THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

Cite This: J. Phys. Chem. C 2019, 123, 14610−14618

pubs.acs.org/JPCC

Accelerated Discovery of High-Refractive-Index Polyimides via *First-Principles* Molecular Modeling, Virtual High-Throughput Screening, and Data Mining

Published as part of The Journal of Physical Chemistry virtual special issue "Young Scientists".

Mohammad Atif Faiz Afzal,*[,†] Mojtaba Haghighatlari,[†] Sai Prasad Ganesh,[†] Chong Cheng,[†] and Johannes Hachmann*[,†,‡,§]

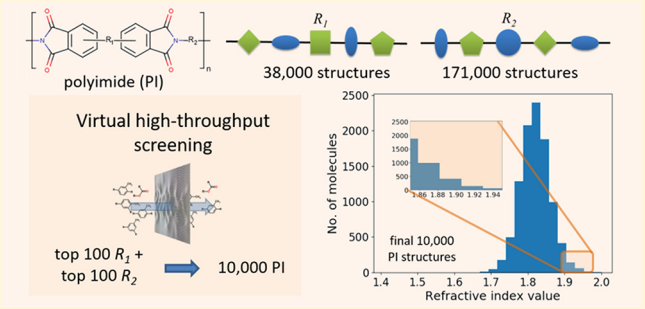
[†]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260, United States

[‡]Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, New York 14260, United States

[§]New York State Center of Excellence in Materials Informatics, Buffalo, New York 14203, United States

Supporting Information

ABSTRACT: We present a high-throughput computational study to identify novel polyimides (PIs) with exceptional refractive index (RI) values for use as optic or optoelectronic materials. Our study utilizes an RI prediction protocol based on a combination of *first-principles* and data modeling developed in previous work, which we employ on a large-scale PI candidate library generated with the *ChemLG* code. We deploy the virtual screening software *ChemHTPS* to automate the assessment of this extensive pool of PI structures in order to determine the performance potential of each candidate. This rapid and efficient approach yields a number of highly promising lead compounds. Using the data mining and machine learning program package *ChemML*, we analyze the top candidates with respect to prevalent structural features and feature combinations that distinguish them from less promising ones. In particular, we explore the utility of various strategies that introduce highly polarizable moieties into the PI backbone to increase its RI yield. The derived insights provide a foundation for rational and targeted design that goes beyond traditional trial-and-error searches.



1. INTRODUCTION

Polyimides (PIs), shown in Figure 1, are synthesized by polycondensation of R_1-containing dianhydride with R_2-based diamine or diisocyanate.[1] They are an appealing class of organic materials due to their exceptional thermal stability and easy processability.[2−4] These properties are complemented by mechanical stability, flexibility, light weight, low cost, as well as flame and radiation resistance, and they thus hold much promise for a range of applications.[5,6]

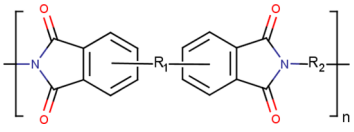


Figure 1. Polyimide (PI) core structure with residues R_1 and R_2.

However, their generally low index of refraction (RI) undermines their utility for use in many optic and optoelectronic devices,[7−11] such as (image) sensors,[12,13] displays,[14] and light sources (including organic light-emitting diodes),[15] in which organic materials can be deployed *in situ* as microlenses,[16] waveguides,[17] microresonators,[18] interferometers,[19] antireflective coatings,[20] optical adhesives,[21] and substrates.[22] Most of these applications demand large RI values, often upward of 1.7 or 1.8. Typical carbon-based polymers—PIs included—only exhibit values in the range of 1.3−1.5.[11] This situation provides a strong incentive to pursue novel high-RI PIs that are suitable for the aforementioned applications.

As the properties of organic materials can be tailored and tuned by controlling their molecular structure, they are a prime target for combinatorial search or rational design attempts.[23−25] The addition of highly polarizable moieties that do not have extensive π-electron conjugation can increase the RI of polymers without negatively affecting other optical properties. (Significant π-conjugation can lead to large optical dispersion and

Received: February 4, 2019
Revised: May 7, 2019
Published: May 15, 2019

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

birefringence,[26] as well as poor transparency and coloration.) For example, the incorporation of small aromatic rings, halogens, metals, and in particular sulfur has shown promise for this purpose.[3,27−31] In 2007, Ueda et al. developed PI films with relatively high RI values of up to 1.75, but they, unfortunately, exhibited large birefringence.[32] In recent years, these findings have been improved upon by increasing the sulfur content of the PIs, yielding RI values of up to 1.76 and smaller birefringence.[33]

In this paper, we present a computational and data-driven approach to study the RI values of PIs and rapidly identify promising lead compounds. We investigate different ways that introduce highly polarizable moieties into the PI framework in order to overcome the technical limits of existing compounds. Given the encouraging results from the earlier work discussed above, we put a focus on incorporating sulfur into the PIs and, in doing so, create a new class of high-RI polymers. In section 2, we detail the methods employed in this work; i.e., we provide background on our data-driven *in silico* approach (section 2.1), introduce our RI modeling protocol (section 2.2), discuss the molecular design space we consider (section 2.3), and outline our data mining and analysis techniques (section 2.4). Section 3 presents and discusses the results of our study, and our findings are summarized in section 4.

2. BACKGROUND, METHODS, AND COMPUTATIONAL DETAILS

2.1. Data-Driven *In Silico* Approach. The development of new materials such as PIs has traditionally been an experimentally driven, trial-and-error process, guided by experience, intuition, and conceptual insights. This approach is, however, often costly, slow, biased toward certain domains of chemical space, and limited to relatively small-scale studies, which may easily miss promising leads (both on individual compounds as well as compound classes).

The study at hand instead embraces a data-driven *in silico* research paradigm that has gained considerable interest in the past few years[34,35] for its promise to address the inherent complexities of structure−property relationships and the vastness of chemical space more efficiently (see, e.g., refs 36−43). Our work brings together molecular modeling, high-throughput computational screening, and machine learning as well as a corresponding software ecosystem to support data-driven discovery and rational design.[44,45] It has its greatest utility as part of integrated research pipelines with experimentalist partners, where it provides guidance for experimental efforts and mitigates some of their before-mentioned shortcomings.

