FROM VIRTUAL HIGH-THROUGHPUT SCREENING AND MACHINE LEARNING TO THE DISCOVERY AND RATIONAL DESIGN OF POLYMERS FOR OPTICAL APPLICATIONS

by

Mohammad Atif Faiz Afzal

May 2018

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Doctor of Philosophy

Department of Chemical and Biological Engineering

The dissertation of Mohammad Atif Faiz Afzal was reviewed by the following:

Johannes Hachmannn
Assistant Professor of Chemical and Biological Engineering
Dissertation Advisor, Chair of Committee

Jeffrey R. Errington
Professor of Chemical and Biological Engineering
Committee Member

Edward P. Furlani
Professor of Chemical and Biological Engineering
Committee Member

Dedication

This dissertation is dedicated to my parents, Farhat Yasmeen and Mohammed Afzal, and my eldest brother Mohammed Asim Hafeez, for their love, constant support and encouragement.

Acknowledgments

I am deeply grateful to my advisor and mentor, Prof. Johannes Hachmann, for his continuous support, patience, and encouragement. His guidance helped me in all aspects of my career; from developing written and oral communication skills and networking, to coding and research practices. All of these have played a part in helping me write this dissertation. I could not have imagined having a better advisor to look up to for my Ph.D.

I would like to thank Prof. Jeffrey Errington and Prof. Edward Furlani for serving on my committee, and for their insightful comments and encouragement. Additionally, I am grateful to Prof. Michel Dupuis, Prof. Chong Cheng, and Dr. Andrew Schultz, for being the source of sound advice. They all have inspired me as great teachers and communicators of science.

In the course of my graduate studies and research, I was fortunate to work with a number of remarkable colleagues and friends: Mojtaba Haghighatlari, William Evangelista, Yudhajit Pal, Aditya Sonpal, Sai Prasad Ganesh, Vigneshwar Kumaran Sudalayandi Rajeswari, Shirish Shivaraj, and Edward Donowick. My thanks also go to the rest of the Hachmann Group, in particular, Ching-Yen Shih, Bryan Moore, Dana Havas, Gaurav Vishwakarma, and Yujie Tian. They

have all made my time in Buffalo cheerful and rewarding.

I thank my friends and fellow Ph.D candidates at UB, Nikhil Paliwal, Prakhar Jaiwal, Pavan Kumar Behera, and Karnesh Jain, for all the fun we had during our graduate studies. I would also like to thank Tamera Knight for her support and encouragement throughout my dissertation writing.

I offer a special thanks to Prof. Stelios Andreadis, Prof. Gary Dargush, Prof. Mark Swihart, Prof. Carl Lund, Dean Liesl Folks, Dr. Christine Human, and Christopher Conner for counseling me and for having faith in me. My Ph.D dissertation would not have been possible without their support.

Lastly, I would like to thank my family: my parents and my brothers for supporting me throughout my education and my life in general.

Table of Contents

Abstract

This dissertation is concerned with the application of materials discovery framework developed in our group to discover high-refractive-index polymers. Development and application of the framework includes four key parts.

In the first part, we present a method to accurately predict the refractive index (RI) of polymers using a combination of *first-principles* and data modeling. We validated the model with experimental RI values of polymers (Chapter 2). We further benchmark our results using different model chemistries to optimize the tradeoff between the accuracy and computation time (Chapter 3).

The second part covers the development of a molecular library generator (*ChemLG*) and a virtual high-throughput screening (*ChemHTPS*) infrastructure. We demonstrate the applicability of these software suites by providing examples (Chapter 4).

In the third part, we apply *ChemLG* and *ChemHTPS* to generate a library of polyimides and compute their RI values, respectively. Using the data generated in this work, we identify structure-property relationships *via* hypergeometric

distribution analysis (Chapter 5).

Finally, we present the application of machine learning to accelerate the process of property prediction. We construct efficient machine learning models to accurately predict the packing density, polarizability, and RI values of organic molecules and characterize them on a massive scale (Chapter 6).

Chapter 1

Introduction

1.1 Organic Polymers for Optical Applications

Organic polymers are emerging materials that feature many attractive properties compared to conventional inorganic materials. Devices made out of organic polymers are generally flexible, mechanically stable on impact, light-weight, and inexpensive to produce. This has led to increased efforts in utilizing these compounds in many different application domains, which include optic and optoelectronic devices such as organic light-emitting diodes [1], complementary metal oxide semiconductor [2], photovoltaics [3], field-effect transistors [4], displays, and image sensors [5]. In these devices, they can be introduced *in situ* as microlenses, waveguides, microresonators, interferometers, anti-reflective coatings, optical adhesives, and substrates (see Fig. 1.1). Most of these applications require materials with superior optical properties (such as high RI values). While typical carbon-based polymers exhibit poor optical properties when compared to inorganic materials [6], an important advantage of organic materials is

that their properties can be tuned readily and significantly by controlling their molecular structure. Thus, organic materials are a prime example for a rational design target.

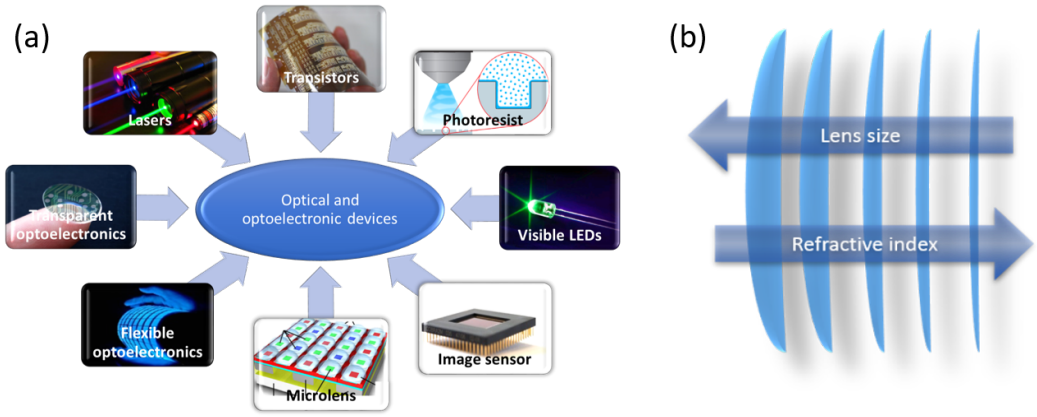


Figure 1.1. **(a)** Application range of organic polymers in optoelectronics. **(b)** Relationship of RI value and lens thickness.

The two principal challenges in creating new chemistry and materials are that their behavior is governed by complicated structure-property and structure-activity relationships [7, 8, 9], and that chemical space is practically infinite [10, 11, 12]. Traditional experiment-driven trial-and-error approaches are increasingly ill-equipped to meet these challenges on their own, in particular since advanced systems require more and more intricate property profiles [13, 14, 15]. Progress thus tends to be slow and incremental, in particular for advanced materials systems, which require more and more intricate property profiles. However, chemical and materials research has been undergoing a significant transformation in recent years that can alleviate many of these shortcomings: After decades of continuous advances in methods, algorithms, and computer hardware, the fields of modeling and simulation have reached a tipping point, and they are finally at a stage where they can make accurate

predictions for systems that are both realistic and relevant. Progress is now increasingly driven by computational studies, which have become crucial assets in the pursuit of next-generation materials and chemistry. By making guiding predictions, they can significantly boost the efficiency of research endeavors, and uncover promising targets for investigations in the laboratory (see, e.g., Ref. [16, 17, 3, 18, 19, 20, 21, 22]). The White House Materials Genome Initiative (MGI) [23] underscores the value of integrated joint ventures between experimentalists and theoreticians in tackling complex discovery and design challenges and delivering revolutionary new materials. That being said, the usual focus on individual compounds has so far been limiting the utility of computational research. While there is obvious value in characterizing particular systems of interest, the insights gained in these small-scale studies cannot easily be transferred or generalized.

Our group is developing a data-driven and rational design framework for accelerating the discovery and design of new chemicals/materials. In this dissertation, I deploy this framework to identify new polymer systems with superior optical properties, specifically the high-refractive-index polymers (HRIPs).

1.2 Data-Driven Design of Chemical Systems and the Exploration of Chemical Space

The shift towards a data-driven discovery and rational design paradigm (cf. Fig. 1.2) promises to mitigate many of the inefficiencies and shortcomings that are still prevalent in contemporary chemical research. There is now a growing agreement on the value of incorporating modern data science – the 4th pillar

of science – into chemical research, and this development has been recognized by high-profile funding programs such as the MGI [23]. Yet, despite impressive pioneering efforts (e.g., [24, 25, 26, 27, 28]), there is still a distinct disconnect between the promise of this approach and the realities of every-day research in the chemistry community, where data-driven work does not yet play a significant role.

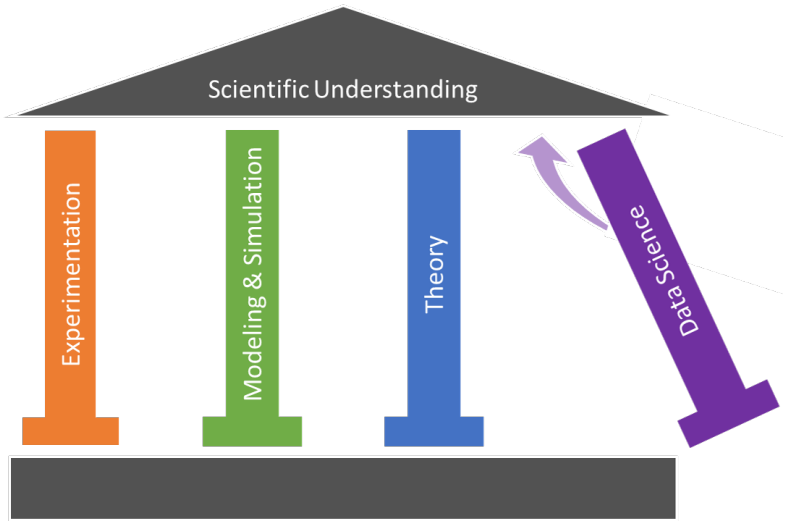


Figure 1.2. The rise of data science as the 4th pillar of science.

1.2.1 Group's Approach for Data-Driven *In Silico* Research

We have developed a basic template for data-driven *in silico* research, which addresses the inherent challenges in the discovery and design of new chemistry, in particular as part of integrated joint ventures with experimentalists. It provides the foundation and framework for our research program, and its rationale can be summarized as following:

- Using computational modeling and simulations, we can rapidly and efficiently assess the properties, behavior, and performance potential of

candidate compounds, materials, and/or chemical transformations for a given problem setting [18, 29, 30].

- By combining modeling and simulation with high-throughput screening techniques, we can characterize candidates on a massive scale. These studies naturally lead to big data scenarios [31, 3, 20, 32, 33].

- Using modern database technology, we can readily store and access the resulting data sets, e.g., to identify candidates with desired property combinations for on-demand applications [34, 35].

- In addition to the immediate information obtained for these thousands or even millions of candidates, we can mine the generated data in its entirety. Using machine learning, we can gain insights into the mechanisms that determine their characteristics and cast these findings into predictive models [36, 37, 38].

- By identifying these design rules as well as high-value moieties, building blocks, structural patterns, or more general features, we can accomplish the *de novo* design of next-generation candidates [39, 40, 41]. Using the predictive models, we can conduct hyperscreenings, i.e., screenings based on data-derived models that typically surpass the scale of the original screenings (based on physics-derived models) by several orders of magnitude.

- Experimentalist partners can pursue the top candidates from the (hyper-)screening and/or *de novo* design. This guidance allows the experimentalists to focus on highly promising targets and avoid wasted efforts on unpromising ones [17]. Additional in-depth modeling and simulations

can contribute further insights to the experimental findings, which allow for the advanced optimization of lead candidates.

- The experimental results can be included as training data in the machine learning approaches. In addition, they can be used to validate, benchmark, calibrate, and potentially improve the physics-based modeling and simulation protocols, which closes the design loop [42].

1.2.2 Software Ecosystem for Materials Discovery

Our group has identified four components that are critical for the efficient discovery of materials with targeted properties. These four components are being developed in the group as four different software packages (see Fig. 1.3):

1. ***ChemLG*: Screening library generator [43].** A prerequisite for the high-throughput exploration of chemical space is access to suitable, large-scale screening libraries. We have developed a corresponding generator for compound and material candidate libraries. We discuss in detail the underlying concepts of *ChemLG* in the Chapter 4 and provide case studies where *ChemLG* was successfully applied. I developed *ChemLG* primarily for the creation of high RI polymer candidates.

2. ***ChemHTPS*: Virtual high-throughput screening infrastructure [44].** *ChemHTPS* is a suite for performing large-scale simulations on high-performance computing clusters. This suite supports various quantum chemistry and molecular modeling codes including Q-Chem, ORCA, GROMACS, and LAMMPS. I co-developed *ChemHTPS* infrastructure along with William Evangelista.

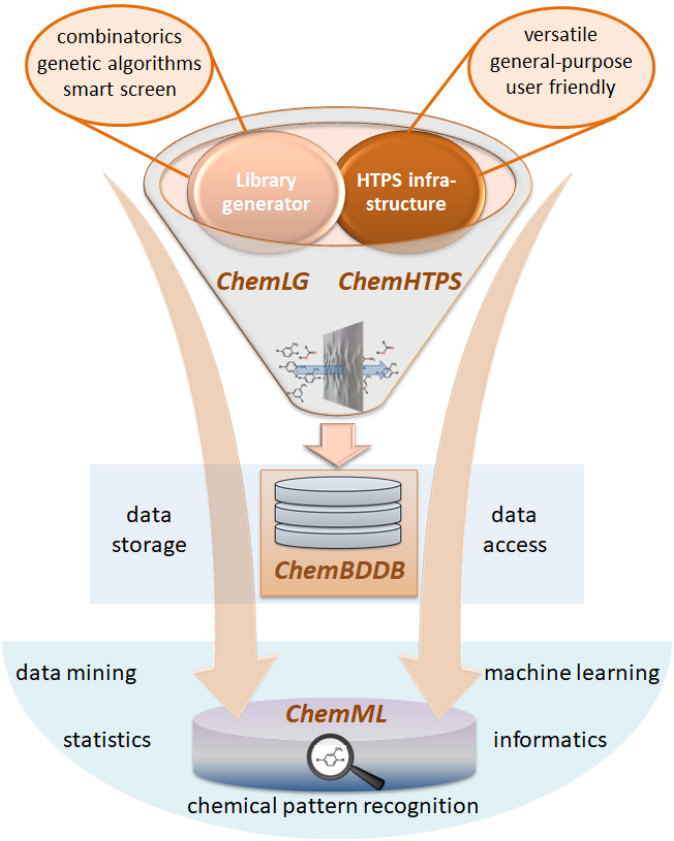


Figure 1.3. Schematic and connectivity of the software ecosystem comprised of the *ChemLG*, *ChemHTPS*, *ChemBDDB*, and *ChemML* codes.

3. ***ChemBDDB*: Database infrastructure [45].** The use of modern databases is of particular importance in the context of data-intensive research. Despite their great utility and despite being essential for projects that accumulate large data sets, they are still rarely employed in chemical research. Our group has developed a database platform to store and provide access to the various results in a centralized fashion. The database also serves as the focal point for the information exchange within the project team and potential external collaborators, as well as for the coordination of the different project components. The developers of *ChemBDDB* are Aditya Sonpal and Shirish Sivaraj.

4. ***ChemML*: Data analysis, mining, and modeling infrastructure [46].** *ChemML* suite includes data analysis, mining, and modeling capabilities that allow us to apply state-of-the-art machine learning and informatics methodology to chemical and materials data sets. Using *ChemML*, we can identify underlying structure-property relationships which are key to create new and more targeted candidate libraries. The primary developer of *ChemML* is Mojtaba Haghighatlari.

1.2.3 Discovery of High-Refractive-Index Polymers

Harnessing the potential of organic polymers in optical and optoelectronic industries has been quite limited because these applications require materials to have RI greater than 1.7, while carbon-based polymers inherently have RIs between 1.3 and 1.5. As a result, there is an incentive to discover new high-refractive index polymers for the aforementioned applications. According to the Lorentz-Lorenz equation, the incorporation of substituents with a high molar refractivity and low molar volume can increase the RI values of these materials. This suggests that the ability to tailor the molecular structure of polymers is the key to increasing their RI values [6]. On the other hand, tailoring the structure using different functional groups could potentially lead to an infinite number of molecular candidates. It is impractical to empirically characterize a large number of candidates, whereas computational analysis allows greater exploration at a mere fraction of the time and cost. The current work is concerned with creating fast and accurate predictive models for the optical properties of organic polymers, which will guide our experimentalist partners and allow them to target the most promising candidates. The application of above mentioned

rational design framework in the current work has allowed us to accelerate the discovery of polymers with high RI values materials.

We have developed an accurate model for the prediction of RI values of polymers based on *first-principles* and molecular modeling techniques. The model is based on the Lorentz-Lorenz equation and thus includes the calculation of polarizability and number density values for the candidate structures. In this scheme, we compute the molecular polarizability using *first-principles* electronic structure theory (DFT) and the number density using molecular modeling (MD). The synergistic combination of DFT and MD resulted in a successful and economical model for RI prediction. We validated the RI model using experimental RI values of 112 non-conjugated polymers, which shows that the model is in a good agreement (R^2=0.94) with the experimental results [47]. Further details of the model and its validation will be discussed in Chapter 2.

In the RI prediction model, we apply computationally expensive DFT computations. While the choice proved remarkably successful, it represents the bottleneck step in our RI protocol due to computational inefficiency. It thus limits the utility of the overall approach, particularly in the context of virtual high-throughput screenings of large-scale candidate libraries. Therefore, we have systematically benchmarked several DFT model chemistries to find one that optimizes the balance between accuracy and efficiency in the target compounds space. We further compare the results for non-conjugated and conjugated polymers, offer guidance for method selection, and analyze the errors that propagate into the RI predictions. We will discuss the details of the benchmarking study in Chapter 3.

In search of candidates with high RI values, we created a library of poly-

mers using the *ChemLG* software. In this work, we selected polyimides as a model polymer and created a library of novel polyimides based on the building blocks suggested by our experimental collaborators. We provide an exposition of *ChemLG* software package in chapter 4. The resultant library of polyimides consists of 270,000 candidates. Application of our rational design framework resulted in more than 2000 polyimides with RI values greater than 1.8. The results from the study are presented in Chapter 5.

A large-scale exploration of chemical space will not only identify exceptional molecular targets but also allow us to identify underlying patterns and global trends. These structure-property relationships can aid in the inverse design of materials with specific properties. To achieve this, we created a massive library of 1.5 million small organic molecules and evaluated their properties. Characterization of candidates on such a large scale was possible by the application of advance machine learning techniques. Using the big data obtained form this work, we identified design rules as well as high-value building blocks, and structural patterns that correlate with optical properties. Additionally, we uncovered regions in chemical space where we can maximize the optical properties of organic molecules. These guidelines allow us to target specific molecular motifs and create the next generation of materials with exceptional properties. We present the results from this large-scale exploration in Chapter 6.

This dissertation demonstrates that the developed rational design framework is a powerful tool. The exploration of high-refractive-index polymers is one example for applications problems and I spearheaded this project as part of my dissertation.

Chapter 2

Accurate Prediction of the Refractive Index of Organic Polymers

We present an efficient computational protocol for the accurate prediction of RI values in polymers to facilitate *in silico* studies that can guide the discovery and design of next-generation high-RI materials. Our protocol is based on the Lorentz-Lorenz equation and is parametrized by the polarizability and number density values of a given candidate compound. In the proposed scheme, we compute the former using *first-principles* electronic structure theory and the latter using an approximation based on van der Waals volumes. The critical parameter in the number density approximation is the packing fraction of the bulk polymer, for which we have devised a machine learning model. We demonstrate the performance of the proposed RI protocol by testing its predictions against the experimentally known RI values of 112 optical polymers. Our approach to combine *first-principles* and data modeling emerges as both a success-

ful and highly economical path to determining the RI values for a wide range of organic polymers.

We thank Prof. Chong Cheng for helpful discussions on the scope of high RI polymers and synthetic feasibility of new polymers. The results of this study were published in M. A. F. Afzal, C. Cheng, J. Hachmann, J. Chem. Phys. 128 (2018), 144101 [47], and this chapter is based on our exposition in this paper.

2.1 Introduction

Organic small molecules, oligomers, and polymers are emerging materials and of significant interest for numerous fields of application due to their unique or otherwise desirable properties [48]. Unlike most conventional inorganic materials, they are generally flexible, light-weight, mechanically stable on impact, easy to process, and inexpensive to produce [49, 50]. Perhaps most importantly, their properties can be tailored towards specific demands by controlling their molecular structure [6]. A particular area of interest is the application of organic materials in optic and optoelectronic devices [51], such as (image) sensors [5, 52], displays [53], and light sources (including organic light-emitting diodes) [54], in which they can be introduced *in situ* as microlenses [55], waveguides [56], microresonators [57], interferometers [58], anti-reflective coatings [59], optical adhesives [60], and substrates [61]. Some of the optical properties that are relevant for these applications are the refractive index (RI), Abbe number, birefringence, absorption spectrum, and color [62].

The RI value dictates the shape and size of many optical components, in particular those with lens function. Most of the aforementioned applications require materials with large RI values (i.e., larger than 1.7), and there are several

applications that require very large ones (i.e., larger than 1.8) [63]. Unfortunately, the vast majority of organic polymers only offer RI values ranging from 1.3 to 1.5 [6] (compared to inorganic materials, which can feature values up to $\sim$4). The development of high-RI polymers has thus gained attention, and several approaches have been proposed to overcome the RI-value limitations of typical organic polymers. They include the notion to incorporate highly polarizable moieties, such as rigid aromatic fragments [64], heteroatoms [65, 66], or organometallics [67], into the polymer scaffold. Another strategy that has been pursued is to reinforce the polymer matrix with metal alkoxides (e.g., TiO_2, Fe_3O_4) [68, 69] or other high-RI molecules (e.g., ZnS, diamondoids) [70, 71]. While these approaches have resulted in a few systems with RI values between 1.6 and 1.8, most of them are of limited utility for practical applications due to a variety of materials, processability, or preparation issues [6]. Increasing the RI values of organic polymers beyond 1.8 has remained a completely elusive task and continues to be an important challenge in synthetic chemistry [65].

The traditional, experimentally focused discovery process for new materials is very time-, labor-, and resource-intensive, which limits the number and diversity of candidate compounds that can be explored. Progress thus tends to be slow and incremental, in particular for advanced materials systems, which require more and more intricate property profiles. However, chemical and materials research has been undergoing a significant transformation in recent years that can alleviate many of these shortcomings: After decades of continuous advances in methods, algorithms, and computer hardware, the fields of modeling and simulation have reached a tipping point, and they are finally at a stage where they can make accurate predictions for systems that are both realistic and relevant. Progress is now increasingly driven by computational studies,

which have become crucial assets in the pursuit of next-generation materials and chemistry. By making guiding predictions, they can significantly boost the efficiency of research endeavors, and uncover promising targets for investigations in the laboratory (see, e.g., Ref. [16, 17, 3, 18, 19, 20, 21, 22]). The White House Materials Genome Initiative [23] underscores the value of integrated joint ventures between experimentalists and theoreticians in tackling complex discovery and design challenges and delivering revolutionary new materials. A prominent example in the context of optical materials is the work by Ramprasad and co-workers [72, 73, 74], which we will use as a reference point.

A prerequisite for the computationally-driven development of new materials is access to suitable (i.e., accurate and efficient) computational protocols for the target property within a compound space of interest. This paper presents such a protocol for the prediction of RI values of organic polymers. One of the distinctive feature of this protocol compared to prior work by others [72, 73, 75, 76] is that it fuses *first-principles* and data modeling.

In Sec. 3.2, we introduce the physical foundations of the proposed protocol (Sec. 2.2.1), motivate a number of assumptions and approximations that are used (Sec. 2.2.2 and 2.2.3), and discuss the details of the employed computational approach (Sec. 2.2.4). In Sec. 3.3, we present and discuss results for the different components that comprise the protocol (Sec. 2.3.1, 2.3.2, and 2.3.3) as well as the overall protocol itself (Sec. 2.3.4). In each case, we evaluate the predictive performance of our model by comparing its results with data from a validation set of experimentally known compounds. Sec. 2.3.5 provides a discussion of the interplay between the physical parameters of our model. Our findings are summarized in Sec. 2.4.

2.2 Background and Methods

2.2.1 Lorentz-Lorenz Equation

The RI value (n_r) is defined as the ratio between the speed of light in vacuum (c_0) and in a given material (c). For non-magnetic materials, the RI is thus the square root of its relative permittivity or dielectric constant (ϵ_r), i.e.,

$$n_r = \frac{c_0}{c} = \sqrt{\epsilon_r}.$$

The permittivity is a function of the polarizability (α) and using the Lorentz local field approximation, it can be written as

$$\epsilon_r = \frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0},$$

where N is the number density, i.e., the number of molecules per volume. It follows that the RI is

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

which is a version of the Lorentz-Lorenz equation (equivalent to the Clausius-Mossotti relation). It follows, that the Lorentz-Lorenz equation connects the macroscopic RI value of a bulk material to the electronic polarizability α and number density N of its molecular constituents. The Lorentz-Lorenz equation thus offers a route to calculating the RI value of a material *via* α and N, and we use it as the physical basis for the proposed computational protocol (for a detailed discussion of this approach see App. A).

2.2.2 *First-Principles* Molecular Polarizability Calculations

The polarizability α of a compound can be obtained from quantum chemical linear response calculations. An array of electronic structure methods has been used to determine the polarizability values of various materials [77, 78, 79, 80, 81, 82, 83], including organic polymers [84, 85, 86]. Polarizability is generally a frequency-dependent (i.e., dynamic) property as shown in Fig. 2.1. Frequency-dependent polarizability is relatively hard to compute, as it formally involves solving the time-dependent Schrödinger equation and/or scanning through the range of relevant frequencies [87]. Consequently, only relatively few studies consider the polarizability dispersion in organic polymers [88, 89].

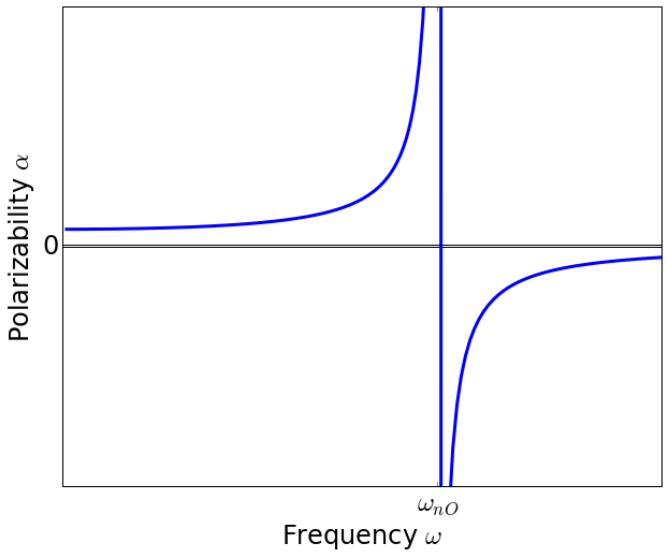


Figure 2.1. Dispersion characteristic of the polarizability: Excited states are marked by singularities, and in the frequency range below, the polarizability converges asymptotically towards the constant static polarizability value.

From the frequency dependence of the polarizability follows that the RI is also a frequency-dependent property. However, the variation of both polarizability and RI in the visible frequency region is in fact often relatively small [71], as long as low-lying excited states are absent. Since the latter would render a

material unsuitable for optical applications anyways, we generally do not consider materials that exhibit them. Large variations can be observed in the ultraviolet region where resonances with the excited state manifold become a dominant feature, but since stability considerations prohibit organic polymers to be used for high-energy applications, this is also not a relevant concern. Below the frequency range of the excited states, the polarizability and RI value taper off monotonically (cf. Fig. 2.1). In most cases, they become essentially constant throughout the visible and infrared range and converge to an asymptotic value [71]. The asymptotic RI corresponds to the value that can be obtained from the static polarizability. The latter can be computed much more easily than the frequency-dependent value. It only requires a single linear response calculation without explicit time dependence, and is thus much less demanding in terms of computing time and numerical stability. We can conclude that the RI values obtained from static polarizability calculations form a close lower bound for the frequency-dependent values in the relevant spectral range. This approach has been used in the past and has given very good agreement with experimental results [90, 84, 91, 76, 85].

Another challenge is to perform the polarizability calculations for quasi-infinite polymers. Realistic systems are amorphous and may thus not be well represented by periodic boundary condition calculations, while non-periodic calculations on long-chain oligomer models are generally cost-prohibitive. However, for systems with a relatively short correlation length (e.g., due to finite conjugation and delocalization of the π-electron backbone), we can expect an early onset of extensivity in the optical properties. We can exploit this behavior through an extrapolation scheme. In this scheme, we perform a series of relatively simple monomer and small oligomer calculations until we observe

a linear trend in the polarizability results, based on which we can project to the polymer limit.

The molecular polarizability calculations in the proposed protocol utilize Kohn-Sham density functional theory (DFT) for its advantageous trade-off of cost and accuracy [92]. The former includes its low-order polynomial scaling with system size and its relatively modest basis set demands (compared to high-level wavefunction methods). Given the molecular-level disorder in amorphous polymers, we forgo the expensive *first-principles* optimization of idealized geometries of our candidate compounds in favor of an inexpensive molecular mechanics approach. A simple, yet efficient way to identify (and exclude) compounds with potentially low-lying excited states is to assess the gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). The HOMO–LUMO gap is a first approximation for the lowest excitation, and it is readily obtained in DFT at no additional cost.

2.2.3 Data Model for the Number Density

The number density N for amorphous polymers is typically computed using classical molecular dynamics simulations. However, this approach is relatively cumbersome and computationally demanding. As an alternative, we pursue an approach based on the molecular volume (approximated by the van der Waals volume V_{vdW}), i.e.,

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

where K_p is the packing fraction in the bulk polymer, which has shown good agreement with experimental results in other work [93, 94]. There are a number of ways to compute the van der Waals volume, ranging from complicated elec-

tronic structure calculations with subsequent partitioning of the electron density to simplistic fragment methods [95, 96]. A benchmark study that we will detail elsewhere has shown that the differences in results from different methods are generally small. For the present work, we thus adopt the latter, i.e., we calculate V_{vdW} by adding tabulated atomic values [97] and subtracting off the overlap in the bonding region. The average packing fraction K_p for organic polymers is given in the literature as 0.68 [98], however, the actual values of different polymers show a significant spread and are known to range at least from 0.5 to 0.8. (K_p is generally also a function of the degree of polymerization, but except for shorter oligomers, this only plays a minor role.) As the average value of this critical parameter is thus essentially meaningless, we have devised a machine learning model to correlate the polymer structure with its packing fraction. Due to the relatively small volume of available training data, we chose a comparatively inflexible (but exceedingly fast) support vector regression (SVR) approach to avoid overfitting. Modern support vector machines were introduced in 1992 for supervised classification problems and have been a popular machine learning technique since their inception [99]. A version for regression analysis was added in 1996 [100, 101]. Non-linear SVR prediction models are generated by projecting the training data into a high-dimensional kernel-induced feature space where they become linear regression problems subject to cost functions that penalize prediction errors.

2.2.4 Computational Details

The polarizability calculations of the proposed protocol use an all-electron, restricted DFT framework with the PBE0 hybrid functional [102] in combination

with the triple-ζ quality def2-TZVP basis set by the Karlsruhe group [103]. We include Grimme's D3 correction [104] to account for dispersion interaction. The proof-of-principle study shown in the following section was carried out using the ORCA 3.0.2 quantum chemistry program package [105] with default settings. We optimized the geometries of all monomers and oligomers using the universal force field (UFF) [106] as implemented in the OpenBabel software [107]. We calculated the van der Waals volumes using Slonimskii's method detailed in Ref. [108], for which we implemented a Python script. We generated the packing fraction model using SVR within a feature space of 43 constitutional descriptors on a training data set of 84 polymers with experimentally known K_p values compiled from the literature. The available data was divided into 80% training and 20% test set for cross-validation. The data modeling was performed using *ChemML* [46], our program suite for machine learning and informatics in chemical and materials research. In this work, *ChemML* employed the scikit-learn 0.17 SVR library [109] and descriptors from Dragon 7 [110]. The proof-of-principle study involved about 450 individual calculations, which we performed using *ChemHTPS* [44], our program suite for automated virtual high-throughput screening in chemical and materials research.

2.3 Results and Discussion

We developed the proposed RI protocol on two common non-conjugated polymers – polyethylene (PE) and polystyrene (PS) – as prototype systems. Subsequently, we performed a study of 112 non-conjugated polymers for which experimental RI values are known in order to validate the predictive performance of the RI protocol as well as its individual components.

2.3.1 Polarizabilities

The PBE0/def2-TZVP-D3 polarizability results for PE and PS from monomer to heptamer are shown in Fig. 2.2. The linear trend with respect to the number of monomer units n (due to extensivity) is easily recognized. The correlation coefficient R^2 for the linear regression is $\gg 0.99$. For all cases studied in this work, extensivity was observed for very short oligomer sequences, and we based our extrapolation scheme on the linear regression slope obtained from the monomer to tetramer results.

