1 Introduction

MQSPR refers to the application of the science of quantitative structure-property relationship (QSPR) modeling to materials informatics. QSPR/QSAR has been successfully used in drug discovery studies. The motivation of MQSPR study is to build up a time-efficient and low-cost screening method for desired materials, relating structural information (descriptors) to macroscopic properties. By the application of the tools of statistical learning, MQSPR has already shown exciting performance on the prediction of glass transition temperature, polarizabilities and HOMO-LUMO gap of some polymers and functional groups. [1] In this study, a new set of descriptors designed specially for polymers (Infinite Chain Descriptors) has been introduced. MQSPR models for the dielectric constants, dielectric loss (1kHz and 60Hz), band gap and glass transition temperature were built and validated. A web tool integrating all the available models was also developed.

2 Materials and Methods

2.1 Source function

A natural question that arises when performing non-periodic computations on polymer systems is how many atoms or repeat units need to be included in the computations in order to provide a reasonably accurate representation of the infinite chain. This reduces to a question of the chemical transferability of repeat units and groups between different molecules or polymers of different length. Bader[2] has argued that fragments bounded by surfaces of zero flux in the electron density \((\nabla \rho(r)).n = 0\) maximize such transferability. A sensitive measure of this transferability of fragments is provided by the source function is[3], which is a much more sensitive index than is the atomic energy or the integrated electron population.

The local source function is given by the expression[3, 4]:

\[
L_S(r, r') = \frac{-1}{4\pi} \frac{\nabla^2 \rho(r')}{|r-r'|}
\]  

(1)

While the Laplacian of the electron density \(\nabla^2 \rho(r)\) maps out regions of concentration or depletion of electron density[2], the local source represents the effectiveness of the concentration (or depletion)
∇^2 \rho(r') at r' in functioning as a source (or sink) for the electron density at r modulated by the Green’s function \((4\pi|\mathbf{r}-\mathbf{r}'|)^2\). Thus the Laplacian also serves as the generator of the electron density distribution, by virtue of Poisson’s equation:

\[
\rho(r) = \frac{-1}{4\pi} \int \mathrm{d}r' \frac{\nabla^2 \rho(r')}{|r-r'|} \tag{2}
\]

Equation (1) may be rewritten as:

\[
LS(r, r') = \frac{-1}{\pi} \frac{2G(r') + V(r')}{|r-r'|} \tag{3}
\]

Any region where the electron density is locally concentrated \((\nabla^2 \rho(r') < 0)\) in a molecule, and where the potential energy dominates the kinetic energy, acts as a source for the electron density at other points in the molecule, while a region where the electron density is locally depleted \((\nabla^2 \rho(r') > 0)\) and where the kinetic energy dominates acts as a sink, removing electron density from \(r\).

The integral of the local source over the basin \(\Omega\) of an atom or functional group is the source function \(S(r, \Omega)\) contribution from that atom or functional group to \(\rho(r)\):

\[
S(r, \Omega) = \int_{\Omega} LS(r, r') \mathrm{d}r' \tag{4}
\]

Thus

\[
\rho(r) = \int_{\Omega} LS(r, r') \mathrm{d}r' + \sum_{\Omega' \neq \Omega} \int_{\Omega'} LS(r, r') \mathrm{d}r' = S(r, \Omega) + \sum_{\Omega' \neq \Omega} S(r, \Omega') \tag{5}
\]

The electron density at any point in an Atom in a Molecule is thus decomposed into a contribution arising from sources within the basin of the atom and a contribution arising from sources external to the atom. The source function is a measure of the relative contribution of an atom or group to the density at any point.

Perfect transferability of group/repeat unit properties requires not just the transferability of the corresponding electron density, but also that the sum of the contributions to this density from the remaining groups or repeat units in the molecule or polymer be constant. Transferability of functional groups within alkyl chains has been demonstrated[5], but there are no corresponding studies on diverse polymer systems in the available literature. The source function is thus the appropriate quantity to study to establish the convergence of polymer properties with chain length.

In this study the source function was calculated using AIMALL[6] from the wave function file generated from Gaussian09[7]. The calculation was done using PBEPBE/6-31G*.

### 2.2 Dataset

The experimental data on dielectric constants for 58 polymers and of Tg for 162 polymers were obtained from literature[8]. The electronic and ionic components of dielectric constant and band gap of 152 polymer structures were calculated by DFT methods (Dr. Ramprasad’s group), treating polymers as infinite chains. Dielectric loss data on 21 polymers were obtained from Dr. Sotzing’s group.