Our research approach and its rationale can be summarized as following: We first establish a computational modeling protocol that can make sufficiently accurate and fast predictions for the target property in the compound class(es) of interest—in the study at hand for the RI values of PIs (cf. section 2.2). We then create a large-scale virtual library of candidates within that compound class (cf. section 2.3), on which we apply this modeling protocol to evaluate the performance potential of each candidate compound. In addition to obtaining information about each individual compound, we can mine the screening results in their entirety to reveal underlying structure−property relationships (cf. section 2.4). Our approach can thus identify both specific lead compounds as well as high-value molecular patterns and features for the *de novo* design of tailored PIs or the creation of additional screening libraries.

2.2. Refractive Index Modeling Protocol. According to the Lorentz−Lorenz equation, the RI (n_r) is a function of the polarizability (α) and the number density (N), i.e.,

$$n_r = \sqrt{\frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0}}$$

In previous work,[46,47] we developed a modeling protocol that allows us to accurately and efficiently predict the RI values of polymers within the Lorentz−Lorenz equation. This protocol is based on a combination of *first-principles* quantum chemistry calculations and data modeling. We compute (static) polarizability values α using the coupled-perturbed self-consistent field equations within the Kohn−Sham density functional theory (DFT) framework with the PBE0 hybrid functional[48] and the double-ζ-quality def2-SVP basis set by the Karlsruhe group.[49] We employ an all-electron, closed-shell approach and include Grimme's D3 correction[50] to account for dispersion interactions. The polarizability calculations are performed on geometries optimized using the universal force field (UFF)[51] following a 3D conformational screening as implemented in the OpenBabel software.[52] All DFT calculations are carried out using the ORCA 3.0.2 quantum chemistry package.[53] We compute the number density N via the van der Waals volume V_{vdW} and packing factor K_p of the amorphous bulk polymer as

$$N = \frac{K_p}{V_{vdW}}$$

We obtain V_{vdW} using Slonimskii's method detailed in ref 54 and K_p using the support vector regression (SVR) machine learning model introduced in ref 46 (except during the prescreening of individual residues R_1 and R_2 discussed in the following section, for which we simply employ a constant K_p of 0.75 that is typically found in PIs[55]). We obtain the asymptotic limit for the polymers via a linear extrapolation scheme as outlined in ref 46. [In exploratory calculations, we already observe near-perfect extensivity based on the monomer units, which we explain with the large size of the PI monomers (∼150 atoms) and the finite correlation length exhibited in them. There is thus no need to perform expensive oligomer-sequence calculations.] We previously demonstrated on a set of 112 nonconjugated polymers with experimentally known RI values that this protocol can make rapid and accurate predictions,[46,47] and thus enable the high-throughput screening study at hand. We execute the latter using our automated virtual screening code *ChemHTPS*[56] which creates inputs, executes and monitors the calculations, parses and assesses the results, and extracts and postprocesses the information of interest. Where applicable, we report the mean absolute error (MAE) and the root mean squared error (RMSE) as well as their percentage errors (MAPE and RMSPE, respectively) compared to available experimental data.

2.3. Molecular Design Space. In Figure 2, we show the RI heat map in the α/N parameter space and mark the positions of the 112 polymers we previously studied.[46,47] The figure emphasizes the relative importance of the counteracting parameters and their feasibility in real-world compounds. In order to design high-RI systems, we in principle have to pursue compounds that simultaneously feature both large polarizability and number density values. The positioning of the known compounds with high RI leans more toward high polarizability and relatively small number density rather than vice versa. One path to narrowing down compound space is to maximize the polarizability for systems from the same compound family, for

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

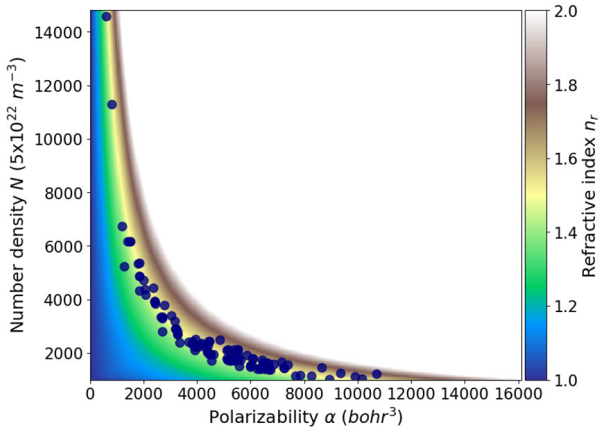


Figure 2. Refractive index (RI) heat map showing the dependence on polarizability and number density expressed in the Lorentz−Lorenz equation. The dots mark 112 experimentally known polymers that were used to benchmark the employed RI model.

which the number density is similar. Given that this is known to be the case for PIs, we pursue this path and focus our search on highly polarizable PI candidates.[55]

For this, we create a PI library based on 29 building blocks and bonding rules shown in Figure 3, which we select following

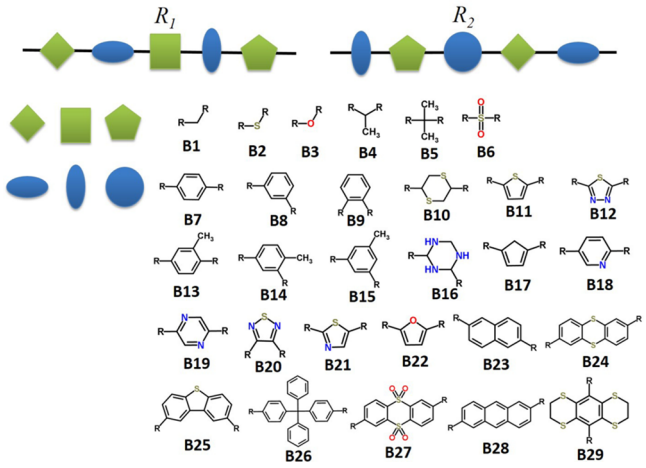


Figure 3. Molecular building blocks used to create the residues R_1 and R_2 of the PI screening library studied in this work. The "R" on the building blocks denote allowed sites for linking. The building blocks marked in green (**B1−B6**) are linkers, and the blue ones (**B7−B29**) are (hetero)aromatic moieties.