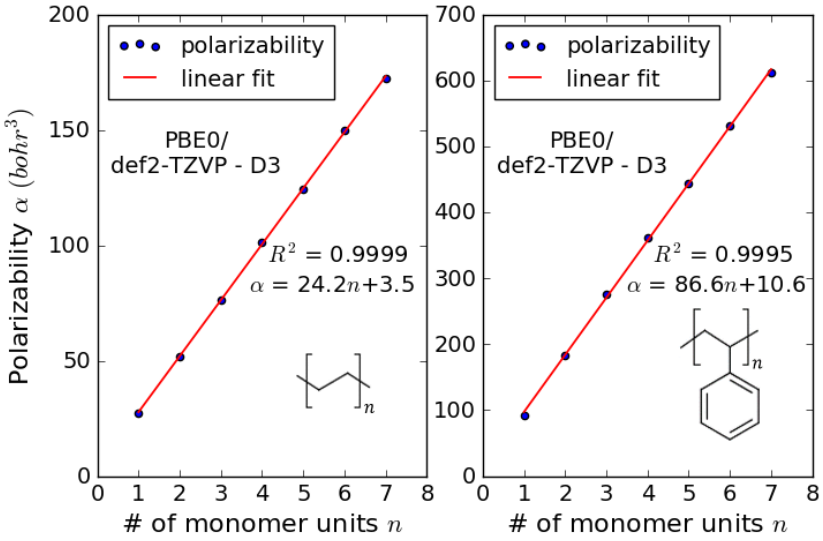


Figure 2.2. Linear relationship between number of monomer units and polarizability for polyethylene (PE) and polystyrene (PS) as prototypes for non-conjugated polymers.

Note that for conjugated polymers with longer correlation lengths, the onset of extensivity can occur at significantly longer chain length, i.e., values for a sequence of shorter oligomers will not show a linear trend [111, 112]. An example is given in Fig. 2.3. The extrapolation scheme can still be used in these cases, but it requires the calculation of longer oligomer sequences until a linear trend for extrapolation is found.

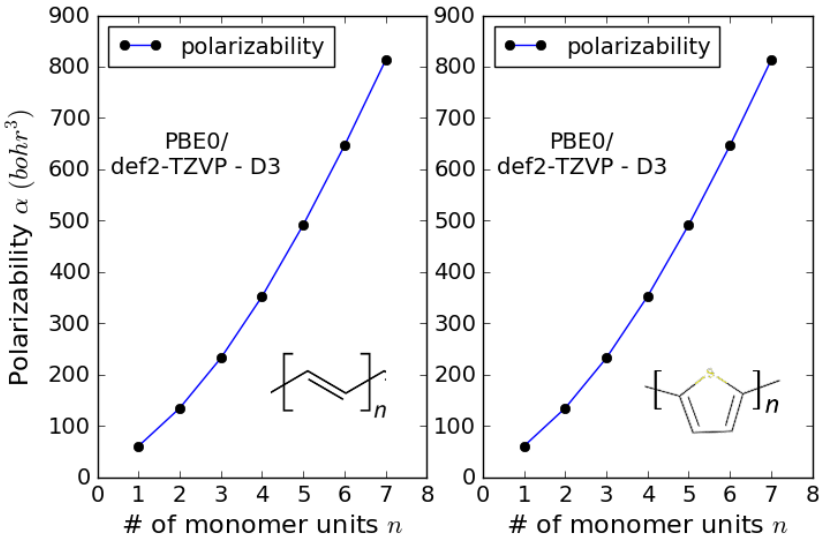


Figure 2.3. Non-linear relationship between number of monomer units and polarizability for polyacetylene and polythiophene as prototypes for conjugated polymers.

2.3.2 Number Densities and Densities

Using Slonimskii's method we could readily compute the van der Waals volumes V_{vdW} for the prototype systems PE and PS. Assuming the average packing fraction of $K_p = 0.68$, we obtain the number density values N as a function of the number of monomer units n shown in Fig. 2.4. The plots illustrate that N decreases monotonically with increasing number of monomer units, and the inverse $1/N \propto V_{vdW}$ is evidently extensive.

Note that we can use this approach to compute another property of interest in organic materials research, i.e., the density ρ of amorphous polymers in the bulk (*via* $\rho = N \cdot N_A / M$ with the Avogadro number N_A and molecular weight M). The density results for our prototype systems are presented in Fig. 2.5. The plots show that ρ increases and ultimately convergences towards asymptotically constant values. This finite size effect due to the terminal groups is typically of limited magnitude. The results of the oligomers with $n = 50$ offer a

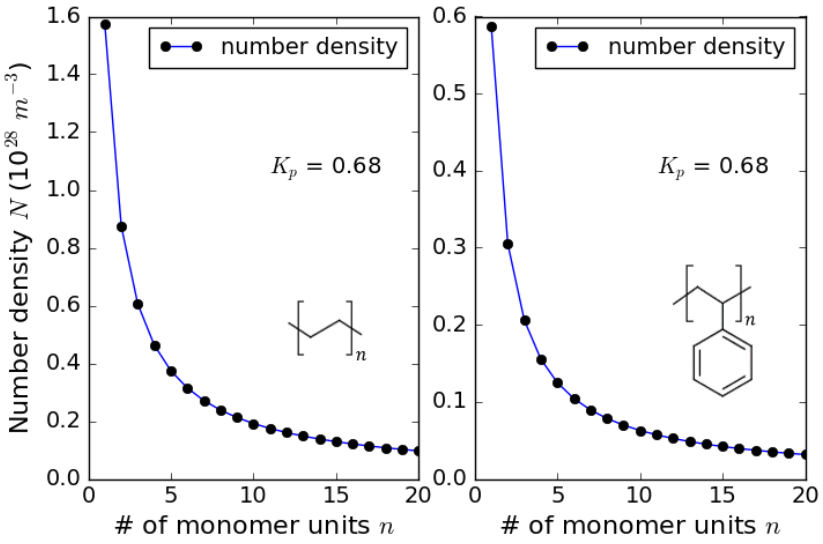


Figure 2.4. Change of number density values for increasing degree of polymerization for the PE and PS prototype systems.

good representation of the polymer limit and can thus be used as the default for determining ρ. The resulting densities are in very good agreement with experimentally known values [113]. We can also use ρ to work backwards and obtain the actual K_p values. For PE and PS we obtain 0.64 and 0.66, respectively, i.e., using the average packing fraction of 0.68 happened to be a valid assumption in these particular cases.

2.3.3 Packing Fractions

As K_p is generally not known and the average value referenced in the literature [98] is of limited utility, we have devised an SVR data model that correlates the polymer structure with the packing fraction as outlined in Sec. 2.2.3 and 2.2.4. Fig. 2.6 displays the range and distribution of K_p values for the 84 polymers for which we found experimental results. This data – ranging from 0.53 to 0.79 with an average value of 0.67 – formed the basis for our data-derived K_p prediction

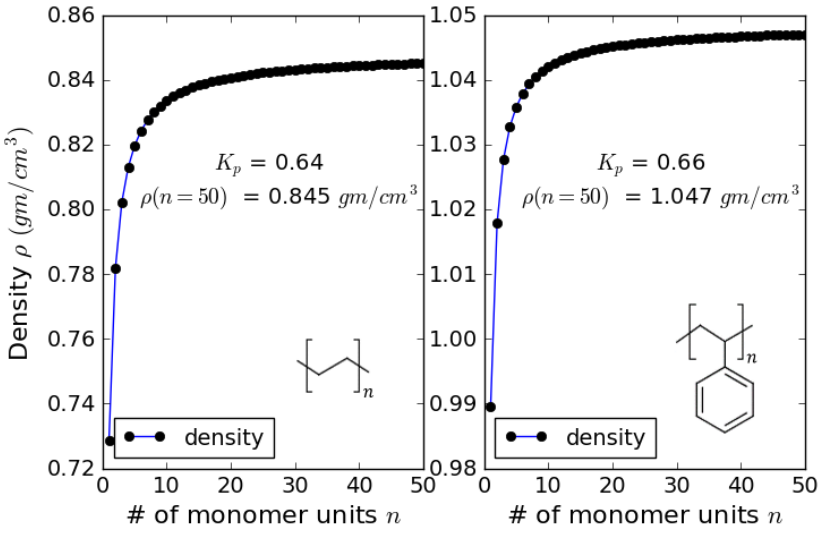


Figure 2.5. Change of density for increasing degree of polymerization for the PE and PS prototype systems. Note the characteristic asymptotic convergence to a constant value.

model. (Note that the average K_p for our data set is nearly identical with the average $K_p = 0.68$ cited in Ref. [98]).

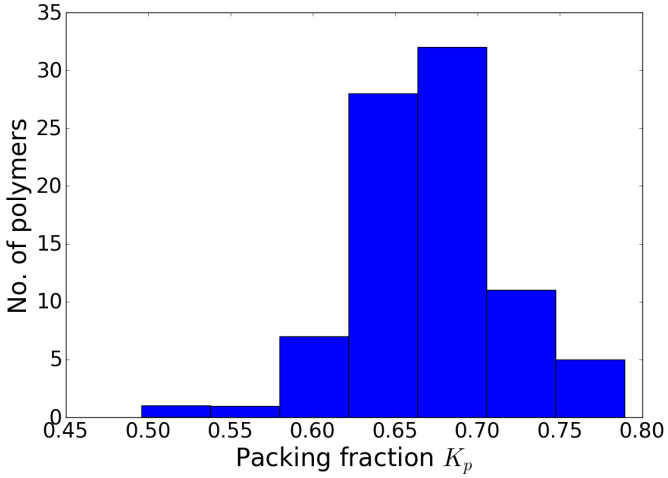


Figure 2.6. Range and distribution of experimental packing fraction values for 84 polymers used in the creation of our data-derived prediction model.

The model gives an R^2 of 0.97 for the training and 0.87 for the test set. The performance drop for the latter is reasonable and acceptable given the small size of the available data set. The computational demand for the K_p prediction

model is negligible and results for even large-scale compound libraries can be obtained in minutes on a single processor.

2.3.4 Refractive Indices

Given the modeling protocols and resulting data for α, V_{vdW}, K_p, and N, we use the Lorentz-Lorenz equation to make RI predictions. Using the α and N values obtained for our PE and PS prototype systems as shown in Figs. 2.2 and 2.4, we calculate the RI values and their variation with the number of monomer units n given in Fig. 2.7. The RI increases for longer oligomers before reaching a plateau for $n = 20$ to 30.

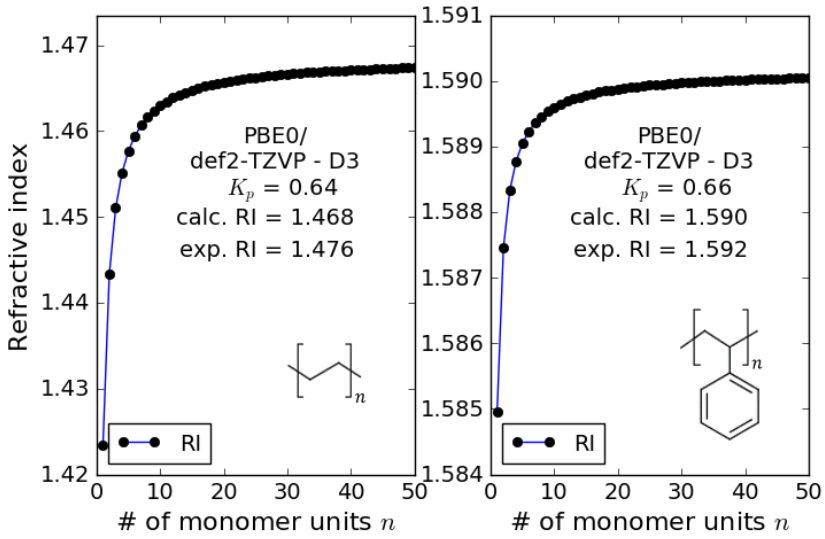


Figure 2.7. Change of refractive index (RI) with increasing degree of polymerization for the PE and PS prototype systems. Note the characteristic asymptotic convergence to the constant value at the polymer limit, that is in excellent agreement with the experimental data.

Our modeling protocol predicts the RI values for PE and PS to be 1.468 and 1.590, respectively, which is in outstanding agreement with the experimental RI values of 1.476 and 1.592, respectively [113]. We further validate our modeling

protocol by predicting the RI values of 112 polymers for which we could find experimental data for comparison (see Fig. 2.8).

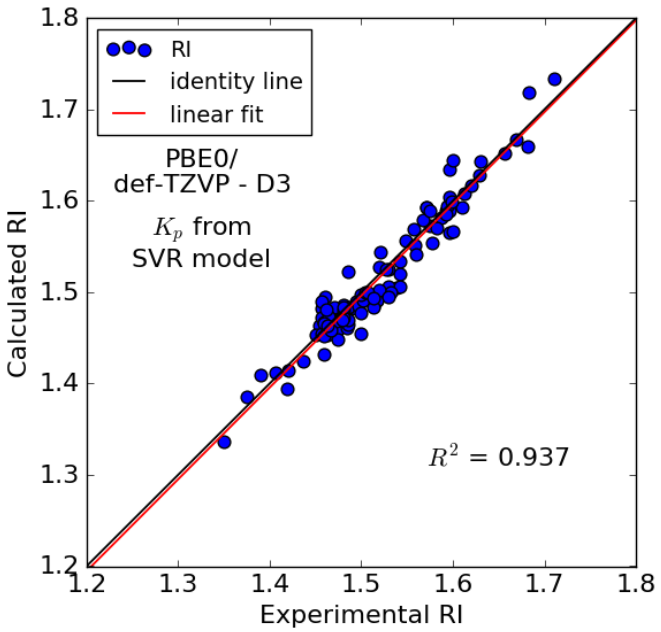


Figure 2.8. Validation of the proposed RI prediction model (based on the data-derived model for the packing fraction) through comparison with 112 experimental data points.

The R^2 of 0.94 shows that the model is in very good agreement with the experimental RI values. The benchmark comparison gives a mean absolute deviation (MAD) of 0.010 (0.9%), a root mean square deviation (RMSD) of 0.018 (0.1%), and a maximum deviation (MaxD) of 0.045 or 3.0%, respectively, i.e., our modeling protocol is quite accurate and affords at least semi-quantitative predictions (in particular since typically only two decimals in the RI values are considered as significant). The average deviation (AD) is very small with +0.004 (+0.3%), i.e., our model is not significantly biased towards systematic over- or under-predictions. A result of particular importance for studies that focus on candidate rankings rather than quantitative predictions for individual candidates is that the trends in the data are generally well captured. We stress that

the experimental RI data is independent of the data used for the creation of the SVR model for N, i.e., the SVR model was not biased towards providing good RI results.

For comparison, using the average packing fraction value of 0.68 – as is oftentimes cited in related work – instead of our SVR model leads to the results shown in Fig. 2.9. This model is considerably worse, as can be seen from the R^2 of 0.78, MAD of 0.019 (1.9%), RMSD of 0.026 (0.3%), MaxD of 0.139 and 6.9%, and AD of -0.009 (-0.6%).

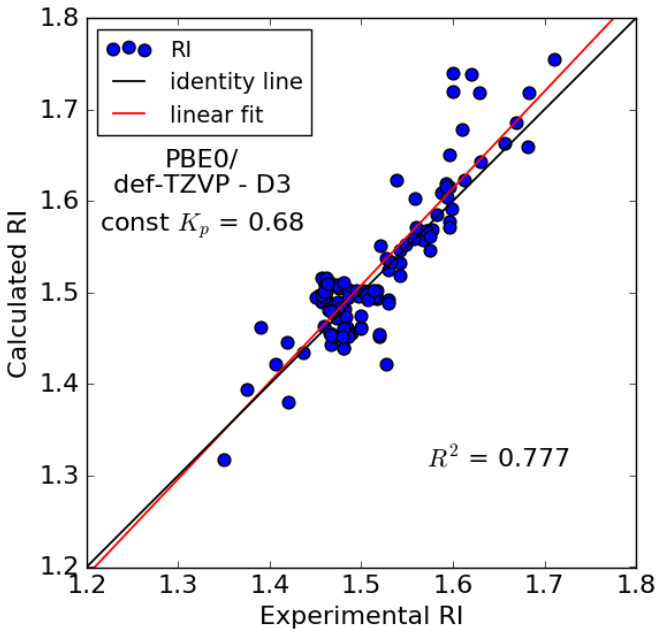


Figure 2.9. Validation of an RI prediction model based on a constant average packing fraction.

The before-mentioned pioneering work by Ramprasad *et al.* offers another instructive reference point. It creates a machine learning prediction model for the dielectric constant ($\epsilon_r = {n_r}^2$) of organic and organometallic compounds that is based on training data from plane-wave DFT for crystalline systems [72, 73]. Ramprasads database includes 7 experimental data points that can be used for

the validation of the underlying DFT model [74]. The DFT predictions show a correlation of only $R^2 = 0.78$ with respect to the experimental data. This suggests that the use of DFT crystal structures may be the cause of some loss in predictive performance, which can propagate into the derived machine learning models. This observation underscores the importance of accurately accounting for the amorphous bulk structure, which our approach achieves.

It is worth stressing that while our discussion focused on non-conjugate polymer examples, our computational protocol can also be employed on conjugate systems provided the extensive regime is considered in the extrapolation schemes. The SVR prediction model for K_p was trained on typical organic polymers and may start to fail for unusual or extreme cases that were not represented in the training set. In such a situation it would be advisable to retrain the model with more targeted training data.

2.3.5 Interplay between Polarizability and Number Density

As the Lorentz-Lorenz equation relies on α and N as input parameters, we analyze their interplay for the 112 polymers in our validation and benchmark data set. Fig. 2.10 shows the calculated α and N values as well as the contour lines for the resulting RI values in this parameter space.

To achieve high RI values, a candidate compound must feature both large number density and polarizability values (as is also apparent from the structure of the Lorentz-Lorenz equation). Optimizing both properties simultaneously is a challenging task as extensivity couples α and N, i.e., the longer a polymer, the larger α (as α increases with the number of contributing monomer units), but the smaller N (as fewer molecules fit into a given volume element). A design

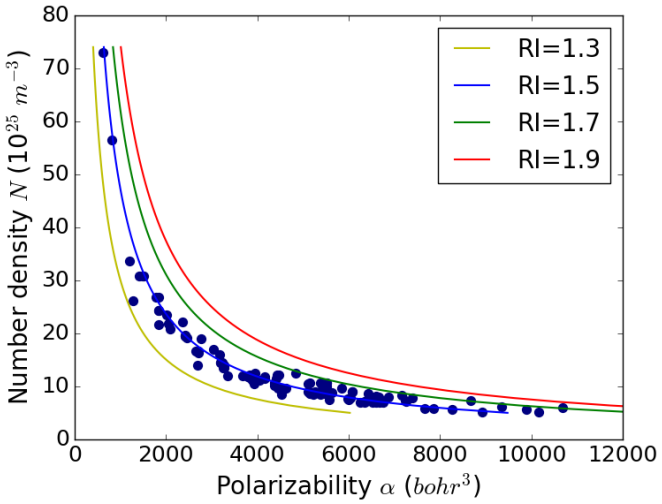


Figure 2.10. Parameter space of polarizability α and number density N as well as resulting RI value domains with examples from our validation and benchmark data.

strategy that can be derived from this notion is to incorporate highly polarizable moieties that have a limited effect on the number density. The data set at hand is tilted towards a larger spread in α while N is more clustered. It is worth noting that the RI regions in the α *vs* N parameter space are relatively narrow and most compounds in the data set group around the contour line for RI=1.5. The high-RI examples primarily stand out for large polarizability values rather than large number densities, which supports the before-mentioned design strategy.

2.4 Conclusions

We have successfully developed a modeling protocol for the accurate prediction of the RI values of organic polymeric materials and validated it against the experimentally known RI values of 112 compounds. The current work is an example for the benefits of fusing physical and data models, with the former providing the general structure of the approach (i.e., the Lorentz-Lorenz equation)

and a significant part of the required input parameters (i.e., polarizabilities from *first-principles* quantum chemistry and van der Waals volumes from Slonimskii's method), while the latter provides rapid access to input data that is otherwise not readily available (i.e., an SVR model for the packing fraction and an extrapolation scheme for molecular results towards the polymer limit). A subset of our RI protocol can also be used to predict the density of amorphous polymers. Our work is furthermore an example for the great promise of applying machine learning and modern data science in chemical research. In later chapters, we will utilize this new RI protocol to conduct virtual high-throughput screening studies on large-scale candidate libraries with the goal of accelerating the discovery of novel organic materials with unprecedented RI values.

Chapter 3

Benchmarking DFT Approaches for the Calculation of Polarizability Inputs for Refractive Index Predictions in Organic Polymers

In chapter 2, we introduced a computational protocol to accurately predict the index of refraction (RI) of organic polymers using a combination of *first-principles* and data modeling. This protocol is based on the Lorentz-Lorenz equation and involves the calculation of static polarizabilities and number densities at the polymer limit. We chose to compute the former within the density functional theory (DFT) framework using the PBE0 functional and def2-TZVP basis set along with D3 dispersion correction. While this choice proved remarkably successful, it is also relatively expensive from a computational perspective and represents the bottleneck step in our RI modeling protocol. It thus limits the utility of the overall approach, in particular in the

context of virtual high-throughput screenings of large-scale candidate libraries where efficiency is essential. In the work presented here, we systematically benchmark DFT model chemistries to identify approaches that optimize the balance between accuracy and efficiency for this target application. We compare results for conjugated and non-conjugated polymers, analyze the errors that propagate into the RI predictions, and offer guidance for method selection.

We thank Prof. Michel Dupuis for helpful discussions on the polarizability calculation of conjugated polymers.

3.1 Introduction

Organic materials with high index of refraction (RI) have gained considerable attention in recent years as they hold tremendous potential for applications in optical and optoelectronic devices [48, 114, 115, 51, 62]. However, the vast majority of carbon-based polymers has relatively low RI values (typically in the range of 1.3 to 1.5) [6, 116], which limits their utility. The discovery and design of compounds with high and very high RIs (greater than 1.8) has thus been an active area of research [63, 65]. The key to increasing the RI values of organic polymers is our ability to tailor their molecular structure [6, 117, 118, 65]. The number of compounds that results from considering even only a modest collection of building blocks is, however, practically infinite. Experimental efforts are too time-, labor-, and resource-intensive to effectively survey the massive chemical space of this problem setting (and many others in the molecular sciences).

Computational high-throughput screening studies have emerged as a way to rapidly characterize and assess candidates, and to identify lead compounds for further in-detail investigations. In the context of optical materials with large

dielectric constants (and thus large RI values), the work by Ramprasad and co-workers [72, 73, 74] is particularly noteworthy. The foundation for *in silico* screening approaches are suitable modeling protocols for the properties and compound classes of interest. For use in large-scale studies, these protocols not only have to produce sufficiently accurate predictions, but they also have to be fast.

A number of modeling approaches for the RI values of polymers have been introduced in the past [72, 73, 119, 76, 75, 120, 121, 122], each with distinct advantages and disadvantages in the areas of accuracy, reliability, robustness, cost, and range of applicability. As discussed in chapter 2, we introduced a new protocol based on a synergistic combination of *first-principles* and data modeling [47]. In this protocol, we calculate RI values using the Lorentz-Lorenz equation with the number density and polarizability of a given candidate compound as input parameters. We obtain the former using the van der Waals volume and packing factor of the compound, and the latter from quantum chemistry. We compute the van der Waals volumes using Slonimskii's method [108] and for the packing fraction of the bulk polymer, we introduced a support vector regression [100, 101] (i.e., machine learning) model. For the polarizabilities, we employ Kohn–Sham density functional theory (DFT) with the PBE0/def2-TZVP-D3 model chemistry. We tested the RI predictions of this protocol on 112 non-conjugated polymers and the results show very good agreement with the experimental values ($R^2 = 0.94$). The protocol is overall economical and suitable for high-throughput *in silico* studies, but the polarizability calculations nonetheless stand out as the bottleneck that limits its efficiency.

In this paper, we present a systematic benchmark study of several DFT model chemistries to identify approaches that deliver a more favorable balance

of accuracy and efficiency for polarizabilities in the context of large-scale RI studies. In addition, we demonstrate the performance of extrapolation schemes from small-oligomer calculations to the polymer limit for both conjugated and non-conjugated systems. We provide an analysis of how the errors in the polarizability results propagate into the RI value predictions.

3.2 Background and Methods

3.2.1 Benchmarking Setup

As mentioned before, our RI protocol introduced in Ref. [47], employs the PBE0 hybrid functional [102], atom-centered def2-TZVP basis set [103], and D3 dispersion correction [104] for the closed-shell, all-electron calculation of the static polarizabilities that serve as input for the Lorentz–Lorenz equation. The target systems are amorphous, quasi-infinite polymers, and we obtain the polarizability results at the polymer limit through an extrapolation scheme from a sequence of small-oligomer calculations. This scheme is based on the finite correlation length in these systems, which typically leads to an early onset of extensivity in the response properties.

In this study, we benchmark the accuracy of several DFT functionals and basis sets. As a reference, we chose the double hybrid functional B2PLYP [123] and def2-TZVP basis set. The calculated RI values from different methods are compared with the experimental RI values of 112 polymers.

Extrapolation schemes for both conjugate and non-conjugated polymers are presented in this work. For non-conjugated polymers, polyethylene (PE) is selected as an example polymer, whereas for conjugated polymers, polyacetylene

(PA) is selected. For comparison between conjugated and non-conjugated polymers, two polymers polythiophene (PT) and poly(1,4-phenylene) (PB) are selected as examples. The conjugation of these polymers is broken by introducing non-planarity in the polymer chain. Non-planar chains are obtained by constraining consecutive rings perpendicular to each other (see fig. 3.1).

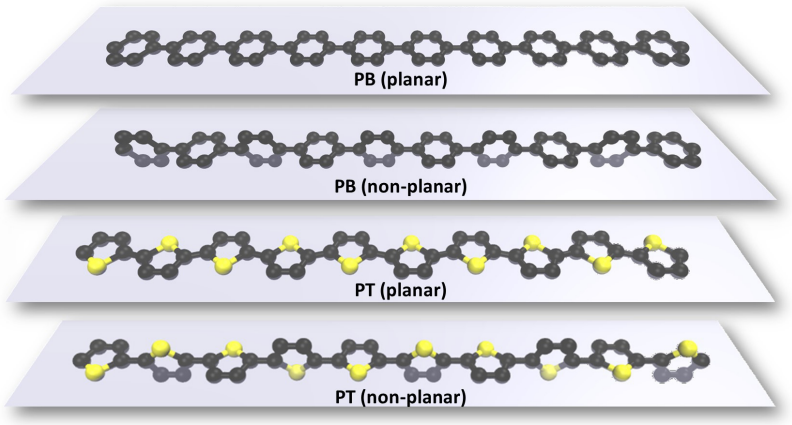


Figure 3.1. Representation of planar/non-planar structures of PB and PT with a chain length of 10.

The values of polarizability are generally dependent on DFT approximation and different model chemistries will give different results. In our previous work, we used PBE0/def2-TZVP to calculate polarizability of polymers, but a benchmarking study is necessary to determine a more rational method towards polarizability calculation. Therefore, in this study, polarizability of the polymers is calculated using six different functionals including BP86, PBE0, TPSSh, M06-2X, B2PLYP, and B3LYP. Each of these functionals was used with two def2 basis sets by the Karlsruhe group [103], def2-SVP and def2-TZVP, which are abbreviated as DZ and TZ respectively. Geometry optimization of oligomers is performed at B3LYP/DZ. All the single point calculations as well as the geometry optimizations were performed with D3 correction [104]. All the quan-

tum computations are performed using the ORCA quantum chemistry package [105].

The RI of polymers is calculated based on the Lorentz-Lorenz equation, which includes the calculation of two different properties, polarizability and number density. The former is calculated by the method mentioned above, whereas the latter is calculated based on van der Waals volume and packing fraction. The van der Waals volume of the molecules is calculated using Slonimskii's methods [108]. However, the critical parameter in number density calculation is the packing fraction of the bulk polymer. The most efficient way of calculating packing fraction is to perform molecular dynamics study, but this approach is computationally expensive, therefore, not a viable option for high-throughput screening studies. In our RI prediction model, we established a machine learning approach using a support vector machine on a training data set compiled from the literature to correlate the polymer structure with their packing fraction. The same method is also used in this work for RI calculation.

The RI values from different model chemistries is validated by comparing with experimental RI values of 112 polymers. The experimental values of these polymers have been taken from Bicerano, polymerdatabase.com, chemicalbook.com, and scientificpolymer.com [113].

In this work, 112 polymers polarizability values are calculated using 12 different methods. For each polymer, four individual calculations are performed from monomer to tetramer. This leads to a total of 5376 calculations. We performed all these calculations in an automated fashion using *ChemHTPS*.

The following abbreviations are used for error analysis terms: MAPE (mean absolute percentage error), RMSE (root mean squared deviation), AE (average error), MaxE (maximum error), SPRE (difference of maximum and minimum

error), MAE (mean absolute error), MARE (mean absolute relative error) and RMSRE (root mean squared relative error)

3.3 Results and Discussion

In our previous studies, we showed that the polarizability of non-conjugated polymers, calculated at PBE0/TZ level, varies linearly with increasing chain length. Here, we calculate the polarizabilities of polyethylene (PE) using different methods. The plot in fig: 3.2 shows the comparison between different methods. Blue color is for DZ and red for TZ.

Observations:

i. In all methods, the polarizability values have a linear relationship with the PE oligomer chain length. The initial decrease in the values from monomer to trimer is due to the end effects of the hydrogen.

ii. The polarizability per oligomer of PE converge to a constant value right after trimer. This suggests that for non-conjugated systems, it is sufficient to calculate polarizability until tetramer.

iii. The calculated polarizability values in the decreasing order of BP86 $>$ B3LYP $>$ TPSSh $>$ PBE0 $>$ M06-2X $>$ B2PLYP

It is well known that LGA and GGA functionals significantly overestimate the polarizability values, which is also depicted in these studies. The key to solve the problem of overestimation is to include an exact exchange treatment [124]. Therefore, the functionals B3LYP, TPSSh, PBE0 and M06-2X, which include HF treatment predict lower polarizability values. Further, the functionals which include perturbation theory along with HF, e.g., double hybrid functional B2PLYP, predict much lower polarizability values. This benchmarking

study will assist in understanding the degree of over-prediction for various lower level methods.

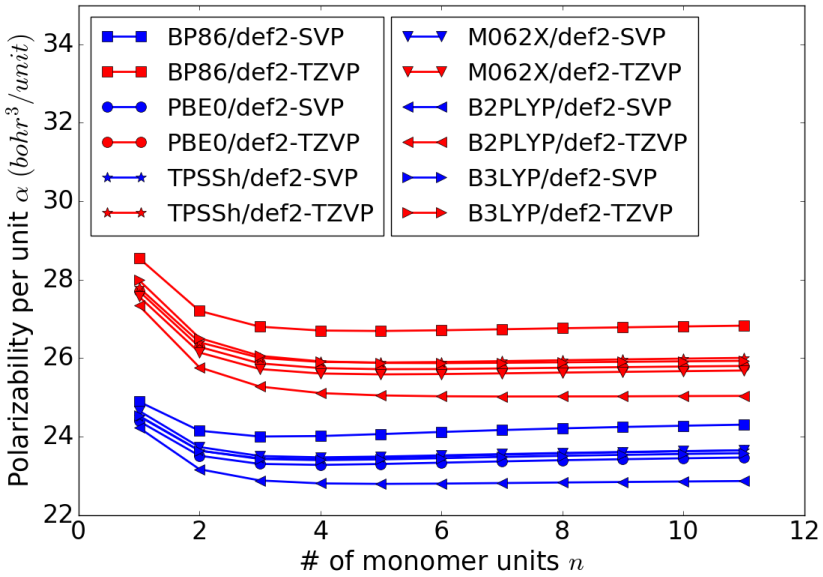


Figure 3.2. Polarizability per oligomer unit of polyethylene with varying chain length calculated from different model chemistries.

For conjugated polymers, polyacetylene (PA) is selected as test polymer. Polarizability with increasing chain length is calculated using different functionals and basis set (see fig: 3.3). The variation of polarizability follows a non-linear behavior with the increasing oligomer length. This is because, as the length is increasing, the conjugation in the molecule also increases leading to increased polarizability. Non-linear correlation can be used to fit the polarizability variation of conjugated polymers.

