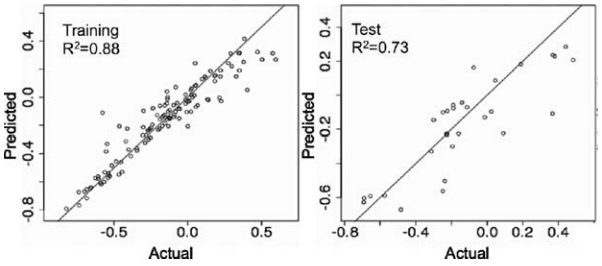


FIGURE 4 Model for the ionic component of the dielectric constant, left: Training set (125 data points), right: Test set data (30 data points).

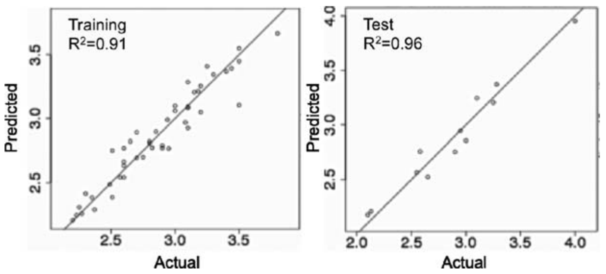


FIGURE 5 Model for dielectric constant from experiments compiled by Ref. 16, left: Training set (125 data points), right: Test set data (30 data points).

the total dielectric constant (average ~15%). For the total dielectric constant, the model shows very good performance (RMSE = 0.52).

A predictive model was also built directly using the experimental data from Bicerano's book.[16] The best model was found by setting the colinearity cutoff to 0.85, with seven iterations of PLS feature selection and 80% of the descriptors retained in each iteration, as shown in Figure 5. Since the data set is quite small, the linear PLS method was used instead of kernel SVM. Linear models with low capacity are less susceptible to overfitting on small datasets. The descriptor set used is the same as in previous models (2D, TAE and GAP_inf3_inv). The R^2 values for training and test were 0.91 and 0.96, respectively, with RMSE values of 0.11 in both cases. The cross-validation R^2 was 0.86.

Dielectric Loss Tangent

The dielectric loss tangent quantifies the energy dissipated by the material during operation. It is defined as the ratio of the imaginary part (dielectric loss, ε'') and the real part (dielectric constant, ε') of the relative permittivity.

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} \tag{4}$$

Dielectric loss results from the mismatch of the polarization rate in material and rate of the oscillating applied electric field. A high loss material transfers large amount of input energy into heat, thus in energy-storage applications, low loss material is desired. QSPR studies for predicting $\tan\delta$ are rare in the literature. Bicerano[16] gave a regression model relating the loss tangent to the refractive index, static dielectric constant and several structural parameters. Alternatively, Yu et al.[50] used quantum descriptors together with measurement frequency to predict the loss tangent of vinyl polymers. However, structural similarities between the polymers in this dataset may lead to a limitation in the domain of applicability for the resulting heuristic models. It should be recognized that local models on similar structures always perform better than universal models, since the latter must generalize at the expense of accuracy. Consequently, the choice between the two types of models depends largely on the requirements of the synthesis planning process. In this project, universal models were sought to allow virtual experiments to be performed on large portions of chemical/materials space.