empirical expertise with regard to promise and synthetic feasibility. The building blocks can be divided into linkers (**B1−B6**) and (hetero)aromatic moieties (**B7−B29**). Using our *ChemLG* molecular library generator code[57] with a systematic combinatorial linking scheme, we initially generate 38,619 R_1 structures and 171,172 for R_2. Combining all R_1 and R_2 to form PIs would lead to a total of 6.6 billion compounds. To restrict the search space to a more manageable number of candidates, we select only the most promising R_1 and R_2 groups (which ultimately distinguish the PI candidates) on the basis of the computed RI values of the individual R_1 and R_2. After prescreening the residues, we develop PI candidates by introducing the top 100 R_1 and 100 R_2 structures into the PI core motif, as shown in Figure 1. We subsequently screen the 10,000 resulting PI candidates. *ChemLG* keeps a record of the list of building blocks and connections that are used to create a particular compound, and we use this information for the analysis of structure−property relationships.

2.4. Data Mining and Pattern Analysis. In addition to identifying the best candidates from our high-throughput screening, i.e., those with the highest predicted RI values, we further analyze the collected data to develop a better understanding of the structural patterns that lead to high-RI PIs. For instance, we conduct a hypergeometric distribution analysis, in which we compute the *Z*-scores (Z_i) of each building block i used in the creation of the screening library as

$$Z_i = \frac{k_i - m\frac{K_i}{M}}{\sigma_i}$$

with

$$\sigma_i = \left[\frac{mK_i}{M} \times \left(\frac{M - K_i}{M}\right) \times \left(\frac{M - m}{M - 1}\right)\right]^{1/2}$$

where M is the total number of compounds in the entire library, m is the subset of compounds that is considered (e.g., the top compounds), K_i is the number of occurrences of building block i in M molecules, and k_i its occurrences in the subset of m molecules. A large *Z*-score thus indicates a statistical over-expression of the associated building block in the high-RI candidates relative to the overall screening library. By applying the *Z*-score analysis, we can thus identify the most important building blocks and the degree to which they correlate with large RI values. We perform a similar analysis of building block combinations to reveal synergistic effects similar to those known from push−pull or donor−acceptor copolymers. In addition, we present an analysis of *Z*-score trends for each building block in ranked candidate subsets as well as of the average RI values of the candidates derived from each building block. All data mining work was conducted using our *ChemML* code.[58]

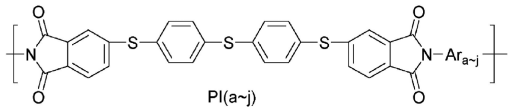
3. RESULTS AND DISCUSSION

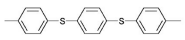
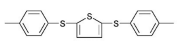
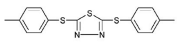
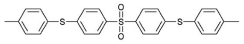
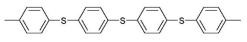
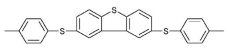
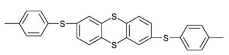
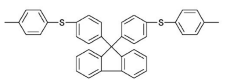
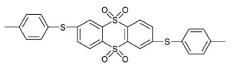
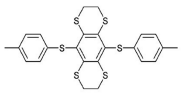
We previously tested the RI modeling protocol described in section 2.2 for the before-mentioned set of 112 nonconjugated polymers and established its predictive performance. However, this benchmark set did not contain any PIs. To prove our protocol's validity for this particular compound class, we perform additional calculations on 10 PIs with experimentally known RI values. We compare the experimental and computational results in Table 1. The MAE (MAPE) and RMSE (RMSPE) values for this comparison are 0.021 (1.2%) and 0.025 (1.4%), respectively, which suggests that our protocol can accurately predict the RI values of PIs as well.

The results of the R_1 and R_2 residue prescreening are summarized in parts a and b of Figure 4, respectively, and the results of the subsequent PI screening are given in Figure 4c. The plots show histograms of the computed RI results. Most of the candidates for R_1 and R_2 have RI values between 1.5 and 1.7 with an average of (a) 1.600 and (b) 1.627, respectively. However, the PI candidates derived from the top R_1 and R_2 residues have significantly larger RI values with an average of 1.843. We note that either of these averages is well above the range of typical organic polymers (i.e., 1.3−1.5), which indicates a good choice of building blocks. As per the objectives of this work, we are able to identify a sizable number of compounds with RI values greater than 1.7, i.e., (a) 2851 (7.4%), (b) 22388 (13.1%), and (c) 9985

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

Table 1. Comparison of RI Results from the Employed Prediction Model with the Experimental Values of 10 Known High-RI PIs[a]



Ar_{a-j}	Structure	Experiment value[11]	Calculated value	Error
a		1.746	1.738	-0.008 (-0.5%)
b		1.753	1.739	-0.013 (-0.7%)
c		1.749	1.707	-0.042 (-2.4%)
d		1.748	1.751	0.003 (0.2%)
e		1.733	1.760	0.027 (1.6%)
f		1.758	1.779	0.021 (1.2%)
g		1.760	1.735	-0.025 (-1.4%)
h		1.726	1.741	0.015 (0.9%)
i		1.737	1.743	0.005 (0.3%)
j		1.769	1.724	-0.045 (-2.5%)

[a]We report errors and percentage errors in each case. For the entire set, we obtain a mean absolute error (MAE) and root mean squared error (RMSE) of 0.021 (1.2%) and 0.025 (1.4%), respectively. (The values in parentheses are the corresponding percentage errors, i.e., MAPE and RMSPE.)

(99.8%). At or above the critical threshold of 1.8, we still find (a) 131 (0.3%), (b) 1252 (0.7%), and (c) 6961 (69.6%) compounds. The shift to considerably larger RI values also demonstrates the success of our strategy to prescreen the R_1 and R_2 individually and build PI candidates on the basis of the top residues. (Details of the prescreening and screening results are provided in the Supporting Information.)

