Determining influential descriptors for polymer chain conformation based on empirical force-fields and molecular dynamics simulations

Ruimin Ma, Dezhao Huang, Teng Zhang, Tengfei Luo*

Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

Abstract

Many properties of bulk polymers are closely related to the molecular chain conformation. It is ideal if guidance exists to link molecular features to polymer chain conformation to significantly narrow down the design space. Here, we analyze a series of molecular descriptors derived from empirical force-fields and relate them to single chain conformation characterized by radius of gyration calculated from molecular dynamics simulations. Using data analyses and machine learning techniques, we identify that the weakest dihedral angle, characterized by the energy constants, along the chain backbone is the most influential descriptor determining the single chain radius of gyration.

1. Introduction

Identifying and designing polymers with superior properties is of great interest to both scientific research and industrial applications. Polymer properties, such as mechanical strength, flexibility, electrical and thermal conductivity, are closely related to the molecular chain conformation. Simulations and experiments have been performed to explore the relationship between polymer chain conformation and its mechanical properties [1–3]. For example, by controlling the content ratio and additive in polymer-fullerene films, fine nanomorphology is found to increase the stiffness and strength, reducing the ductility of the films [1]. Chain entanglements in low-bandgap conjugated polymers are found to significantly increase the elastic modulus and toughness by molecular dynamics (MD) simulations of different chain conformations (melt-quenched and self-aggregated morphologies) [2]. A review summarizes that the electronic properties of conjugated polymers depend sensitively on the spatial conformation of the polymer chains [4]. Thermal conductivity is also found closely related to the polymer chain conformation [5–11]. For example, due to mechanical drawing, polymer chains can be aligned to significantly improve orientation order and thermal conductivity [9].

Radius of gyration ($R_g$), which describes the extension of polymer chains in space, is an important parameter to characterize the chain conformation [12]. $R_g$ can be manipulated by environmental factors, such as pH [13], temperature [14], and the size of inorganic additives in polymers [15,16]. Previous works mostly focus on how to enhance the $R_g$ in the bulk polymers via different techniques and thus to influence their properties. However, there are very limited studies to correlate the molecular features (e.g., atom type, bond strength and etc.) to the $R_g$ of polymer chains. $R_g$ can be both calculated using MD simulations and measured experimentally. However, to screen a large amount of bulk amorphous polymers by calculating their chain conformation is impractical due to the large parameter space involved in sample preparation, which makes simulation intractable. The same challenge can be encountered for experimental screening of polymer conformation. However, if an easily computable molecular descriptor that has a strong correlation with the chain conformation (e.g., $R_g$) can be identified, it is possible to screen a large number of polymer molecules in a high-throughput manner or design polymeric materials based on machine learning techniques to identify those with desirable properties. Using easily computable material descriptors, such as partial charge, molecular polarizability [17,18], the location of the $d$-band center [19], or even a combination of many descriptors [20], material explorations using high-throughput screening and machine learning have been carried out for applications like organic photovoltaics [17], catalysts [19], and thermoelectrics [20]. Some of these key material descriptors are determined based on mining large data [17], while others are from empirical knowledge [19,20].

In materials informatics, theory about molecular descriptors has been well developed [21,22], and with the development of artificial intelligence [23], significant interest has arisen surrounding machine learning for predicting materials with desirable properties [24]. Combining molecular descriptors with machine learning techniques, reduced order models for property prediction can also be derived [25]. Such a method has a remarkable advantage compared to traditional predictive models like MD simulations and

---

* Corresponding author.
E-mail address: tluo@nd.edu (T. Luo).

https://doi.org/10.1016/j.cplett.2018.05.035
0009-2614/© 2018 Elsevier B.V. All rights reserved.
first-principles density functional theory (DFT) calculations, which can be time-consuming and computationally expensive.