From the viewpoint of a materials scientist, the level of loss is often more important than the exact value. Consequently, classification models for dielectric loss at 100 and 1000 Hz (As we found are the most frequently reported and representative frequencies in literatures) were developed using the procedures described in Section 2. A total of 143 descriptors were originally included, which were pared down to 139 descriptors for use in the principal component analysis (PCA) after feature selection. In the 100 Hz model, four principal components were ultimately used for the training set of 42 polymers, the training accuracy is 88% and cross-validation accuracy is 76%. Six polymers were used for the blind test. In this example, it was found that only one polymer was misclassified. Similar results were found for the 1 kHz model. Four principal components were used in that case as well. A training accuracy of 82% and a cross-validation accuracy of 82% were found for 38 polymers. One of the six blind test data was found to be misclassified. Tables (1–4) show the confusion matrices and e results for the test set. In the 1 kHz model, the misclassified polymer was 1,2-DAE, with the value of 0.0076, which is very close to the arbitrary boundary defined between class Medium and class High. Such misclassification is actually consistent with the overall model performance, since the boundaries for each class are set to only indicate the data range and in nature the property should be continuous in value. The misclassified polymer in the 100 Hz model was methocel. The error may be due to the uncertainty of the structure, since for methocel, the R group position could not be set with certainty. For classification models built using support vector machines, the importance of descriptors cannot be easily accessed, however, based on the analysis of the principal components that are used in the two models, it is observed that descriptors related to partial charge distribution and structural flexibility are highly involved.

Apart from Bicerano's work, this may be the first time that a dielectric loss tangent for polymers from various categories has been successfully modeled and predicted. This model can aid in the search for new dielectric materials by providing prospective predictions of structures likely to exhibit unacceptable levels of dielectric loss.

Band Gap

Band gap is defined as the energy difference between the conduction band and valence band. Large band gap implies

TABLE 1 Confusion Matrix for Loss Tangent at 100 Hz

		Actual		
		Low	Medium	High
Predicted	Low	4	1	0
	Medium	1	15	2
	High	0	1	18

TABLE 2 Blind Test Result for Loss Tangent at 100 Hz

Polymer	Loss Tangent	Actual Class	Predicted Class
1,3-DAP	0.0117	H	H
PMDA-D230	0.0037	M	M
DieG	0.0149	H	H
Poly(vinyl acetate)	0.0049	M	M
Methocel	0.128	H	M
Poly(vinyl toluene)	0.0007	L	L

better insulating capability since it requires more energy to excite an electron into the conduction band. It was found that band gap could be used as an indicator of a polymer's breakdown strength.[51] The breakdown strength is usually described by the Weibull distribution,[52] which is related not only to chemical structure, but also to the thickness of the test sample, defects and temperature. It was also observed that breakdown strength shows significant change at T_g and the breakdown process may involve different mechanisms below and above T_g. Frohlich's electron avalanche theory[53] applies when the temperature is low, while other theories are usually used to describe the situation near and beyond T_g. Ku and Liepins' book[26] provides a good review on the theories.

In concept the bulk band gap is related to the HOMO-LUMO gap for the repeat unit. Thus using the HOMO-LUMO gap calculated by some semiempirical method seems to work. However the HOMO-LUMO gap decreases with the growth of the chain and it can still be time-consuming to determine the convergence limit. In this study, we used the ICD topological descriptor with the extrapolated HOMO-LUMO gap (GAP_inf3) to model the band gap.

Figure 6 shows the MQSPR model for band gaps. Partial least square (PLS) regression was used, with colinearity cutoff 0.75. Five iterations of recursive feature selection were used with 80% of the overall 144 descriptors retained in each iteration. Nine descriptors were left after feature selection. The R^2 values for training and test sets are 0.84 and 0.88, respectively, and RMSE values are 0.37 and 0.44. The cross-validated R^2 is 0.79. The most important descriptors were found to be GAP_inf3, polyb_chi1v_C, b_singleR, and polyq_b_double. GAP_inf3 on its own shows a correlation coefficient of 0.7 with the band gap. polyb_chi1v_C is the normalized chi1v_C originally designed by Kier and Hall[54]. It describes the extent of conjugation of the structure. b_singleR and polyq_b_double also describe similar properties with different emphasis: b_singleR is the number density of the single bond count and polyq_b_double is the double bond count normalized by the repeat unit displacement q.