Figure 5 shows the RI distribution of (a) R_1, (b) R_2, and (c) PI candidates containing each constituent building block (cf. the PI candidate library construction illustrated in Figure 3). We find that R_1 and R_2 candidates containing building blocks **B28** (anthracene), **B25** (dibenzothiophene), and **B24** (thianthrene) have the highest RI values, while those containing building blocks **B17** (cyclopentadiene) and **B16** (1,3,5-triazinane) show the lowest RI. The ranking of the building blocks is very similar for both R_1 and R_2 structures, which is unsurprising considering their general similarity. The PI candidates do not contain building blocks **B10** (1,4-dithiane) and **B14** (toluene with linking in the 2,4-position), as these were missing in the top 100 structures of R_1 and R_2 that were used in their construction. The average RI values for all of the other building blocks—with the exception of **B26** (tetraphenylmethane)—are greater than 1.8. However, it should be noted that the number of PI candidates containing a specific building block is very variable. The red plot in Figure 5c shows the count of each building block in the 10,000 PI structures. The most common building blocks in the PI library are **B1** (CH_2-linker), **B25**, **B2** (S-linker), **B28**, and **B3** (O-linker), with **B28** and **B2** occurring in almost all PI candidates. Building blocks **B1**, **B2**, and **B3** do not exhibit particularly large average RI values; however, given the construction template for the residues introduced in Figure 3, they are statistically more likely to occur. The average RI value alone is thus not a sufficient metric to gauge the potential impact of each building block on the performance of PI candidates.

In the following, we analyze the contribution of the building blocks in the high-RI candidates of R_1 and R_2, using the hypergeometric distribution analysis detailed in section 2.4. We focus on the top 10% of the R_1 and R_2 candidates (with RI values greater than 1.687 and 1.711, respectively). We do not include PI candidates in this analysis, as the selective construction scheme with its biased building block counts makes it less meaningful. Figure 6 shows the resulting Z-scores, which identify the over- or underrepresentation of each building block in the high-RI R_1 and R_2 candidates compared to a random sample. The results point to **B28** and **B25** as the most prevalent building blocks in the top residues, and thus the most promising ones to consider for the design of high-RI polymers. This finding is in good agreement with the prior analysis of the RI averages and distribution. The rankings of the building blocks in both R_1 and R_2 are again largely the same, suggesting that the effect of the building blocks is similar for the somewhat different sequences in R_1 and R_2.

The above Z-score analysis only yields insights into the pervasiveness of building blocks in the top 10% candidates. In order to gain a more comprehensive picture, we now evaluate the Z-score of each building block in each 10% segment of the R_1 and R_2 libraries. For this, we sort the R_1 and R_2 libraries by increasing RI value, divide them into 10 subsets, and perform a

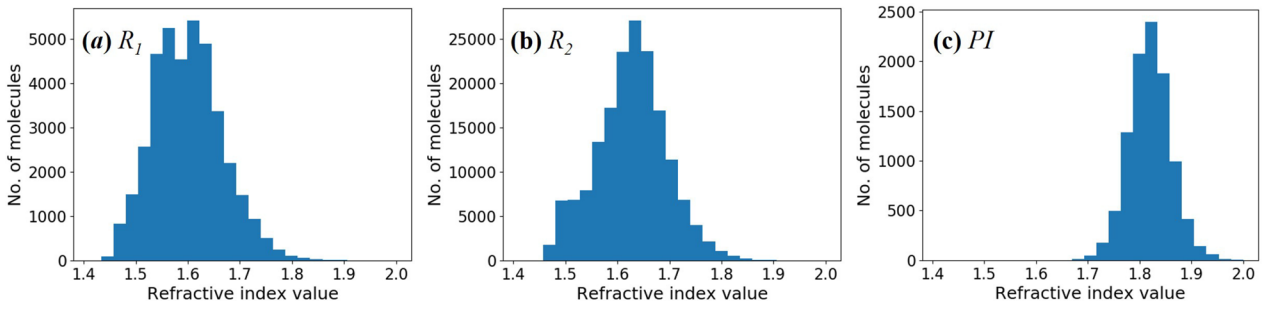


Figure 4. RI distribution histograms of (a) the 38,619 individual R_1 residues; (b) the 171,172 individual R_2 residues; and (c) the 10,000 PI structures resulting from the top R_1 and R_2 residues. We observe a distinct shift toward higher RI values in part c.

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

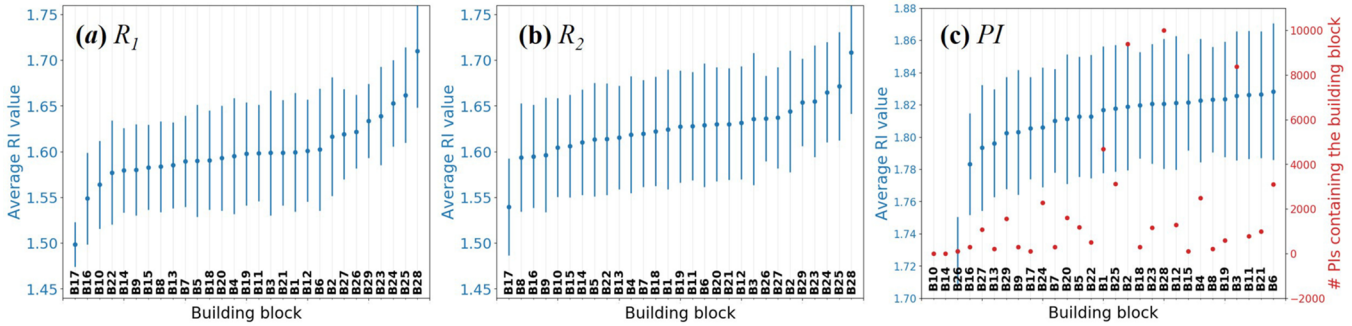


Figure 5. RI value distribution around the respective average RI values (blue points) of the (a) R_1 residues, (b) R_2 residues, and (c) PI candidates containing each building block. The blue bands refer to one standard deviation. The red points in part c show the counts of PI candidates containing a specific building block.