### 2.3 Infinite chain descriptors
Descriptors are numerical values containing certain structural information, which are used to search for a correlation to the target properties. The descriptors used in this study are redesigned from traditional descriptors that were previously used in (M)QSPR for polymers and small molecules, retaining the chemical meaning of these descriptors. Since properties for polymers usually converge as the chain grows, a reasonable procedure to employ is to directly use descriptors for the infinite chain instead of a local structural representation, which either lack information on the connection environment or are time-consuming if using oligomers. It is also unacceptable for modeling if monomer and dimer give different predictions, since they represent the same polymer. So we seek descriptors that are scale-insensitive. The problem of lack of information on 3D conformations is a stochastic issue. So primarily 2D descriptors were considered, which are conformation-insensitive. The descriptors can be divided into three categories:

- **Topological descriptors**: containing the information of shape and structural flexibility, based on classic topological descriptors like Kier’s shape descriptors[9] and BalabanJ[10].

- **Partial charge descriptors**: based on PEOE 2D partial charge algorithm[11], containing information on partial charge distributions and electrostatic interactions.

- **Electronic TAE descriptors[12]**: derived from transferable electron density distributions, and containing information such as local ionization potential and other electronic properties.

Three types of treatment have been used to convert those descriptors into the infinite chain version:

- **Separate side chains from backbones and calculate primarily descriptors on the side chains.** This is based on the observation that side chain and backbone usually have distinct structural effects. Descriptors with prefix “sg” or “s” are calculated in this way. A special version of Kier descriptor (pKier) also takes the distance between side chains into consideration.

- **Redefine the connection condition for connecting atoms (head and tail) in the descriptor algorithm.** All TAE and PEOE descriptors are calculated in this way.

- **After redefining the connection condition, normalize the descriptor value by either the total number of bonds or atoms (depend on the meaning), with prefix “polyb”, or the repeat unit displacement (the “distance” from head to tail), with prefix “polyq”.** The latter can be seen as a measure of the extension of the polymer chain in the direction perpendicular to the backbone.

Fig 1 shows the idea of redefining the connection environment and how it facilitates the calculation. A total number of 228 the descriptors were calculated in MOE®[13].

![Fig 1. Left: redefine the connection environment according to the structure. Right: descriptor value (TAE descriptor: RECON_FDel(G)NA1) in convergence condition can be easily obtained.](image-url)
2.4 SVR, cross-validation and feature selection

Support vector machine regression (SVR) is a machine learning technique that utilizes the kernel trick in order to find linear relationships in 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 for outliers in the training set. K-fold cross-validation employs an iterative approach, whereby a subsection of training data is withheld, models are built on the remaining data, and the withheld data are used assess the generalizability of the model. In this study, the training data are randomly assigned to one of 10 cross-validation groups. The randomization is performed 10 times. Feature selection is used to reduce the number of descriptors in order to avoid over-fitting. In this study, objective and PLS feature selection were used.

2.5 Network measures in polymer space

2-D and 3-D descriptors for polymers were computed from their repeat units, followed by Principal Component Analysis (PCA) using the MOE® software package. Euclidian distances in PCA space were then used to generate network graphs[14], the nodes of which represent distinct polymers and the edges represent similarity relations between them. The degree distributions of nodes, clustering coefficients and betweeness centrality measures were determined on these networks. Centrality Indices are quantitative measures of the relative importance of vertices in a given network. The clustering coefficient describes the correlation between high degree nodes and provides information on the hierarchical character of a network. Such measures can identify hubs in polymer space and provide information on local as well as global network behavior, revealing hitherto unnoticed correlations within the data set. Use of these network measures is inspired by the need to generate predictive models for polymer properties, while investigating hitherto unexplored regions of polymer space.

3 Results and discussion

3.1 Source function and convergence behavior

Fig 2. The contribution of the electron density from each repeating unit to the “head” hydrogen (square). The convergence of HOMO-LUMO gap calculated using PM3 method for oligomers (triangle). Left: Polyethylene. Right: Polytetrafluroethylene.
Fig 2 shows the convergence of the electron density contribution from each repeat unit to the first hydrogen atom for polyethylene and polytetrafluoroethylene (PTFE), and the change in HOMO-LUMO gap with chain growth, calculated by the PM3 method. Convergence behavior can be clearly observed, both for the electron density distribution and for physical properties.

3.2 MQSPR models

Fig 3 shows the MQSPR models for dielectric constants, band gap and Tg. 20% of the data were retained as a test set and not used in the model training. For each model, the test set results are generally good.
Fig 3. The MQSPR models for dielectric constants, Tg and band gap. The dielectric constants calculated by DFT methods consist of two parts: electronic component ($\varepsilon_e$) and ionic component ($\varepsilon_i$). The test set $r^2$ for each models are: 0.86 ($\varepsilon(exp)$), 0.75 (Tg), 0.78 ($\varepsilon_e$), 0.62 ($\varepsilon_i$) and 0.77 (band gap).

