5-5-2017

Rational Design of Polymer Dielectrics Using First Principles Computations and Machine Learning

Arun Kumar Mannodi Kanakkithodi
University of Connecticut, arun.mannodi_kanakkithodi@uconn.edu

Rational Design of Polymer Dielectrics Using First Principles Computations and Machine Learning

Arun Kumar Mannodi Kanakkithodi, PhD

University of Connecticut, 2017

While intuition-driven experiments and serendipity have guided traditional materials discovery, computational strategies have become increasingly prevalent and a powerful complement to experiments in modern day materials research. A novel approach for efficient materials design is “rational co-design”, where high-throughput computational screening is used synergistically with experimental synthesis and testing. In this Thesis, the utility and promise of such an approach was demonstrated for the design of advanced polymer dielectrics for electrostatic energy storage applications. Density functional theory computations were applied to study the structural, electronic and dielectric properties of polymers, based on which targeted synthesis and property measurements were carried out for promising candidates. These co-design efforts led to the identification of potential replacements for present day “standard” dielectrics (such as biaxially oriented polypropylene) not only by new organic polymer candidates within known generic polymer subclasses (e.g., polyurea, polythiourea, polyimide), but also by organometallic polymers,

a hitherto untapped but promising chemical subspace. Further, the prospects of significantly accelerating the materials design process using state-of-the-art machine learning techniques were explored. Vast computational data generated as part of this work was *mined* for the development of accurate 'instant prediction' and 'design' models for the relevant properties of polymers. These models were converted into user-friendly polymer design tools, and along with the computational and experimental data, compiled in the form of a web-based application (http://khazana.uconn.edu/polymer_genome/) to facilitate the rapid design and discovery of polymer dielectrics.

Rational Design of Polymer Dielectrics Using First Principles Computations and Machine Learning

Arun Kumar Mannodi Kanakkithodi

Bachelor of Technology, Indian Institute of Technology Roorkee, India, 2012

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2017

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Rational Design of Polymer Dielectrics Using First Principles Computations and Machine Learning

Presented by
Arun Kumar Mannodi Kanakkithodi, B.Tech.

Major Advisor

_____

Rampi Ramprasad

Associate Advisor

_____

Gregory Sotzing

Associate Advisor

_____

Yang Cao

University of Connecticut

2017

Dedicated to everyone who believed in science,

and everyone who will continue to do so.

ACKNOWLEDGEMENTS

I look back upon my decision to pursue a Ph.D. in materials science from the University of Connecticut very fondly, thanks to a host of people who helped make most of my pursuits a success. Prof. Rampi Ramprasad is the best advisor any graduate student could ask for, and his influence on my life goes well beyond academic progress. He taught me to challenge the scientific *status quo*, to always wonder if there is a better way of tackling a problem, and most of all, to conduct myself with utmost scientific integrity. Working and learning with him has been a genuine pleasure, and his unbridled joy and enthusiasm for scientific discovery is something I hope to inherit, *from master to protégé*.

One of the most rewarding things for me during my Ph.D. was being a part of the Multi-Disciplinary University Research Initiative (MURI) project (funded by the Office of Naval Rsearch, ONR) headed by Prof. Ramprasad, concerned with designing novel polymeric dielectrics for capacitive energy storage. I got the opportunity to work with people from various disciplines, such as polymer synthesis, electrical characterization, and computer science, which added many necessary and unique dimensions to my computational studies. I would particularly like to acknowledge the guidance and regular advice of Prof. Gregory Sotzing and Prof. Yang Cao, who led their respective teams at UConn as a part of this project. Further, I worked closely with several outstanding students, scientists and researchers, including Dr. Huan Tran, Dr. Chenchen Wang, Dr. Chiho Kim, Dr. Aaron

Baldwin, Dr. Rui Ma, Gregory Treich and Shamima Nasreen; I owe every one of them a great debt of gratitude.

During my Ph.D., I got the fantastic opportunity to visit Los Alamos National Laboratory (LANL) for a research internship, and I consider the time I spent there as some of the formative months of me as a scientist. Not only was Dr. Turab Lookman in the Theory Division an excellent mentor to me, I was able to work with several amazing scientists and researchers like Dr. Ghanshyam Pilania and Dr. James Gubernatis.

I would further like to thank Prof. George Rossetti and Prof. Avinash Dongare for many fruitful discussions and valuable advice regarding my research (and beyond). I would be remiss not to mention the many friendships I forged over the years, with my fellow group members (past and present) Dr. Venkatesh Botu, Dr. Satyesh Yadav, Dr. Vinit Sharma, Dr. Hom Sharma, Dr. Yenny Quintero, Dr. Anand Chandra, Lihua Chen, Rohit Batra, Dr. Sridevi Krishnan, James Chapman, Erik Nykwest and Deepak Kamal, and with all the other remarkable people I met during this five year journey, including Sriram Vijayan, Danielle Heichel, Amit Joshi, Ameya Akkalkotkar, Sumit Suresh, Matt Janish, Austin McDannald and Drew Clearfield.