TABLE 3 Confusion Matrix for Loss Tangent at 1 kHz

		Actual		
		Low	Medium	High
Predicted	Low	4	3	0
	Medium	0	11	2
	High	0	2	16

TABLE 4 Blind Test Result for Loss Tangent at 1 kHz

Polymer	Loss Tangent	Actual Class	Predicted Class
1,2-DAE	0.0076	M	H
EDR	0.0429	H	H
DieG	0.0188	H	H
PVDF	0.016	H	H
Poly(methyl-*p*-xylene)	0.0025	M	M
Polystyrene	0.0005	L	L

Glass Transition Temperature

Glass transition temperature (T_g) is one of the most important and widely studied properties of polymers. At T_g, dramatic changes in dielectric constant, dielectric loss and breakdown strength occur. Below the T_g, polymer chains are restricted to local motion and polymers are brittle and glassy, and above it, polymers become rubber-like. Even though studies on T_g prediction have been reported for several decades, the physics behind it is still not fully understood. Qualitative descriptions of factors that influence T_g, like the rigidity of the main chain, flexibility of the side chain, size of the side chain and electrostatic interaction have been discussed widely. Many QSPR models have been built in the past based on small data sets or lacking modern statistical validation. As mentioned above, properties like T_g actually converge with increasing molecular weight, but different polymers may have very different convergence limits[55]. Theories have been proposed to explain the influence of molecular weight, such as the free volume theory[56], but the difference between the converged value of T_g and the values for polymers with relatively large molecular weight is usually small. Furthermore, polymers with small molecular weight are usually not feasible as dielectric materials; so converged T_g prediction can provide enough instructive information for synthesis.

The most popular methods are still based on group contributions[17] or partially include some specific description of substructures[16]. For the discovery of new materials, a universal model

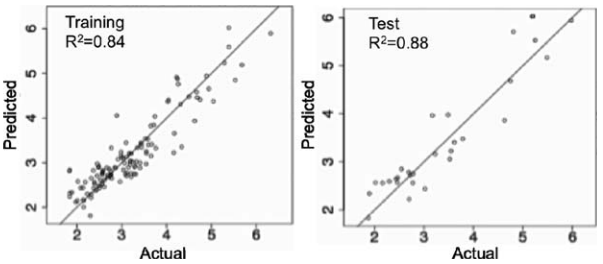


FIGURE 6 Model for the band gap of polymers computed using DFT. left: Training set (125 data points), right: Test set data (30 data points). Unit: eV.

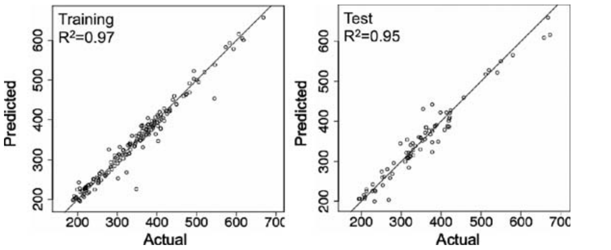


FIGURE 7 Model for glass transition temperature trained using experimental data compiled in ref. 16 Left: Training set (190 data points. Right: Test set (80 data points). Unit: K.

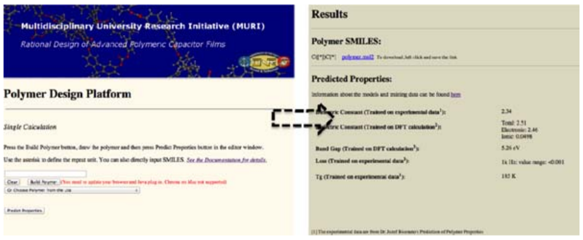
with minimum dependence on specific substructures is required, since the prediction will fail if some new fragments are used in the design that are missing from the parameter library. In the synthesis of new dielectric materials, one should avoid materials with T_g in the working temperature zone; so robust, accurate and universal models are required.

A regression model was built with only the structural descriptors, including the topological descriptors and PEOE partial charge descriptors, resulting in models with high accuracy.