A better way to compare conjugated and non-conjugated polymers is to break the conjugation by introducing an aliphatic carbon between aromatic rings. For this, two conjugated polymers, PT and PB, are selected and conjugation is broken by introducing non-planarity in the molecule as shown in fig: 3.1. The polarizability per oligomer unit increases with chain length for

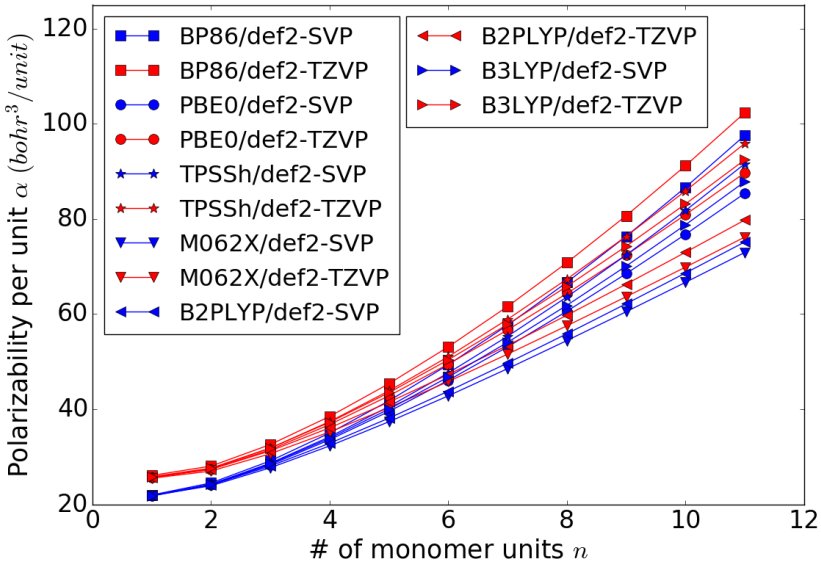


Figure 3.3. Polarizability per oligomer unit of polyacetylene with varying chain length calculated from different model chemistries.

planar PB and PT polymers (see fig: fig: 3.4 and fig: 3.5). For the non-planar PT and PB, the increase in polarizability per oligomer is significantly less. This suggests that the conjugation plays an important role in the polarizability of the polymers. The slight increase of polarizability in non-planar PT and PB shows that there exists a weak conjugation between the aromatic rings. This observation can be better understood by looking at the HOMO and LUMO of PT and PB with a chain length of 5 (see fig: 3.6). The HOMO for planar PB and PT is alternating, suggesting a strong conjugation in the molecule, whereas for non-planar PB and PT, the HOMO is slightly alternating which shows evidence of a weak conjugation. The LUMO distribution of planar PB and PT extends along the full length of molecule, suggesting that the excited electrons would have a clear path to flow.

The variation of polarizability per oligomer unit for conjugated polymers initially increases linearly and then asymptotically converges. A mathematical

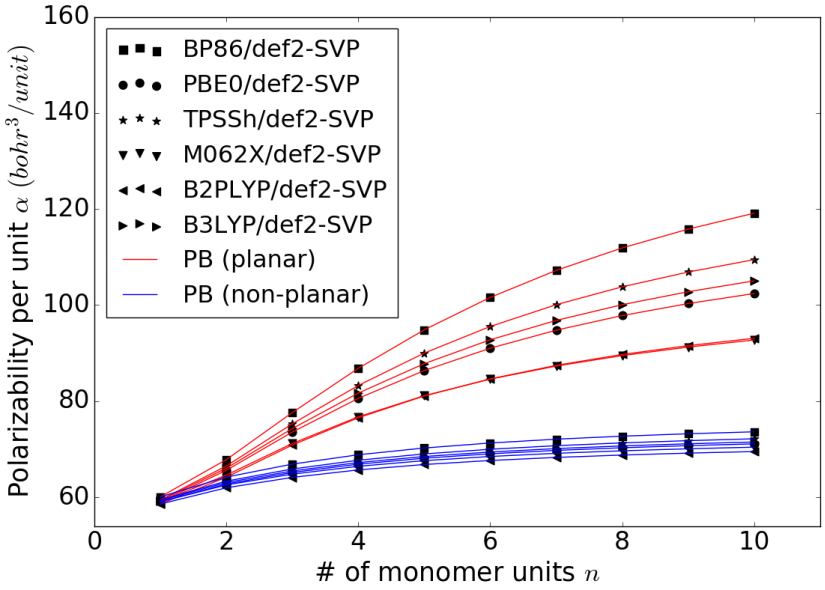


Figure 3.4. Polarizability per oligomer unit of planar and non-planar poly(1,4-phenylene) (PB) with varying chain length calculated from different model chemistries.

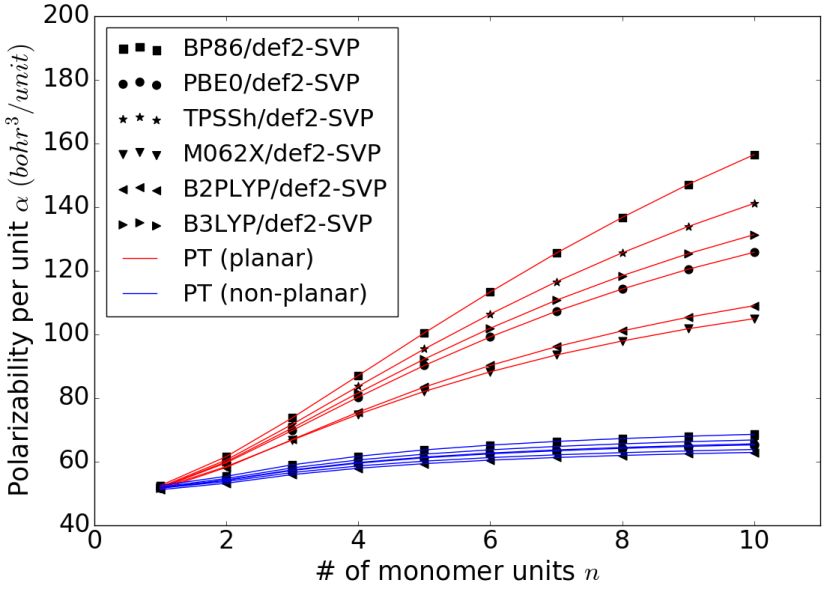


Figure 3.5. Polarizability per oligomer unit of planar and non-planar polythiophene (PT) with varying chain length calculated from different model chemistries.

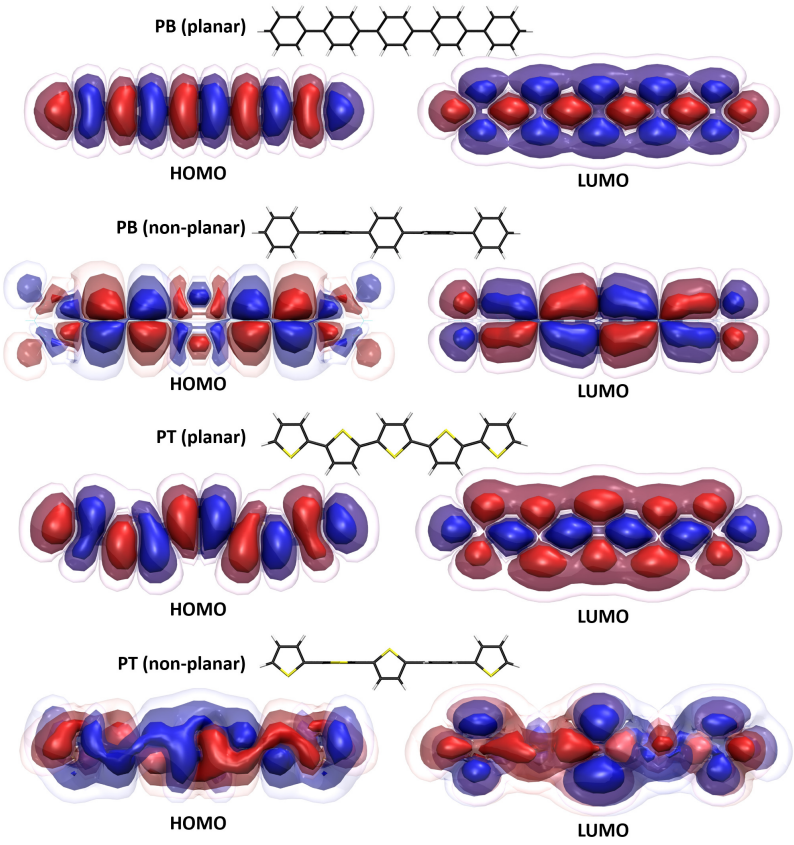


Figure 3.6. HOMO (B2PLYP/TZ) and LUMO (B2PLYP/TZ) of planar/non-planar poly(1,4-phenylene) (PB) and polythiophene (PT).

expression was proposed in literature as shown in the eqn: 3.1 [125]. However, this expression does not fit well for long oligomers (see fig: 3.7). An extra term is added to the expression as shown in eq: 3.2, such that the variation fits well for longer chains (see fig: 3.8). This expression is valid for all the three studied conjugated polymers PA, PB and PT. Thus, using this expression, the polarizability of polymer limit can be calculated based on the calculations for small chain oligomers.

$$\log(\alpha) = a + \frac{b}{N} + \frac{c}{N^2} \tag{3.1}$$

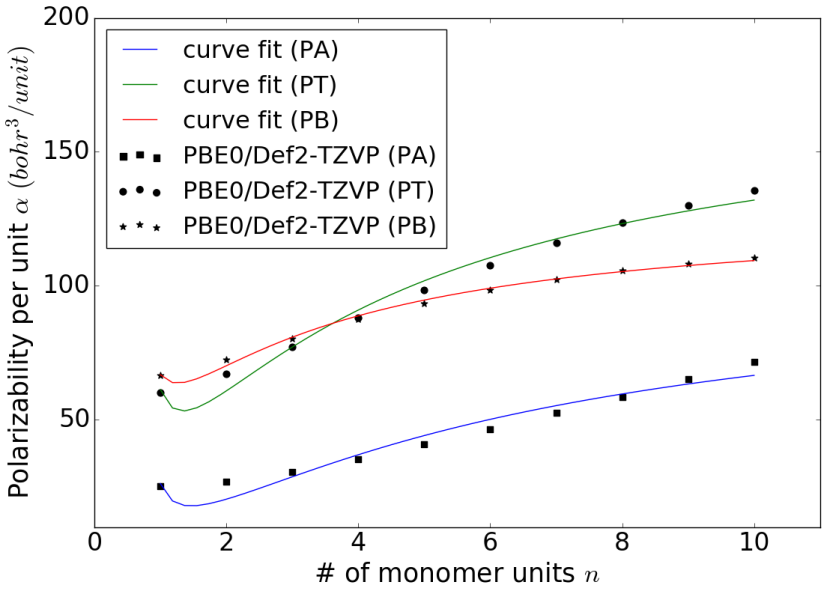


Figure 3.7. Curve fitting using polarizability expression from [125] for the conjugated polymers polyacetylene (PA), poly(1,4-phenylene) (PB) and polythiophene (PT). Polarizability is calculated using PBE0/TZ.

$$\log(\alpha) = a + \frac{b}{N} + \frac{c}{N^2} + \frac{d}{N^3} \tag{3.2}$$

To check the validity of the expression, oligomers up to a chain length of 50 were created. It should be noted that these oligomers were not geometry optimized. The bond lengths and angles were selected based on the optimized geometry of 10 oligomer chain length. We observed that the expression is valid for very long chain lengths (see fig: 3.9). Using these extrapolation schemes, the polarizability of polymers can be calculated quickly. Further, these schemes are implemented in the high throughput screening framework to accelerate polarizability for large number of polymers. Using the *ChemHTPS* framework, the polarizability values of 112 polymers are determined from 12 different methods.

Fig: 3.10 shows the performance of different model chemistries in comparison to B2PLYP/TZ. Observations from these comparisons are as follows:

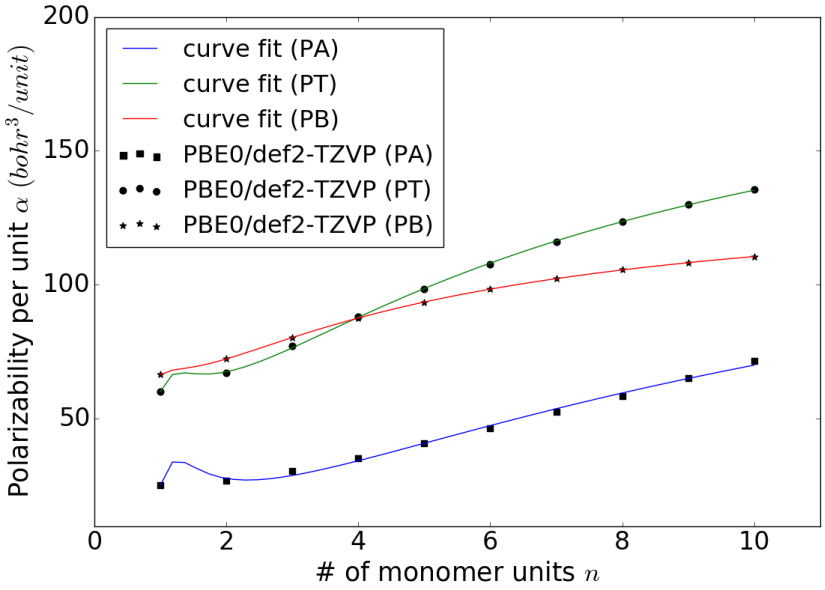


Figure 3.8. Curve fitting using new polarizability expression for the conjugated polymers polyacetylene (PA), poly(1,4-phenylene) (PB) and polythiophene (PT). Polarizability is calculated using PBE0/TZ.

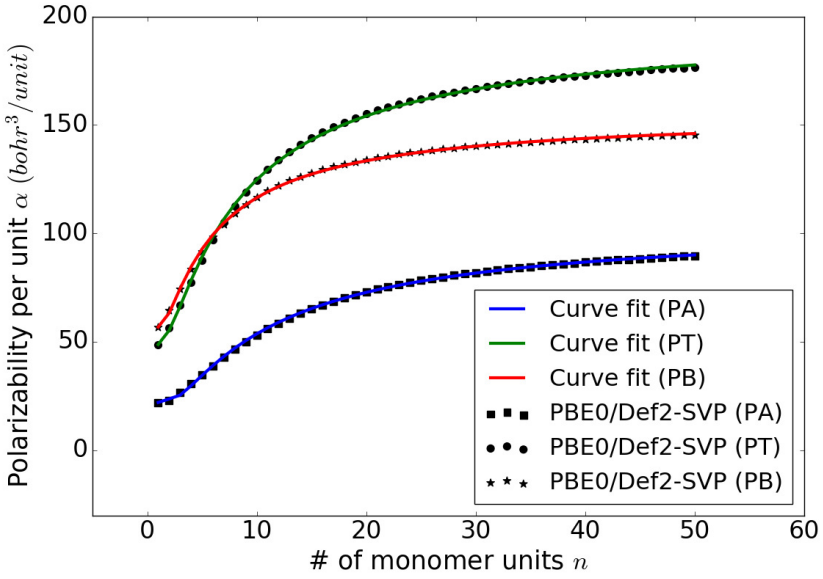


Figure 3.9. Validation of the polarizability expression for long chain lengths of polyacetylene (PA), poly(1,4-phenylene) (PB) and polythiophene (PT). Polarizability calculated using PBE0/DZ

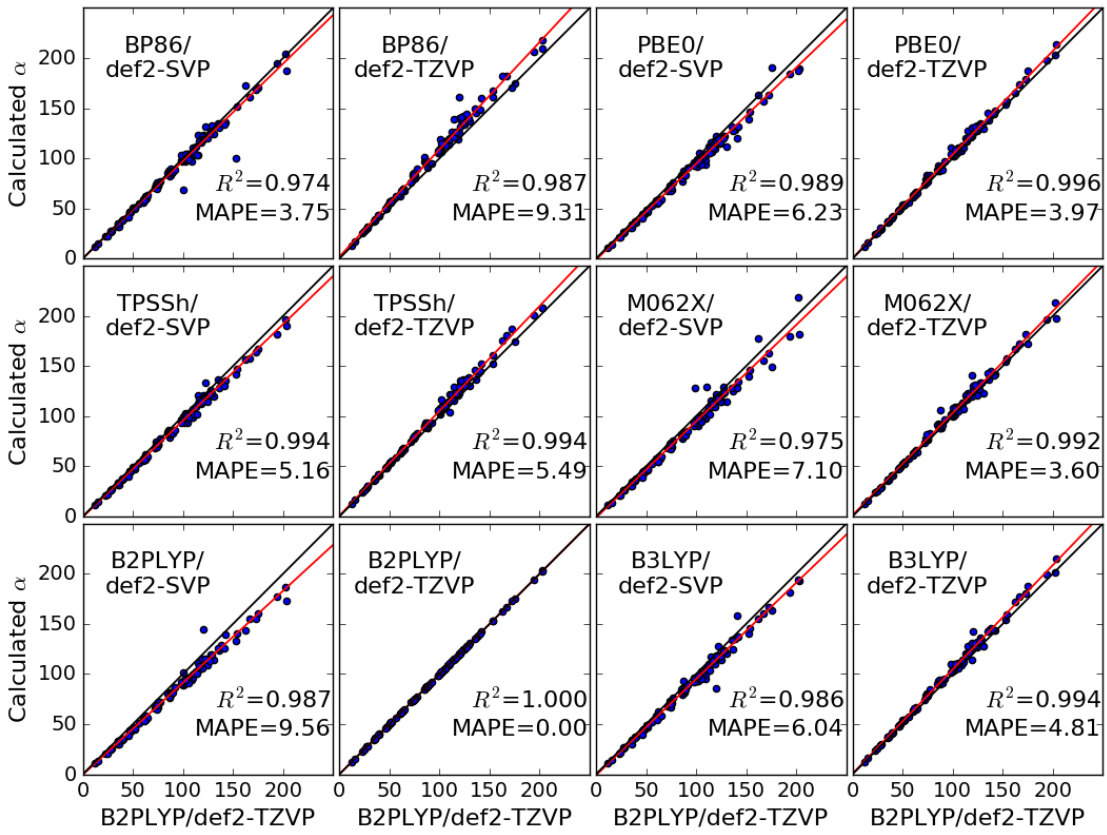


Figure 3.10. Comparison of polarizability from different methods for 112 polymers with B2PLYP/TZ as a reference.

- All the methods have a good correlation (>0.975) when compared with B2PLYP/TZ. However, few methods have high MAPE values (>5) which is caused due to an offset in the calculation as can be seen from the linear fits of the plots (red lines). For example, the linear fit of B2PLYP/DZ method below the line of symmetry (black line), as most of the values calculated from this method are lower than B2PLYP/def-TZVP. In most of the methods, the offset in the polarizability is larger for higher polarizability values. Therefore, for higher polarizability values, more accurate methods are required.

- In all TZ calculations, the linear fit is always above the line of symmetry,

whereas for all DZ calculations the linear fit is below the line of symmetry.

- The correlation for all TZ calculations is better when compared with the DZ calculations.
- Among the functionals, PBE0 has the best correlation with B2PLYP/TZ and also has a lower MAPE value. Therefore, PBE0 would be a good functional to use for polarizability calculations.
- The MAPE of B2PLYP/DZ, when compared to B2PLYP/TZ, is 9.56%.

The RI (n) values are calculated based on the Lorentz-Lorenz equation (see eq: 3.3), which includes the calculation of two different properties, the polarizability (α) and number density (N). To understand the error propagation from polarizability and number density to RI values, we take a logarithm on both sides of the equation and differentiate to obtain the eq: 3.4.

$$\frac{(n^2-1)}{(n^2+2)} = \frac{4\pi}{3} N\alpha \tag{3.3}$$

$$\frac{dn}{n} = \frac{(1-\frac{1}{n^2})(1+\frac{2}{n^2})}{6}(\frac{dN}{N} + \frac{d\alpha}{\alpha}) \tag{3.4}$$

$$error factor(E) = \frac{(1-\frac{1}{n^2})(1+\frac{2}{n^2})}{6} \tag{3.5}$$

From eq: 3.5, we observe that the error factor (E) is dependent only on the RI value. For the RI value ranging from 1 to 2, the value of E ranges from 0 to 0.187 respectively (see fig: 3.11). Therefore, when the number density error is ignored, the error in RI calculation is less than 20% of the error in polarizability calculation. For example, if a lower level method, BP86/def-SVP (MAPE

3.75% compared to B2PLPY/TZ), is used to calculated polarizability instead of B2PLYP/TZ, the additional error in RI calculation would be 0.75%. This suggests that use of lower level methods for polarizability calculations would not affect RI values significantly.

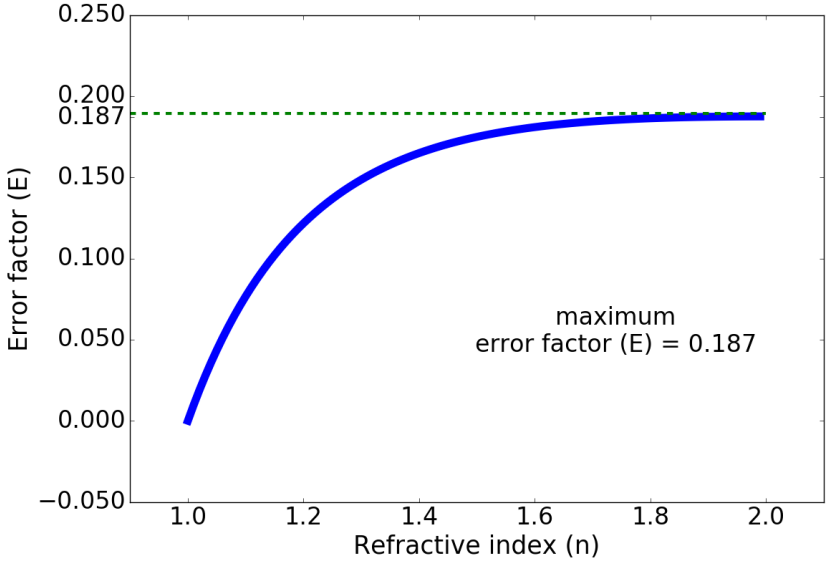


Figure 3.11. Error factor (E) for RI values ranging from 1 to 2.

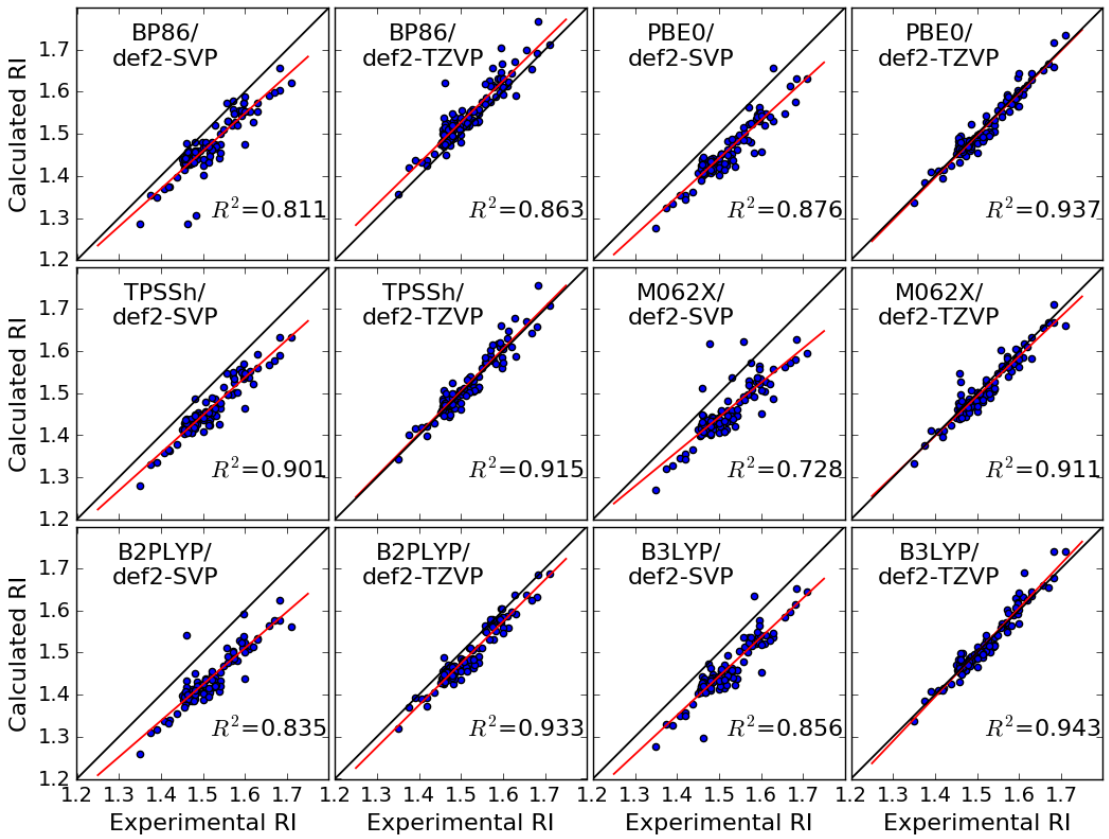


Figure 3.12. Comparison of RI values calculated from different methods for 112 polymers with experimental RI values.

Table 3.1. Performance of different model chemistries in comparison to experimental values

Functional	**BP86**		**PBE0**		**TPSSh**		**M06-2X**		**B2PLYP**		**B3LYP**	
Basis set	**DZ**	**TZ**	**DZ**	**TZ**	**DZ**	**TZ**	**DZ**	**TZ**	**DZ**	**TZ**	**DZ**	**TZ**
R^2	0.811	0.863	0.876	0.937	0.901	0.915	0.728	0.911	0.835	0.933	0.856	0.943
AE	0.042	-0.027	0.06	0.004	0.054	-0.005	0.06	0.008	0.078	0.026	0.057	-0.001
MaxE	0.177	0.159	0.142	0.045	0.137	0.073	0.149	0.088	0.161	0.068	0.166	0.077
SPRE	0.195	0.197	0.17	0.089	0.144	0.115	0.29	0.137	0.242	0.096	0.218	0.111
MAE	0.043	0.029	0.06	0.014	0.054	0.016	0.065	0.017	0.079	0.027	0.058	0.013
MARE	0.028	0.019	0.04	0.009	0.036	0.011	0.042	0.011	0.052	0.018	0.038	0.008
RMSD	0.052	0.038	0.064	0.018	0.058	0.021	0.07	0.022	0.082	0.031	0.063	0.018
RMSRE	0.042	0.031	0.052	0.014	0.047	0.017	0.056	0.018	0.067	0.025	0.051	0.014
Slope	0.9	0.98	0.91	1	0.9	1	0.82	0.95	0.86	0.99	0.93	10.05
Intercept	0.11	0.06	0.07	-0.01	0.1	0	0.21	0.07	0.13	-0.02	0.05	-0.08

The RI values for the 112 polymers is calculated using six functionals (BP86, PBE0, TPSSh, M06-2X, B2PLY and B3LYP) and two basis sets (DZ and TZ). The correlation with experimental values and the error analysis of each of these methods is shown in the table 3.1.

Observations:

i. In all the cases, TZ calculations are better than the DZ calculations.

ii. The functionals B2PLYP, B3LYP and PBE0 have relatively better performance, whereas the M06-2X functional has the worst performance.

iii. Comparing all the methods, the R^2 value for B3LYP/TZ is the highest. This method also has the lowest AE, MAE and MARE. However, the lowest maximum error is observed for PBE0/TZ.

iv. All the methods showed a maximum error less than 0.2, with a lowest error of 0.045 for PBE0/TZ. Further, the SPR error for this method is 0.089, which suggests that this method is accurate for RI calculation.

v. If only DZ is considered, then PBE0 and TPSSh functionals have the better performance compared to other functionals.

vi. The slope and intercept of the linear correlations for the methods TPSSh/TZ, B2PLYP/TZ and PBE0/TZ are 1 and 0 respectively, suggesting that the linear correlation is exact. This can be observed in Fig. 3.12, where the linear fit (red line) overlaps the line of symmetry (black line). Thus, the calculated values do not have to be corrected by a factor for these methods. For all the other methods except B3LYP/TZ, the slope is less than 1, which shows that these methods under-predict the RI of polymers with high RI values.

vii. Considering the time of calculations v/s the accuracy of the methods, PBE0/DZ and TPSSh/DZ show the best promise for RI calculation.

Now that we have developed a robust protocol for RI prediction of poly-

mers, our next step is to cast these into our automated high-throughput framework, and employ in on an extensive candidate screening library. We plan to apply this framework to discover high RI polyimides and polyenes. The candidate library of these polymers will be created from promising building blocks using our combinatorial library generator. We will also pursue the development of a smart, responsive, and thus more efficient scheme in the course of the project, which will avoid the problems of exponential growth. The different screening phases will successively reduce the pool of viable candidates. The resulting data will be analyzed, mined, and modeled using machine learning and informatics techniques to extract structure-property relationships.

3.4 Conclusions

We have successfully created 270,000 novel PI candidates, and evaluated the RI values of these candidates by casting into *ChemHTPS*. From the screening studies, we found PI candidates that possess RI values greater than 1.8. Z-score analysis showed that prevalence of certain building blocks (e.g., building blocks 25 and 28) in a top candidates. In addition to identifying individual building blocks that are promising for high RI polymers, promising building block pairs are also identified. Z-score value. We observed that the combination of building block 28 with blocks 2 and 3 to be highly promising in the design of HRIPs. Thus, we successfully identified favorable building blocks and synergy between building blocks combinations for developing high RI PIs. We demonstrated that our cyberinfrastructure is a powerful tool and has shown to be highly promising for identifying polyimides with exceptional RI values.

Chapter 4

Molecular Library Generator and Virtual High-Throughput Screening Framework

The discovery of new compounds, materials, and chemical reactions with exceptional properties is the key to progress in chemistry. This process can be dramatically accelerated by means of the virtual high-throughput screening of large-scale candidate libraries. This approach has been extensively used for many years in the drug discovery community and has more recently been applied to the search for energy materials. The key challenge is that chemical space is practically infinite, and any approach to survey it or enumerate certain of its domains has to address the problem of combinatorial complexity.

Therefore, we have developed a general-purpose suite, *ChemLG*, a generator for compound and material candidate libraries. In this chapter, we primarily discuss the algorithms implemented in *ChemLG*, and give examples of its successful application in various domains. We also briefly discuss the other three

software packages in our group's software ecosystem, *ChemHTPS*, *ChemBDDB*, and *ChemML*.

I am the primary developer of *ChemLG* code. The code *ChemHTPS* was co-developed with William Evangelista. The developers of *ChemBDDB* are Aditya Sonpal and Shirish Sivaraj, whereas the primary developer of *ChemML* is Mojtaba Haghighatlari.

4.1 Introduction

A prerequisite for the high-throughput exploration of chemical space is access to suitable, large-scale screening libraries. The key for a successful library generation approach is to balance the ambition for a systematic and exhaustive enumeration of the combinatorial search space, with the need for an efficient and responsive scheme. The generation of combinatorially exhaustive libraries is relatively straightforward, but it rapidly becomes impractical for screening purposes due to its exponential growth. For instance, the largest small-molecule library, GDB-17, contains 166.4 billion molecules, which were generated using only up to 17 atoms (C, N, O, S, and halogens) [126]. Most currently available codes are thus limited to generating small, drug-like molecules.

ChemLG extends and generalizes the molecular library generation to identify lead candidates in various applications such as functional polymers, optoelectronics, and catalysis. It offers a multitude of options to customize and restrict the scope of the enumerated chemical space and thus tailor it for the demands of specific applications. To streamline the non-combinatorial exploration of chemical space, we incorporate genetic algorithms into the framework. Genetic algorithms have shown to be efficient in optimizing chemical structures

and generating useful compounds for different target applications. In addition to implementing smarter algorithms, we also focus on the ease of use, workflow, and code integration to make this technology more accessible to the community.

4.2 Methodology

There are two different approaches for enumeration of compound space: product-based approach and reaction-based approach. The former one is based on the application of Markush structures, which expands the library by attaching the functional groups to various sites on the scaffold whereas the latter approach is based on chemical transformation [4]. Combinatorial libraries can be obtained by using three different algorithms: the first is the grow algorithm where the growth in the molecule starts from a seed point, second is the fragment-link algorithm and the third algorithm is the sampling approach where the growth of the molecules is based on random selection [5]. In our generator, we use the fragment-link method which follows a reaction transform approach. The reaction can be obtained by two methods: first method is the reshuffling of the SMILES strings and the second method follows an actual reaction scheme. Both of these methods are discussed in the following subsections.

4.2.1 SMILES Rearrangement

SMILES, which stands for simplified molecular-input line-entry system, is a string notation for describing the structure of chemical species [127]. It is a widely used notation as it is human readable and can be imported by most

molecule editors for molecular visualization. In the SMILES rearrangement method, the SMILES of the linking molecules is written such that the linking atoms are positioned at the extreme end of the string. The linkage can then be obtained by directly combining the SMILES strings. A schematic representation of this method is shown in the Fig. 2.8. In this method, it is not necessary to specify reaction sites or have to use any reaction algorithm. This way of linkage can generate all possible combinatorial structures.

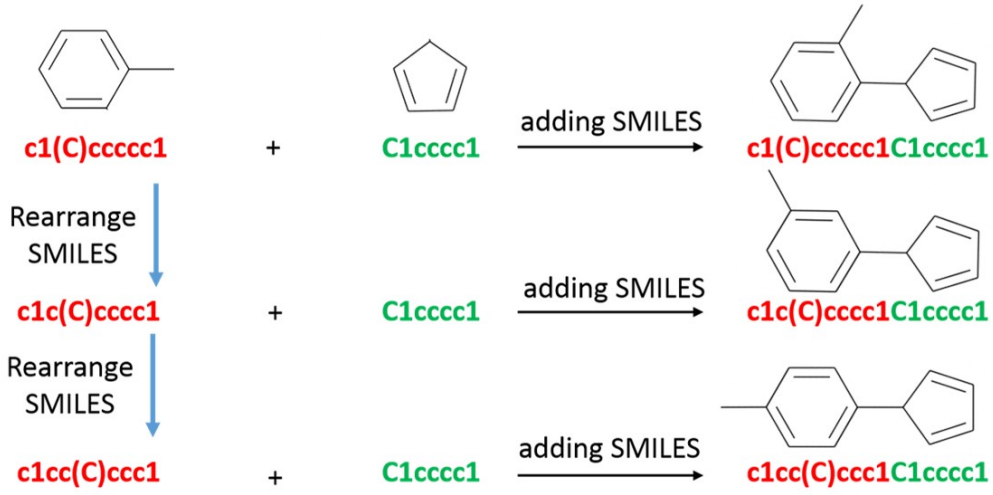


Figure 4.1. Fragment linkage based on SMILES rearrangement.