In this study, we exhaust all the parameters in the polymer consistent force field (PCFF) for 16 different polymers and analyze their correlations with the conformation property, $R_g$, of single molecular chains, which is calculated using MD simulations. Using pair-wise plots, Pearson correlation analysis, and machine learning techniques, such as Decision Tree and Random Forests, we are able to identify that the smallest dihedral angle energy constants of the polymer molecule backbone is the most influential descriptor determining $R_g$. This study may provide the first step to the high-throughput screening of polymer chemistry to identify compositions with desirable bulk properties.

2. Result & discussion

In materials informatics, there are well-established algorithms [26,27] to extract the molecular descriptors based on the chemical composition and molecular structure of the molecules. These descriptors are then utilized to study the property-structure relation, but such a practice usually requires large amount of existing or easily computable property data [25,28]. However, polymer chain conformation information is not readily available in large amount, which prevents us from effectively utilizing existing molecular descriptors to explore the underlying correlation between molecular features with our property of interest – chain conformation. Using MD simulations to generate data is a potential solution, but the availability of force-fields and their accuracy can limit the usefulness of the generated data and thus any conclusion derived from studying these data. As a first step towards the ultimate determination of the polymer chemistry-conformation relation, we try to understand how the interatomic interactions, which are more fundamental than molecular chemistry and composition, impact the spatial conformation of polymer chains. To achieve this, we randomly select a moderate number of polymer molecules that can be modeled using the well-established PCFF potentials and use MD simulations to calculate their $R_g$ (see the Section 4 for details). Interaction parameters in the force-fields are considered as candidate descriptors (see supporting information (SI) Section 1 for all force-field parameters). We then combine data analysis, machine learning and physical reasoning to identify the descriptor that influences $R_g$ the most.

After researching the literature and the Materials Studio software, 16 polymer molecules that can be modeled by the PCFF are selected. The molecules selected are shown in Table 1. As a starting point, we take the following descriptors shown in Table 2 and try to evaluate their correlations with $R_g$. More detailed descriptions of these descriptors are included in Section 1 of the SI.

We start from the average energy constants for bond, angle, dihedral, intra-chain vdW and Coulombic interactions. Fig. 1 shows the pair-plots between different descriptors and the target property, $R_g$, where we can qualitatively observe the correlation between any pair of parameters. If the data fall along the diagonal line in a plot, the corresponding pair of parameters have strong correlation. Otherwise, if the data are scattered randomly all over the plot, the two are not correlated. Since our objective is to identify the correlations between different descriptors and $R_g$, we focus on the top row of plots in Fig. 1. We also note that one can potentially identify correlations among descriptors from these pair plots. Visually, we cannot see any obvious correlations between the descriptors and the target property, $R_g$, but it seems that the average dihedral angle energy constant, $DS_{mean}$, has relatively stronger correlation with $R_g$ among all the descriptors. $DS_{mean}$ is calculated by averaging the dihedral energy constant $K_n$ of all the dihedral angles in a polymer chain. Dihedral energy is calculated via the formula below in the PCFF model [29].