Fig 4 shows the MQSPR models for dielectric loss at 60Hz and 1kHz. Due to the limited available data, we did not use a test set here. But only 3 (1kHz) and 4 (60Hz) descriptors are used for the 21 polymers. The responses were found to correlate well with single descriptors. The two descriptors found important by feature selection for the two models, polyq_b_rotN and polyb_chi0, can both be related to flexibility of the polymer structures. In addition, descriptors sharing similar chemical information were also found to be correlated, as expected.

The results show significant correlation between the infinite chain descriptors and the polymer properties. The design of infinite chain descriptors guarantees a broad domain of applicability of the MQSPR models. Among all the properties of interest, the model for the ionic component of dielectric constant is relatively less satisfactory, but since in practice, the ionic component only accounts for around 10% of the total dielectric constant, the larger variation of predicted $\varepsilon_i$ would not make much difference for high throughput screening.

3.3 The web tool “Polymer Design Platform”

Based on the MQSPR models, a web tool “Polymer Design Platform” was developed in order to allow users to make use of the results and search for desired polymer structures, which usually involve trade-offs between multiple properties. Users can either input the SMILES strings of polymer repeat units or upload a list of polymers repeat units in .mdb format, whereupon the polymer properties are calculated and the results displayed. Fig 7 shows the user interface and results page.
4 Future directions

4.1 Supplementing the TAE Atom types

The Transferable Atom Equivalent (TAE) library used in our group[15-17] is well populated with organic functional groups, since these are the most important atom types for drug design applications for which the TAE RECON descriptors were developed. However the TAE library is poor in heavier elements and metals. As we explore polymer systems containing atoms lower in the periodic table than found in typical organic molecules, it becomes increasingly important to populate the TAE library with more atom types representative of such systems. Adding new atom types is generally done using the Atomgen program[15, 16], which uses the Atomtyper algorithm[15, 16] to identify the new atom types required, constructs small molecules containing these atom types, then runs the Gaussian program suite (Gaussian) to compute molecular wavefunctions and makes calls to Bader’s Atoms-in-Molecules[2] algorithm PROAIM, which identifies the zero-flux ($\nabla \rho \cdot n = 0$) atomic bounding surfaces and integrate the electronic properties within these atomic boundaries to compute the TAE properties. The atom types in our present TAE library are computed at the Hartree-Fock level of theory using all-electron basis sets. However such basis sets become increasingly cumbersome for heavier atoms and it is preferable to employ effective core potentials for such atoms. Heavier atoms and effective core potentials cause problems with the original PROAIM code, with the integrations often failing (leaving large residual differences between the total integrated electron density and the total number of electrons). Other more efficient codes, such as AIMAll[6] by Todd Keith, are now available, which are capable of handling these complications. An order for an AIMAll Professional license has been placed, that will allow us to integrate AIMAll into the Atomgen work flow. When this code is acquired and

Fig 5. The user interface and result page of “Polymer Design Platform”. Left: Single structure mode; Right: Batch mode
testing the implementation is complete, several atom types for elements such as germanium and tin, in various chemical bonding environments, will be generated and added to the TAE library, allowing rapid evaluation of TAE descriptors for any polymers containing these atom types.

### 4.2 Further Electronic Descriptors

Several important electronic descriptors employed by our group for predicting polymer properties are derived from density functional theory, and there is still scope for further exploration. The Hohenberg-Kohn energy functional:

$$E[\rho(r)] = F[\rho(r)] + \int v(r) \rho(r) dr$$  \hspace{1cm} (6)

and the Hohenberg-Kohn theorem[18] give an expression for arbitrary variations in the density $\rho'(r)$, with external potential $v(r)$ fixed:

$$\delta \{E[\rho'(r)] - \mu N[\rho'(r)]\} = 0$$  \hspace{1cm} (7),

where the chemical potential $\mu$, the Lagrange multiplier for the normalization constraint on the density, is the functional derivative of the energy functional with respect to the density:

$$\mu = \frac{\delta E[\rho(r)]}{\delta \rho(r)}$$  \hspace{1cm} (8),

and has been identified[19, 20] as the negative of the electronegativity familiar to chemists:

$$\chi = -\mu = -(\partial E/\partial N)_v \approx \frac{1}{2}(I+A)$$  \hspace{1cm} (9),

where $I$ is the ionization energy and $A$ the electron affinity. Parr's electronegativity formula reduces to Mulliken's electronegativity definition[21] in the finite difference approximation. The equilibrium condition requires equalization of chemical potentials in a molecule, termed the principle of equal orbital electronegativities[11].