Figure 7 shows the result for the T_g modeling. The SVM model was built with colinearity cutoff 0.85, 143 2D descriptors, seven iterations of SVM recursive feature selection with 90% of descriptors being retained in each iteration. This resulted in 26 descriptors used in modeling. The R^2 and RMSE for the training set were 0.97 and 18.05, respectively and 0.95 and 23.32 for the test set. The R^2 for cross-validation was 0.88. By sensitivity analysis, polyb_zagreb_mod, polyb_zagreb, polyb_Polar, polyb_apol, polybRothN_Weight and polyb_BBrotN_Weight were found to be the top 6 most important descriptors. Descriptors polyb_zagreb_mod and polyb_zagreb represent the number density of the square of the vertex degree only on heavy (non-hydrogen) atoms, the difference being that the former does not consider the terminal heavy atoms (like Cl in PVC) while the latter take all heavy (non-hydrogen) atoms into consideration. Descriptor polyb_Polar is the number density of polar atoms. Polyb_apol is the number density of atomic polarizability. polybRotN_Weight is the molecular weight per rotatable bond for the whole structure. The important descriptors are consistent with the physical meaning of the glass transition temperature. Partial charge descriptors and other descriptors related to the presence of polar groups were also found in the selected descriptor list. From sensitive analysis, they were found to be not as important as those descriptors describing structure flexibility, indicating that the polar-polar interaction may be a secondary mechanism of T_g at least for this training set.

Web Tool Implementation

All the models were implemented in a web tool named Polymer Design Platform (reccr.chem.rpi.edu/polymerdesign). The intention was to make all the models accessible to synthesis groups and other QSPR groups if comparison with other models is desired. Two types of computations are supported: the user can either input a SMILES string with asterisks (*) to indicate the head and tail atoms, or upload a MOE mdb file for a batch computation. Figure 8 shows the user interface and the example results page. We encourage people to use our web tools, report bugs and tell us if any other function is desired.


Multidisciplinary University Research Initiative (MURI)

Rational Design of Advanced Polymeric Capacitor Films

Polymer Design Platform

Single Calculation

Press the Build Polymer button, draw the polymer and then press Predict Properties button in the editor window.

Use the asterisk to define the repeat unit. You can also directly input SMILES. See the Documentation for details.

Results

Polymer SMILES:

C([*])C[*] polymer.mol2 To download, left click and save the link

Predicted Properties:

Information about the models and training data can be found here

Dielectric Constant (Trained on experimental data[1]): 2.34

Dielectric Constant (Trained on DFT calculation[2]): Total: 2.51 Electronic: 2.46 Ionic: 0.0498

Band Gap (Trained on DFT calculation[2]): 5.26 eV

Loss (Trained on experimental data[3]): 1k Hz: value range: <0.001

Tg (Trained on experimental data[1]): 193 K

[1] The experimental data are from Dr. Jozef Bicerano's Prediction of Polymer Properties

FIGURE 8 Polymer design platform. Left: input page; right: results page.

CONCLUSIONS AND FUTURE DIRECTIONS

The new infinite chain descriptors provide a time-efficient and physics-related option for heuristic modeling using the MQSPR techniques on repeat units with any arbitrary size. Models for glass transition temperature, electronic and ionic component of the polarizability computed from DFT, experimental dielectric constant, dielectric loss tangent and band gap were developed and show good predictive performance.

The practical use of QSPR models in materials design workflows requires that some estimate of prediction uncertainty be available. While confidence intervals of linear models have been used for this purpose, there is no consensus on best practices for estimating prediction accuracy for nonlinear machine learning models. This is a rich area of current and future research. Techniques such as bootstrapping[57] and Bayesian estimations[58] provide potential solutions, and may prove valuable in materials informatics applications.

Other future directions include three areas: 1. to develop solubility models to facilitate the design of new polymers; 2. to improve the model for loss tangent with higher resolution; 3. to utilize all the models in the search for promising dielectric polymer materials. It is expected that this program will lead to the development of a general method with broad domain of applicability for modeling and prediction of diverse polymer properties and to computer-aided design of polymeric materials with specific desired physical, chemical, structural and electronic properties.