$$E_d = \sum_{n=1}^{3} K_n [1 - \cos(n\phi - \phi_n)]$$

Table 1

<table>
<thead>
<tr>
<th>Polymer chemical structures.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kevlar</td>
</tr>
<tr>
<td>Nylon 6,6</td>
</tr>
<tr>
<td>PK</td>
</tr>
<tr>
<td>PPP</td>
</tr>
<tr>
<td>PE</td>
</tr>
<tr>
<td>PET</td>
</tr>
<tr>
<td>PB</td>
</tr>
<tr>
<td>PMMA</td>
</tr>
<tr>
<td>PANI</td>
</tr>
<tr>
<td>PPy</td>
</tr>
<tr>
<td>PVDF</td>
</tr>
<tr>
<td>PIB</td>
</tr>
<tr>
<td>PPE</td>
</tr>
<tr>
<td>PS</td>
</tr>
<tr>
<td>POAIA</td>
</tr>
<tr>
<td>PNDAPDA</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Descriptors used in Fig. 1.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DS_{mean}$</td>
<td>average value of dihedral energy constant</td>
</tr>
<tr>
<td>$BS_{mean}$</td>
<td>average value of bond energy constant</td>
</tr>
<tr>
<td>$AS_{mean}$</td>
<td>average value of angle energy constant</td>
</tr>
<tr>
<td>$Intra_{vdW}$</td>
<td>intra-chain van der Waals (vdW) energy in a polymer chain</td>
</tr>
<tr>
<td>$Intra_{coul}$</td>
<td>intra-chain Coulombic energy in a polymer chain</td>
</tr>
</tbody>
</table>
where $E_d$ is the dihedral energy, $K_n$ is the dihedral energy constant, $\Phi$ is the dihedral angle, and $\Phi_n$ is the equilibrium dihedral angle. We note that in PCFF, $K_1$ and $K_3$ are zeros except for a few dihedral angles, and thus $K_2$ is usually the leading term for the dihedral energy. For the bond and angle interactions, $K_2$ is the leading term energy constant (see Section 1 in SI).

To better quantify the correlations in Fig. 1, we calculate the Pearson correlation coefficient (PCC), which is being adopted to quantify the correlation between two random variables [30]. PCC closer to 1 or −1 indicates strong correlations, while that approaching 0 suggests no correlation. Details of the PCC calculation is included in the Section 4. Fig. 2 summarizes the PCC values between different descriptors corresponding to the pair plots in Fig. 1. We can see that among all the investigated descriptors, $DS_{mean}$ shows the largest PCC value in correlation with the property of interest, $R_g$.

Physically, polymer chains with stronger dihedral energy will be less likely to have segmental rotations [31]. When such segmental rotations happen in the backbone, there will be kinks along the backbone, and thus the spatial extent of the polymer chain will be

![Fig. 1.](image1.png) The pair plots between different descriptors and the target property $R_g$.

![Fig. 2.](image2.png) Pearson correlation coefficient map between different molecular descriptors and the property $R_g$, $R_g$ (Radius of gyration), $BS_{mean}$ (average value of bond energy constant $K_2$), $AS_{mean}$ (average value of angle energy constant $K_2$), $DS_{mean}$ (average value of dihedral energy constant $K_2$), $Intra_{vdW}$ (intra-chain vdW energy in a polymer chain), $Intra_{coul}$ (intra-chain coul energy in a polymer chain).
smaller, which is reflected by the smaller \( R_g \) (see Fig. 3). One may also expect that intra-molecular non-bond interaction, characterized by \( \text{Intra}_{\text{vdW}} \) and \( \text{Intra}_{\text{coul}} \), to have influences on \( R_g \) since attracting force between different segments of the polymer may lead to self-collapse of the chain and thus reduce \( R_g \). From the data in Fig. 2, however, this is not the case since the PCC of these descriptors with \( R_g \) are low. It is also understandable that bond strength \( (B_S_{\text{mean}}) \) and angle strength \( (A_S_{\text{mean}}) \) both do not have strong impacts on \( R_g \), since the degrees of freedom associated with them (i.e., bond length and angle change) would not influence the chain segmental rotation and thus do not impact the spatial extent of the chains.

According to the above analyses, \( R_g \) shows the strongest dependence on \( DS_{\text{mean}} \). However, this descriptor characterizes the averaged dihedral angle stiffness for both the backbone and side branches. More detailed analysis is then performed on this descriptor. First, it is reasonable to argue that the dihedral angle strength for segments along the backbone is much more important in determining \( R_g \) compared to that of the side branches since it directly impacts the rotation of the backbone segments. It is thus natural to examine the backbone dihedral angle strength as a descriptor. Fig. 4 shows the pair plots of \( DS_{\text{NB\_mean}} \), \( DS_{\text{NB\_max}} \), \( DS_{\text{NB\_min}} \), \( DS_{\text{B\_mean}} \), \( DS_{\text{B\_max}} \), \( DS_{\text{B\_min}} \) and \( R_g \) and the corresponding PCCs.