We developed a code to represent the SMILES of a molecule with all possible string notations in which the chemical handles are positioned on the extremes. This method is extremely powerful as it can produce a large molecular library spontaneously. Further, this method is more suited for the generation of polymer libraries. We apply this method to generate a library of polyimides for the exploration of high-refractive-index polymers discussed in Chapter 5. However, there are a few limitations to this approach. For example, a large number of duplicates will be generated in the process and the removal of duplicates will be computationally expensive. Additionally, narrowing the chemical space will

be difficult as applying constraints on string notations is challenging. Further, the generation of fused molecules is not possible using this approach.

4.2.2 Fragment Link and Fusion

Fragment link or fusion of fragments is performed by casting into a reactor. The reactor in *ChemLG* takes the fragments as arguments along with combination type and generates a linked or fused molecule. A schematic representation for the link and fusion of molecules is shown in the Fig. 4.2. In this scheme, the reaction sites on the fragments are denoted as X. In case of linking, once the reaction occurs, there is a bond formed between the fragments and X is removed. While, in fusion, the molecules lose two carbons and a hydrogen along with X.

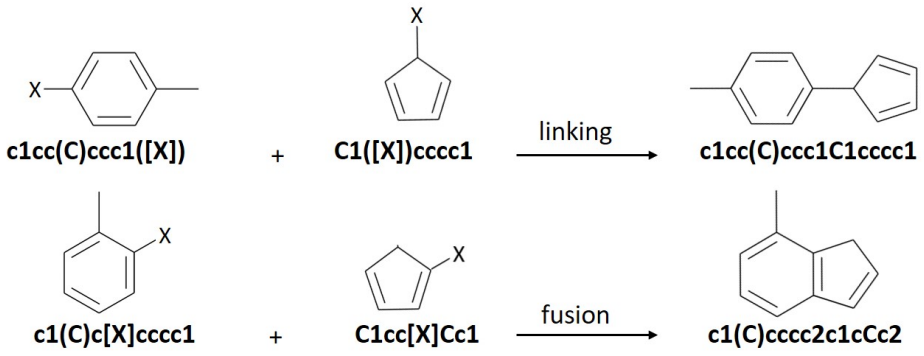


Figure 4.2. Schematic for the link and fusion of molecules in *ChemLG*.

4.3 Generation Constraints

We find that instead of exploring a limited chemical domain exhaustively, it is often more useful to bias the search in directions where candidates are most promising and synthetically viable. In our development of *ChemLG* , we have

thus augmented the combinatorial schemes by a number of 'smart' modules that make use of additional input. To address the concern that virtual candidates may not be accessible or desirable (e.g., from a synthetic perspective), we have introduced a constrained-growth scheme that continually prunes the generation process. In this scheme, molecules are rejected at every generation to constrain the growth as shown in the Fig. 4.3. It accepts user-defined constraints, e.g., to exclude certain structural patterns or substructures, fingerprint matching, building block combinations, or sequences; to limit size or chemical makeup; or to enforce symmetries or other rules (e.g., Lipinski's rule).

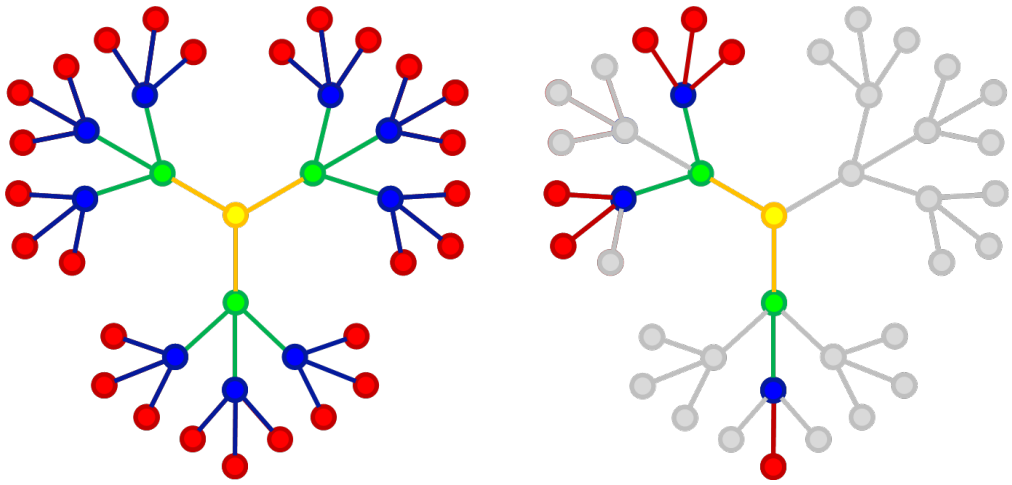


Figure 4.3. Schematic for constrained growth of molecular library. The left scheme shows an exhaustive enumeration, whereas the right scheme represents a targeted enumeration.

4.4 Smart Algorithm

Even applying above mentioned generation constraints might lead to a large number of unwanted molecules in the library. We need to only target the molecules that are of interest and discard the rest. Application of smart algorithms can narrow down the chemical space to a more specified regions. A

schematic for the pruning of molecular libraries is shown in Fig. 4.4. *ChemLG*'s genetic algorithm module allows us to optimize candidate pools and the chemical structures they contain for specific target applications.

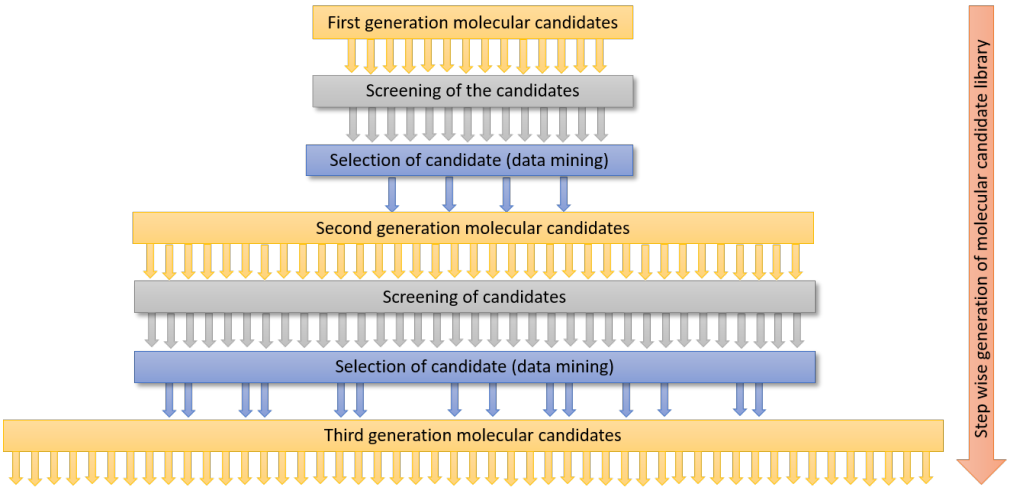


Figure 4.4. Pruning of molecular library at every generation by use of smart algorithms.

We employ an on-the-fly prescreening through rapid candidate assessment *via* DFT, MD, or data-derived prediction models. These models also serve as fitness functions in *ChemLG*'s genetic algorithm module, which allows us to optimize candidate pools and the chemical structures they contain for specific target applications. In genetic algorithm approach, at every generation of molecular library, best candidates are retained and certain operations are performed to generate new promising candidates. These operations include crossover and mutation (see Fig. 4.5). In the crossover operation, the structures from top candidates are split and exchanged with each other. Whereas in mutation, randomly selected substructure is replaced with other building blocks from the initial list. Following this scheme for several generations will result in a library that is tailored for a targeted application. As an example, this scheme is applied to create a library of molecules with similar structure to a targeted molecule with fin-

gerprint matching as fitness function (see Fig. 4.6). We observe that at every generation we obtain molecules that match better to the target molecule.

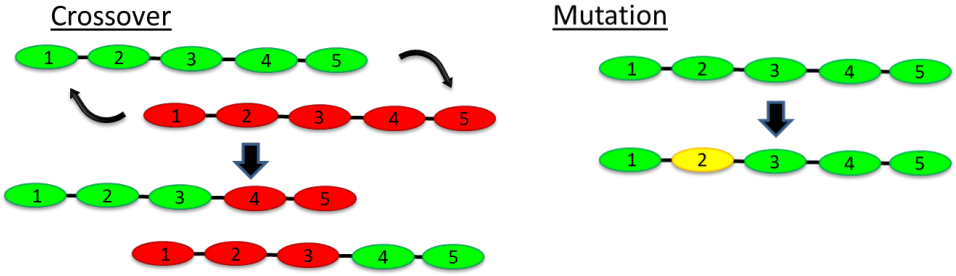


Figure 4.5. Genetic algorithm operations: crossover and mutation.

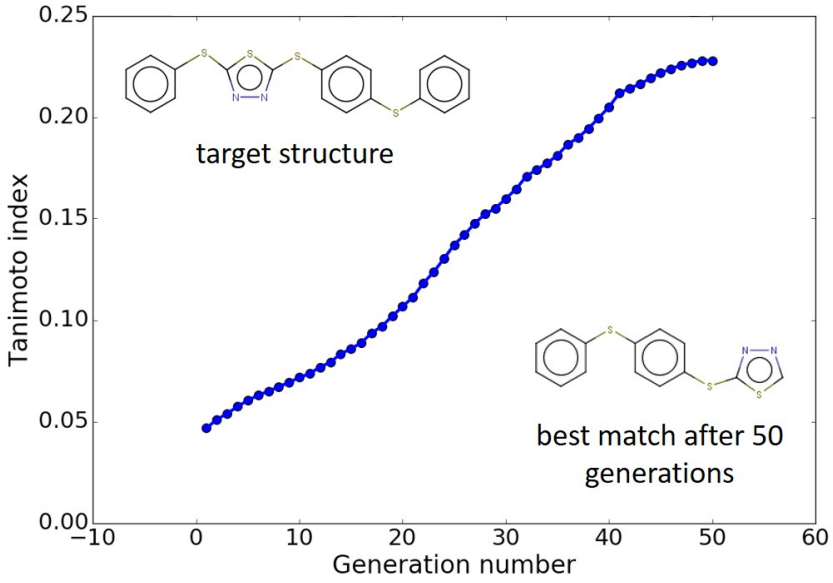


Figure 4.6. Implementation of genetic algorithm to find similar structures.

ChemLG's 'smart' modules can thus facilitate self-regulating growth of the candidate libraries and ultimately a self-optimizing traversal of chemical space. It offers options for the various approaches mentioned above to provide the most suitable solution for different problem settings. We have successfully employed *ChemLG* to produce screening libraries for a number of application projects.

4.5 Parallel Implementation of *ChemLG*

An important feature of *ChemLG* is that the code is massively parallel. The parallelization is obtained using MPI4Py library and it is implemented at various levels to accelerate the library generation process (see Fig. 4.7). This allows us to create a library of millions of molecules in just a few seconds. The bottleneck step is the removal of duplicates at every level. There are several algorithms for efficient removal of duplicates, but they are designed to run in a serial fashion. Parallelizing these algorithms is not efficient as there is significant communication overhead in the process. For efficiently scraping duplicated in molecular library, we exploit the fact that no two molecules with different molecular weights are same. We group the molecules based on their molecular weights by keeping tags and scatter the groups to different processors for duplicates removal. This process is efficient as it involves no communication between the processors. The speedup and parallel efficiency of the code for the generation of a million candidates is shown in Fig. 4.8. The performance of *ChemLG* scales very well with increasing number of processors. Due to a high-level of parallelization, *ChemLG* performs well on a high performance computing (HPC) platform. This is helpful as it integrates well with *ChemHTPS* infrastructure, where the molecular library can be generated and readily applied for HTPS.

4.6 Example Applications of *ChemLG*

ChemLG has been successfully applied in various applications. In the current dissertation, it is applied to make a library of polyimides and small organic molecules for the exploration of chemical systems with high RI values. Further

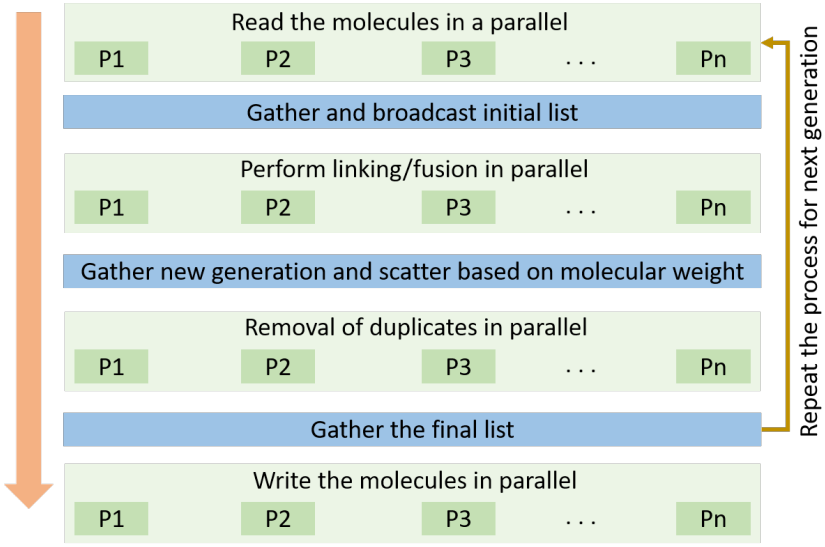


Figure 4.7. Levels of parallelization in *ChemLG* package.

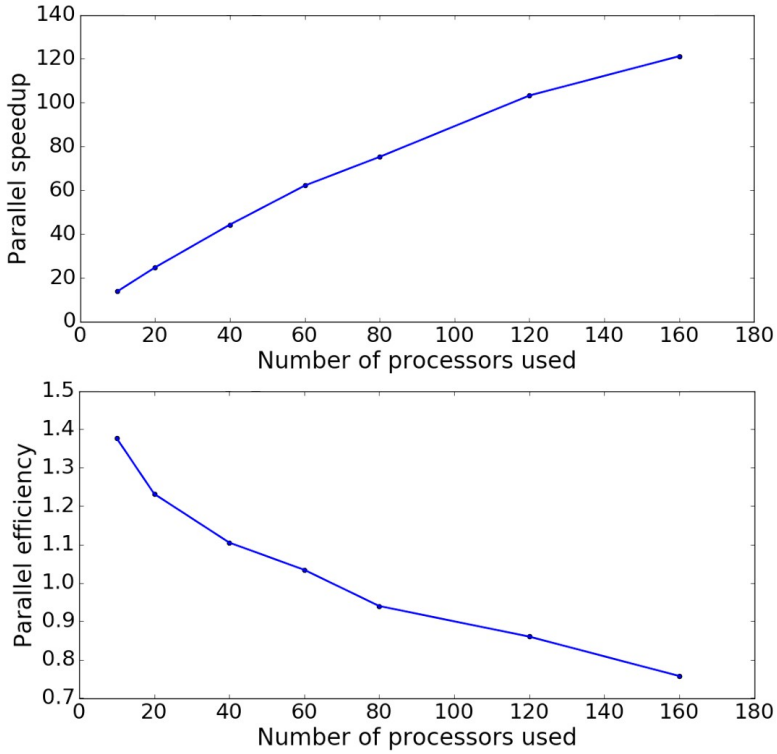


Figure 4.8. Parallel speedup and efficiency observed in *ChemLG* package.

details of these two libraries are discussed chapter 5 and chapter 6, respectively.

The generator was also used for creating a library of polyesters in a quest to design polymers with superior degradation behavior. We found that the activation energy for the degradation of polyesters is dependent on the functional groups that are attached to the β-carbon of the ester bond. This results motivated us to generate a library functional groups that can ba attached to β-carbon. This project is being led by Vigneshwar Kumaran Sudalayandi Rajeswari.

In another project, internship work at ExxonMobil Chemical Company, *ChemLG* was used to create a library of novel polyolefins [128]. The library was built by linking saturated carbon atoms in a combinatorial fashion. We used methane, ethane, and butane as the building blocks while considering all the hydrogens as connection sites. Using these rules, the library was built until three generations, which resulted in 275 monomer structures. These monomers were used to build corresponding oligomers and subsequently screened them in a virtual high-throughput fashion to evaluate their coil dimensions.

ChemLG has also been instrumental in formulating chemical libraries for compounds with application as deep eutectic solvents (DES) and photosensitizing (PS) catalysts. Candidate compounds were obtained by relying on a common molecular structure, for e.g., alcohols and amides as hydrogen bond donors in DES systems, and corroles and porphyrins for PS systems. Different building blocks were attached to these molecular backbones at desired linking sites to get hundreds of thousands of candidate molecules in each case. These libraries are part of a strategy to exhaustively access chemical spaces of interest for these applications. Yudhajit Pal is taking a lead on these two projects.

4.7 Other Parts of Group's Cyberinfrastructure

4.7.1 Virtual High-Throughput Screening Infrastructure

A prerequisite for conducting computational high-throughput screening studies is a software infrastructure that can facilitate the execution of thousands or even millions of modeling calculations. For this task, we have been developing the *ChemHTPS* program suite. A schematic for a typical use of HTPS is represented in Fig. 4.9. *ChemHTPS* is designed to streamline and automatize the setup of project environments and directory structures, the generation of job pools based on user-defined candidate libraries (e.g., from *ChemLG*) and modeling protocols, their submission to available hardware in an orderly and load-balanced fashion, job monitoring, error handling, as well as the parsing and bookkeeping of returning results. These processes follow generalized workflow templates that we have been developing from abstraction in the course of different application projects. It currently supports the ORCA [105], Q-Chem [129], and GROMACS [130] modeling packages, and bindings to other quantum chemistry, molecular dynamics, and solid state physics codes are planned for the future. Given the required user input for the candidate library and modeling protocols, we can now set up and launch a high-throughput *in silico* screening project like the Clean Energy Project [3, 18, 19, 20, 21, 22] from scratch in a few minutes, which is a dramatic reduction from its original lead time of several months. We have been using *ChemHTPS* in a number of studies, searches for new high RI polymers, deep eutectic solvents for supercapacitors and battery electrolytes, molecular hydrolysis catalysts for solar water splitting and fuel cells, doped and defect nanographene anode materials for lithium ion batteries, polyvinyl-based

biodegradable polymers for biomedical plastics, liquid organic hydrogen carriers for the hydrogen economy, and organic semiconductors for photovoltaics and other applications [3, 18, 19, 20, 22, 47, 131, 132, 133]. In the case of high RI polymers, we apply this framework to automate the calculations for the polarizability (see Fig. 4.10) and packing density. Contributions to *ChemHTPS* were made by William Evangelista.

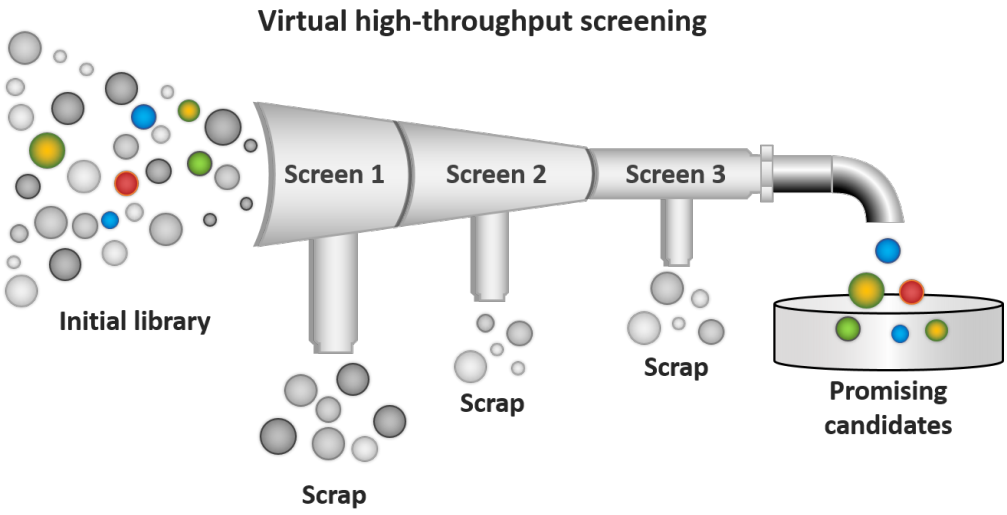


Figure 4.9. Schematic for the screening of a library of molecules to identify promising candidates. At every level of screening, a higher level of theory is applied while narrowing down the chemical space.

4.7.2 Database Infrastructure

The use of modern database technology is of particular importance in the context of data-intensive research. Despite their great utility and despite being essential for projects that accumulate large data sets, databases are still rarely featured in chemical research. Our group is developing the *ChemBDDB* code to simplify and streamline the use of databases and thus make them more accessible to non-expert users in the chemistry community. *ChemBDDB* provides

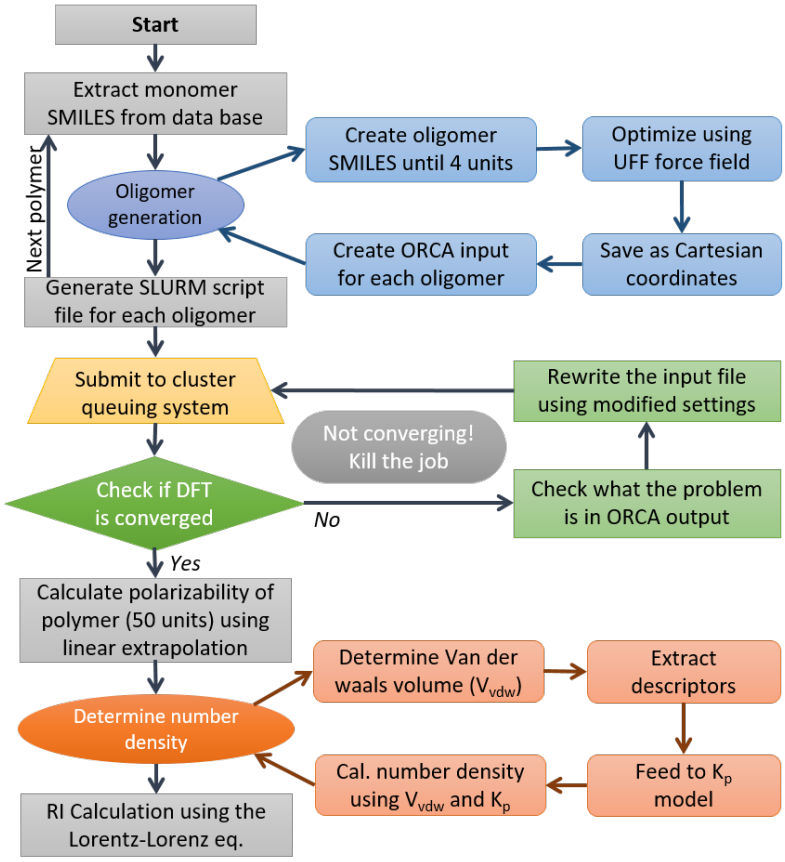


Figure 4.10. Work-flow implemented for the automation of polarizability values of candidate libraries.

an automated database setup, a data model template that can readily be customized, and the necessary tools to access and manipulate the database. As in *ChemHTPS* , we have been developing the corresponding workflows by abstracting our experience from real-life application projects with flexibility and reusability in mind. All the data generated from the virtual high-throughput screening of high RI polymers is incorporated into a database using *ChemBDDB*. Screenshots of the database are shown in Fig. 4.11, which were provided by Shirish Sivaraj.

Developers of *ChemBDDB* are Aditya Sonpal, Shirish Sivaraj and Supriya Agrawal.

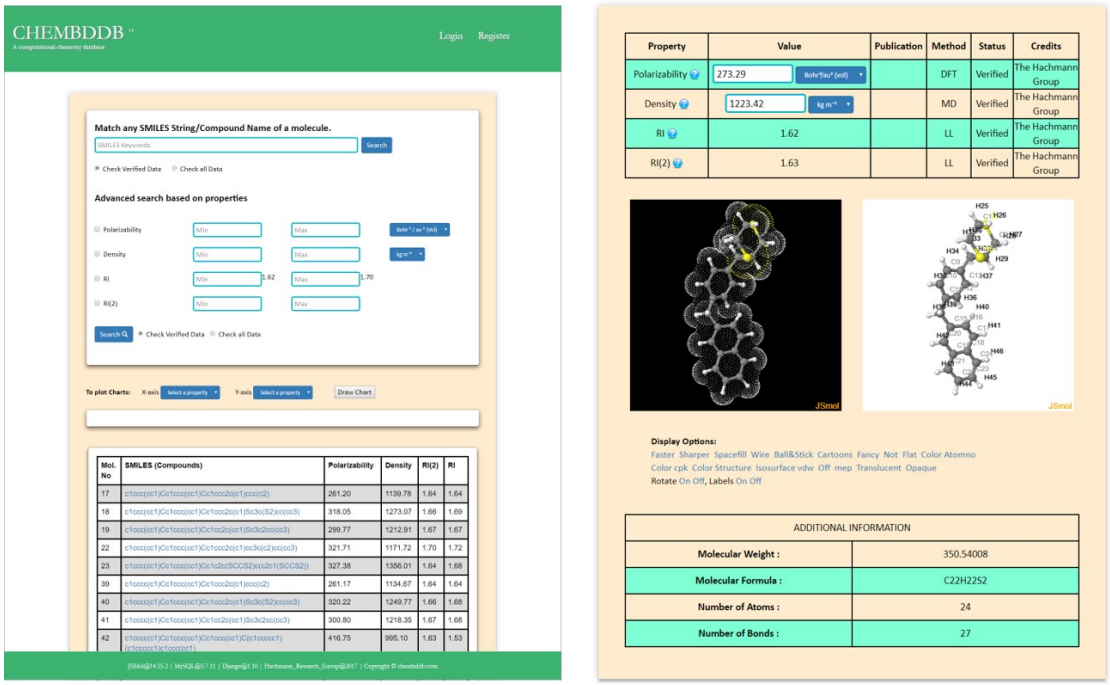


Figure 4.11. Screenshots from the HRIP database.

4.7.3 Data Analysis, Mining, and Modeling Infrastructure

We have been developing the *ChemML* program suite to establish data analysis, mining, and modeling capabilities that allow us to apply state-of-the-art machine learning and informatics methodology to chemical and materials data sets. *ChemML*'s principal tasks are the creation of predictive regression models and chemical pattern recognition/classification [134].

The resulting mapping functions (i.e., data-derived prediction models) are usually much easier to evaluate than physical models (e.g., the Schrödinger equation), so using them allows us to dramatically accelerate the characterization of chemical systems and it thus enables the hyperscreening of chemical space. A typical *ChemML* workflow encompasses a number of distinct steps that can be categorized in six main tasks: (1) input/output, (2) preprocessing, (3) learning, (4) validation, (5) evaluation, and (6) visualization.

ChemML provides facilities for all these tasks *via* classes of methods. These are either accessed from advanced third-party libraries and stand-alone programs (e.g., scikit-learn [135], Tensorflow [136], Keras [137], Dragon [110], OpenBabel [138], RDKit [139]), if these represent the state of the art for certain tasks, or from *ChemML Library* where we compile our new/original contributions as well as existing methods that are not otherwise accessible. The feature representation and the machine learning approach are two particularly important aspects in the generation of data-derived models, and they determine a model's predictive performance. *ChemML* can readily call many standard machine learning (as well as preprocessing and visualization) methods. Examples of available algorithms include multivariate regression [140], support vector machines [141], artificial neural networks [142, 28], and deep learning [143]. To support chemistry applications, *ChemML* interfaces the core learning algorithms with domain-specific tools, such as the feature space of molecular descriptors [144, 145], topological fingerprints [146, 147], or more recent developments such as the Coulomb matrix [36] and bag-of-bonds [148] descriptors. These feature spaces are the abstract 'basis set' in terms of which a machine learning model for a particular structure-property relationship is numerically expressed [149], and *ChemML* provides a comprehensive collection of existing schemes. While our current work focuses on supervised learning, we plan to broaden our scope to unsupervised and reinforcement learning in the future.

The primary developer of *ChemML* is Mojtaba Haghighatlari. Other contributors include Ramachandran Subramanian, Bhargava Urala, Gaurav Vishwakarma, Aditya Sonpal, Po-Han Chen, and Srirangaraj Setlur.

4.8 Software Ecosystem

The four program packages discussed in the previous sections (*ChemLG*, *ChemHTPS*, *ChemBDDB*, and *ChemML*) are loosely connected, i.e., they can either be employed as a comprehensive unit (see Fig. 1.3), or in combination with drop-in replacements (e.g., with a different library generator or a custom database engine), or as standalone applications. The development work on our software ecosystem includes conceptual work, the design and assessment of protocols and workflows, the formulation of guidelines and best practices, and the implementation of both glue code and new methods. The key cyberinfrastructure challenges include the robust abstraction of workflows for general-purpose applications, scaling issues (e.g., associated with expensive data generation and the combinatorial nature of chemical space), code sustainability, as well as platform and distribution issues. While we emphasize black-box automation to reach non-expert users, we allow full customization of all settings (in particular in *ChemML*). We continually extend and refine the features and capabilities of all four codes. These improvements are driven by feedback from application projects both inside and outside our group (cf. Sec. 4.6). This input from different real-world application problems is a key to making this cyberinfrastructure as resourceful as possible. All codes are open and freely shared with the community under 3-clause BSD license [43, 44, 45, 46]. All the repositories on GitHub are maintained by Hachmann Group.

There are a number of exciting, high-profile software development efforts along similar lines by others in the field (e.g., [32, 150]). However, despite great popular demand, there is currently no cyberinfrastructure for data-driven *in sil-*

ico research that is accessible to the community, applicable to a wide range of chemical problems, and that offers a level of comprehensiveness, automation, and integration comparable to that of modern computational chemistry program packages. Our contribution pursues this niche, which stands out in its scope and prospective utility.

4.9 Conclusions

The developed massively parallel generator, *ChemLG*, and the program suite for automated, virtual high-throughput screening studies, *ChemHTPS*, offers a multitude of options to customize and restrict the scope of the enumerated chemical space and thus tailor it for the demands of specific applications. We incorporate genetic algorithms into the framework to streamline the non-combinatorial exploration of chemical space. Genetic algorithms have shown to be effective in optimizing chemical structures and generating useful compounds for different target applications. The parallel implementation of these codes and their integration with HPC systems allows us to create and setup screening of millions of molecular candidates in few seconds to minutes.

Chapter 5

Accelerated Discovery of High-Refractive-Index Polyimides

Polyimides have attracted much attention due to their exceptional thermal stability and ease of processability. Further, polyimides possess mechanical stability, good flexibility, flame resistance, radiation resistance and low dielectric constant, thus holding great promises for various applications. However, these polymers have low refractive index (RI) values which limit their use in optical and optoelectronic applications. In this study, we present a computational approach to discover novel high RI polyimides (PI). We use an RI prediction model, developed in our previous work, to calculate the RI values of large candidate library of PIs, created from building blocks provided by our experimental collaborators. To accelerate the development process and effectively screen the relevant high RI PIs, we cast the RI model into *ChemHTPS*: a high-throughput screening, materials informatics, and rational design framework software developed in our group. We prove that *ChemHTPS* is promising for rapidly identifying PIs with exceptionally high RI values. We explore various paths that

introduce highly polarizable moieties into PI backbones to increase RI. We also identify monomer building blocks within PIs that are prevalent in high RI PIs, and discover building block combinations that result in the same. Additionally, we provide insights into the relationship between the structure and the RI values of polyimides, thus allowing us to target most promising candidates. We thank Prof. Cheng for providing us with building blocks and generation rules for synthetic feasibility. We acknowledge Sai Prasad Ganesh for his contributions in this work. He contributed to understanding the dependence of the polarizability on the geometry of molecules.

5.1 Introduction

Organic small molecules, oligomers, and polymers are emerging materials that feature many attractive properties compared to conventional inorganic materials. Devices made out of organic polymers are generally flexible, mechanically stable on impact, light-weight, and inexpensive to produce. This has led to increased efforts in utilizing these compounds in many different application domains, including optic and optoelectronic devices such as organic light-emitting diodes [1], complementary metal oxide semiconductor [2], photovoltaics [3], field-effect transistors [4], displays, and image sensors [5], in which they can be introduced *in situ* as microlenses, waveguides, microresonators, interferometers, anti-reflective coatings, optical adhesives, and substrates. However, most of these applications require materials with a refractive index (RI) greater than 1.7 or larger, while typical carbon-based polymers only exhibit values in the range of 1.3-1.5 [6]. This provides an incentive to discover or design new high-refractive index polymers (HRIPs) for the aforementioned applications. As the

properties of organic polymers can be tailored by controlling their molecular structure, they are a prime example for a rational design target.

In recent years, polyimides (PIs) have been shown to have favorable electronic and mechanical properties that could form potential HRIP candidates. Despite showing inherently low RI values leading to a lack of present applicability, PIs have other attractive properties [151, 152]. PIs are strong potential candidates due to their exceptional thermal stability, and ease of processability [153, 154, 48]. These properties are complemented by their favorable mechanical stability, flexibility, flame resistance, radiation resistance and their sufficiently high molecular polarizability: properties which would allow for potential use in optoelectronics [155, 156].