I would have never got where I am today if not for the love and support of my parents, Sarada Ravindran and M.K. Ravindran, and my brother, Robin. It means the world to me

that they are proud of my achievements, and I hope to continue making them proud for a long time to come.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1. Motivation and Overarching Perspectives

1.1.1. Philosophy of Materials Design and Discovery

Throughout human history, every age and every culture has perhaps been best defined by the materials they used. Prehistoric humans carved tools out of bone and wood and used them for hunting. The stone age, which started nearly 3 million years ago and lasted till around 3000 BCE, was known for the use of stone in collecting food and building shelters. While much of the less advanced parts of the world remained in the stone age for a long time, the advent of metallurgy started the bronze-age in eastern and southern Asia around 7000 BCE, before it made its way to Europe. Iron in its native metallic state was already being used during the bronze age, but the true iron age is said to have started around 1000 BCE as humans found the means to smelt iron ore. With metal-working now commonplace, the next 2000 or so years saw marked improvements in production and processing of metals and alloys, woodworking, paper, glasses, ceramics and polymers, ultimately leading to the industrial revolution in Europe in the 18^{th} century.

The pace of progress in the 20th century was more dramatic than ever before, thanks to the accumulation of years of documented knowledge and vast swathes of data pertaining to failed and successful experiments. The advent of high-powered special purpose machinery and mass factory production saw stainless steel become mainstream, and incredible advances in transportation, building and communication. One area where iterative experiments and past data majorly benefited materials design was alloys: it was realized that with additions of different amounts of carbon, chromium, nickel, manganese and molybdenum, the properties of steel can be tailored [1]. Solid solutions of aluminium with copper found applications in the aeronautical industry [2], and NiTi-based alloys found amazing shape-memory applications [3]. This period also saw the development of some of the most important phenomenological models in materials science, such as the Hume-Rothery rules [4] and the Hall-Petch relationship [5], which emerged from experimental documentation on solid solutions and mild steels, respectively.

In the latter half of the 20th century, materials research was taken over by a romantic notion: that of designing materials on a computer before a single laboratory experiment is performed. The accumulation of data via experiments, while invaluable, was seen to be time intensive and prone to human observational errors. Today, massive parallel supercomputers with thousands of processors are being used the world over in weather forecasting, oil and gas exploration, and molecular modelling. The advent of supercomputers along with theoretical advancements in classical mechanics [6] [7] and quantum mechanics [8] [9], formulations of force-field simulations and molecular

dynamics [10] [11] [12], and the development of quantum mechanics based methods like density functional theory (DFT) [13] [14] formally kick-started the era of computational materials science [15].

Quantum mechanics, which provided a fundamental look at the structure and properties of materials in the smallest available length and time scales, made for accurate (but computationally expensive) solutions of many materials science problems. Perhaps the most popular approach in this regard is DFT, where the Schrodinger's equation is solved for a many electron system by converting it into an effective one electron problem. The accuracy of DFT in investigating the electronic structure of atoms, molecules and condensed phases has been well demonstrated, and it is being widely used today to study the mechanical, electronic, dielectric and thermodynamic properties of metals, inorganic compounds, molecules and polymers [16] [17] [18] [19] [20] [21] [22]. One of the significant transformations that computational materials science underwent over the last 50 years is the evolution of methods like DFT from being merely post hoc (i.e., being applied to study materials and explain observations post-experiments) to driving rational materials design by eliminating guesswork from experiments [23]. In the literature, many glittering examples can be found of DFT-driven experiments leading to the accelerated design of new materials, such as the identification of new cathode materials for Li batteries [24], the design of novel NiTi shape-memory alloys [25] [26], and the discovery of previously unknown ABX type thermoelectrics and conductors [27].