ACKNOWLEDGMENTS

The present work was supported by a Multi-University Research Initiative (MURI) grant from the Office of Naval Research, under award number N00014100944.

REFERENCES AND NOTES

1 N. Sukumar, M. Krein, Q. Luo, C. Breneman, *J. Mater. Sci.* **2012**, *47*, 7703–7715.

2 C. M. Breneman, L. C. Brinson, L. S. Schadler, B. Natarajan, M. Krein, K. Wu, L. Morkowchuk, Y. Li, H. Deng, H. Xu, *Adv. Funct. Mater.* **2013**, 5746–5752.

3 X. Hao, *J. Adv. Dielectr.* **2013**, *3*, 1330001.

4 X. L. Yu, X. Y. Wang, X. B. Li, J. W. Gao, H. L. Wang, *Macromol. Theory Simul.* **2006**, *15*, 94–99.

5 C. Camacho-Zuniga, F. A. Ruiz-Trevino, *Ind. Eng. Chem. Res.* **2003**, *42*, 1530–1534.

6 A. R. Katritzky, P. Rachwal, K. W. Law, M. Karelson, V. S. Lobanov, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879–884.

7 M. G. Koehler, A. J. Hopfinger, *Polymer (Guildf).* **1989**, *30*, 116–126.

8 X. L. Yu, X. Y. Wang, H. L. Wang, A. H. Liu, C. L. Zhang, *J. Mol. Struct.* **2006**, *766*, 113–117.

9 C. C. Cypcar, P. Camelio, V. Lazzeri, L. J. Mathias, B. Waegell, *Macromolecules* **1996**, *29*, 8954–8959.

10 T. T. M. Tan, B. M. Rode, *Macromol. Theory Simul.* **1996**, *5*, 467–475.

11 W. Brostow, R. Chiu, I. M. Kalogeras, A. Vassilikou-Dova, *Mater. Lett.* **2008**, *62*, 3152–3155.

12 X. L. Yu, B. Yi, X. Y. Wang, Z. M. Xie, *Chem. Phys.* **2007**, *332*, 115–118.

13 C. Duce, A. Micheli, R. Solaro, A. Starita, M. R. Tine, *J. Math. Chem.* **2009**, *46*, 729–755.

14 W. Q. Liu, *Polym. Eng. Sci.* **2010**, *50*, 1547–1557.

15 Q. Luo, N. Sukumar, C. M. Breneman, K. Bennett, M. J. Embrechts, *Abstr. Pap. Am. Chem. Soc.* **2002**, *224*, U497–U497.

16 Bicerano, J. Prediction of Polymer Properties; CRC Press, **2002**.

17 Van Krevelen, D. W.; Te Nijenhuis, K. Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions; Access Online via Elsevier, **2009**.

18 A. R. Katritzky, S. Sild, M. Karelson, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171–1176.

19 A. J. Holder, L. Ye, J. D. Eick, C. C. Chappelow, *QSAR Comb. Sci.* **2006**, *25*, 905–911.

20 W. Q. Liu, P. G. Yi, Z. L. Tang, *QSAR Comb. Sci.* **2006**, *25*, 936–943.

21 J. Xu, L. Wang, G. J. Liang, L. X. Wang, X. L. Shen, *Polym. Eng. Sci.* **2011**, *51*, 2408–2416.

22 T. Le, V. C. Epa, F. R. Burden, D. A. Winkler, *Chem. Rev.* **2012**, *112*, 2889–2919.

23 T. Dipaolo, L. B. Kier, L. H. Hall, *Mol. Pharmacol.* **1977**, *13*, 31–37.

24 B. E. Mattioni, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 232–240.

25 V. Sharma, C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs, R. Ramprasad, *Nat. Commun.* **2014**, *5*, 4845.

26 Ku, C. C.; Liepins, R. Electrical Properties of Polymers: Chemical Principles; Hanser Publishers, **1987**.

27 R. G. Lorenzini, J. A. Greco, R. R. Birge, G. A. Sotzing, *Polymer (Guildf).* **2014**, *55*, 3573–3578.

28 R. G. Lorenzini, W. M. Kline, C. C. Wang, R. Ramprasad, G. A. Sotzing, *Polymer (Guildf).* **2013**, *54*, 3529–3533.

29 A. F. Baldwin, R. Ma, T. D. Huan, Y. Cao, R. Ramprasad, G. A. Sotzing, *Macromol. Rapid Commun.* **2014**, *35*, 2082–2088.