\( DS_{\text{NB\_mean}} \), \( DS_{\text{NB\_max}} \), \( DS_{\text{NB\_min}} \) are respectively the average, maximum and minimum dihedral energy constants \( K_2 \) of the non-backbone dihedrals. \( DS_{\text{B\_mean}} \), \( DS_{\text{B\_max}} \), \( DS_{\text{B\_min}} \) are respectively the average, maximum and minimum dihedral energy constants \( K_2 \) of the backbone dihedrals. It is interesting that \( DS_{\text{B\_min}} \) is a descriptor strongly correlated to \( R_g \) with a high PCC of 0.8. The understanding is that the segmental rotation along the chain backbone will most likely happen around the weakest dihedral angle in the backbone. In other words, it is the “weak point” in the backbone that determines the likelihood of segmental rotation, which in turn influences \( R_g \). We note that one can calculate the dihedral energy landscape using DFT calculations and fit the potential curve using Equation (1) to obtain descriptor \( DS_{\text{B\_min}} \). With this descriptor identified, one can potentially screen a large number of polymer molecules by using accurate DFT calculations.

The above descriptor identification is based on exploratory data analysis. Here, we show that the same conclusion can be obtained via machine learning technique. We use random forest (RF) to process the descriptor data and calculate the feature importance. The Decision Tree (DT) algorithm is the fundamental component of RF. The DT algorithm splits the dataset into branches using provided features until each leaf node of the tree contains highly homogeneous entries. Fig. 5a illustrates the DT algorithm for our classification task. The intrinsic-regression task is converted into the classification task for feature importance measurement function nested in the Scikit-learn [32]. The samples with \( R_g \) value above the average are assigned label 1, and those with \( R_g \) value below the average are assigned label 0. We use the “gini” attribute to measure the impurity of a node in the tree. If all training instances belong to the same class, then “gini” equals to 0. Simple DT algorithm is prone to overfitting and missing global optimum. Thus, a more robust method, RF, which splits the training set into subsets with group of DTs, is adopted in our study (Fig. 5b). The final results are averaged over the DTs to decrease the variance and increase the accuracy. Besides performing the classification and regression task, RF also measures feature importance. The number of the tree nodes that use certain feature to reduce the impurity is used to define the measurement [33]. The measurement is a
weighted average across all trees in the forest by the associated training sample number in each node. We use Python packages Pandas [34], NumPy [35], Scikit-learn [32], and Matplotlib [36] for data manipulation, machine learning and visualization. Table 3 shows the calculated feature importance scores of descriptors discussed in Figs. 1 and 4. For those descriptors studied in Fig. 4, $\text{DS}_{\text{B}_\text{min}}$ shows the largest feature importance score of 0.254. These feature importance measurements correspond well with the PCC results discussed previously in Figs. 2 and 4.

### 3. Conclusion and perspective

In summary, we analyze a number of molecular descriptors derived from empirical force-field PCFF and relate them to single molecular chain conformation characterized by the radius of gyration, which is calculated from MD simulations. Using data analysis techniques including pair-wise plots, Pearson correlation analysis, and machine learning algorithms, such as Decision Tree and Random Forests, we are able to eventually identify that the smallest dihedral energy constant along the chain backbone is the most influential descriptor determining the single chain radius of gyration. While this work might present a meaningful first step towards the high-throughput screening of polymer chemistry to identify compositions with desirable bulk properties, there are still lots to be done to eventually realize this goal. Follow-up research needs to be done to establish reliable force-fields and extend their applicability to more polymer molecules so that more data can be generated. Analysis of datasets much larger (>10,000 data samples) than the one used in the present study is also needed to consolidate the conclusion from this work and potentially identify other influential descriptors. The chain conformation studied here is only an intermediate property that is more easily computable using MD simulations compared to bulk amorphous polymer properties which should be the ultimate properties of interest. Research is needed to either quantify the relation between the intermediate property and bulk polymer property or to directly link the fundamental descriptors to these bulk properties. However, these are currently limited by the amount of data available, which should be a focus in future studies. With larger datasets, one may establish accurate neural network models that can be used to predict the polymer properties based on the identified influential descriptors, which will eventually help chemists to design/synthesize polymers with desirable properties for different applications.