As previously mentioned, the optical properties of PIs are significantly inferior compared to conventional metal oxides currently used in optical and optoelectronic devices [157, 6]. However, PI optical properties can be improved upon by several methods [158, 159, 160, 161, 162]. One such technique is to control the chemical structure of PIs to allow for precise tuning of optical properties, in particular to increase their RI values [154, 161, 163, 164]. In our study, we present a computational approach to study the RI of PIs and explore techniques that introduce highly polarizable moieties into polyimides framework to create a new class of high RI PIs.

Typically, HRIPs exist in the form of aromatic polyamides, and aromatic heterocyclic ring polymers, and certain conjugated aromatic polymers. However, these HRIPs struggle when it comes to optical implementation due to their large optical dispersion and large birefringence. The large birefringence is caused due to their aromatic, and conjugated pi-electron structures, which leads to poor transparency and coloration. The addition of highly polarizable

moieties, which do not have significant pi-electron conjugation, can aid in increasing the RI of polymers. For example, small aromatic rings, halogen atoms, metals and particularly sulfur atoms have shown to be promising for this purpose [165, 166, 154, 167]. Previous experimental studies have demonstrated the ability of sulfur infused PIs to overcome potential unfavorable properties. In particular, high thermal stability, a low birefringence, an optical transparency in the visible light region, and high RI values have been demonstrated. In 2007, Ueda *et. al.* developed PI films which were shown to have high RI values, but had unfavorably high birefringence in the range of 0.012 [168]. However, in recent years, experimental work has shown improvement in terms of birefringence and RI, with RI in the range of 1.76 and birefringence of 0.009 attained in particular, due to the high sulfur content of the PIs [169]. These encouraging results have led to our study being primarily concerned with sulfur incorporated PIs, and in doing so, generate a new class of HRIPs that does away with the technical limits of existing HRIPs.

Most of the PIs developed for high RI applications are based on the intuition of empirical observations. Therefore, there is a high possibility that there are potential HRIP candidates that are not studied experimentally simply because there are too many possible candidates to be feasibly studied. This motivation has led to our approach in generating a large library of promising PI candidates created by our molecular library generator. The library generator operates based on combinatorial linking, which could possibly to lead to an astronomical number of candidates beyond the scope of affordable computational studies. To counteract this, we narrowed the candidates generated from the library based on keen observations from past research and generation rules based on the input provided from our experimental collaborators. We created our PI

library by casting these rules into the molecular library generator. We used our RI prediction model to evaluate the RI values of generated PI candidates [47].

To facilitate RI evaluation of our large pool of candidates in a timely manner we use our virtual high-throughput screening framework, *ChemHTPS* [44]., that draws inspiration from The Harvard Clean Energy Project, which was successfully screened millions of organic molecules for photovoltaic applications [3, 20, 18]. *ChemHTPS* creates inputs, executes and monitors the calculations, parses and assesses the results, extracts and post-processes the information of interest, inserts the key outcomes into the project database, and archives all other data. Using this *in silico* methodology we created a large number of novel PI candidates and characterized these candidates at a fraction of the time and cost of traditional studies.

5.2 Methods

In our previous work, we developed a model for the prediction of RI of polymers, which was validated against experimental RI values of 112 non-conjugated polymers [47]. The RI prediction model is based on a synergistic combination of *first-principles* quantum chemistry calculations and data modeling. In this scheme, we calculate RI values using the Lorentz-Lorenz equation, which involves two critical parameters, the number density and the polarizability. We calculate the former using van der Waals volume and packing factor of the polymer, while for the polarizability calculations, we use *first-principles*.

We developed a library of PIs using our molecular library generator, *ChemLG* [43]. The library is based on 29 building blocks and bonding rules (see Fig. 5.1, which were selected based on the suggestions provided by our experimental

collaborator, Cheng's group. Based on these generation rules, we created about 50 thousand and 230 thousand possible structures for R_1 and R_2, respectively. We saved the structures in the form of SMILES string along with a code which contains the information on the list of building blocks and connections that were used to make a particular structure. We subsequently used *ChemHTPS* to screen the PI candidates to evaluate their RI values, and collected the data in our project database.

For each R_1 and R_2 SMILES, we generated fifty different 3D conformations using OpenBabel software [107]. We selected the lowest energy conformation for each structure and further optimized the geometry using the universal force field (UFF) [106] as implemented in the OpenBabel software. We use a packing factor of 0.75 to calculate the number density of PIs, as previous experimental studies have shown that PIs typically have a packing factor of 0.75 [170]. We calculated the van der Waals volumes using Slonimskii's method detailed in Ref. [108]. For the polarizability calculations, we use an all-electron, restricted DFT method with the PBE0 hybrid functional [102] in combination with the double-ζ quality def2-DZVP basis set by the Karlsruhe group [103]. We include Grimme's D3 correction [104] to account for dispersion interaction. We carried out all the quantum computations using the ORCA 3.0.2 quantum chemistry package [105].

In addition to identifying the best candidates from our high-throughput screening studies, we further analyze the collected data to understand structure-property relations. By applying Z-score analysis, we identify the building blocks that contribute the most to the RI values. We compute the

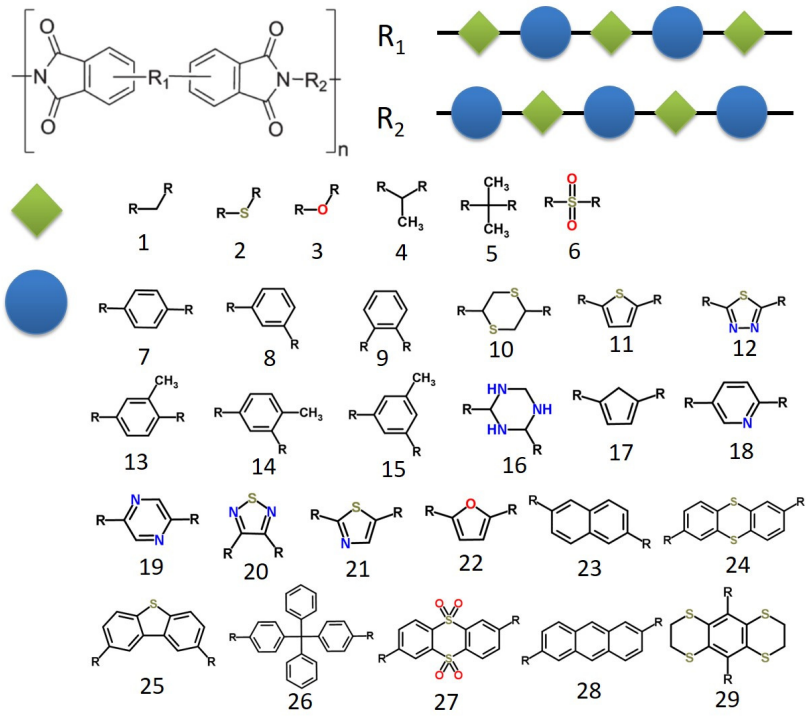


Figure 5.1. Building blocks used to create the library of PIs.

Z-score (Z_i) of the candidates by

$$Z_i = \frac{k - n\frac{K}{N}}{\sigma}, \; \sigma = \left[\frac{nK}{N} \times \left(\frac{N-K}{N}\right) \times \left(\frac{N-n}{N-1}\right)\right]^{\frac{1}{2}}$$

where N is the total number of molecules, n is the subset of molecules that are considered, K number of occurrences of building block i in N molecules and the k is the occurrences of building block i in n subset. We perform similar calculations to calculate the Z-score of the building block connections to identify the synergistic combinations that lead to the high RI values.

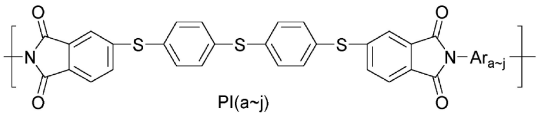
5.3 Results and Discussion

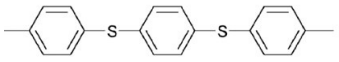
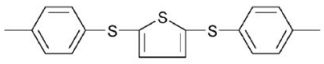
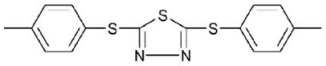
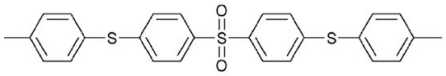
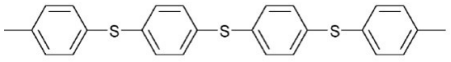
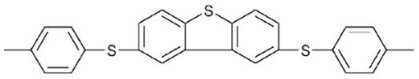
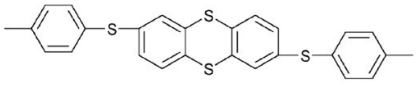
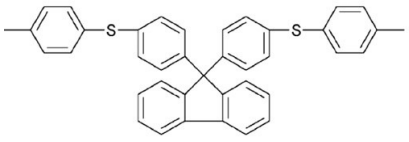
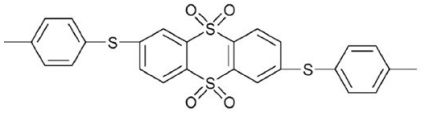
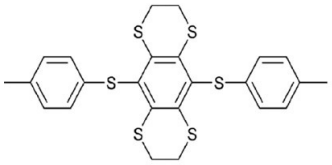
According to the Lorentz-Lorenz equation, the RI of a material is dependent on its polarizability and number density. Using our previous studies conducted on a library of 112 non-conjugated polymers [47], we generated a contour plot to elucidate the relationship between the polarizability and the Number density (see Fig. 5.2). The contour plot demonstrates an inverse proportionality relationship between the number density and the polarizability. Our preferable high RI region happens when both the polarizability and number density are sufficiently high. However, there is a tendency for the number density to be low in highly polarizable materials. Therefore, in order to attain desirable optical properties, it is necessary to maximize both these parameters at the same time. One approach could be to restrict the compound space in a constant number density region and explore highly polarizable compounds in that region. Going forward with this logic, we choose the structure of PI as shown in Fig. 5.1. Given that the densities of these PIs are fairly similar, we can now search for highly polarizable PI candidates [170].

As mentioned in the methods section, we calculate the RI of PIs using a model that involves the calculation of the polarizability and the number density values. We previously tested this model against 112 non-conjugated polymers, however, there were no PI candidates in this list. Therefore, to check the validity of our RI model on the PI structures, we compared the model with experimental RI values of 10 PIs shown in table 5.1. The RMSD of the model is 0.024, suggesting that our model is accurate for calculating the RI of PIs.

In search of high RI PIs, we created about 50 thousand and 220 thousand structures of R_1 and R_2, respectively, using 29 promising building blocks as

Table 5.1. Comparison of RI values from RI prediction model with the experimental values of 10 PI candidates



Ar_{a-j}	Structure	Experiment value [6]	Calculated value	Error
a		1.746	1.738	-0.008
b		1.753	1.739	-0.013
c		1.749	1.707	-0.042
d		1.748	1.751	0.003
e		1.733	1.760	0.027
f		1.758	1.779	0.021
g		1.760	1.735	-0.025
h		1.726	1.741	0.015
i		1.737	1.743	0.005
j		1.769	1.724	-0.045

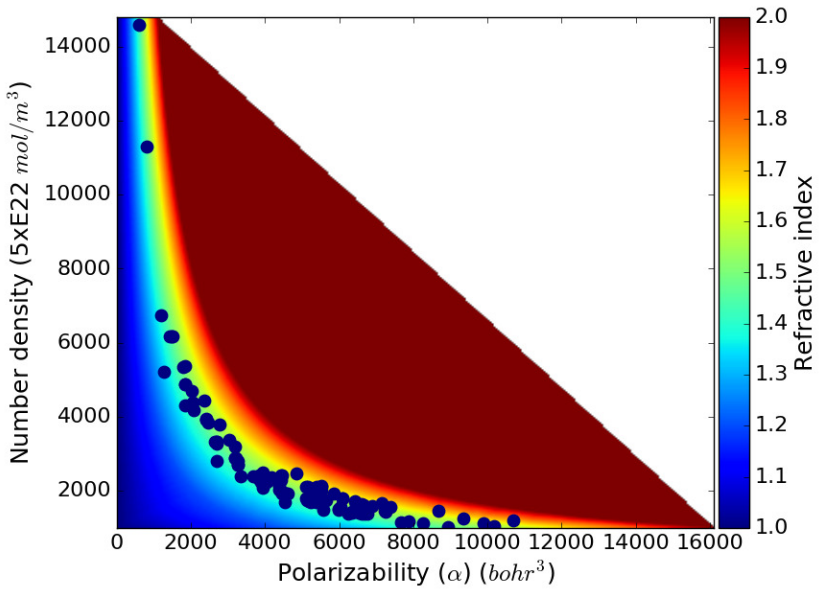


Figure 5.2. Contour plot representing the dependence of RI on the polarizability and number density. Blue dots refer to the 112 polymers used to develop RI model.

shown in Fig. 5.1. Combining all these R_1 and R_2 structures to form PI could lead to a total of 11 billion PIs. However, virtual high-throughput screening of such an astronomical number of structures is not a viable option. Therefore, we narrowed down the number of possible PI candidates by picking the best of the R_1 and R_2 structures. We did so by computing the RI values of the individual R_1 and R_2 structures. After the screening of these structures, we developed the most promising PI candidates by combining the top R_1 and R_2 structure as shown in 5.1.

In this paper, we discuss the results of R_2 structures. Figure 5.3 shows the RI distribution of R_2 structures. We observe a Gaussian type distribution for the RI values of R_2 structures. Most of the candidates have RI values between 1.5 and 1.7, which suggests that there is a strong possibility of obtaining molecules with such RI values using empirical approaches. However, using a computational approach, we we were able to find the outliers, i.e., the candidates with

RI values greater than 1.7.

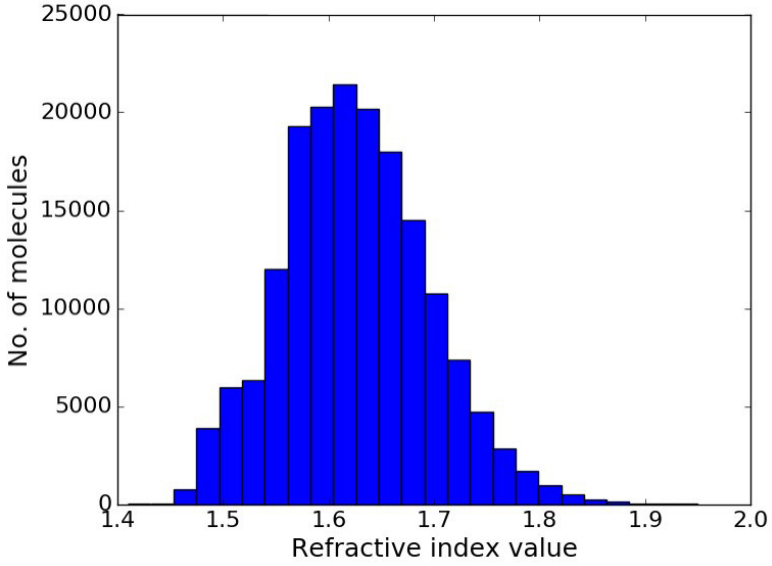


Figure 5.3. RI distribution of the R_2 structures

The top 100 structures of R_1 type and top 1000 structures of R_2 type are selected to form 100,000 PI monomers. We casted these PI structures into our HTPS framework to evaluate the RI values.

Besides identifying the potential HRIP candidates, understanding the underlying structure-property relationships would enable us to discern candidates with optimal RI values. This would help us create a special subset of candidates for our experimental collaborators to further explore. To do so, we tried to recognize the contribution of each building block towards a targeted property to aid in identifying favorable synergies in building block combinations, with an aim to narrow our chemical search space. We decided to implement this into our studies by looking at two different scoring systems for the building blocks: i) the average RI value of the candidates which contain each building block and ii) the Z-score of each building block in the top candidates. We plotted the

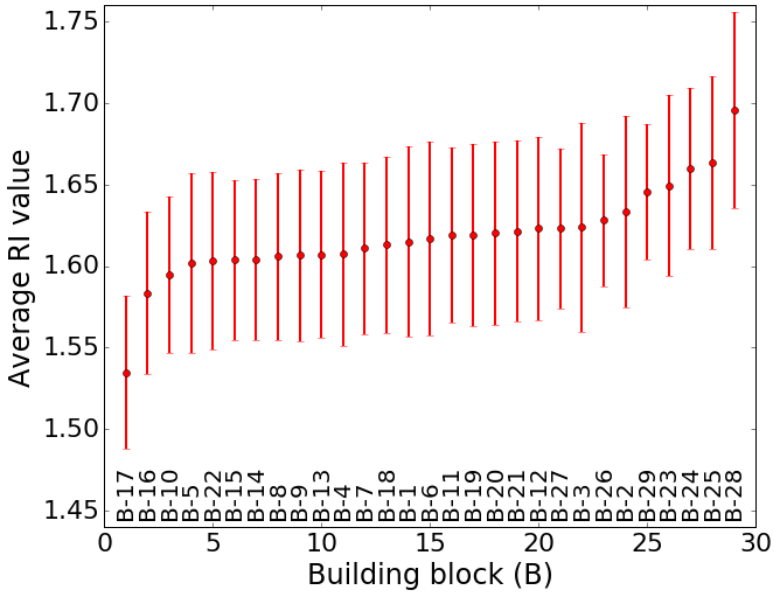


Figure 5.4. Average RI of the structures containing each building block in the top 10% candidates of R_2.

results obtained from the first method in Fig. 5.4, which presents the RI distribution of candidates containing a particular building block. Here, we found that for the R_2 structures, candidates containing building blocks 24, 25 and 28 have the highest RI values, while the candidates containing building blocks 17, 16 and 10 showed the lowest RI values. Figure 5.5 on the other hand shows the z-scores. This technique is favored as a higher z-score would mean a larger prevalence of a particular building block in a HRIP. In our case, we picked the top 10% of candidates in terms of their RI values and evaluated the Z-scores of individual building blocks in this subset. The green color in the Fig. 5.5 represents a positive Z-score, whereas the red color represents a negative Z-score. This technique confirmed our results obtained by the first method as the Z-score values also suggest that the building blocks 24, 25 and 28 are the most promising for developing HRIPs.

The above Z-score analysis only gives the prevalence of building blocks in

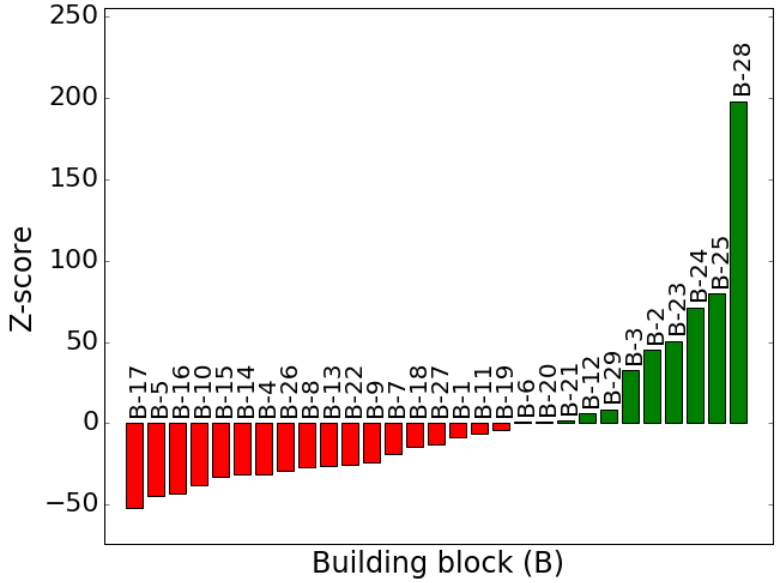


Figure 5.5. Z-score of each building block in the top 10% candidates of R_2.

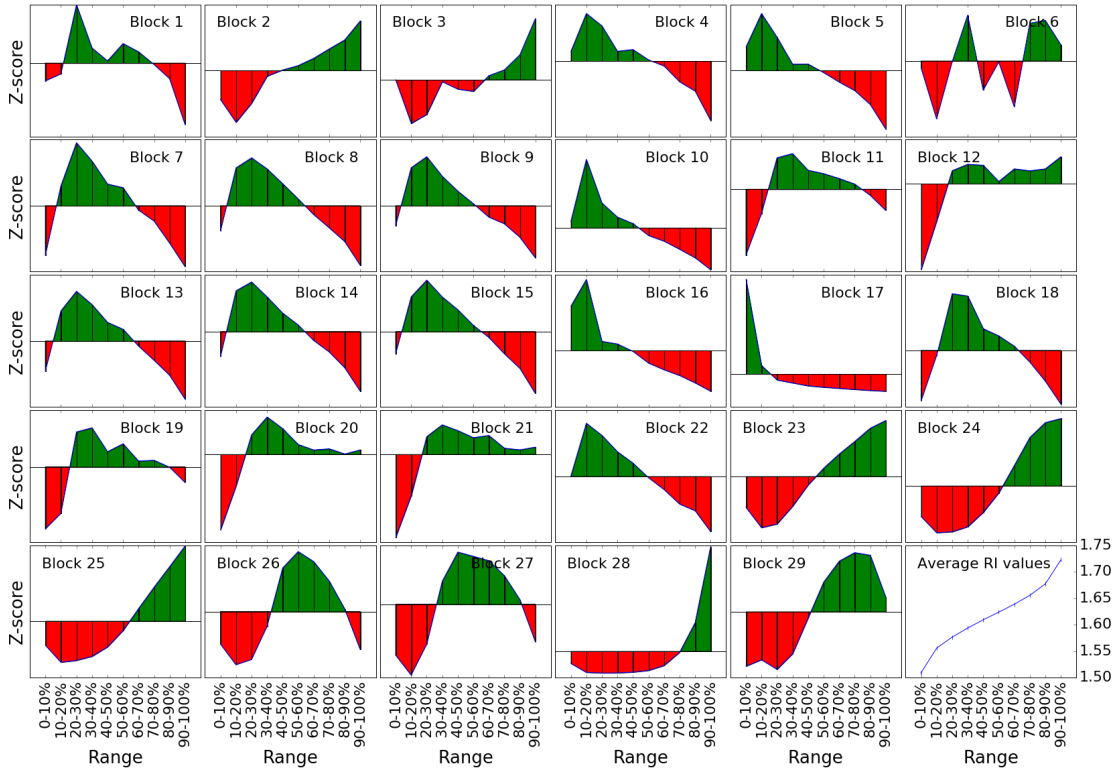


Figure 5.6. Z-score of each building block in all the R_2 structures.

the top 10% candidates, but not the remaining ones. In order to generate a more comprehensive structure-property relationship, we evaluated the Z-score of building blocks in the full spectrum of the library. For this, the library was divided into 10 subsets ordered by increasing RI values. We evaluated the Z-score of the building blocks in each of these 10 subsets and plotted them with increasing RI values (see Fig. 5.6). The green color represents a positive Z-score and the red represents a negative Z-score. We observed certain trends in this plot:

1. The Z-score of building blocks 2, 3, 23, 24, 25, 28 and 29 increase with increasing RI values.

2. The Z-score of building blocks 4, 5, 10, 16, 17 and 22 decreases with increasing RI values.

3. The Z-score of building blocks 7, 8, 9, 11, 13, 14, 15, 18, 26 and 27 change from a negative value to a positive value before becoming negative again.

4. The building blocks 1, 6, 12,19, 20 and 21 do not show a clear trend, therefore appear to have less impact on the RI values

In addition to building blocks contribution to RI values, it is also important to note the impact of various building block combinations to further understand the structure-property relationship. Therefore, we calculated the Z-scores of all the possible building block combinations in the top candidates. Fig. 5.7 illustrates which building blocks combinations are promising for high RI polymers. The size of the circle in the plot represents the magnitude of the Z-score value. The green color of the circle indicates a positive Z-score, whereas the red circle indicates a negative value. Using this figure, we can see that the combination of

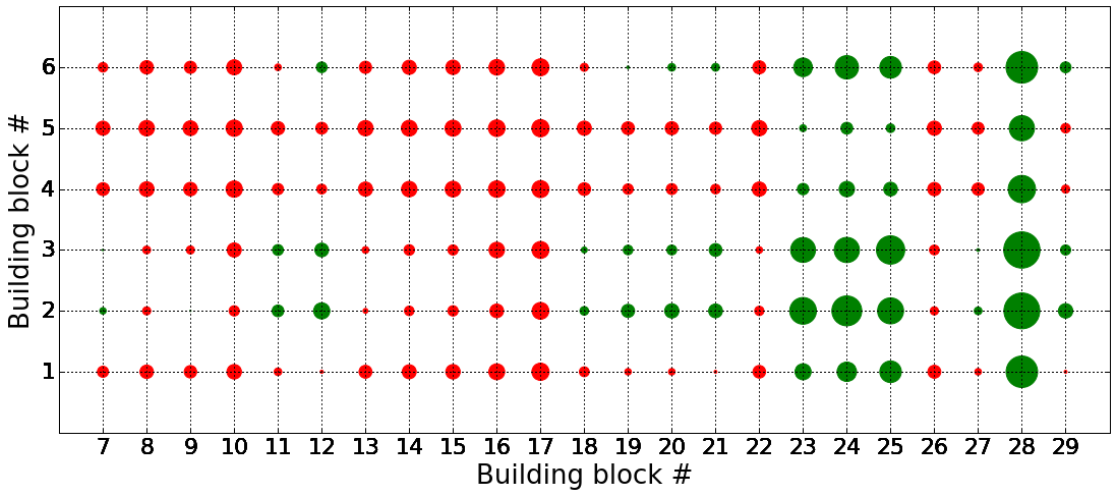


Figure 5.7. Z-score of building block combinations in the top 10% candidates of R_2.

building block 28 with blocks 2 and 3 occur very frequently in top candidates given by the large size of the circle. The combinations of building blocks 23, 24, 25 and 28 with building blocks 1, 2 and 3 appear to be promising for the creating HRIPs.

The above structure-property relationships demonstrate the best building blocks and the combinations of these building blocks which are highly promising in developing HRIPs. We will use these guiding principles to create the next generation of HRIPs, and further validate these candidates by synthesizing them and testing their RI values. The best candidates from this study are currently being synthesized by our experimental collaborators, which we will be publishing soon.

5.4 Conclusions

We demonstrated the developed RI model is successful in calculating the RI values of polyimides. We successfully created a molecular library of 270,000

PIs from promising building blocks, suggested by our experimental collaborators. We identified the best candidates by screening the library using our virtual high-throughput screening framework. The screening study resulted in more than 2000 PIs with RI values greater than 1.8. In addition to identifying high RI candidates, we applied materials informatics and data mining techniques to understand the relationship between the molecular structure and the RI values. Using these techniques, we identified building blocks and combinations of building blocks that are prominent in the top RI candidates. The developed rational design framework is successful for the accelerated discovery of polyimides with exceptional RI values.

Chapter 6

Neural Networks for the Prediction of RI of 1.5 Million Organic Molecules

One of the most important and non-trivial properties of these molecules is the molecular packing. The packing density of organic molecules is highly dependent on the molecular structure. Organic molecules with a targeted density can be achieved by controlling the chemical structure. We can understand more about the correlation between the molecular structure and packing density of molecules by studying a large number of molecular candidates. Using our in-house molecular generator, we generated a molecular library of 1.5 million small organic molecules. The most accurate way to evaluate the density of these molecules is by performing molecular dynamics (MD) simulations. Since MD simulations are computationally intensive, performing simulations for a large library of molecular candidates is not a viable option. Therefore, we selected hundred thousand molecules from the library and evaluated their density

values. Using this data as training set, we developed an exceedingly efficient (R^2=0.98) neural network model to correlate molecular structural descriptors with their properties. Using this correlation, we were able to quickly compute the density values of 1.5 million organic molecules. We mine this huge data to obtain insights into the correlation between the molecular structure and density of materials. Additionally, we evaluate the learning curve for the density prediction to obtain a correlation between the model efficiency and size of the data. We demonstrate that the developed machine learning approach is a powerful tool and has shown to be highly promising for rapidly identifying molecular candidates with a targeted density.

We thank Dr. Andrew Schultz for helpful discussions on the molecular modeling of organic molecules. Aditya Sonpal performed the literature review for the packing density of materials. We also thank Mojtaba Haghighatlari for helpful discussions on machine learning techniques.

6.1 Introduction

In the current century, there is a pressing demand for discovering new materials with superior mechanical, optical and other bulk properties and a negligible negative impact on the environment. The traditional trial-and-error approach for experimental research has proved to be time consuming and resource intensive. The advent of stable and efficient computational power has paved the way for computational and data-driven research to take the center stage in materials discovery. Machine learning has been established as a reliable approach to accelerate prediction of material properties and materials discovery [18]. The ability to predict the properties of novel materials prior to synthesis, and to

understand the relationships between the microscopic properties of molecular components and the macroscopic materials properties, would be of substantial benefit to materials discovery [20]. There is a strong need for machine learning methods that can generate robust, predictive models linking these microscopic properties. Artificial Neural Networks have been widely accepted as an efficient computational model used in machine learning. Multi-layer NNs have been used to predict properties like RI [119], dielectric constant [74], atomization energy [171], chemical reactivity [172], melting point [173], viscosity [174], solubility [175], etc.

Focusing on the increasing demand for complex and intricate optical appliances, it has become imperative to find easily processable materials with superior optical properties, good mechanical integrity, eco-friendly nature and economically viability. A holistic materials approach has brought to light the viability of organic High Refractive Index Polymers (HRIPs) for such optoelectronic devices [48, 176]. Since linear optical properties are dependent on packing density and polarizability [94], an accurate model for its calculation would play an important role in the computationally driven quest to discover suitable HRIPs. Both, packing density and polarizability can be improved by controlling the molecular or chemical structure and an efficient model to calculate these would allow us to tune various properties to develop suitable HRIPs in accordance with their targeted area of use [177]. In addition to HRIPs, the packing of particles in a given volume has garnered the interest of researchers from a variety of disciplines for over a century [178]. Packing density, directly impacts ionic conductivity [179] or mobility in solvents [180], optical properties [77, 94] and other applications in physical organic chemistry [181].

The goal is to select suitable candidates from a large library of molecules.

For this, it is primarily essential to establish a protocol for accurately calculating packing density and polarizability. We employed three different methods to calculate the density of commonly used small organic molecules and the results are compared to their corresponding experimental values. This comparison displays the supremacy of molecular dynamics (MD) as the most accurate method for density calculation. For polarizability, DFT calculations are performed at DFT level. As the DFT and MD computations of a large library of molecules are computationally expensive, we develop a NN models to accelerate the property predictions. We apply *ChemHTPS* framework to automate the evaluation of density and polarizability values of a hundred thousand molecules and use the resulting data as the training set for learning the NN model. This enables us to accurately and swiftly predict the values for the larger data set of 1.5 million molecules.

In addition to identifying the best candidates from our high throughput screening studies, we analyze the collected data to further understand structure-property relations. The building blocks that contribute the most towards high density, polarizability, and RI values are identified using Z-score analysis. Additionally, the building blocks that are prominent in a particular property interval are also identified using Z-score mapping for the entire property range.

6.2 Methods

6.2.1 Library Generation using *ChemLG*

Controlling density of materials with different functional groups by tailoring the chemical structure is rather difficult. It could potentially lead to an infi-

nite number of candidate molecules. Therefore it is impractical to empirically characterize a large number of candidates. On the other hand, computational analysis allows greater exploration at a fraction of the time and cost. A large candidate library of 1.5 M small organic molecules is generated using our in-house molecular library generator, *ChemLG* [43]. The library is generated using combinatorial linking of 15 building blocks (see Fig. 6.1) for 4 generations, while enforcing constraints like molecular weight within the range of 150 to 400 and setting the maximum number of rings to 4. The complete library can be found in our database.

We use Performing DFT calculations and MD simulations for 1.5 M molecules is not a viable option. Therefore, we select a small subset of the library to perform these calculations with the aim of using the data to build machine learning models. We randomly select 100,000 molecules and evaluate the properties using our HTPS framework.

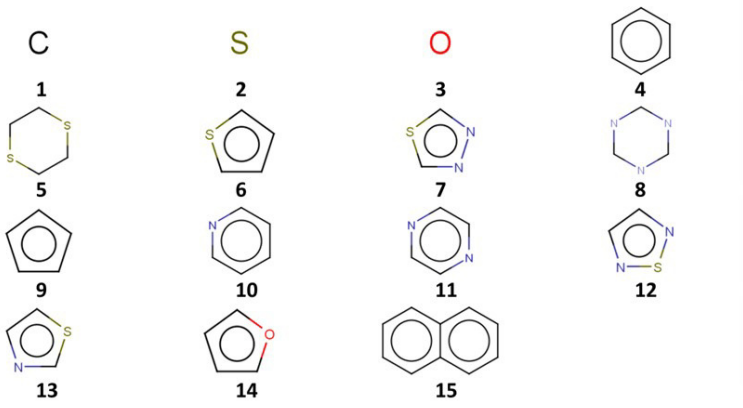


Figure 6.1. Building blocks used to build a library of 1.5 million molecules.