30 N. M. O'Boyle, *J. Cheminformatics* **2012**, *4*, 22.

31 L. B. Kier, *Quant. Struct. Relationships* **1989**, *8*, 221–224.

32 A. T. Balaban, *Pure Appl. Chem.* **1983**, *55*, 199–206.

33 A. Graovac, I. Gutman, N. Trinajstić, T. Živković, *Theor. Chim. Acta* **1972**, *26*, 67–78.

34 J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.

35 C. M. Breneman, M. Rhem, *J. Comput. Chem.* **1997**, *18*, 182–197.

36 C. E. Whitehead, C. M. Breneman, N. Sukumar, M. D. Ryan, *J. Comput. Chem.* **2003**, *24*, 512–529.

37 N.Sukumar, C. M. Breneman, In The Quantum Theory of Atoms in Molecules; Wiley-VCH Verlag GmbH & Co. KGaA, **2007**; pp. 471–498.

38 T. S. Balaban, A. T. Balaban, D. Bonchev, *J. Mol. Struct.* **2001**, *535*, 81–92.

39 Molecular Operating Environment (MOE) **2012**.

40 H. Kuhn, *J. Chem. Phys.* **1949**, *17*, 1198.

41 S. S. Zade, M. Bendikov, *Org. Lett.* **2006**, *8*, 5243–5246.

42 Stewart, J. J. P. MOPAC2012 **2012**.

43 A. J. Smola, B. Schölkopf, *Stat. Comput.* **2004**, *14*, 199–222.

44 M. Krein, T.W. Huang, L. Morkowchuk, D. K. Agrafiotis, C. M. Breneman, *Stat. Model. Mol. Descriptors QSAR/QSPR.* **2012**, *2*, 33–64. Vol.

45 C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273–297.

46 Hyvärinen, A.; Hurri, J.; Hoyer, P. In *Natural Image Statistics SE - 5*; Computational Imaging and Vision; Springer: London, **2009**; Vol. *39*, pp. 93–130.

47 C. J. C. Burges, *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.

48 H. Ruuska, E. Arola, K. Kannus, T. T. Rantala, S. Valkealahti, *J. Chem. Phys.* **2008**, *128*, 064109.

49 D. B. Turner, P. Willett, *Eur. J. Med. Chem.* **2000**, *35*, 367–375.

50 X. L. Yu, B. Yi, F. Liu, X. Y. Wang, *React. Funct. Polym.* **2008**, *68*, 1557–1562.

51 L.M. Wang, In *2006 25th International Conference on Microelectronics.* 576–579. IEEE; p

52 D. Fabiani, L. Simoni, *IEEE Trans. Dielectr. Electr. Insul.* **2005**, *12*, 11–16.

53 H. Frohlich, *Proc. R. Soc. A Math. Phys. Eng. Sci.* **1947**, *188*, 521–532.

54 L. H. Hall, L. B. Kier, *J. Pharm. Sci.* **1977**, *66*, 642–644.

55 Y. Ding, A. Kisliuk, A. P. Sokolov, *Macromol.* **2004**, *37*, 161–166.

56 J. S. Vrentas, C. M. Vrentas, J. L. Duda, *Polym. J.* **1993**, *25*, 99–101.

57 T. J. Diciccio, B. Efron, *Stat. Sci.* **1996**, *11*, 189–228.

58 K. De Brabanter, J. De Brabanter, J.A. Suykens, B. De Moor, *IEEE Trans. Neural Netw.* **2011**, *22*, 110–120.