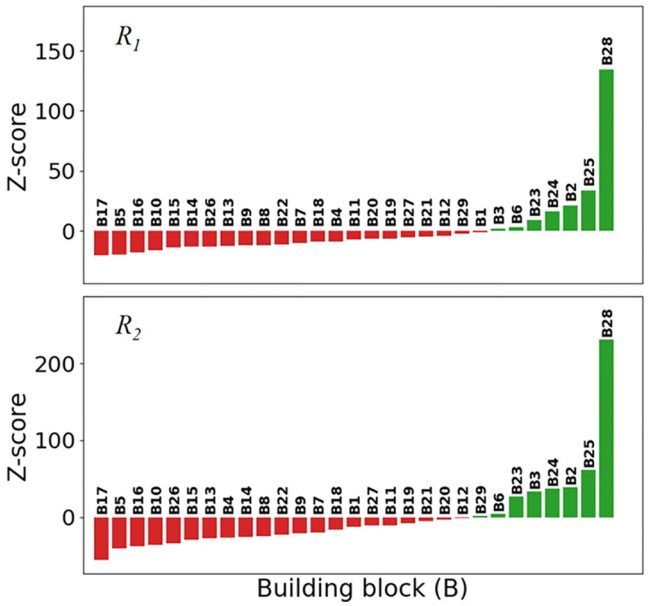


Figure 6. *Z*-score of each building block in the top 10% R_1 and R_2 candidates compared to the entire R_1 and R_2 candidate pool, respectively. Green color represents a positive *Z*-score, and negative ones are shown in red.

hypergeometric distribution analysis for each building block in each subset. The results are shown in Figure 7. We observe certain general trends for the building blocks making up R_1 and/or R_2:

1. The *Z*-score of some building blocks increases with increasing RI values, indicating a direct correlation (e.g., for **B28**, **B25**, **B24**, and **B23**).
2. The *Z*-score of some building blocks decreases with increasing RI values, indicating a negative correlation (e.g., for **B17**, **B22**, **B10**, and **B16**).
3. The *Z*-score of some building blocks goes through a maximum for intermediate RI values, indicating a corresponding correlation with average candidates (e.g., for **B11**, **B26**, **B19**, **B18**, **B20**, and **B7**).
4. The *Z*-score of some building blocks does not show a clear trend, indicating a lesser impact on and correlation with the RI values (e.g., for **B1** and **B6**.)

In the bottom right corner of Figure 7, we plot the average RI values of the R_1 and R_2 candidates in each of the 10 subsets. We observe that the R_2 structures have somewhat higher RI values in comparison to R_1 structures. The principal difference in R_1 and R_2 candidates is the number of aromatic building blocks, i.e., two vs three, and the higher content of these moieties correlates with higher RI values.

In addition to analyzing the influence of individual building blocks on the RI values, we also study the potential impact of building block pairs. For this, we calculate the joint *Z*-scores of all possible building block combinations in the top 10% candidates. The results for R_1 and R_2 are shown in Figure 8. This analysis reveals some dependence on particular building block combinations. For instance, **B23**, **B24**, and **B25** perform significantly worse when paired with **B4** and **B5** but exhibit large *Z*-scores in combination with **B2** and **B3**. However, overall, we find the impact of individual building blocks to be the dominant factor. For instance, **B28** has the largest positive *Z*-score, regardless of its counterpart.

Overall, we find anthracene (**B28**) to be the most promising moiety, a surprising finding and somewhat contradictory to our initial hypothesis considering that it does not contain sulfur. However, it is in line with (and independently confirms) other efforts in the community that aim to integrate anthracene into high-RI optical polymers (see e.g., ref 59). The other outstanding moieties are the S-heterocycles dibenzothiophene (**B25**) and thianthrene (**B24**). Naphthalene (**B23**) shows some promise as well. The O- and S-linkers (**B2** and **B3**, respectively) and to a lesser degree SO_2 (**B6**) outperform the carbon-based linkers.

4. CONCLUSIONS

In this study, we demonstrated that the data-driven *in silico* approach that is being advanced by us and others can rapidly and efficiently assess the properties and performance potential of high-RI candidate compounds, identify numerous leads for next-generation PIs, and elucidate structure−property relationships that form the foundation for rational design rules. By combining our RI prediction model with virtual high-throughput screening techniques, we characterized candidates on a large scale at a fraction of the time and cost of traditional studies. We identified high-value building blocks (e.g., anthracene, dibenzothiophene, thianthrene) and structural patterns (e.g., S- and O-linkers) that correlate with large RI values. Correspondingly, we identified regions in chemical space in which we can hope to maximize the RI values of PIs. These guidelines allow us to target specific molecular motifs and create polymers with exceptional optical properties. In future experimental work, we will utilize these guidelines and pursue the promising candidates that have emerged.