6.2.2 Density Prediction

6.2.2.1 Bondi's Method

As per Bondi's method, we obtain the van der Waals volume (V_{vdw}) from the contribution of individual atoms [97] and assume the packing fraction to be a constant. Applying correctional factors to the sum of atomic volumes, we use equation 6.1 to obtain V_{vdw}, where V_i is the volume of atom i, N_B is the number of bonds, N_{ar} is the number of aromatic rings and N_{non-ar} is the number of non-aromatic rings in the molecule [96]. We calculate the molecular volume from V_{vdw} by using Slonimskii's method as shown in equation 6.2, where K_p is the packing constant of the material [108]. Once the molecular volume is obtained, we compute the density of the material (ρ) by using the equation 6.3.

$$V_{vdw} = \Sigma V_i - 5.92N_B - 14.7N_{ar} - 3.8N_{non-ar} \tag{6.1}$$

$$V_{mol} = \frac{V_{vdw}}{K_p} \tag{6.2}$$

$$\rho = \frac{M_W}{N_A V_{mol}} \tag{6.3}$$

6.2.2.2 Wavefunction Method

Similar to the previous method, the wavefunction approach also calculates the density from the van der Waals volume using equation II.2 and II.3. The only difference lies in the fact that the van der Waals volume here is obtained from

the wavefunction of molecules using the Monte Carlo technique. We obtain the wavefunction of molecules from quantum calculations using the GAMESS quantum chemistry code [182]. In the Multiwfn software, we then define a box of volume L that fits the entire system [95]. Using the Monte Carlo principle we randomly distribute N particles in the box and if n particles are present within the van der Waals region, then the van der Waals volume of the system is given by $n/N * L$.

6.2.2.3 Molecular Dynamics Method

The third method that we use to calculate density, is the Molecular Dynamics method. We use OpenBabel to create a file representing the 3D structure of the molecule from their SMILES code. We then pre-optimize molecular structures in the MMFF94s force field using the steepest descent algorithm [107]. We compute the packing density of the molecules using the general amber forcefield (GAFF) [183]. We use the Antechamber toolkit within AmberTools to generate GAFF parameters for the molecules in an automated fashion [184]. Our simulations are carried out using GROMACS (GROningen MAchine for Chemical Simulations) [185], which is a molecular dynamics package developed by the Biophysical Chemistry department of the University of Groningen. We use the solvate tool within GROMACS to create a simulation box of 10nm and fill it with pre-optimized molecules. The system is first subjected to a minimization step to minimize the internal energy associated with construction of bonds, bond angles and bond dihedrals. This is followed by NVT and NPT equilibration for 100 ps and 240 ps respectively. Both NVT and NPT ensembles use a Nose-Hoover thermostat at 298.15 K for temperature control, while the NPT ensemble uses the Parinello-Rahman barostat for pressure control. We conclude the MD pro-

cess with a final NPT production run which lasts 40 ps. We compute the final density by averaging out the density values of the system at intervals of 0.2 ps in the production run.

6.2.3 Polarizability Prediction

The molecular polarizability calculations utilize Kohn-Sham density functional theory (DFT) for its advantageous trade-off of cost and accuracy [92]. We applied three different methods to compute the polarizability of the molecules (see Table 6.1). In all these three methods, the single point calculations use an all-electron, restricted DFT framework with the PBE0 hybrid functional [102] along with basis sets by the Karlsruhe group [103] and include Grimme's D3 correction [104] to account for dispersion interaction. In the first method, we optimized the geometries of all molecules using the MMFF94s forcefield as implemented in the OpenBabel software [107]. The selection of molecules for the screening is discussed in the following section. The DFT calculations are carried out using the ORCA 3.0.2 quantum chemistry program package [105] with default settings.

Table 6.1. Different methods used to compute the polarizability of molecules.

Method no.	Geometry optimization	Polarizability method	Molecules characterized
1	MMFF94s	PBE0/def2-SVP - D3	100,000
2	BP86/def2-SVP	PBE0/def2-SVP - D3	100,000
3	BP86/def2-SVP	PBE0/def2-TZVP - D3	10,000

6.2.4 Neural Networks

We generated the NN models within a feature space of 197 molecular descriptors on a training data set of 100,000 molecules with density values computed using molecular dynamics. For the initial model evaluation, the data set of 100,000 molecules was randomly divided into 80% training and 20% test set for testing. The NN modeling was performed using *ChemML* [46], our program suite for machine learning and informatics in chemical and materials research. In this work, *ChemML* employed the scikit-learn 0.18.2 multi-layer perceptron (MLP) regressor 1.17.1 [109] and descriptors from Dragon 7 [110]. We applied grid search method to optimize the hyper-parameters of NN model. The final NN model included two hidden layers having 100 neurons each. We constructed the model using rectified linear unit as the activation function, 'adam' solver for weight optimization, adaptive learning rate, and an L2 regularization parameter of 0.0001.

To evaluate the learning curve, we increased the training set size in increments from 0.5% to 100%. For every training set size, we applied bootstrapping method to evaluate the model performance. This method includes randomly selecting the training set from the complete data set and testing on the remaining data set. The process is repeated 50 times by replacing the previous test set, i.e. all the 50 repetitions are independent of each other. The mean and deviation of R^2 value from these 50 repetitions for each training size is used for plotting the density learning curve.

6.2.5 Virtual High-Throughput Screening using *ChemHTPS*

To facilitate density evaluation of the large library that was generated in a timely and efficient way, we use our virtual high throughput screening framework *ChemHTPS*. *ChemHTPS* draws inspiration from the Harvard Clean Energy Project which successfully screened millions of organic molecules for photovoltaic applications. Not only does it create inputs, executes and monitors the calculation but this virtual framework also parses and assesses the results. In the end, it extracts and processes the information of interest, inserts the key outcomes into the project database and archives all the other data. Having said this, it cannot be ignored, that performing MD simulations on 1.5 M molecules will be extremely time consuming. Therefore a small subset of 100,000 molecules is selected from the library to perform MD calculations with the aim of using the resulting data to train a NN model to predict the density of the entire library.

6.3 Results and Discussion

6.3.1 Molecular Methods

Method	R^2	MAE [1] (Kg/m^3)	MAPE[2]
Bondi $(const.\ K_P)$	0.820	45.97	5.62
Wavefunction $(const.\ K_P)$	0.856	29.31	4.91
Molecular dynamics [76]	0.990	7.50	1.11

Table 6.2. Results for all the molecular methods. [1]Mean absolute error, [2]Mean absolute percentage error.

The comparison between the three methods to calculate density was made using the experimental density value of 22 small organic molecules. From the

table 1, we observe that the wavefunction method is better than the Bondis method, but the MD method is significantly better compared to the first two methods. This shows that MD would be the best choice to evaluate more accurate density values. Therefore, we developed a high-throughput screening (HTPS) framework using the density protocol to automate the process of calculating the density of large library of organic molecules. To test the accuracy of the developed MD protocol for calculating the density, we collected experimental density values of 175 small organic molecules with densities varying from 600 kg/m^3 to 2000 kg/m^3 [186]. We observed a good agreement (R^2=0.95) between the calculated density and the experimental density.

6.3.2 Neural Networks

Using our in-house molecular library generator, we generated 1.5 million small organic molecules. We randomly selected 100,000 molecules from the 1.5 million library and applied the MD protocol to evaluate the density of these molecules. This was done by casting the MD protocol into our ChemHTPS framework. These computations took a total of 3 months wall time on 200 compute nodes each equipped with 16 cores. The density values of randomly chosen 80% molecules from these 100,000 molecules were then used to train a NN model. The resultant model was exceedingly efficient with R^2 (training set) = 0.98, R^2 (test set) = 0.97. The efficiency of the model can be seen in Fig. 6.2. The benchmark comparison for the test set gives a mean absolute deviation (MAD) of 7.96 kg/m^3 (0.95%), a root mean square deviation (RMSD) of 10.05 kg/m^3 (1.20%), and a maximum deviation (MaxD) of 58.82 kg/m^3 (6.18%), respectively, i.e., the NN model is quite accurate for the prediction of density. In addition to the

quantitative comparison, we observe, qualitatively (from the Fig. 6.2), that the deviation in the distribution of training set is similar to the test set.

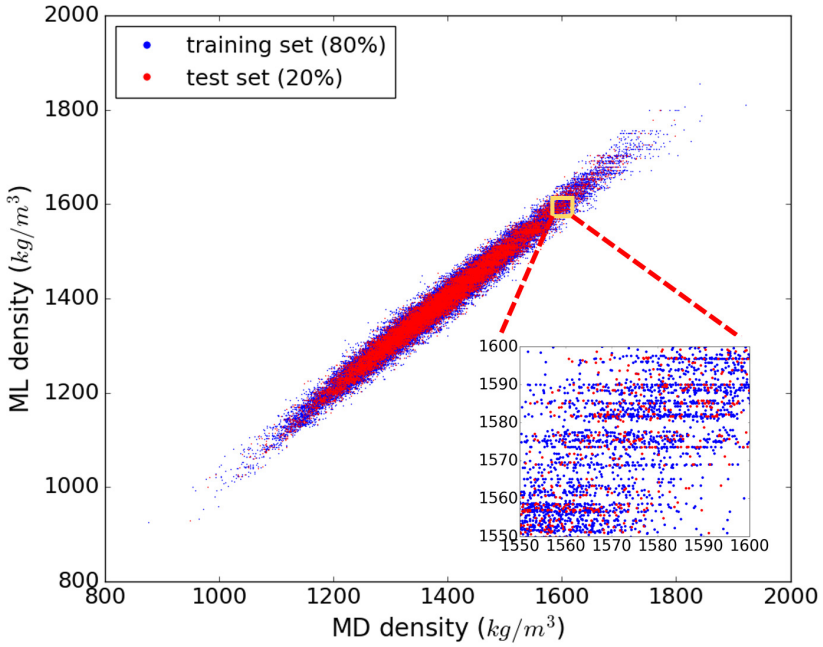


Figure 6.2. Comparing the calculated density (MD) to the predicted density (NN)

Using the developed NN model we were quickly able to calculate the packing density of remaining 1.41 million molecules in the library. The density of these molecules range from 864 kg/m^3 to 1645 kg/m^3 with an average value of 1261 kg/m^3. The distribution of the density of the molecules show a Gaussian distribution with a sigma of 0.084 and variance of 0.007 (see Fig. 6.3). A low variance in the data indicates that most of the candidates in the library have density values in the range 1200-1600 kg/m^3. The number of molecules with density greater than 1600 kg/m^3 are very few. Therefore, by a combination of large-scale HTPS and Machine learning, candidates with high density and low density values were identified, which would otherwise not be feasible through empirical studies. Compounds with such extreme packing density values would cater to

a specific type of applications, such as, in discovering materials with superior optical properties like refractive index.

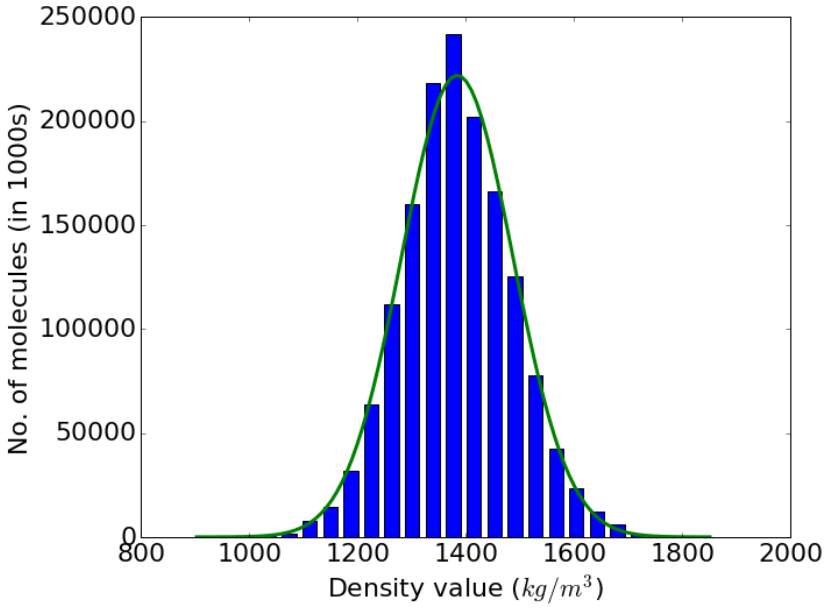


Figure 6.3. Density vs No. of Molecules

6.3.3 Analyzing Relationship between Molecular Structure and Density

When considering screening results from a large candidate library, we are not only in a position to assess a large number of compounds, but we can also learn patterns from the data set in its entirety. Thus, in addition to identifying candidates with a targeted density value, we evaluated structure-property relationships by understanding the effect of building blocks on the density of the molecules.

We computed the average density values of all the candidates that contain a particular building block and then ranked the building blocks. The molecules containing building blocks 7 and 12 have the highest average density values,

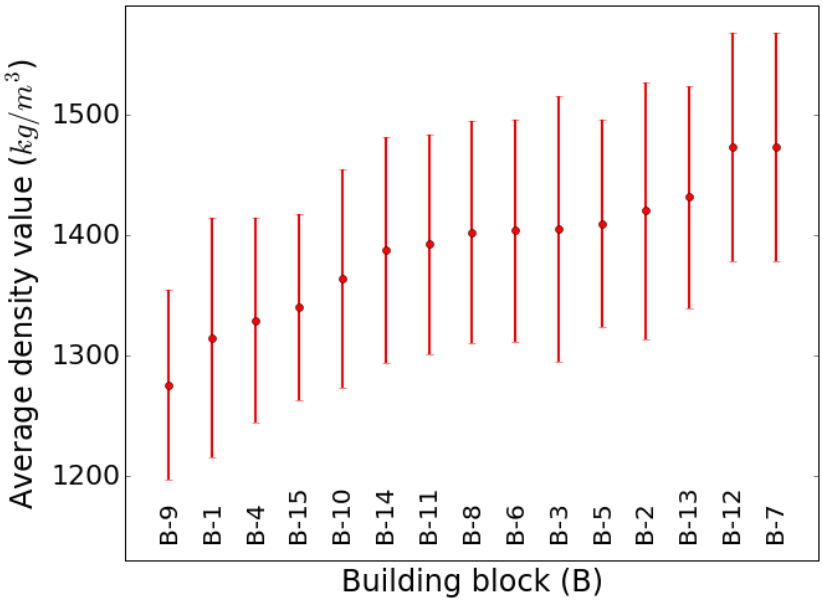


Figure 6.4. Average density for each building blocks

while molecules containing building blocks 1 and 9 have lowest average density values (see Fig. 6.4). The average density values shows the cumulative performance of a particular building block, however, it does not present information on the distribution and occurrences of building blocks in a particular subset. Therefore, to generate a more comprehensive structure-property relationship, we evaluated the Z-score of each building block. The Z-score is an indicator of the occurrence of a building block in a subset of the library. The Z-score (Z_i) of the building block is calculated by

$$Z_i = \frac{k - n\frac{K}{N}}{\sigma}, \; \sigma = \left[\frac{nK}{N} \times \left(\frac{N-K}{N}\right) \times \left(\frac{N-n}{N-1}\right)\right]^{\frac{1}{2}}$$

where N is the total number of molecules, n is the subset of molecules that are considered, K is the number of occurrences of building block i in N molecules and k is the occurrences of building block i in the subset. A large Z-score indi-

cates that a building block appears more frequently in that subset compared to rest of the library. For example, we evaluated the Z-score of building blocks in the 10% candidates with high density values as shown in the Fig. 6.5. We observe that building blocks 7 and 12 have high Z-score values, i.e. these building blocks are more prevalent in high density molecules, which is in good agreement with the results from the average density rankings of building blocks shown in Fig. 6.4.

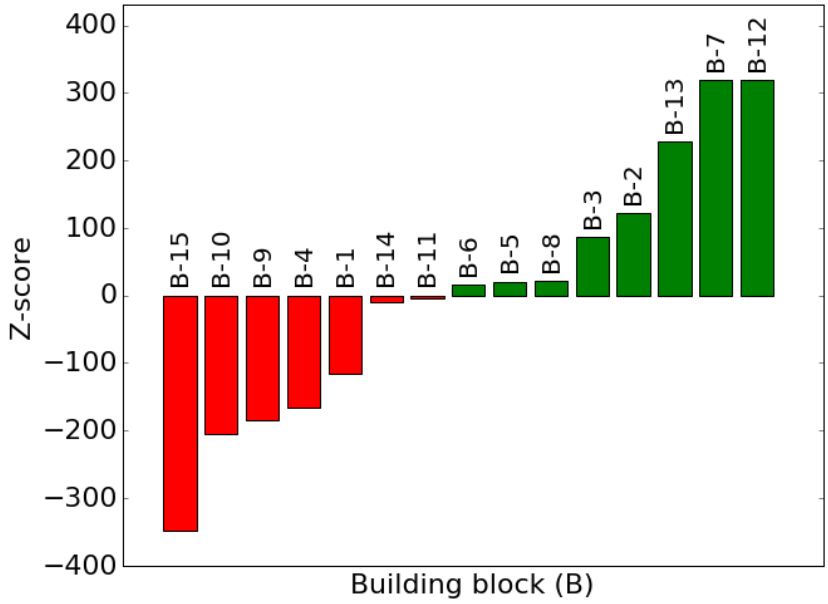


Figure 6.5. Z-score values for each building block

In addition to the top 10% candidates, we use the Z-score analysis method to identify the prevalence of building blocks in the complete spectrum of density values. For this, we sorted the molecular library with increasing density values and then divided into ten equal subsets. Subsequently, the Z-score values of each building block were evaluated in each of these ten subsets and plotted in Fig. 6.6. The green color in the plot represents a positive Z-score and the red color represents a negative Z-score. Based on the data, we identified certain

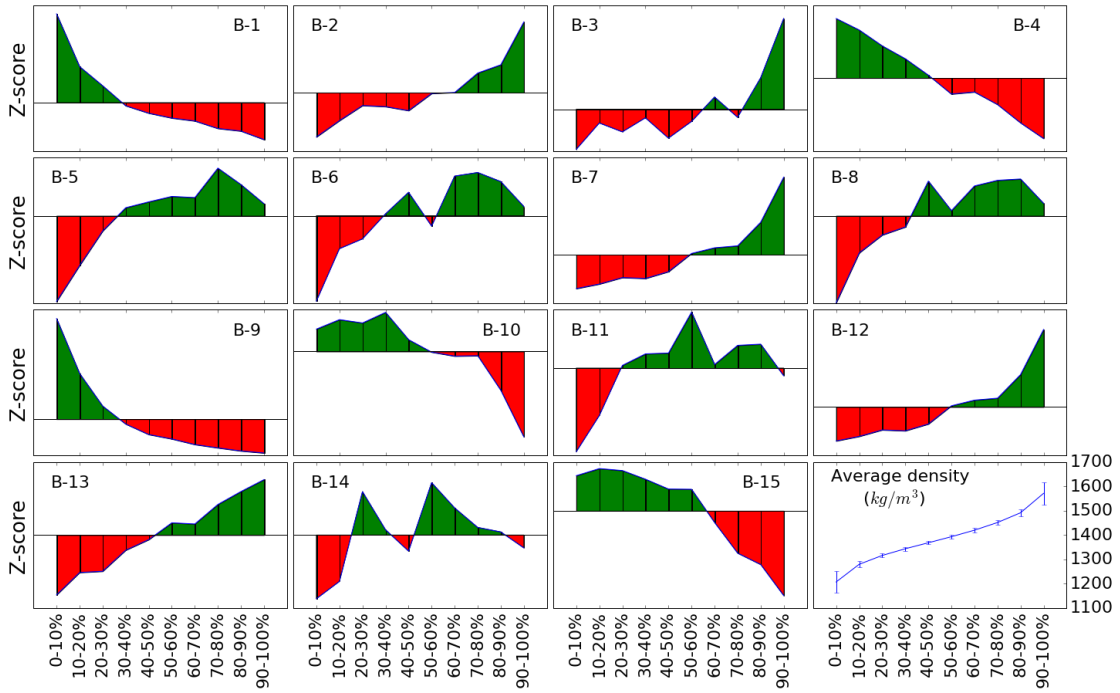


Figure 6.6. Z-score of all the building blocks in the subsets with increasing density values.

trends in the performance of individual building blocks. For example, building blocks 2,3,7,12, and 13 would be ideal candidates to develop organic molecules with high density, whereas building blocks 1, 4, 9, 10 and 15 would be better to generate molecules with lower density.

6.3.4 Learning Curve Analysis

One of the most challenging questions in applying machine learning to learn material properties is what amount of data is required to produce an efficient model. For example, in the current studies we selected 100,000 molecules to learn the density property. However, the question is if we can learn the density from lesser number of molecules. This will allow us to invest less computational resources to perform expensive MD calculations. For this, we varied the size of

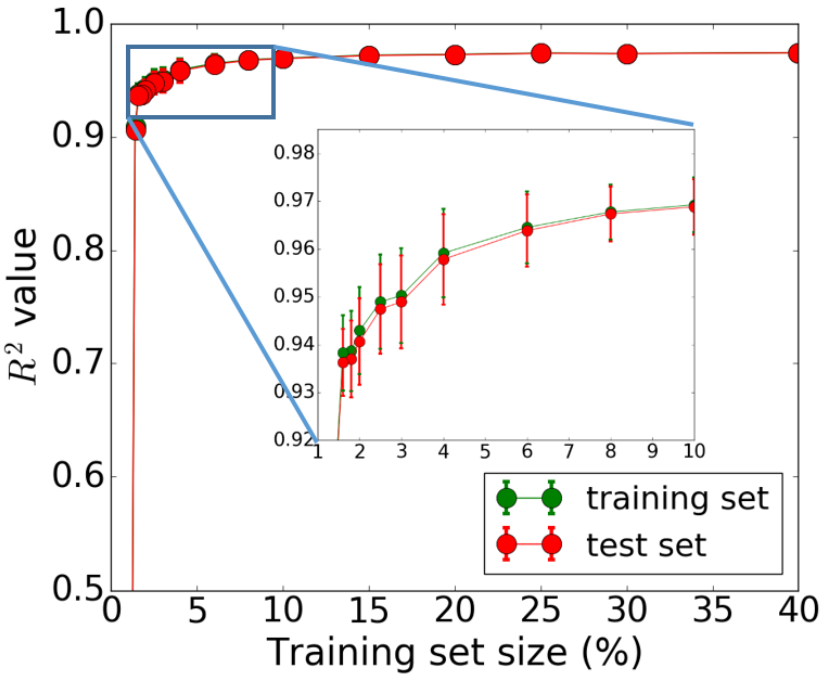


Figure 6.7. Dependence of model accuracy (R^2) on the training set size.

the training set from 100 to 80,000 molecules (see Fig. 6.7). For all the data sets less than 1800, the model was not able to learn. However, an efficient model can be obtained from using only 2000 molecules. The R^2 value keeps increasing until 6000 data sets and then plateaus to a constant value with increasing training size. Thus, we do not require a large data set of 100,000 molecules to learn the density of organic molecules. We can develop an exceedingly efficient machine learning model just by using 2-6 thousand molecules. This has significant implications on the amount of computational power required to predict the properties of large library of materials.

6.3.5 Descriptors

In the inset of Fig. 6.2, it can be observed that clusters of the molecules of have similar predicted density value (suggested by the horizontal lines). This might

be due the incomplete description of molecules in the selected descriptors. We believe more accurate prediction models can be developed by selecting more comprehensive descriptors such as hashed topological torsion (HTT), Morgan fingerprints, HAP, MACCS, 3D descriptors from Dragon etc. [110]. In addition to improving features we are also working to implement different deep neural network architecture to improve the learning efficiency. This framework of using machine learning to predict the density of molecules can be extended to learn other material properties. We are currently working on implementing these techniques to predict the optical properties of organic molecules and we will be publishing the results soon.

6.3.6 High RI Candidates

We performed the similar analysis for the polarizability calculations of all the 1.5 million candidates. Using the density and polarizability values in the Lorentz-Lorenz equation, we obtained the RI value of all the candidates. This allowed us to identify candidates with high RI values and extract underlying structural patterns for optimizing RI values. We can visualize the regions in the chemical space where we can maximize the RI value by mapping of 1.5 million data points as shown in the Fig. 6.8.

6.4 Conclusions

We successfully generated a molecular library of 1.5 million small organic molecules using our in-house molecular library generator. We performed DFT and MD simulations on 100,000 compounds to evaluate their polarizability and

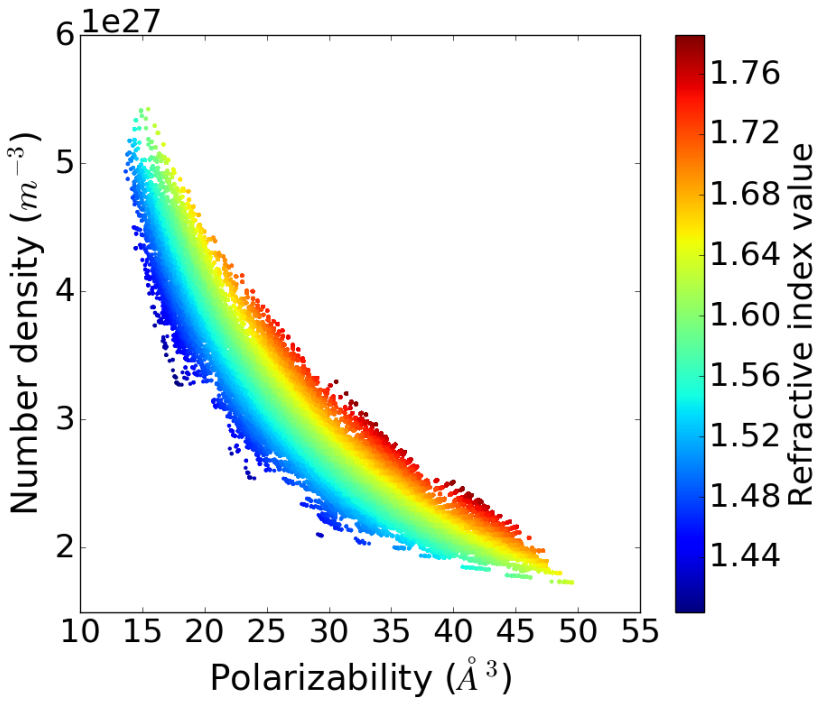


Figure 6.8. Mapping of 1.5 million molecules to identify regions of high RI values.

packing density values, respectively. Using the data from this high-throughput screening, we successfully developed an exceedingly efficient (R^2=0.98) neural network model to correlate molecular structural descriptors with their packing density. Using this correlation, we were able to accelerate the density computations of 1.5 million molecules. We mined this huge data and obtained insights into the correlation between the molecular structure and density of materials. Additionally, we evaluated the learning curve for the density prediction to obtain correlation between the model efficiency and size of the data. We performed similar analysis for the polarizability values, thus, allowing us to identify targets that can maximize the optical properties.

We demonstrate that by combining computational modeling and machine learning, we can rapidly and efficiently assess the properties and performance potential of candidate compounds. By combining molecular modeling with vir-

tual high-throughput screening techniques and machine learning, we characterized candidates on a massive scale. We identified design rules as well as high-value building blocks, and structural patterns that correlate with the packing of molecules. These guidelines allow us to target specific molecular motifs and create next generation materials with targeted properties.

Chapter 7

Summary and Outlook

7.1 Conclusions

We have successfully developed an *in silico* modeling protocol which accurately and efficiently predicts RI values of organic polymeric materials. This has also demonstrated its superior performance by faithfully reproducing the experimentally known RI values of 112 compounds. We have benchmarked DFT approaches for calculation of polarizability, which require RI calculations as an input. Our work provides an example of the synergistic benefits of fusing physical and data models. Furthermore, it is a clear embodiment of the promising nature of machine learning and modern data science in chemical research.

We utilized this new RI protocol to conduct virtual high-throughput screening studies on large-scale candidate libraries to discover novel organic polymers with high RI values. We demonstrated that computational modeling can rapidly and efficiently assess the properties and performance potential of candidate compounds. By combining our RI prediction model with virtual high-throughput screening techniques, we characterized polyimide candidates on a

massive scale. We identified design rules as well as high-value building blocks, and structural patterns that correlate with the RI values. Additionally, we identified regions in the chemical space where we can maximize the RI values of organic polymers. These guidelines allow us to target specific molecular motifs and create next-generation polymers with exceptional optical properties.

Using our in-house molecular library generator, *ChemLG*, we have generated a molecular library of 1.5 million small organic molecules. We performed DFT and MD simulations on a subset of this library, 100,000 compounds, to evaluate their polarizability and packing density values, respectively. By using the data from these simulations as a training set, we developed exceedingly efficient neural network models to correlate molecular structural descriptors with the target properties. Application of neural network models resulted in the acceleration of the property prediction of 1.5 million molecules. We mined this huge data and obtained insights into the correlation between the molecular structure and density of materials. Additionally, we evaluated the learning curve for the density prediction to obtain a correlation between the model efficiency and size of the data. We performed a similar analysis for the polarizability values, thus, allowing us to identify targets that can maximize the optical properties.

The developed rational design framework is a powerful tool and has shown to be highly promising for rapidly identifying molecular candidates with exceptional RI values as well as discovering design rules for advanced materials. This dissertation serves as a proof-of-principle for our software ecosystem, which recognizes the great opportunities that are appearing with the shift towards a data-driven *in silico* research paradigm in chemistry, materials science, and the corresponding engineering disciplines. We have shown *via* this project that this approach indeed offers a path to overcome some of the prevalent limitations

of traditional trial-and-error approaches. Our aim is to extend the capabilities of our cyberinfrastructure to tackle complex discovery and design challenges, increase the rate and quality of innovation, improve our understanding of the associated molecular and condensed matter systems, and democratize the tools that make these developments possible. The long-term objective of our work and related efforts by others is to help pioneer a fundamental transformation of the discovery process in chemistry, to make data science an integral part of the chemical enterprise, to shape the transition towards a data-driven discovery and rational design paradigm, and to spearhead a broad move by the community along those lines.

7.2 Challenges for Organic Materials in Optical Applications

In addition to the RI property, it is also very important to evaluate other optical properties such as transparency, Abbe number, and birefringence.

- **Optical transparency:** Although most of the highly polarizable moieties increase the RI of the organic polymers, it is possible that the optical transparency of the material is affected. For example, in case of the polyimides, there is a possibility of formation of intermolecular charge-transfer complexes leading to the coloration of the material [187]. The complexes are formed due to the charge transfer between electron accepting dianhydride and electron accepting diamine moiety. Thus, overcoming this trade-off between the optical transparency and RI is crucial.

- **Abbe number:** The high RI polymers that are being developed by incor-

porating highly polarizable moieties typically have low Abbe numbers. Abbe number is inversely proportional to the optical dispersion and therefore materials having low Abbe numbers will have high optical dispersions. Such materials are generally not preferable in optical devices because of reduced chromatic aberration [71]. One of the approaches for improving the Abbe number is to include saturated moieties into the polymer as a side chain [188]. For example, a compact and saturated structure like diamondoid can be incorporated as a side chain to tune the Abbe number of the polymer [189]. Diamondoids not only increase the Abbe number, but as they have high molar refraction they also concomitantly increase the RI of the polymers [71, 190]. However, the RI of these polymers is not quite high to be used in optical applications where high ($>$1.7) RI materials are required. Thus, it a challenging topic to optimize the structure of the polymers that have both high RI and high Abbe number values.

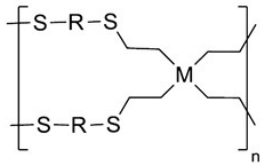
- **Birefringence:** Another vital property that should be considered in developing organic optical polymers is the birefringence parameter. If a material has high birefringence value, the polarization state of the incident light can be degraded resulting in a poor performance of the optical device, hence it is critical to consider the birefringence when designing materials for optical devices. The method of developing high RI polymers by introducing highly polarizable moieties often also results in the polymers with high birefringence. Thus, it is a challenge to design polymers having high RI and at the same time having low birefringence. Numerous efforts are being made to lower the birefringence of organic polymers. Recently, it was reported that polyamides show lower birefringence com-

pared to polyimides while both have similar RI [117]. Quite recently, Seto *et al.* have attempted to develop aromatic polyesters and polycarbonates which are shown to have low birefringence along with having moderately high RI [64]. Although these polymers have very attractive birefringence values, the RI of these polymers is still not high enough for optical applications. Further improvement of the RI of such polymers while keeping low birefringence is possible by optimizing the structure of the polymers in a favorable fashion.

Cross-linked structures (hyper-branched polymers): Many optical applications not only require materials with superior optical properties but also need them to have good mechanical integrity. Cross-linking of polymer networks has shown to increase the mechanical stability of the polymer materials [65, 191, 192, 193, 194]. It has been observed that the cross-linking of high RI polymers concomitantly results in a further increase of RI [65]. In addition to this, cross-linking can also result in materials that have high optical transparency, low optical dispersion, low birefringence, and increased thermal stability [194, 193]. For example, a very recent study suggests that the cross-linked phenylthiophenyl silicone structures not only showed high RI values but also exhibited high thermal stability [192]. Thus, developing cross-linked polymer networks can lead to materials with very promising properties such as good mechanical and thermal stability along with high RI, low birefringence, high optical transparency and low optical dispersion.