The functional derivative of the chemical potential with respect to the density is the chemical hardn or band gap[22]:

$$\eta = \delta \mu/\delta \rho(r)|_{\n=r} = \delta^2 E[\rho(r)]/\delta \rho(r)^2|_{\n=r} = (\partial^2 E/\partial N^2)_v \approx \frac{1}{2}(I-A)$$  \hspace{1cm} (10),

which is a descriptor of crucial importance for capacitors. The other functional derivatives have been identified as the Fukui function[23-26]:

$$f(r) = \delta \mu/\delta v(r)|_{\n=r} = \delta^2 E[\rho(r)]/\delta \rho(r)\delta v(r) = \partial \rho(r)/\partial N|_v$$  \hspace{1cm} (11),

which measures propensity for chemical reactivity, and the response function $\delta \rho(r)/\delta v(r)|_{\n=r} = \delta^2 E[\rho(r)]/\delta v(r)\delta v(r)|_{\n=r}$. The Fukui function is approximated using Koopmans theorem as:

$$f'(r) = \left[ \frac{\partial \rho(r)}{\partial N} \right]_{v(r)} \approx \rho_{\text{HOMO}}(r)$$  \hspace{1cm} (12)$$
$$f'(r) = \left[ \frac{\partial \rho(r)}{\partial N} \right]_{v(r)} \approx \rho_{\text{LUMO}}(r)$$  \hspace{1cm} (13)$$

and$$f_0'(r) \approx \frac{1}{2}(\rho_{\text{HOMO}}(r) + \rho_{\text{LUMO}}(r))$$  \hspace{1cm} (14),

which describe reactivity towards electrophilic, nucleophilic and radical attack, respectively. $f'(r)$ and $f_0'(r)$ are implemented in the TAE RECON algorithm[15-17]. Applications of these descriptors derived from density functional theory for drug design have been reviewed in[17, 27]. Descriptors derived from higher functional derivatives of the energy functional, such as the hyper-hardness[28].
\[ \gamma = \eta^{(2)} = \delta^3 E [\rho(r)] / \partial N^3 \mid_v = \delta \eta(r) / \partial N \mid_v \approx (\partial I / \partial N) \mid_v - (\partial A / \partial N) \mid_v \]

(15),

and the dual descriptor[29]

\[ f^{(2)}(r) = \partial f(r) / \partial N \mid_v = \delta \eta(r) / \partial v \mid_N \approx [\rho_{LUMO}(r) - \rho_{HOMO}(r)] \]

(16),

have recently been discussed: hyper-hardness[29] gives a comparison between the variation of the ionization potential and the electron-affinity when the number of electrons in the system changes. Systems with positive hyper-hardness have been found to exhibit high stability, while large negative values of \( \gamma \) are generally found for chemically reactive systems. The dual descriptor[29] has been touted as a complement to the molecular electrostatic potential to explain local interactions. It has been employed to rationalize covalent interactions that cannot be explained by the molecular electrostatic potential, which focuses primarily on ionic interactions. However applications of these descriptors in prospective predictions of properties of materials have not been seriously explored.

Another descriptor derived from the electron density and implemented in the TAE RECON algorithm is the Average Local Ionization Potential, PIP(\( r \)), introduced by Politzer et al.[30-32] : 

\[ \overline{\text{PIP}}(\mathbf{r}) = \sum \frac{\rho_i(\mathbf{r}) e_i}{\rho(\mathbf{r})} \]

(17),

where the summation runs over the occupied orbitals, \( e_i \) are the orbital energies and the molecular surface is thus encoded with energy-weighted orbital densities \( \rho_i(\mathbf{r}) \). PIP(\( \mathbf{r} \)) identifies hard regions of the electron density (represented by maxima of the local average ionization potential on the molecular surface) and soft, electron-donor or hydrogen bond acceptor regions (corresponding to minima of PIP on the surface). PIP(\( \mathbf{r} \)) has been found to be important in several applications. Analogous to PIP(\( \mathbf{r} \)), an extension of Koopmans’ theorem can be used to define the local electron affinity \( EA_i(\mathbf{r}) \)[33] (better described as a local index of the acceptor properties of the molecule) by summing over only the unoccupied or virtual orbitals in equation (17). Another local property descriptor is the local polarizability, an occupation-weighted sum of the orbital polarizabilities[33]. The utility of such descriptors for polymer property prediction will be explored in future work.

4.3 Modeling beyond chemical structure

Even though the major contribution to the dielectric properties is from the chemical structure, it is also realized that information “beyond chemical structure” such as the crystal structure and inter-chain non-covalent interaction also play a role either as perturbations on the basic value or in a more important manner. It would be especially interesting to investigate this in future work.

4.4 Polymer Structure Generator

Finally the most practical application of MQSPR modeling would be to provide recommendations on the polymer structures for synthesis. We address this need by introducing the tool Polymer Structure Generator. Borrowing from the idea of genetic algorithms and the knowledge of polymer chemists, large numbers of new and synthesizable polymer structures can be generated from molecular fragment libraries being constructed. This software will also be available as a web tool in due course.