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

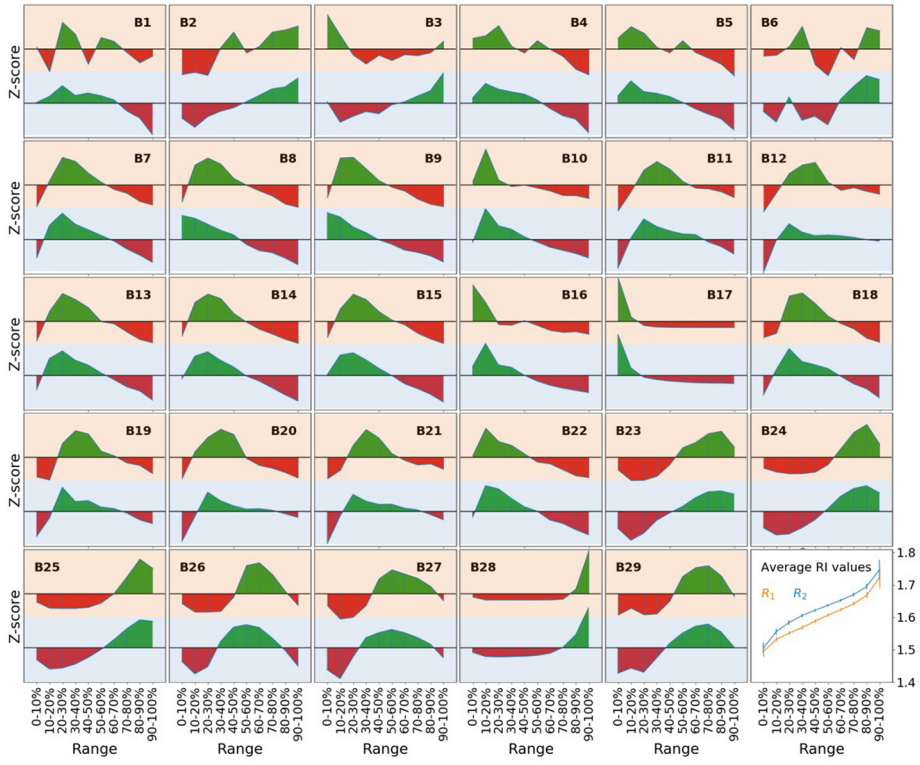


Figure 7. *Z*-score of each building block in all segments of the residual libraries. The top plot in each cell corresponds to R_1, and the bottom plot, to R_2. Green color indicates positive *Z*-scores and red negative values. The last cell shows the average RI values in each segment.

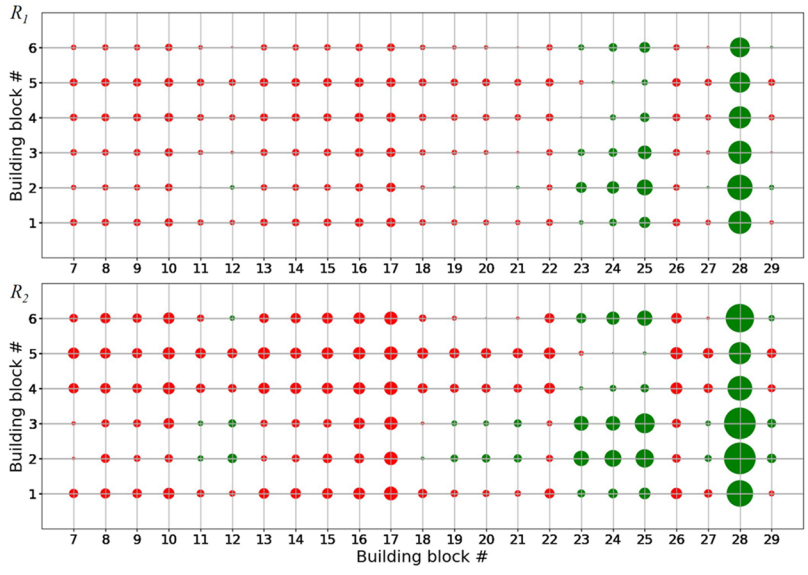


Figure 8. *Z*-scores of building block combinations in the top 10% candidates of R_1 and R_2. The size of the circles corresponds to the magnitude of the *Z*-score value. Green circles indicate positive *Z*-scores and red circles negative values.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcc.9b01147.

Details of the computational data displayed in the figures throughout this paper or that were used in the statistical analysis and also detailed definitions of all statistical metrics used in this work (XLSX).

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: m27@buffalo.edu
*E-mail: hachmann@buffalo.edu

ORCID

Mohammad Atif Faiz Afzal: 0000-0001-8261-2024
Chong Cheng: 0000-0002-4637-1320
Johannes Hachmann: 0000-0003-4501-4118

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

Notes

The authors declare no competing financial interest.

Biography

Johannes Hachmann is an Assistant Professor of Chemical Engineering at the University at Buffalo (UB), a Core Faculty Member of the UB Computational and Data-Enabled Science and Engineering graduate program, and a Faculty Member of the New York State Center of Excellence in Materials Informatics. He earned a Dipl.-Chem. degree (2004) after undergraduate studies at the universities of Jena and Cambridge and M.Sc. (2007) and Ph.D. (2010) degrees in Chemistry from Cornell University, and he conducted postdoctoral research at Harvard University before joining the UB faculty in 2014. The research of the Hachmann Group fuses (*first-principles*) molecular and materials modeling with virtual high-throughput screening and modern data science (i.e., the use of database technology, machine learning, and informatics) to advance a data-driven discovery and rational design paradigm in the chemical and materials disciplines. One of the centerpieces of the group's efforts is the creation of an open, general-purpose software ecosystem for the data-driven design of chemical systems and the exploration of chemical space. This work was recognized with a 2018 NSF CAREER Award.

■ ACKNOWLEDGMENTS

This work was supported by start-up funds provided through the University at Buffalo (UB), the National Science Foundation (NSF) CAREER program (Grant No. OAC-1751161), and the New York State Center of Excellence in Materials Informatics (Grant No. CMI-1148092). Computing time on the high-performance computing clusters "*Rush*", "*Alpha*", "*Beta*", and "*Gamma*" was provided by the UB Center for Computational Research (CCR). The work presented in this paper is part of M.A.F.A.'s Ph.D. thesis.[60] M.H. gratefully acknowledges support by Phase-I and Phase-II Software Fellowships (Grant No. ACI-1547580-479590) of the NSF Molecular Sciences Software Institute (Grant No. ACI-1547580) at Virginia Tech.[61,62]