Recently, Bhagath et al. reported high RI thiol-ene polymers which have high packing density [191, 195, 196, 197]. The RI of these polymers was increased by incorporating functional moieties like aromatic groups, sulfur and other highly

polarizable atoms like Si and Sn as shown in Fig. 7.1. These polymers show high RI (>1.7) as well as have very good mechanical strength. Further, the processing of these polymers is relatively easier compared to other polymers. We performed a few preliminary studies on these polymers (see App. B). Our plan for the future is to apply our materials discovery framework to discover new thiol-ene polymers with exceptional RI values.



R = m-C_6H_4, M = Si,
R = m-C_6H_4, M = Sn,

Polymer	Density (gm/cm^3)	Refractive index (589 nm)
M=Si	1.43	1.687
M=Sn	1.24	1.703

Figure 7.1. Thiol-ene polymers with high packing density and high RI [191].

7.3 Improving our Cyberinfrastructure

We are currently working on the following aspects of our cyberinfrastructure to further improve its applicability:

- *ChemLG*: The library generator currently implements the genetic algorithm for linear molecules. Ongoing work in this area aims to extend this algorithm to traverse molecules in multiple dimensions. Further, we plan to implement other smart algorithms such as Monte Carlo search and Simulated Annealing.

- *ChemHTPS*: Currently, *ChemHTPS* is setup to run on SLURM queuing systems. We plan to extend *ChemHTPS* to other queuing systems such as PBS. It currently supports the ORCA [105], Q-Chem [129], and GROMACS

[130] modeling packages, and bindings to other quantum chemistry, molecular dynamics, and solid state physics codes are planned for the future.

- *ChemBDDB*: The future work in *ChemBDDB* includes the integration of external databases to extract information on existing molecules or obtain data using cheminformatics from the websites such as ChemSpider.

- *ChemML*: We plan to extend *ChemML* capabilities by implementing new advances in machine learning, such as convolutions, transfer learning, active learning, and reinforcement learning. Mojtaba Haghighatlari is taking a lead on adding these functionalities as well as applying them to improve the packing density, polarizability, and RI prediction models

In addition to implementing new algorithms and extending our cyberinfrastructure's applicability, we continuously focus on the ease of use, workflow, and code integration to make this technology more accessible to the community.

Appendix A

A.1 Polarizability

The presence of an external electric field causes the electron distribution of a molecule (or an atom) to rearrange from its ground state configuration and the molecular geometry to distort. A typical source of an electric field is light (i.e., electromagnetic radiation), and a common situation for the above process is light-matter interaction. The change of the electronic structure results in an induced dipole moment within the molecule. The quantum mechanical description of this process involves the addition of the electric field interaction to the molecular Hamiltonian and the resulting Schrdinger equation is commonly solved employing a perturbation approach. The response of the electronic structure following the perturbation of the original Hamiltonian can be considered using response theory. If the applied electric field is weak, the induced dipole moment (μ_I) varies linearly with the applied electric field strength (ε) and the constant of proportionality is the polarizability (α). As the electric field strength increases, higher order terms in the field strength have to be included in the description of the induced dipole moment as shown in eqn (1). The total dipole moment (μ) is the sum of permanent (μ_0) and induced (μ_I) dipole moment. The

coefficients of the higher order terms are the hyperpolarizabilities of different order (β,γ,). The latter are the source of important non-linear optical (NLO) effects.

$$\mu = \mu_0 + \alpha\varepsilon + \frac{1}{2}\beta\varepsilon^2 + \frac{1}{6}\gamma\varepsilon^3 + ... \tag{A.1}$$

The relationship between the polarizability and energy of a molecular system can be obtained *via* a Taylor expansion of the energy in the electric field strength as shown in eqn (2). The derivative of the energy with respect to the electric field is simply the negative of the dipole moment as shown in the equation (3). Comparing eqns (1) and (3), we can readily obtain the expressions for permanent dipole moment, polarizability, and hyperpolarizabilities as shown in eqn (4). These expressions give the link between the properties we are interested in and the energy, which can be calculated using the perturbation theory.

$$E = E(0) + \left(\frac{dE}{d\varepsilon}\right)\varepsilon + \frac{1}{2!}\left(\frac{d^2E}{d\varepsilon^2}\right)\varepsilon^2 + \frac{1}{3!}\left(\frac{d^3E}{d\varepsilon^3}\right)\varepsilon^3 + ... \tag{A.2}$$

$$\mu = -\left(\frac{dE}{d\varepsilon}\right)\varepsilon - \left(\frac{d^2E}{d\varepsilon^2}\right)\varepsilon - \frac{1}{2}\left(\frac{d^3E}{d\varepsilon^3}\right)\varepsilon^2 - ... \tag{A.3}$$

$$\mu_0 = -\left(\frac{dE}{d\varepsilon}\right),\ \alpha = -\left(\frac{d^2E}{d\varepsilon^2}\right),\ \beta = -\left(\frac{d^3E}{d\varepsilon^3}\right),\ \gamma = -\left(\frac{d^4E}{d\varepsilon^4}\right) \tag{A.4}$$

A.1.1 Static Polarizability

The time-independent perturbation expression for the energy is given in eqn (5). The zeroth-order Hamiltonian is the original molecular Hamiltonian, the first-order Hamiltonian is the dipole moment operator $\varepsilon\mu$, and higher-order Hamiltonians do not have to be considered in this problem, so that the energy can be rewritten as eqn (6). Using this expression and eqn (4), the polarizability can be written as eqn (7). Using this expression, the mean (isotropic) electric polarizability can be written as eqn (8).

$$E = E_0^{(0)} + \langle 0|H^{(1)}|0\rangle + \sum_n \frac{\langle 0|H^{(1)}|n\rangle \langle n|H^{(1)}|0\rangle}{\Delta E_{n0}} + ... \tag{A.5}$$

$$E = E_0^{(0)} - \langle 0|\mu|0\rangle \varepsilon + \sum_n \frac{\langle 0|\mu|n\rangle \langle n|\mu|0\rangle}{\Delta E_{n0}} \varepsilon^2 + ... \tag{A.6}$$

$$\alpha = -2 \sum_n \frac{\langle 0|\mu|n\rangle \langle n|\mu|0\rangle}{\Delta E_{n0}} \tag{A.7}$$

$$\alpha = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right) = \frac{2}{3} \sum_n \frac{|\mu_{n0}|^2}{\Delta E_{n0}} \tag{A.8}$$

The SI unit of polarizability is $C^2m^2J^{-1}$. The polarizability volume (α') is also sometimes used instead of the polarizability as this can be a more convenient way to represent the polarizability (eqn (9)). The polarizability volume is

expressed generally in Å^3.

$$\alpha' = \frac{\alpha}{4\pi\epsilon_0} \tag{A.9}$$

A.1.2 Dynamic Polarizability

The eqn (5), which is the time-independent perturbation expression for energy, cannot be used to determine the dynamic polarizability. Thus, in order to determine the dynamic polarizability, $\alpha(\omega)$, we need to solve the time-dependent perturbation theory. Using the time-dependent wavefunction, the mean dynamic polarizability as shown in eqn (10) can be determined. It should be noted that as $\omega \rightarrow 0$, the expression will be reduced to a static polarizability equation. It can also be noted that as $\omega \rightarrow \infty$, the polarizability goes to zero which is because the field changes so rapidly for the electrons to respond to the changing field. The electrons cannot contribute to the induced dipole moment if the applied field is changing very fast.

$$\alpha(\omega) = \frac{1}{3}\left(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}\right) = \frac{2}{3\hbar}\sum_n \frac{\omega_{n0}\left|\mu_{n0}\right|^2}{{\omega_{n0}}^2 - \omega^2} \tag{A.10}$$

A.1.3 Relative Permittivity and the Electric Susceptibility

The electric susceptibility (χ_e) of a medium is defined in eqn (11), where P is the polarization and ϵ_0 is the vacuum permittivity. In presence of an external electric field, a molecule experiences a local field (ε^*), rather than the applied one. The local electric field is the combination of external field and the field

resulting from the electric dipoles present in the system. The local electric field is related to polarization as shown in eqn (12), where N is the number density of molecules.

$$\chi_e = \frac{P}{\epsilon_0 \varepsilon} \tag{A.11}$$

$$\varepsilon^* = \frac{P}{\alpha N} \tag{A.12}$$

The relationship between the local electric field and the applied electric field is given in the eqn (13), which assumes that the medium is a continuous dielectric. This is called as Lorentz local field expression. Using the eqn (12) and eqn (13), the polarization can be expressed as eqn (14). Comparing this expression with the eqn (11), the susceptibility can be written as eqn (15). The relative permittivity (ϵ_r) of the medium is related to the susceptibility as shown in the eqn (16). Thus, the relation between the permittivity and polarizability can be determined as shown as eqn (17).

$$\varepsilon^* = \varepsilon + \frac{P}{3\epsilon_0} \tag{A.13}$$

$$P = \left(\frac{3\alpha N}{3\epsilon_0 - \alpha N}\right)\epsilon_0 \varepsilon \tag{A.14}$$

$$\chi_e = \left(\frac{\alpha N/\epsilon_0}{1 - \alpha N/3\epsilon_0} \right) \epsilon_0 \varepsilon \tag{A.15}$$

$$\epsilon_r = \chi_e + 1 \tag{A.16}$$

$$\epsilon_r = \frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0} \tag{A.17}$$

A.1.4 Refractive Index

Refractive index (RI; n_r) can be related to ϵ_r using the Maxwells equations. The Maxwells equations are shown in the eqn (18), where E is the electric field strength, is the charge density, D is the electric displacement, H is the magnetic field strength and B is the magnetic induction. The units for these equations are given in SI units. The relations between these fields can written as shown in eqn (19), where ϵ_r is the electric permittivity and μ_r is the magnetic permeability.

$$\nabla D = \rho,\ \nabla B = 0,\ \nabla \times E = -\frac{\partial B}{\partial t},\ \nabla \times H = \frac{\partial D}{\partial t} \tag{A.18}$$

$$D = \epsilon_r \epsilon_0 E,\ B = \mu_r \mu_0 H \tag{A.19}$$

The propagation of light through any medium can be determined by the wave equation, which can be derived by considering the eqn (20). The left hand side of this eqn can be simplified as shown in eqn (21), whereas the right hand side can be simplified as shown in eqn (22). Comparing these simplified eqns yields the eqn (23). Thus, the velocity of light in the medium can be written as shown in the eqn (24).

$$\nabla \times (\Delta \times E) = -\nabla \times \left(\frac{\partial B}{\partial t}\right) \tag{A.20}$$

$$\nabla \times (\nabla \times E) = \nabla \times (\nabla E) - (\Delta^2 E) = -\nabla^2 E \tag{A.21}$$

$$\nabla \times (\frac{\partial B}{\partial t}) = \frac{\partial}{\partial t}(\nabla \times \mu_r \mu_0 H) = \mu_r \mu_0 \frac{\partial^2}{\partial t^2} D = \mu_r \mu_0 \epsilon_r \epsilon_0 \frac{\partial^2}{\partial t^2} E \tag{A.22}$$

$$\nabla^2 E = \mu_r \mu_0 \epsilon_r \epsilon_0 \frac{\partial^2}{\partial t^2} E \tag{A.23}$$

$$C = \frac{1}{\sqrt{\mu_r \mu_0 \epsilon_r \epsilon_0}} \tag{A.24}$$

As the refractive index is defined as the ratio between velocity of light in

vacuum (C_o) and the velocity of light in the medium (C), the refractive index can be written as eqn (25). Assuming that the material is non-magnetic ($\mu_r = 1$), it follows that the refractive index is the square root of the relative permittivity of the medium. Thus, using the eqn (17), the refractive index of a medium can be written as eqn (26). If the refractive index does not differ much from 1, the refractive index can be represented as eqn (27).

$$n_r = \frac{C_0}{C} = \sqrt{\mu_r \epsilon_r} \tag{A.25}$$

$$n_r = \left(\frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0} \right)^{1/2} \tag{A.26}$$

$$n_r \approx 1 + \frac{\alpha N}{2\epsilon_0} \tag{A.27}$$

Because refractive index of material is a property dependent on the frequency of the oscillating electric field, we have to use the dynamic polarizability, $\alpha(\omega)$, of the molecule. Thus, using the eqn (10) for the dynamic polarizability and eqn (27), the refractive index can be determined as a function of the frequency as

The dependence of refractive index with the frequency can be seen in the Fig. A.1. It can be seen that as the frequency goes to zero, i.e. at infinite wavelength, the refractive index becomes a constant value. A singularity can be observed at the resonance ($\omega = \omega_{n0}$), this is because the perturbation theory breaks down

close to this point. At higher frequencies ($\omega > \omega_{n0}$), the refractive index is shown to have a value less than 1. In my project, I will be calculating the refractive index at infinite wavelength which can be determined using the mean static polarizability values. This is because it is relatively easy to calculate the refractive index this way. Further, it has been shown that the values of the RI obtained through this way is in good agreement with the experimental values. More details are provided in the next section of this appendix.

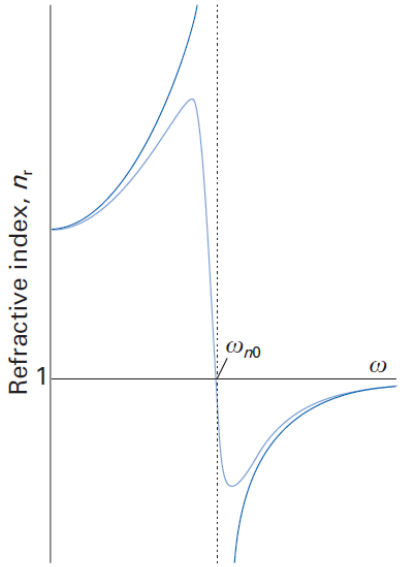


Figure A.1. Dispersion characteristic of the refractive index: Excited states are marked by singularities, and in the frequency range below, the RI values converges asymptotically towards the constant value [198].

A.2 Modeling of the Refractive Index of Polymers

As explained in the previous section, the RI value of a material is connected to the electric polarizability and can thus be obtained from quantum chemical linear response calculations. An array of electronic structure methods has been

used to determine the RI of various materials [77, 78, 79, 80, 81, 82, 83, 199, 200, 188, 64], including organic polymers [84, 85, 86, 122, 201**?**].

It is fairly easy to calculate the RI value given the dynamic molecular polarizability (and the number density) of a material using the Lorentz-Lorentz equation. However, it is generally quite challenging to determine the necessary dynamic polarizability, as it formally involves solving the time-dependent Schrödinger equation and/or scanning through the range of relevant frequencies. Consequently, relatively few studies have considered the polarizability dispersion of organic polymers [88, 89].

Although the RI value is a frequency-dependent property, its variation in the visible region is in fact often relatively small. This assumes the absence of low-lying excited states, which would render a material unsuitable for optical applications in the first place. Large variations can be observed in the ultraviolet region where resonances with the excited state manifold become the dominant feature, however, stability considerations would prohibit organic polymers to be used for high-energy applications anyways. Towards the infrared the RI decreases monotonically and becomes constant. Fig. A.2, shows the experimental results for diamondoid containing polymers, which exemplify this behavior. Fig. A.3 shows the quantum chemical result for amber, which exhibits the same trend: We see significant dispersion in the region below 250 nm where the material starts to absorb the incident radiation and becomes electronically excited. Beyond 250 nm, the RI value tapers off and becomes essentially constant throughout the visible and infrared range. The asymptotic RI value corresponds to the one that can be derived from the static polarizability. The latter can be computed much more easily than the frequency-dependent value. It only requires a single linear response calculation without explicit time de-

pendence, and is thus much less demanding in terms of computer-time and numerical stability. We can conclude that the RI values obtained from static polarizability calculations form a close lower bound for the frequency dependent values in the relevant spectral range. This approach has been used extensively in the past and has given very good agreement with the experimental results [84, 85, 91, 76, 202, 90].

The RI values based on static polarizability calculations are thus a useful indicator to the overall performance of a candidate compound, in particular for the purpose of a large-scale screening of potential candidates. A more detailed analysis of the dispersion characteristic and of other relevant properties (such as the stability, low-energy excitations, permittivity, color, etc) are obviously necessary to come to a full assessment regarding the prospects of a candidate compound. We will attempt to develop heuristic correction and calibration schemes to account for some of these effects.

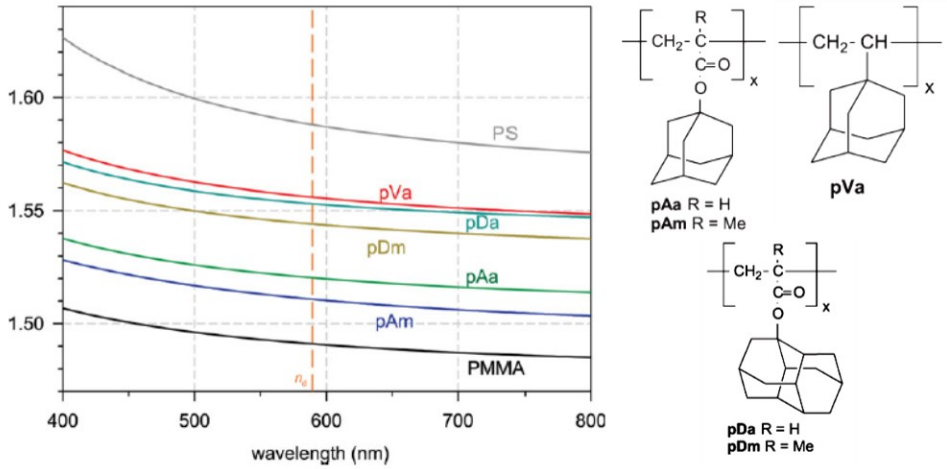


Figure A.2. Experimental dispersion of the RI values for a set of organic polymers [71].

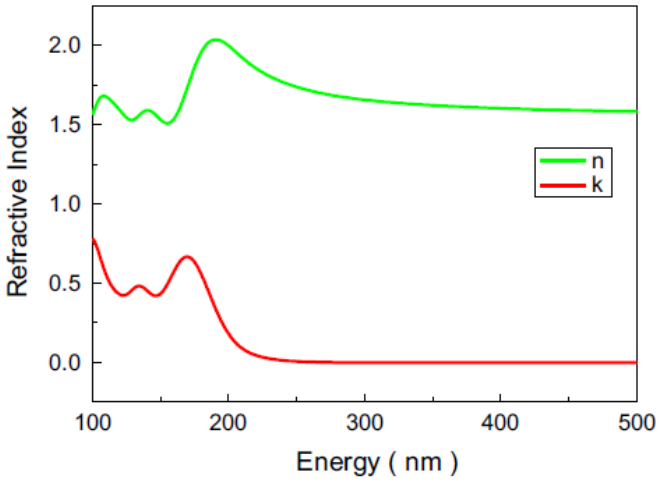


Figure A.3. Wavelength-dependent RI values determined by quantum chemical modeling [87]. (Note: the imaginary component shown in red will not be included in the present discussion as it goes beyond the scope of our current work).

Appendix B

B.1 Effect of Atom Replacement on the Polarizability of Polymers

As discussed Chapter 2, the polarizability is highly dependent on the moieties that are incorporated into the polymer backbone or in the side chain. To examine this dependence, we replaced the atoms in the example polymer (thio-ene polymer) and observed the change in the polarizability. The list of atomic changes and the corresponding change in the polarizability is shown in the Table B.1. We observe that the polarizability is highly dependent on the type of heteroatoms used in the monomer.

Table B.1. Effect of replacing atoms on the polarizability of the thiol-ene monomer

Monomer type	Polarizability (bohr^3)
Original monomer	383.43
S to O	317.20
Si to C	358.75
Si to Ge	393.32
Si to Sn	418.33

B.2 Geometry Dependence of Polarizability

Due to the complicated structure of the thiol-ene polymers, it is vital to understand the dependence of polarizability on the polymer geometry. To check this dependence, we performed a simple test. In this test, we changed the configuration of the aromatic moieties in the thiol-ene monomer as shown in the Fig. B.1(a). The strategy was to fix the angle between specific atoms in the two benzyl rings and then optimize the geometry. In the first case, we fixed the angle between carbon atoms labeled as 1, 2 and 3 (angle 123) and change the angle 234. Both the benzyl rings are aligned parallel in all the runs i.e the sum of angle 123 and angle 234 is 180^0. In the second case, one benzyl ring was fixed while the other benzyl ring orientation was changed. The polarizability changes in both these cases is shown in the Fig. B.1(b). The polarizability does not significantly change with the orientation of the aromatic moieties. Thus, the orientation of benzene rings does not play a significant role in the polarizability calculations of thiol-ene polymers. Further studies are needed to establish a detailed insight on the dependence of polarizability on the geometry.

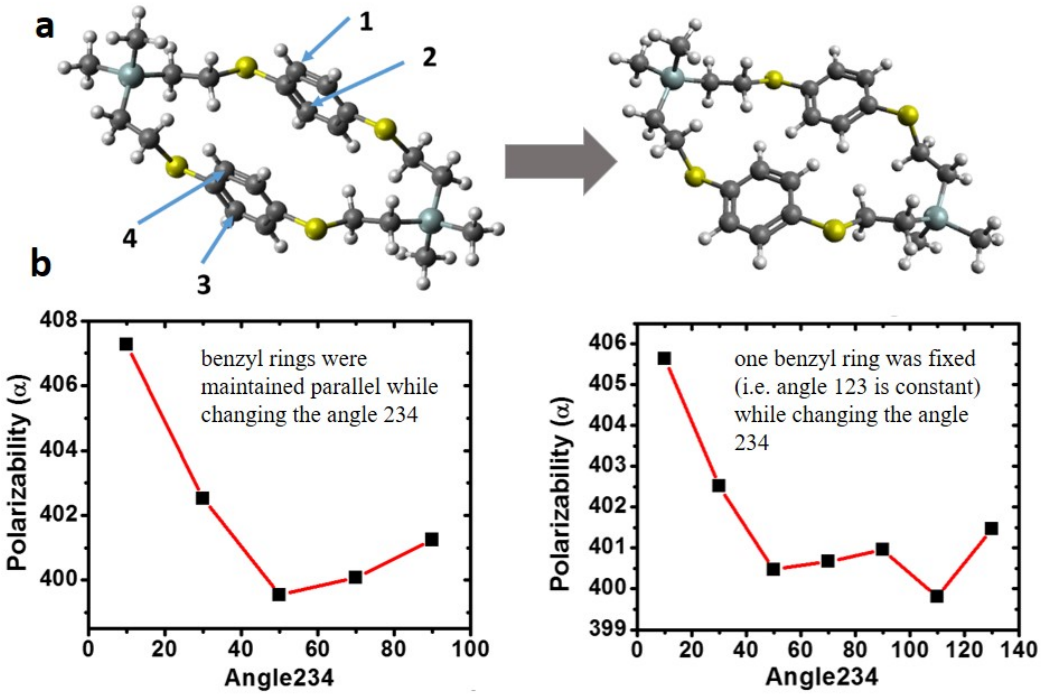


Figure B.1. **(a)** Schematic of the change in the configuration of aromatic moieties in thiol-ene monomer, **(b)** the change in the polarizability with the orientation of aromatic moieties.

Bibliography

[1] N. Thejo Kalyani and S. J. Dhoble, "Organic light emitting diodes: Energy saving lighting technologya review," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2696–2723, 2012.

[2] Y. Nakagawa, T. Ogura, T. Higashihara, and M. Ueda, "Optically transparent sulfur-containing semi-alicyclic polyimide with high refractive index," *Abstracts of Papers of the American Chemical Society*, vol. 240, 2010.

[3] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, "The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid," *The Journal of Physical Chemistry Letters*, vol. 2, no. 17, pp. 2241–2251, 2011.

[4] H. Sirringhaus, "Materials and applications for solution-processed organic field-effect transistors," *Proceedings of the IEEE*, vol. 97, no. 9, pp. 1570–1579, 2009.

[5] M. D. Angione, R. Pilolli, S. Cotrone, M. Magliulo, A. Mallardi, G. Palazzo, L. Sabbatini, D. Fine, A. Dodabalapur, N. Cioffi, and L. Torsi, "Carbon based materials for electronic bio-sensing," *Materials Today*, vol. 14, no. 9, pp. 424–433, 2011.

[6] J.-g. Liu and M. Ueda, "High refractive index polymers: fundamental research and practical applications," *Journal of Materials Chemistry*, vol. 19, no. 47, pp. 8907–8919, 2009.

[7] C. D. Selassie, "History of Quantitative Structure-Activity Relationships," in *Burger's Medicinal Chemistry and Drug Discovery*, pp. 1–48, Hoboken, NJ: John Wiley & Sons, jan 2003.

[8] K.-R. Müller, G. Rätsch, S. Sonnenburg, S. Mika, M. Grimm, and N. Heinrich, "Classifying 'drug-likeness' with kernel-based learning methods," *Journal of chemical information and modeling*, vol. 45, pp. 249–253, jan 2005.

[9] C. Le Bailly de Tilleghem and B. Govaerts, "A review of quantitative structure-activity relationship (QSAR) models," *Technical Report 07027, Universite catholique de Louvain*, 2007.

[10] C. Lipinski and A. Hopkins, "Navigating chemical space for biology and medicine," *Nature*, vol. 432, pp. 855–861, dec 2004.

[11] P. Kirkpatrick and C. Ellis, "Chemical space," *Nature*, vol. 432, p. 823, dec 2004.

[12] C. M. Dobson, "Chemical space and biology," *Nature*, vol. 432, pp. 824–828, dec 2004.

[13] E. Zvinavashe, A. J. Murk, and I. M. C. M. Rietjens, "Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity," *Chemical Research in Toxicology*, vol. 21, pp. 2229–2236, dec 2008.

[14] T. Scior, J. L. Medina-Franco, Q.-T. Do, K. Martinez-Mayorga, J. A. Yunes Rojas, and P. Bernard, "How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review," *Current Medicinal Chemistry*, vol. 16, no. 32, pp. 4297–4313, 2009.

[15] G. Schneider, "Virtual screening: an endless staircase?," *Nature Reviews Drug Discovery*, vol. 9, pp. 273–276, apr 2010.

[16] R. S. Sánchez-Carrera, S. Atahan, J. Schrier, and A. Aspuru-Guzik, "Theoretical characterization of the air-stable, high-mobility dinaphtho[2,3-b:2′3′-f]thieno[3,2-b]-thiophene organic semiconductor," *The Journal of Physical Chemistry C*, vol. 114, no. 5, pp. 2334–2340, 2010.

[17] A. N. Sokolov, S. Atahan-Evrenk, R. Mondal, H. B. Akkerman, R. S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S. C. B. Mannsfeld, A. P. Zoombelt, Z. Bao, and A. Aspuru-Guzik, "From computational discovery to experimental characterization of a high hole mobility organic crystal.," *Nature Communication*, vol. 2, p. 437, jan 2011.

[18] R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt, and A. Aspuru-Guzik, "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics," *Energy & Environmental Science*, vol. 4, no. 12, pp. 4849–4861, 2011.

[19] C. Amador-Bedolla, R. Olivares-Amaya, J. Hachmann, and A. Aspuru-Guzik, "Organic Photovoltaics," in *Informatics Mater. Sci. Eng. Data-driven Discov. Accel. Exp. Appl.* (K. Rajan, ed.), ch. 17, pp. 423–442, 2013.

[20] J. Hachmann, R. Olivares-Amaya, A. Jinich, A. L. Appleton, M. A. Blood-Forsythe, L. R. Seress, C. Roman-Salgado, K. Trepte, S. Atahan-Evrenk, S. Er, S. Shrestha, R. Mondal, A. Sokolov, Z. Bao, and A. Aspuru-Guzik, "Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the harvard clean energy project," *Energy & Environmental Science*, vol. 7, no. 2, pp. 698–704, 2014.

[21] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, and A. Aspuru-Guzik, "What is high-throughput virtual screening? A perspective from organic materials discovery," *Annual Reviews of Materials Research*, vol. 45, pp. 195–216, jul 2015.

[22] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, and A. Aspuru-Guzik, "The Harvard organic photovoltaic dataset," *Sci. Data*, vol. 3, p. 160086, 2016.

[23] National Science and Technology Council, "Materials Genome Initiative for Global Competitiveness," tech. rep., Washington, DC: National Science and Technology Council, 2011.

[24] K. Hansen, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. V. Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies," *Journal of Chemical Theory and Computation*, vol. 9, pp. 3404–3419, 2013.

[25] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, "Accelerated materials property predictions and design using motif-based fingerprints," *Physical Review B*, vol. 92, no. 1, pp. 1–10, 2015.

[26] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, p. 184115, May 2013.

[27] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials cartography: Representing and mining materials space using structural and electronic fingerprints," *Chemistry of Materials*, vol. 27, no. 3, pp. 735–743, 2015.

[28] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Physical Review Letters*, vol. 98, p. 146401, apr 2007.

[29] S. Wen, K. Nanda, Y. Huang, and G. J. O. Beran, "Practical quantum mechanics-based fragment methods for predicting molecular crystal properties," *Physical Chemistry Chemical Physics*, vol. 14, pp. 7578–7590, jun 2012.

[30] V. Stevanović, S. Lany, D. S. Ginley, W. Tumas, and A. Zunger, "Assessing capability of semiconductors to split water using ionization potentials and electron affinities only," *Physical Chemistry Chemical Physics*, vol. 16, pp. 3706–3714, feb 2014.

[31] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: review of state of the art.," *ACS Combinatorial Science*, vol. 13, pp. 579–633, nov 2011.

[32] D. Gunter, S. Cholia, A. Jain, M. Kocher, K. Persson, L. Ramakrishnan, S. P. Ong, and G. Ceder, "Community Accessible Datastore of High-Throughput Calculations: Experiences from the Materials Project," *2012 SC Companion: High Performance Computing, Networking Storage and Analysis*, pp. 1244–1251, nov 2012.

[33] A. A. White, "Big data are shaping the future of materials science," *MRS Bulletin*, vol. 38, no. August, pp. 594–595, 2013.

[34] L. C. Blum, R. van Deursen, and J.-L. Reymond, "Visualisation and subsets of the chemical universe database GDB-13 for virtual screening," *Journal of Computer-Aided Molecular Design*, vol. 25, no. 7, pp. 637–647, 2011.

[35] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2864–2875, 2012.

[36] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," *Physical Review Letters*, vol. 108, p. 058301, jan 2012.

[37] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning.," *Scientific Reports*, vol. 3, p. 2810, jan 2013.

[38] R. A. Mansbach and A. L. Ferguson, "Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics," *The Journal of Chemical Physics*, vol. 142, p. 105101, mar 2015.

[39] S. C.-H. Pegg, J. J. Haresco, and I. D. Kuntz, "A genetic algorithm for structure-based de novo design," *Journal of Computer-Aided Molecular Design*, vol. 15, pp. 911–933, Oct 2001.

[40] E. Proschak, K. Sander, H. Zettl, Y. Tanrikulu, O. Rau, P. Schneider, M. Schubert-Zsilavecz, H. Stark, and G. Schneider, "From Molecular Shape to Potent Bioactive Agents II: Fragment-Based de novo Design," *ChemMedChem*, vol. 4, no. 1, pp. 45–48, 2009.

[41] M. Nakamura, T. Hachiya, Y. Saito, K. Sato, and Y. Sakakibara, "An efficient algorithm for de novo predictions of biochemical pathways between chemical compounds," *BMC Bioinformatics*, vol. 13, p. S8, 2012.

[42] M. Hartenfeller and G. Schneider, "De Novo Drug Design," in *Chemoinformatics and Computational Chemical Biology* (J. Bajorath, ed.), vol. 672, pp. 299–323, Humana Press, 2011.

[43] M. Afzal and J. Hachmann, "ChemLG: A smart and massively parallel molecular library generator." `https://github.com/hachmannlab/chemlg`, 2018.

[44] M. Afzal and J. Hachmann, "Chemhtps: A vitual high-throughput screening infrastructure for the automatation of materials discovery." `https://github.com/hachmannlab/chemhtps`, 2018.

[45] S. Shirish, S. Aditya, M. Afzal, and J. Hachmann, "Chembddb: A big data database program suite for the chemical and materials sciences." `https://github.com/hachmannlab/chembddb`, 2018.

[46] M. Haghighatlari, R. Subramanian, B. Urala, G. Vishwakarma, A. Sonpal, P. Chen, S. Setlur, and J. Hachmann, "Chemml: A machine learning and informatics program suite for the chemical and materials sciences." `https://github.com/hachmannlab/chemml`, 2017.

[47] M. A. F. Afzal, C. Cheng, and J. Hachmann, "Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241712, 2018.

[48] T. Higashihara and M. Ueda, "Recent progress in high refractive index polymers," *Macromolecules*, vol. 48, no. 7, pp. 1915–1929, 2015.

[49] C. L. Lu, C. Guan, Y. F. Liu, Y. R. Cheng, and B. Yang, "PbS/polymer nanocomposite optical materials with high refractive index," *Chemistry of Materials*, vol. 17, no. 9, pp. 2448–2454, 2005.

[50] L. Zimmermann, M. Weibel, W. Caseri, and U. W. Suter, "High refractive-index films of polymer nanocomposites," *Journal of Materials Research*, vol. 8, no. 7, pp. 1742–1748, 1993.

[51] T. Lei, J. Y. Wang, and J. Pei, "Roles of flexible chains in organic semiconducting materials," *Chemistry of Materials*, vol. 26, no. 1, pp. 594–603, 2014.

[52] A. Voigt, U. Ostrzinski, K. Pfeiffer, J. Y. Kim, V. Fakhfouri, J. Brugger, and G. Gruetzner, "New inks for the direct drop-on-demand fabrication of polymer lenses," *Microelectronic Engineering*, vol. 88, no. 8, pp. 2174–2179, 2011.