### 4. Method

#### 4.1. MD methods

In this work, a wide range of molecular dynamics descriptors are screened for strong correlation with $R_g$. 16 different polymers are randomly selected from PolyInfo [37], and modeled using the polymer consistent force field (PCFF). Molecular dynamics descriptors are derived from the force field and chemical structures. $R_g$ is calculated from MD simulations using the large-scale atomic-molecular massively parallel simulator (LAMMPS) [38]. Polymer chains are constructed to have lengths around 50 nm by controlling the degree of polymerization. A single chain of each type of polymer is simulated in vacuum, which is realized by placing the chain in a large simulation box (see Fig. 6). Each system is simulated under the NVT (constant number of atoms, volume, and temperature) ensemble for 1.5 ns, and the last 0.5 ns is used as the production run. A timestep size of 0.25 fs is used for all simulations.

#### 4.2. Pearson correlation coefficient (PCC)

Correlation is a measure of how closely the points lie to the straight line—that is, the strength of a linear association. The PCC, also known as the product moment correlation coefficient, in the

![Fig. 5.](image_url) (a) DT algorithm, splitting the dataset according to present features until Gini equals 0. (b) RF algorithm, bagging many trees and adopting the majority vote to decrease the high variance prediction in DT algorithm.

![Table 3](image_url) Feature importance measurement from the random forests algorithm.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Feature importance</th>
<th>Descriptor</th>
<th>Feature importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS_mean</td>
<td>0.216</td>
<td>DS_NB_max</td>
<td>0.246</td>
</tr>
<tr>
<td>AS_mean</td>
<td>0.154</td>
<td>DS_NB_mean</td>
<td>0.196</td>
</tr>
<tr>
<td>$\text{DS}_{\text{mean}}$</td>
<td>0.300</td>
<td>DS_NB_min</td>
<td>0.110</td>
</tr>
<tr>
<td>Intra_vDW</td>
<td>0.164</td>
<td>DS_B_max</td>
<td>0.030</td>
</tr>
<tr>
<td>$\text{DS}_{\text{coul}}$</td>
<td>0.166</td>
<td>DS_B_mean</td>
<td>0.164</td>
</tr>
<tr>
<td>$\text{DS}<em>{\text{B}</em>\text{min}}$</td>
<td>0.254</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Fig. 6.](image_url) The simulation box of a single polymer chain. The box size was set to be 100 $\times$ 100 $\times$ 100 nm$^3$ and the polymer chain was built to be around 50 nm-long.
population from which the sample was drawn is represented by $\rho$ [39]. For two real-valued random variables $X$ and $Y$, the Pearson correlation coefficient (PCC) is defined as [30]:

$$\rho = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

(2)

The coefficient is dimensionless and takes a value from $-1$ through $0$ to $+1$. If the sign of the correlation coefficient is positive, then a positive correlation exists, and if the sign of the correlation coefficient is negative, then a negative correlation exists. If all the points on the plot lay on a straight line, then a perfect correlation exists (a PCC of $1$ or $-1$). A correlation coefficient of zero indicates no linear association between the two variables—in other words, uncorrelated.

**Acknowledgement**

The authors acknowledge the financial support from the DuPont Young Professor Award program. T.L. would also like to thank the Dorini Family for the endowed professorship in energy studies. This computation was supported in part by the University of Notre Dame, Center for Research Computing, and NSF through XSEDE resources provided by SDSC Comet, TACC Stampede, and NICs Darter under a grant number TG-CTS100078.

**Appendix A. Supplementary material**

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cplett.2018.05.035.

**References**