■ REFERENCES

(1) Odian, G. *Principles of polymerization*, 4th ed.; Wiley: Hoboken, NJ, 2004.
(2) Düsselberg, D.; Verreault, D.; Koelsch, P.; Staudt, C. Synthesis and characterization of novel, soluble sulfur-containing copolyimides with high refractive indices. *J. Mater. Sci.* **2011**, *46*, 4872−4879.
(3) Liu, J.-g.; Nakamura, Y.; Ogura, T.; Shibasaki, Y.; Ando, S.; Ueda, M. Optically transparent sulfur-containing polyimide-TiO2 nanocomposite films with high refractive index and negative pattern formation from poly(amic acid)-TiO2 nanocomposite film. *Chem. Mater.* **2008**, *20*, 273−281.
(4) Higashihara, T.; Ueda, M. Recent progress in high refractive index polymers. *Macromolecules* **2015**, *48*, 1915−1929.
(5) Mathews, A. S.; Kim, I.; Ha, C.-S. Synthesis, characterization, and properties of fully aliphatic polyimides and their derivatives for microelectronics and optoelectronics applications. *Macromol. Res.* **2007**, *15*, 114−128.
(6) Chang, C.-M.; Chang, C.-L.; Chang, C.-C. Synthesis and optical properties of soluble polyimide/titania hybrid thin films. *Macromol. Mater. Eng.* **2006**, *291*, 1521−1528.
(7) Lei, T.; Wang, J.-Y.; Pei, J. Roles of flexible chains in organic semiconducting materials. *Chem. Mater.* **2014**, *26*, 594−603.
(8) Mittal, K. L. *Polyimides: synthesis, characterization, and applications*; Springer Science & Business Media: 2013; Vol. *1*.
(9) Barikani, M.; Mehdipour-Ataei, S. Synthesis, characterization, and thermal properties of novel arylene sulfone ether polyimides and polyamides. *J. Polym. Sci., Part A: Polym. Chem.* **2000**, *38*, 1487−1492.
(10) Butnaru, I.; Bruma, M.; Kopnick, T.; Stumpe, J. Influence of chemical structure on the refractive index of imide-type polymers. *Macromol. Chem. Phys.* **2013**, *214*, 2454−2464.
(11) Liu, J.-g.; Ueda, M. High refractive index polymers: fundamental research and practical applications. *J. Mater. Chem.* **2009**, *19*, 8907−8919.
(12) Angione, M. D.; Pilolli, R.; Cotrone, S.; Magliulo, M.; Mallardi, A.; Palazzo, G.; Sabbatini, L.; Fine, D.; Dodabalapur, A.; Cioffi, N.; et al. Carbon based materials for electronic bio-sensing. *Mater. Today* **2011**, *14*, 424−433.
(13) Voigt, A.; Ostrzinski, U.; Pfeiffer, K.; Kim, J. Y.; Fakhfouri, V.; Brugger, J.; Gruetzner, G. New inks for the direct drop-on-demand fabrication of polymer lenses. *Microelectron. Eng.* **2011**, *88*, 2174−2179.
(14) Ummartyotin, S.; Juntaro, J.; Sain, M.; Manuspiya, H. Development of transparent bacterial cellulose nanocomposite film as substrate for flexible organic light emitting diode (OLED) display. *Ind. Crops Prod.* **2012**, *35*, 92−97.
(15) Xiang, C.; Ma, R. Devices to increase OLED output coupling efficiency with a high refractive index substrate. US Patent 9,640,781, 2017; https://www.google.com/patents/US9640781 (accessed Feb 2, 2019).
(16) Nishiyama, H.; Nishii, J.; Mizoshiri, M.; Hirata, Y. Microlens arrays of high-refractiveindex glass fabricated by femtosecond laser lithography. *Appl. Surf. Sci.* **2009**, *255*, 9750−9753.
(17) Kokubun, Y.; Funato, N.; Takizawa, M. Athermal waveguides for temperatureindependent lightwave devices. *IEEE Photonics Technol. Lett.* **1993**, *5*, 1297−1300.
(18) Wei, H.; Krishnaswamy, S. Direct laser writing polymer micro-resonators for refractive index sensors. *IEEE Photonics Technol. Lett.* **2016**, *28*, 2819−2822.
(19) Rodrıguez, A.; Vitrant, G.; Chollet, P.; Kajzar, F. Optical control of an integrated interferometer using a photochromic polymer. *Appl. Phys. Lett.* **2001**, *79*, 461−463.
(20) Singaravalu, S.; Mayo, D.; Park, H.; Schriver, K.; Haglund, R. Anti-reflective polymernanocomposite coatings fabricated by RIR-MAPLE. *Proc. SPIE* **2013**, 860718.
(21) Kim, J.-B.; Lee, J.-H.; Moon, C.-K.; Kim, S.-Y.; Kim, J.-J. Highly enhanced light extraction from surface plasmonic loss minimized organic light-emitting diodes. *Adv. Mater.* **2013**, *25*, 3571−3577.
(22) Kim, E.; Cho, H.; Kim, K.; Koh, T.-W.; Chung, J.; Lee, J.; Park, Y.; Yoo, S. A facile route to efficient, low-cost flexible organic light-emitting diodes: utilizing the high refractive index and built-in scattering properties of industrial-grade PEN substrates. *Adv. Mater.* **2015**, *27*, 1624−1631.
(23) Terraza, C. A.; Liu, J.-G.; Nakamura, Y.; Shibasaki, Y.; Ando, S.; Ueda, M. Synthesis and properties of highly refractive polyimides derived from fluorene-bridged sulfurcontaining dianhydrides and diamines. *J. Polym. Sci., Part A: Polym. Chem.* **2008**, *46*, 1510−1520.
(24) Carter, K. R.; DiPietro, R. A.; Sanchez, M. I.; Swanson, S. A. Nanoporous polyimides derived from highly fluorinated polyimide/poly (propylene oxide) copolymers. *Chem. Mater.* **2001**, *13*, 213−221.
(25) Yu, D.; Gharavi, A.; Yu, L. Novel aromatic polyimides for nonlinear optics. *J. Am. Chem. Soc.* **1995**, *117*, 11680−11686.