[53] S. Ummartyotin, J. Juntaro, M. Sain, and H. Manuspiya, "Development of transparent bacterial cellulose nanocomposite film as substrate for flexible organic light emitting diode (OLED) display," *Industrial Crops and Products*, vol. 35, no. 1, pp. 92–97, 2012.

[54] C. Xiang and R. Ma, "Devices to increase OLED output coupling efficiency with a high refractive index substrate," May 2 2017. US Patent 9,640,781.

[55] H. Nishiyama, J. Nishii, M. Mizoshiri, and Y. Hirata, "Microlens arrays of high-refractive-index glass fabricated by femtosecond laser lithography," *Applied Surface Science*, vol. 255, no. 24, pp. 9750–9753, 2009.

[56] Y. Kokubun, N. Funato, and M. Takizawa, "Athermal waveguides for temperature-independent lightwave devices," *IEEE Photonics Technology Letters*, vol. 5, pp. 1297–1300, Nov 1993.

[57] H. Wei and S. Krishnaswamy, "Direct laser writing polymer micro-resonators for refractive index sensors," *IEEE Photonics Technology Letters*, vol. 28, pp. 2819–2822, Dec 2016.

[58] A. Rodriguez, G. Vitrant, P. A. Chollet, and F. Kajzar, "Optical control of an integrated interferometer using a photochromic polymer," *Applied Physics Letters*, vol. 79, no. 4, pp. 461–463, 2001.

[59] S. Singaravalu, D. Mayo, H. Park, K. Schriver, and R. Haglund, "Anti-reflective polymer-nanocomposite coatings fabricated by RIR-MAPLE," in *SPIE LASE*, vol. 8607, p. 860718, International Society for Optics and Photonics, 2013.

[60] J.-B. Kim, J.-H. Lee, C.-K. Moon, S.-Y. Kim, and J.-J. Kim, "Highly Enhanced Light Extraction from Surface Plasmonic Loss Minimized Organic Light-Emitting Diodes," *Advanced Materials*, vol. 25, no. 26, pp. 3571–3577, 2013.

[61] E. Kim, H. Cho, K. Kim, T.-W. Koh, J. Chung, J. Lee, Y. Park, and S. Yoo, "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," *Advanced Materials*, vol. 27, no. 9, pp. 1624–1631, 2015.

[62] S.-S. Sun and L. R. Dalton, *Introduction to Organic Electronic and Optoelectronic Materials and Devices (Optical Science and Engineering Series)*. Boca Raton, FL, USA: CRC Press, 2008.

[63] H. Jintoku and H. Ihara, "The simplest method for fabrication of high refractive index polymer-metal oxide hybrids based on a soap-free process," *Chemical Communications*, vol. 50, no. 73, pp. 10611–10614, 2014.

[64] R. Seto, T. Sato, T. Kojima, K. Hosokawa, Y. Koyama, G. I. Konishi, and T. Takata, "9,9 '-spirobifluorene-containing polycarbonates: Transparent polymers with high refractive index and low birefringence," *Journal of Polymer Science Part A-Polymer Chemistry*, vol. 48, no. 16, pp. 3658–3667, 2010.

[65] J. J. Griebel, S. Namnabat, E. T. Kim, R. Himmelhuber, D. H. Moronta, W. J. Chung, A. G. Simmonds, K.-J. Kim, J. van der Laan, N. A. Nguyen, E. L. Dereniak, M. E. Mackay, K. Char, R. S. Glass, R. A. Norwood, and J. Pyun, "New infrared transmitting material via inverse vulcanization of elemental sulfur to prepare high refractive index polymers," *Advanced Materials*, vol. 26, no. 19, pp. 3014–3018, 2014.

[66] Y. Tojo, Y. Arakwa, J. Watanabe, and G. Konishi, "Synthesis of high refractive index and low-birefringence acrylate polymers with a tetraphenylethane skeleton in the side chain," *Polymer Chemistry*, vol. 4, no. 13, pp. 3807–3812, 2013.

[67] W. F. Ho, M. A. Uddin, and H. P. Chan, "The stability of high refractive index polymer materials for high-density planar optical circuits," *Polymer Degradation and Stability*, vol. 94, no. 2, pp. 158–161, 2009.

[68] B. T. Liu and P. S. Li, "Preparation and characterization of high-refractive-index polymer/inorganic hybrid films containing TiO2 nanoparticles prepared by 4-aminobenzoic acid," *Surface & Coatings Technology*, vol. 231, pp. 301–306, 2013.

[69] L. Parke, I. R. Hooper, R. J. Hicken, C. E. J. Dancer, P. S. Grant, I. J. Youngs, J. R. Sambles, and A. P. Hibbins, "Heavily loaded ferrite-polymer composites to produce high refractive index materials at centimetre wavelengths," *APL Materials*, vol. 1, no. 4, 2013.

[70] Q. Y. Zhang, K. Su, M. B. Chan-Park, H. Wu, D. A. Wang, and R. Xu, "Development of high refractive ZnS/PVP/PDMAA hydrogel nanocomposites for artificial cornea implants," *Acta Biomaterialia*, vol. 10, no. 3, pp. 1167–1176, 2014.

[71] D. R. Robello, "Moderately high refractive index, low optical dispersion polymers with pendant diamondoids," *Journal of Applied Polymer Science*, vol. 127, no. 1, pp. 96–103, 2013.

[72] T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, "A polymer dataset for accelerated property prediction and design," *Scientific Data*, vol. 3, p. 160012, 2016.

[73] 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, and R. Ramprasad, "Rational design of all organic polymer dielectrics," *Nature Communications*, vol. 5, p. 4845, 2014.

[74] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," *Scientific Reports*, vol. 6, p. 20952, 2016.

[75] H. Redmond and J. E. Thompson, "Evaluation of a quantitative structure-property relationship (QSPR) for predicting mid-visible refractive index of secondary organic aerosol (SOA)," *Physical Chemistry Chemical Physics*, vol. 13, no. 15, pp. 6872–6882, 2011.

[76] S. S. Park, S. Lee, J. Y. Bae, and F. Hagelberg, "Refractive indices of liquid-forming organic compounds by density functional theory," *Chemical Physics Letters*, vol. 511, no. 4-6, pp. 466–470, 2011.

[77] S. Ando, "DFT calculations on refractive index dispersion of fluorocompounds in the DUV-UV-visible region," *Journal of Photopolymer Science and Technology*, vol. 19, no. 3, pp. 351–360, 2006.

[78] X. Rocquefelte, F. Goubin, H.-J. Koo, M.-H. Whangbo, and S. Jobic, "Investigation of the origin of the empirical relationship between refractive index and density on the basis of first principles calculations for the refractive indices of various TiO2 phases," *Inorganic chemistry*, vol. 43, no. 7, pp. 2246–2251, 2004.

[79] B. Jensen and A. Torabi, "Quantum theory of the dispersion of the refractive index near the fundamental absorption edge in compound semiconductors," *IEEE Journal of Quantum Electronics*, vol. 19, no. 3, pp. 448–457, 1983.

[80] M. Rabah, B. Abbar, Y. Al-Douri, B. Bouhafs, and B. Sahraoui, "Calculation of structural, optical and electronic properties of ZnS, ZnSe, MgS, MgSe and their quaternary alloy $Mg_{1-x}Zn_xS_ySe_{1-y}$," *Materials Science and Engineering: B*, vol. 100, no. 2, pp. 163–171, 2003.

[81] B. Amrani, T. Benmessabih, M. Tahiri, I. Chiboub, S. Hiadsi, and F. Hamdache, "First principles study of structural, elastic, electronic and optical properties of CuCl, CuBr and CuI compounds under hydrostatic pressure," *Physica B: Condensed Matter*, vol. 381, no. 1, pp. 179–186, 2006.

[82] A. H. Reshak and W. Khan, "Electronic structure, optical and thermoelectric transport properties of layered polyanionic hydrosulfate LiFeSO4OH: Electrode for Li-ion batteries," *Journal of Alloys and Compounds*, vol. 591, pp. 362–369, 2014.

[83] S. Azam and A. H. Reshak, "Electronic structure of 1,3-dicarbomethoxy4,6-benzenedicarboxylic acid: Density functional approach," *International Journal of Electrochemical Science*, vol. 8, no. 8, pp. 10359–10375, 2013.

[84] V. Ksianzou, R. K. Velagapudi, B. Grimm, and S. Schrader, "Polarization-dependent optical characterization of poly(phenylquinoxaline) thin films," *Journal of Applied Physics*, vol. 100, no. 6, 2006.

[85] C. D. Zeinalipour-Yazdi and D. P. Pullman, "Quantitative structure - property relationships for longitudinal, transverse, and molecular static polarizabilities in polyynes," *Journal of Physical Chemistry B*, vol. 112, no. 25, pp. 7377–7386, 2008.

[86] X. L. Yu, B. Yi, and X. Y. Wang, "Prediction of refractive index of vinyl polymers by using density functional theory," *Journal of Computational Chemistry*, vol. 28, no. 14, pp. 2336–2341, 2007.

[87] Z. F. Rao and R. F. Zhou, "Electronic structure and optical properties of resin," *Spectrochimica Acta Part a-Molecular and Biomolecular Spectroscopy*, vol. 105, pp. 618–622, 2013.

[88] C. K. Rowan and I. Paci, "Optical properties of Ag/polyvinylidene fluoride nanocomposites: A theoretical study," *Journal of Physical Chemistry C*, vol. 115, no. 16, pp. 8316–8324, 2011.

[89] A. Lenz, H. Kariis, A. Pohl, P. Persson, and L. Ojamae, "The electronic structure and reflectivity of PEDOT:PSS from density functional theory," *Chemical Physics*, vol. 384, no. 1-3, pp. 44–51, 2011.

[90] M. E. Azim-Araghi, J. Baedi, and L. M. Goodarzi, "Electrical and optical properties of an organic semiconductor metal-free phthalocyanine (C32H18N8)," *European Physical Journal-Applied Physics*, vol. 58, no. 3, p. 30201, 2012.

[91] S. Lee and S. S. Park, "Dielectric properties of organic solvents from non-polarizable molecular dynamics simulation with electronic continuum model and density functional theory," *Journal of Physical Chemistry B*, vol. 115, no. 43, pp. 12571–12576, 2011.

[92] F. Neese, "Prediction of molecular properties and molecular spectroscopy with density functional theory: From fundamental theory to exchange-coupling," *Coord. Chem. Rev.*, vol. 253, no. 5-6, pp. 526–563, 2009.

[93] J. G. Liu, Y. Nakamura, T. Ogura, Y. Shibasaki, S. Ando, and M. Ueda, "Optically transparent sulfur-containing polyimide-TiO(2) nanocomposite films with high refractive index and negative pattern formation from poly(amic acid)-TiO(2) nanocomposite film," *Chemistry of Materials*, vol. 20, no. 1, pp. 273–281, 2008.

[94] Y. Terui and S. Ando, "Coefficients of molecular packing and intrinsic birefringence of aromatic polyimides estimated using refractive indices and molecular polarizabilities," *Journal of Polymer Science Part B-Polymer Physics*, vol. 42, no. 12, pp. 2354–2366, 2004. Times Cited: 27 Terui, Y Ando, S 28.

[95] T. Lu and F. Chen, "Multiwfn: A multifunctional wavefunction analyzer," *Journal of Computational Chemistry*, vol. 33, no. 5, pp. 580–592, 2012.

[96] Y. H. Zhao, M. H. Abraham, and A. M. Zissimos, "Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds," *The Journal of Organic Chemistry*, vol. 68, no. 19, pp. 7368–7373, 2003.

[97] A. Bondi, "van der Waals volumes and radii," *The Journal of Physical Chemistry*, vol. 68, no. 3, pp. 441–451, 1964.

[98] A. Askadskiĭ, *Computational Materials Science of Polymers*. Cambridge Int Science Publishing, 2003.

[99] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, (New York, NY), pp. 144–152, ACM, 1992.

[100] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, pp. 155–161, 1997.

[101] A. J. Smola and Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[102] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model," *The Journal of Chemical Physics*, vol. 110, no. 13, pp. 6158–6170, 1999.

[103] F. Weigend, R. Ahlrichs, K. A. Peterson, T. H. Dunning, R. M. Pitzer, and A. Bergner, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," *Physical Chemistry Chemical Physics*, vol. 7, p. 3297, aug 2005.

[104] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," *The Journal of Chemical Physics*, vol. 132, no. 15, p. 154104, 2010.

[105] F. Neese, "The ORCA program system," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 1, pp. 73–78, 2012.

[106] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 114, no. 25, pp. 10024–10035, 1992.

[107] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, pp. 33–33, 2011.

[108] G. Slonimskii, A. Askadskii, and A. Kitaigorodskii, "The packing of polymer molecules," *Polymer Science USSR*, vol. 12, no. 3, pp. 556–577, 1970.

[109] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[110] Talete srl, "DRAGON (Software for Molecular Descriptor Calculation)," 2011. http://www.talete.mi.it/index.htm.

[111] H. Villar, M. Dupuis, J. Watts, G. Hurst, and E. Clementi, "Structure, vibrational spectra, and IR intensities of polyenes from ab initio SCF calculations," *Journal of Chemical Physics*, vol. 88, no. 2, pp. 1003–1009, 1988.

[112] D. H. Mosley, J. G. Fripiat, B. Champagne, and J.-M. André, "Ab initio investigation of the static polarizability of planar and twisted infinite polythiophene chains," *International Journal of Quantum Chemistry*, vol. 52, no. S28, pp. 451–467, 1994.

[113] J. Bicerano, *Prediction of Polymer Properties*. CRC Press, 2002.

[114] G.-S. Liou, P.-H. Lin, H.-J. Yen, Y.-Y. Yu, T.-W. Tsai, and W.-C. Chen, "Highly flexible and optical transparent 6f-pi/tio2 optical hybrid films with tunable refractive index and excellent thermal stability," *Journal of Materials Chemistry*, vol. 20, no. 3, pp. 531–536, 2010.

[115] T. T. Huang, C. L. Tsai, S. Tateyama, T. Kaneko, and G. S. Liou, "Highly transparent and flexible bio-based polyimide/tio2 and zro2 hybrid films with tunable refractive index, abbe number, and memory properties," *Nanoscale*, vol. 8, no. 25, pp. 12793–12802, 2016.

[116] H. P. Liu, I. Blakey, W. E. Conley, G. A. George, D. J. T. Hill, and A. K. Whittaker, "Application of quantitative structure property relationship to the design of high refractive index 193i resist," *Journal of Micro-Nanolithography Mems and Moems*, vol. 7, no. 2, 2008. Times Cited: 6 Liu, Heping Blakey, Idriss Conley, Willard E. George, Graeme A. Hill, David J. T. Whittaker, Andrew K. 6.

[117] A. Javadi, A. Shockravi, M. Kamali, A. Rafieimanesh, and A. M. Malek, "Solution processable polyamides containing thiazole units and thioether linkages with high optical transparency, high refractive index, and low birefringence," *Journal of Polymer Science Part A: Polymer Chemistry*, vol. 51, no. 16, pp. 3505–3515, 2013.

[118] S. Gazzo, G. Manfredi, R. Potzsch, Q. Wei, M. Alloisio, B. Voit, and D. Comoretto, "High refractive index hyperbranched polyvinylsulfides for planar one-dimensional all-polymer photonic crystals," *Journal of Polymer Science Part B-Polymer Physics*, vol. 54, no. 1, pp. 73–80, 2016.

[119] A. Alexandridis, E. Chondrodima, K. Moutzouris, and D. Triantis, "A neural network approach for the prediction of the refractive index based on experimental data," *Journal of Materials Science*, vol. 47, no. 2, pp. 883–891, 2012. Times Cited: 2 0 2.

[120] G. Lisa, S. Curteanu, and C. Lisa, "Artificial neural network for prediction of excess refractive indices of some binary mixtures," *Environmental Engineering and Management Journal*, vol. 9, no. 4, pp. 483–487, 2010. Times Cited: 2 Curteanu, Silvia/D-7347-2011; Lisa, Gabriela/B-9516-2011 Curteanu, Silvia/0000-0002-5281-1265; 0 2.

[121] X. L. Yu, B. Yi, and X. Y. Wang, "Prediction of refractive index of vinyl polymers by using density functional theory," *Journal of Computational Chemistry*, vol. 28, no. 14, pp. 2336–2341, 2007. Times Cited: 20 Yu, Xinliang Yi, Bing Wang, Xueye 22.

[122] A. J. Holder, L. Ye, J. D. Eick, and C. C. Chappelow, "A quantum-mechanical qsar model to predict the refractive index of polymer matrices," *Qsar & Combinatorial Science*, vol. 25, no. 10, pp. 905–911, 2006. Times Cited: 9 Holder, Andrew J. Ye, Lin Eick, J. David Chappelow, Cecil C. 9.

[123] S. Grimme, "Semiempirical hybrid density functional with perturbative second-order correlation," *The Journal of chemical physics*, vol. 124, no. 3, p. 034108, 2006.

[124] P. Mori-Snchez, Q. Wu, and W. Yang, "Accurate polymer polarizabilities with exact exchange density-functional theory," *The Journal of Chemical Physics*, vol. 119, no. 21, pp. 11001–11004, 2003.

[125] G. J. B. Hurst, M. Dupuis, and E. Clementi, "Abinitio analytic polarizability, first and second hyperpolarizabilities of large conjugated organic molecules: Applications to polyenes c4h6 to c22h24," *The Journal of Chemical Physics*, vol. 89, no. 1, pp. 385–395, 1988.

[126] J.-L. Reymond, L. Ruddigkeit, L. Blum, and R. van Deursen, "The enumeration of chemical space," *Wiley Interdisciplinary Reviews-Computational Molecular Science*, vol. 2, no. 5, pp. 717–733, 2012.

[127] E. Anderson, G. D. Veith, and D. Weininger, *SMILES, a line notation and computerized interpreter for chemical structures*. US Environmental Protection Agency, Environmental Research Laboratory, 1987.

[128] M. A. F. Afzal, J. Younker, and G. Rodriguez, "The effect of tacticity and side chain structure on the coil dimensions of polyolefins." Internship Report, 2018.

[129] Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele,

E. J. Sundstrom, H. L. Woodcock, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio, H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," *Molecular Physics*, vol. 113, pp. 184–215, sep 2015.

[130] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, no. Supplement C, pp. 19–25, 2015.

[131] Y. Tian, "Inheritance of molecular orbital energies from monomer building blocks to larger copolymers in organic semiconductors," Master's thesis, University at Buffalo, 2016.

[132] C.-Y. Shih, "Systematic trends in results from different density functional theory models," Master's thesis, University at Buffalo, 2015.

[133] Manuscripts in preparation.

[134] B. R. Kowalski and C. F. Bender, "Pattern recognition. powerful approach

to interpreting chemical data," *Journal of the American Chemical Society*, vol. 94, no. 16, pp. 5632–5639, 1972.

[135] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[136] "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[137] F. Chollet *et al.*, "Keras," 2015. Software available from https://github.com/fchollet/keras.

[138] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, p. 33, Oct 2011.

[139] "RDKit: Open-source cheminformatics." Software available from http://www.rdkit.org.

[140] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[141] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms.," *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, jan 2001.

[142] S. Manzhos and T. Carrington, "A random-sampling high dimensional model representation neural network for building potential energy surfaces," *The Journal of Chemical Physics*, vol. 125, p. 084109, aug 2006.

[143] G. E. Dahl, *Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing*. PhD thesis, University of Toronto, 2015.

[144] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, and H. Timmerman, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.

[145] V. J. Sykora and D. E. Leahy, "Chemical Descriptors Library (CDL): a generic, open source software library for chemical informatics.," *Journal of Chemical Information and Modeling*, vol. 48, pp. 1931–1942, oct 2008.

[146] R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, "Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 27, no. 2, pp. 82–85, 1987.

[147] N. M. O'Boyle and R. A. Sayle, "Comparing structural fingerprints using a literature-based similarity benchmark," *Journal of Cheminformatics*, vol. 8, no. 1, pp. 1–14, 2016.

[148] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *Journal of Physical Chemistry Letters*, vol. 6, no. 12, pp. 2326–2331, 2015.

[149] R. Ramakrishnan and O. A. von Lilienfeld, "Machine Learning, Quantum Chemistry, and Chemical Space," in *Reviews in Computational Chemistry*, vol. 30, pp. 225–256, John Wiley & Sons, apr 2017.

[150] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, p. 16028, 2016.

[151] A. S. Mathews, I. Kim, and C.-S. Ha, "Synthesis, characterization, and properties of fully aliphatic polyimides and their derivatives for microelectronics and optoelectronics applications," *Macromolecular Research*, vol. 15, no. 2, pp. 114–128, 2007.

[152] C.-M. Chang, C.-L. Chang, and C.-C. Chang, "Synthesis and optical properties of soluble polyimide/titania hybrid thin films," *Macromolecular Materials and Engineering*, vol. 291, no. 12, pp. 1521–1528, 2006.

[153] D. Dsselberg, D. Verreault, P. Koelsch, and C. Staudt, "Synthesis and characterization of novel, soluble sulfur-containing copolyimides with high refractive indices," *Journal of Materials Science*, vol. 46, no. 14, pp. 4872–4879, 2011.

[154] J.-g. Liu, Y. Nakamura, T. Ogura, Y. Shibasaki, S. Ando, and M. Ueda, "Optically transparent sulfur-containing polyimide- tio2 nanocomposite films with high refractive index and negative pattern formation from poly (amic acid)- tio2 nanocomposite film," *Chemistry of Materials*, vol. 20, no. 1, pp. 273–281, 2007.

[155] K. L. Mittal, *Polyimides: synthesis, characterization, and applications*, vol. 1. Springer Science & Business Media, 2013.

[156] M. Barikani and S. MehdipourAtaei, "Synthesis, characterization, and thermal properties of novel arylene sulfone ether polyimides and polyamides," *Journal of Polymer Science Part A: Polymer Chemistry*, vol. 38, no. 9, pp. 1487–1492, 2000.

[157] I. Butnaru, M. Bruma, T. Kopnick, and J. Stumpe, "Influence of chemical structure on the refractive index of imide-type polymers," *Macromolecular Chemistry and Physics*, vol. 214, no. 21, pp. 2454–2464, 2013.

[158] N. Fukuzaki, T. Higashihara, S. Ando, and M. Ueda, "Synthesis and characterization of highly refractive polyimides derived from thiophene-containing aromatic diamines and aromatic dianhydrides," *Macromolecules*, vol. 43, no. 4, pp. 1836–1843, 2010.

[159] J. G. Liu, Y. Nakamura, T. Ogura, Y. Shibasaki, S. Ando, and M. Ueda, "Optically transparent sulfur-containing polyimide-tio(2) nanocomposite films with high refractive index and negative pattern formation from poly(amic acid)-tio(2) nanocomposite film," *Chemistry of Materials*, vol. 20, no. 1, pp. 273–281, 2008. Times Cited: 92 Liu, Jin-Gang Nakamura, Yasuhiro Ogura, Tomohito Shibasaki, Yuji Ando, Shinji Ueda, Mitsuru 93.

[160] S. A. Sydlik, Z. Chen, and T. M. Swager, "Triptycene polyimides: soluble polymers with high thermal stability and low refractive indices," *Macromolecules*, vol. 44, no. 4, pp. 976–980, 2011.

[161] C. A. Terraza, J.-G. Liu, Y. Nakamura, Y. Shibasaki, S. Ando, and M. Ueda, "Synthesis and properties of highly refractive polyimides derived from fluorene-bridged sulfur-containing dianhydrides and diamines," *Journal of Polymer Science Part A: Polymer Chemistry*, vol. 46, no. 4, pp. 1510–1520, 2008.

[162] K. R. Carter, R. A. DiPietro, M. I. Sanchez, and S. A. Swanson, "Nanoporous polyimides derived from highly fluorinated polyimide/poly (propylene oxide) copolymers," *Chemistry of materials*, vol. 13, no. 1, pp. 213–221, 2001.

[163] D. Yu, A. Gharavi, and L. Yu, "Novel aromatic polyimides for nonlinear optics," *Journal of the American Chemical Society*, vol. 117, no. 47, pp. 11680–11686, 1995.

[164] N.-H. You, Y. Suzuki, D. Yorifuji, S. Ando, and M. Ueda, "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*, vol. 41, no. 17, pp. 6361–6366, 2008.

[165] C.-L. Tsai, H.-J. Yen, and G.-S. Liou, "Highly transparent polyimide hybrids for optoelectronic applications," *Reactive and Functional Polymers*, vol. 108, pp. 2–30, 2016.

[166] J. Kobayashi, T. Matsuura, Y. Hida, S. Sasaki, and T. Maruno, "Fluorinated polyimide waveguides with low polarization-dependent loss and their applications to thermooptic switches," *Journal of lightwave technology*, vol. 16, no. 6, p. 1024, 1998.

[167] T. Sawada and S. Ando, "Synthesis, characterization, and optical properties of metal-containing fluorinated polyimide films," *Chemistry of materials*, vol. 10, no. 11, pp. 3368–3378, 1998.

[168] J.-g. Liu, Y. Nakamura, Y. Shibasaki, S. Ando, and M. Ueda, "Synthesis and characterization of high refractive index polyimides derived from 4, 4 [variant prime]-(p-phenylenedisulfanyl) dianiline and various aromatic tetracarboxylic dianhydrides," *Polymer journal*, vol. 39, no. 6, p. 543, 2007.

[169] H. Yeo, J. Lee, M. Goh, B.-C. Ku, H. Sohn, M. Ueda, and N.-H. You, "Synthesis and characterization of high refractive index polyimides derived from 2, 5-bis (4-aminophenylenesulfanyl)-3, 4-ethylenedithiothiophene and aromatic dianhydrides," *Journal of Polymer Science Part A: Polymer Chemistry*, vol. 53, no. 7, pp. 944–950, 2015.

[170] V. P. Privalko and A. V. Pedosenko, "Molecular packing density in the crystalline state of semi-rigid chain polymers. i: Polyimides," *Polymer Engineering & Science*, vol. 37, no. 6, pp. 978–982, 1997.

[171] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, and K.-R. Müller, "Learning invariant representations of molecules for atomization energy prediction," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 440–448, Curran Associates, Inc., 2012.

[172] V. Simon, J. Gasteiger, and J. Zupan, "A combined application of two different neural network types for the prediction of chemical reactivity," *Journal of the American Chemical Society*, vol. 115, no. 20, pp. 9148–9159, 1993.

[173] M. Karthikeyan, R. C. Glen, and A. Bender, "General melting point prediction based on a diverse compound data set and artificial neural networks," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 581–590, 2005.

[174] F. Gharagheizi, "Qspr analysis for intrinsic viscosity of polymer solutions by means of ga-mlr and rbfnn," *Computational materials science*, vol. 40, no. 1, pp. 159–167, 2007.

[175] J. Huuskonen, M. Salo, and J. Taskinen, "Aqueous solubility prediction of drugs based on molecular topology and neural network modeling," *Journal of chemical information and computer sciences*, vol. 38, no. 3, pp. 450–456, 1998.

[176] E. K. Macdonald and M. P. Shaver, "Intrinsic high refractive index polymers," *Polymer International*, pp. n/a–n/a, 2014.

[177] N. Tanio and M. Irie, "Refractive index of organic photochromic dye-amorphous polymer composites," *Japanese journal of applied physics*, vol. 33, no. 7R, p. 3942, 1994.

[178] A. Kwan and W. Fung, "Packing density measurement and modelling of fine aggregate and mortar," *Cement and Concrete Composites*, vol. 31, no. 6, pp. 349–357, 2009.

[179] J. Swenson and L. Brjesson, "Correlation between free volume and ionic conductivity in fast ion conducting glasses," *Physical Review Letters*, vol. 77, no. 17, pp. 3569–3572, 1996.

[180] Y. Shen, X. X. He, and F. R. Hung, "Structural and dynamical properties of a deep eutectic solvent confined inside a slit pore," *Journal of Physical Chemistry C*, vol. 119, no. 43, pp. 24489–24500, 2015.

[181] F. Sheu and R. Chern, "Effects of packing density on the gas-transport properties of poly (phenolphthalein phthalate) s," *Journal of Polymer Science Part B: Polymer Physics*, vol. 27, no. 5, pp. 1121–1133, 1989.

[182] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, "General atomic and molecular electronic structure system," *Journal of Computational Chemistry*, vol. 14, no. 11, pp. 1347–1363, 1993.

[183] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of Computational Chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[184] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," *J Mol Graph Model*, vol. 25, no. 2, pp. 247–60, 2006.

[185] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, no. 1, pp. 43–56, 1995.

[186] G. Piacenza, G. Legsa, B. Blaive, and R. Gallo, "Molecular volumes and densities of liquids and solids by molecular mechanicsestimation and analysis," *Journal of Physical Organic Chemistry*, vol. 9, no. 6, pp. 427–432, 1996.

[187] T. Takasaki, Y. Kuwana, T. Takahashi, and S. Hayashida, "Synthesis and optical properties of polyimides," *Journal of Polymer Science Part A: Polymer Chemistry*, vol. 38, no. S1, pp. 4832–4838, 2000.

[188] T. Takano, Y. C. Lin, F. G. Shi, B. Carlson, and S. Sciamanna, "Novel methacrylated diamondoid to produce high-refractive index polymer," *Optical Materials*, vol. 32, no. 5, pp. 648–651, 2010.

[189] T. Namikoshi, T. Hashimoto, Y. Makino, T. Imaeda, M. Urushisaki, and T. Sakaguchi, "Synthesis and properties of poly(2-adamantyl vinyl ether)-based optical plastics," *Polymer Bulletin*, vol. 71, no. 6, pp. 1389–1402, 2014.

[190] M. A. Gunawan, J.-C. Hierso, D. Poinsot, A. A. Fokin, N. A. Fokina, B. A. Tkachenko, and P. R. Schreiner, "Diamondoids: functionalization and subsequent applications of perfectly defined molecular cage hydrocarbons," *New Journal of Chemistry*, vol. 38, no. 1, pp. 28–41, 2014.

[191] S. D. Bhagat, J. Chatterjee, B. H. Chen, and A. E. Stiegman, "High refractive index polymers based on thiol-ene cross-linking using polarizable inorganic/organic monomers," *Macromolecules*, vol. 45, no. 3, pp. 1174–1181, 2012. Times Cited: 17 Bhagat, Sharad D. Chatterjee, Jhunu Chen, Banghao Stiegman, A. E. 17.

[192] D. W. Mosley, G. Khanarian, D. M. Conner, D. L. Thorsen, T. L. Zhang, and M. Wills, "High refractive index thermally stable phenoxyphenyl and phenylthiophenyl silicones for light-emitting diode applications," *Journal of Applied Polymer Science*, vol. 131, no. 3, 2014.

[193] Z. Lin, Y. R. Cheng, H. Lu, L. A. Zhang, and B. Yang, "Preparation and characterization of novel zns/sulfur-containing polymer nanocomposite optical materials with high refractive index and high nanophase contents," *Polymer*, vol. 51, no. 23, pp. 5424–5431, 2010.

[194] C. K. W. Jim, A. J. Qin, J. W. Y. Lam, M. Haussler, J. Z. Liu, M. M. F. Yuen, J. K. Kim, K. M. Ng, and B. Z. Tang, "Facile polycyclotrimerization of "simple" arylene bipropiolates: A metal-free, regioselective route to functional hyperbranched polymers with high optical transparency, tunable refractive index, low chromatic aberration, and photoresponsive patternability," *Macromolecules*, vol. 42, no. 12, pp. 4099–4109, 2009.

[195] S. D. Bhagat, E. B. D. S. Filho, and A. E. Stiegman, "High refractive index polymer composites synthesized by cross-linking of oxozirconium clusters through thiol-ene polymerization," *Macromolecular Materials and Engineering*, vol. 300, no. 6, pp. 580–585, 2015.

[196] S. D. Bhagat, "High refractive index materials," Apr. 6 2017. US Patent App. 15/282,891.

[197] S. D. Bhagat, C. Peroz, V. Singh, and F. Y. Xu, "Monolithic high refractive index photonic devices," Mar. 1 2018. US Patent App. 15/684,530.

[198] P. Atkins and R. Friedman, *Molecular Quantum Mechanics*. OUP Oxford, 2011.

[199] A. Baev, M. Samoc, P. N. Prasad, M. Krykunov, and J. Autschbach, "A quantum chemical approach to the design of chiral negative index materials," *Optics Express*, vol. 15, no. 9, pp. 5730–5741, 2007.

[200] K. Asai, G. I. Konishi, K. Sumi, and K. Mizuno, "Synthesis of silyl-functionalized oligothiophene-based polymers with bright blue light-emission and high refractive index," *Journal of Organometallic Chemistry*, vol. 696, no. 6, pp. 1236–1243, 2011.

[201] E. Ortyl and S. Kucharski, "Refractive index modulation in polymeric photochromic films," *Central European Journal of Chemistry*, vol. 1, no. 2, pp. 137–159, 2003.

[202] C. M. Isborn, A. Leclercq, F. D. Vila, L. R. Dalton, J. L. Brdas, B. E. Eichinger, and B. H. Robinson, "Comparison of static first hyperpolarizabilities calculated with various quantum mechanical methods," *The Journal of Physical Chemistry A*, vol. 111, no. 7, pp. 1319–1327, 2007.