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

(26) Fukuzaki, N.; Higashihara, T.; Ando, S.; Ueda, M. Synthesis and characterization of highly refractive polyimides derived from thiophene-containing aromatic diamines and aromatic dianhydrides. *Macromolecules* **2010**, *43*, 1836−1843.
(27) Tsai, C.-L.; Yen, H.-J.; Liou, G.-S. Highly transparent polyimide hybrids for optoelectronic applications. *React. Funct. Polym.* **2016**, *108*, 2−30.
(28) Kobayashi, J.; Matsuura, T.; Hida, Y.; Sasaki, S.; Maruno, T. Fluorinated polyimide waveguides with low polarization-dependent loss and their applications to thermooptic switches. *J. Lightwave Technol.* **1998**, *16*, 1024.
(29) Sawada, T.; Ando, S. Synthesis, characterization, and optical properties of metalcontaining fluorinated polyimide films. *Chem. Mater.* **1998**, *10*, 3368−3378.
(30) Sydlik, S. A.; Chen, Z.; Swager, T. M. Triptycene polyimides: soluble polymers with high thermal stability and low refractive indices. *Macromolecules* **2011**, *44*, 976−980.
(31) You, N.-H.; Suzuki, Y.; Yorifuji, D.; Ando, S.; Ueda, M. Synthesis of high refractive index polyimides derived from 1, 6-bis (p-aminophenylsulfanyl)-3, 4, 8, 9-tetrahydro-2, 5, 7, 10-tetrathiaanthracene and aromatic dianhydrides. *Macromolecules* **2008**, *41*, 6361−6366.
(32) Liu, J.-g.; Nakamura, Y.; Shibasaki, Y.; Ando, S.; Ueda, M. Synthesis and characterization of high refractive index polyimides derived from 4, 4 [variant prime]-(p-phenylenedisulfanyl) dianiline and various aromatic tetracarboxylic dianhydrides. *Polym. J.* **2007**, *39*, 543.
(33) Yeo, H.; Lee, J.; Goh, M.; Ku, B.-C.; Sohn, H.; Ueda, M.; You, N.-H. Synthesis and characterization of high refractive index polyimides derived from 2, 5-Bis (4-Aminophenylenesulfanyl)-3, 4-Ethylenedithiothiophene and aromatic dianhydrides. *J. Polym. Sci., Part A: Polym. Chem.* **2015**, *53*, 944−950.
(34) National Science and Technology Council. *Materials Genome Initiative for Global Competitiveness*; 2011.
(35) Hachmann, J.; Windus, T. L.; McLean, J. A.; Allwardt, V.; Schrimpe-Rutledge, A. C.; Afzal, M. A. F.; Haghighatlari, M. *Framing the role of big data and modern data science in chemistry*; 2018; NSF CHE Workshop Report.
(36) Sánchez-Carrera, R. S.; Atahan, S.; Schrier, J.; Aspuru-Guzik, A. Theoretical characterization of the air-stable, high-mobility dinaphtho-[2,3-b:2′3′-f]thieno[3,2-b]-thiophene organic semiconductor. *J. Phys. Chem. C* **2010**, *114*, 2334−2340.
(37) Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C. B.; Zoombelt, A. P.; Bao, Z.; et al. From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nat. Commun.* **2011**, *2*, 437−438.
(38) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.
(39) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849−4861.
(40) Amador-Bedolla, C.; Olivares-Amaya, R.; Hachmann, J.; Aspuru-Guzik, A. In *Informatics for materials science and engineering: Data-driven discovery for accelerated experimentation and application*; Rajan, K., Ed.; Butterworth-Heinemann: Amsterdam, The Netherlands, 2013; Chapter 17, pp 423−442.
(41) Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Roman-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698−704.
(42) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What is high-throughput virtual screening? A perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **2015**, *45*, 195−216.
(43) Lopez, S. A.; Pyzer-Knapp, E. O.; Simm, G. N.; Lutzow, T.; Li, K.; Seress, L. R.; Hachmann, J.; Aspuru-Guzik, A. The Harvard organic photovoltaic dataset. *Sci. Data* **2016**, *3*, 160086.
(44) Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. *Mol. Simul.* **2018**, *44*, 921−929.
(45) Haghighatlari, M.; Hachmann, J. Advances of machine learning in molecular modeling and simulation. *Curr. Opin. Chem. Eng.* **2019**, *23*, 51.
(46) Afzal, M. A. F.; Cheng, C.; Hachmann, J. Combining *first-principles* and data modeling for the accurate prediction of the refractive index of organic polymers. *J. Chem. Phys.* **2018**, *148*, 241712.
(47) Afzal, M. A. F.; Hachmann, J. Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers. *Phys. Chem. Chem. Phys.* **2019**, *21*, 4452−4460.
(48) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(49) Weigend, F.; Ahlrichs, R.; Peterson, K. A.; Dunning, T. H.; Pitzer, R. M.; Bergner, A. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.
(50) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(51) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(52) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33−33.
(53) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73−78.
(54) Slonimskii, G.; Askadskii, A.; Kitaigorodskii, A. The packing of polymer molecules. *Polym. Sci. U.S.S.R.* **1970**, *12*, 556−577.
(55) Privalko, V. P.; Pedosenko, A. V. Molecular packing density in the crystalline state of semi-rigid chain polymers. I: Polyimides. *Polym. Eng. Sci.* **1997**, *37*, 978−982.
(56) Hachmann, J.; Evangelista, W. S.; Afzal, M. A. F. *ChemHTPS 0.7* − An automated virtual high-throughput screening program suite for chemical and materials data generation; 2017; https://bitbucket.org/hachmanngroup/chemhtps (accessed Feb 2, 2019).
(57) Hachmann, J.; Afzal, M. A. F. *ChemLG 0.5* − A library generator code for the enumeration of chemical and materials space; 2017; https://hachmannlab.github.io/chemlg (accessed Feb 2, 2019).
(58) Hachmann, J.; Haghighatlari, M. *ChemML 0.10* − A machine learning and informatics program suite for chemical and materials data mining; 2017; https://hachmannlab.github.io/chemml (accessed Feb 2, 2019).
(59) Kudo, H.; Yamamoto, M.; Nishikubo, T.; Moriya, O. Novel materials for large change in refractive index: Synthesis and photochemical reaction of the ladderlike poly(silsesquioxane) containing norbornadiene, azobenzene, and anthracene groups in the side chains. *Macromolecules* **2006**, *39*, 1759−1765.
(60) Afzal, M. A. F. From virtual high-throughput screening and machine learning to the discovery and rational design of polymers for optical applications. Ph.D. thesis, University at Buffalo, 2018.
(61) Krylov, A.; Windus, T. L.; Barnes, T.; Marin-Rimoldi, E.; Nash, J. A.; Pritchard, B.; Smith, D. G.; Altarawy, D.; Saxe, P.; Clementi, C.; et al. Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science. *J. Chem. Phys.* **2018**, *149*, 180901.

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618

(62) Wilkins-Diehr, N.; Crawford, T. D. NSF's inaugural software institutes: The science gateways community institute and the molecular sciences software institute. *Comput. Sci. Eng.* **2018**, *20*, 26−38.

DOI: 10.1021/acs.jpcc.9b01147
J. Phys. Chem. C 2019, 123, 14610−14618