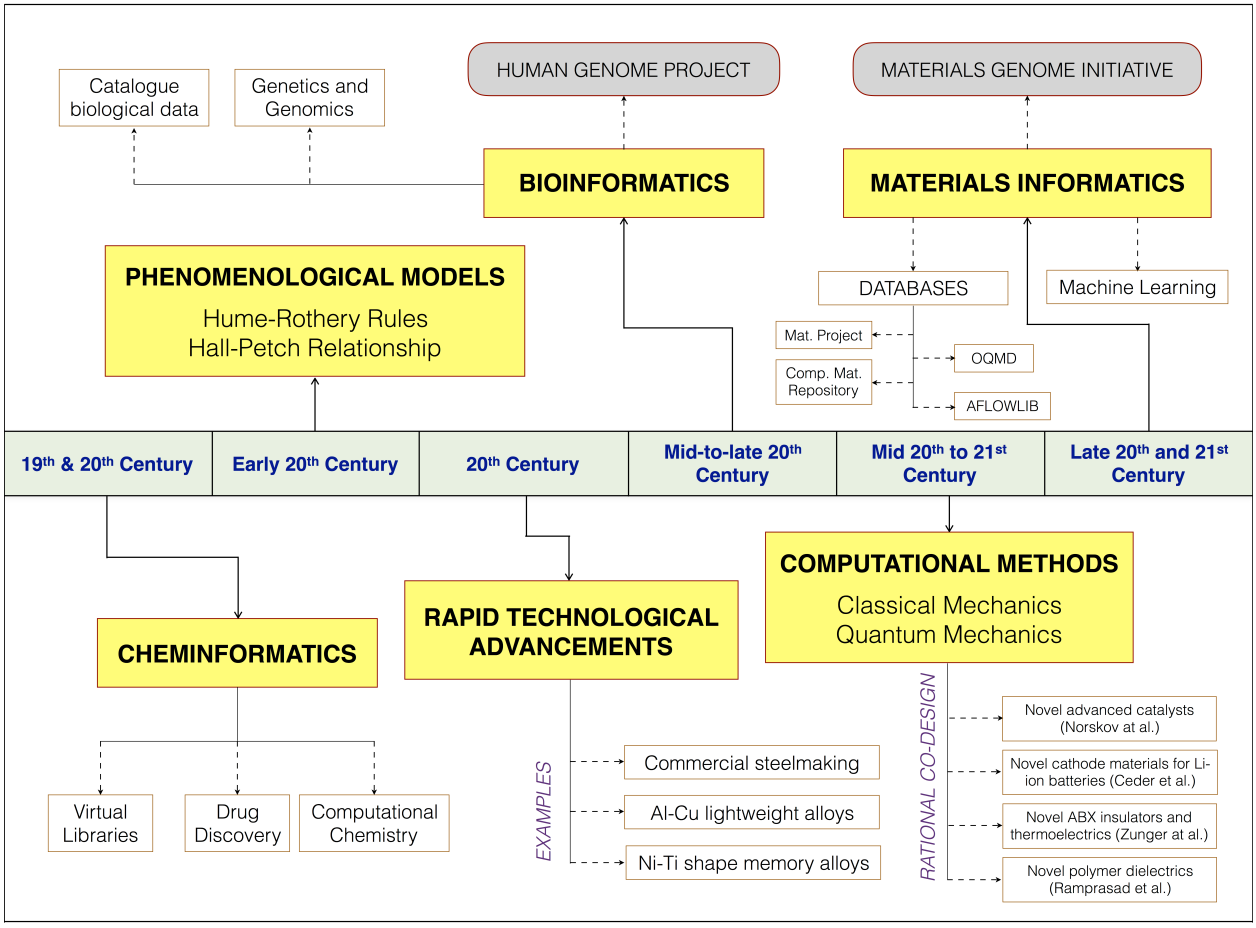


Figure 1.1. *A timeline of major developments in materials science and related fields over the last couple of centuries. Along the same lines as the Human Genome Project (initiated in the 1990s to determine the DNA sequence of the entire human genome), the Materials Genome Initiative was launched a few years ago to accelerate the design and deployment of new and advanced materials [28].*

It must be emphasised that *data* more than anything has been the great ally of the scientist in driving innovation and the discovery of physical and chemical laws. While approximate or phenomenological models enable the quick screening and design of materials, precise theories facilitate the generation of robust materials data which can in

turn lead to newer, more reliable phenomenological models. Indeed, data generation, storage, retrieval and analysis has been of key importance in the fields of cheminformatics [29] and bioinformatics [30] over the last century or so, and in the last few years, in *materials informatics* [31] [32]. The latter is a blossoming field in materials science today, focusing on the development of experimental and computational databases and on the application of modern machine learning or data mining methods that help convert the data into easily accessible models.

Figure 1.1 tries to capture a rough timeline of developments in materials science and related fields over the years, in the form of experiment-driven phenomenological models such as the Hume-Rothery rules, computational theories such as classical and quantum mechanics, and data-driven fields in chemistry (cheminformatics), biology (bioinformatics) and materials science (materials informatics). In recent years, there has been further recognition of the power of computations and databases in guiding the rational experimental design of materials in the form of the *Materials Genome Initiative* [23] (along the same lines as the Human Genome Project [33]), announced by the US government "to discover, manufacture and deploy advanced materials twice as fast, at a fraction of the cost". High-performance computing, efficient computational approaches and machine learning based methods provide great promise in accelerating the pace of discovery and deployment of new materials in practice.

1.1.2. Polymers as Capacitor Dielectrics

Dielectric materials find wide applicability owing to their ability to polarize under applied electric fields. One such application is in capacitors, which are used in electronic devices for energy storage purposes, in pulsed power applications, or as temporary batteries, for instance in car audio and stereo systems. The pervasive utility of capacitors comes from the fact that not only can they can store large amounts of electrical energy, they can discharge it all in a single flash.

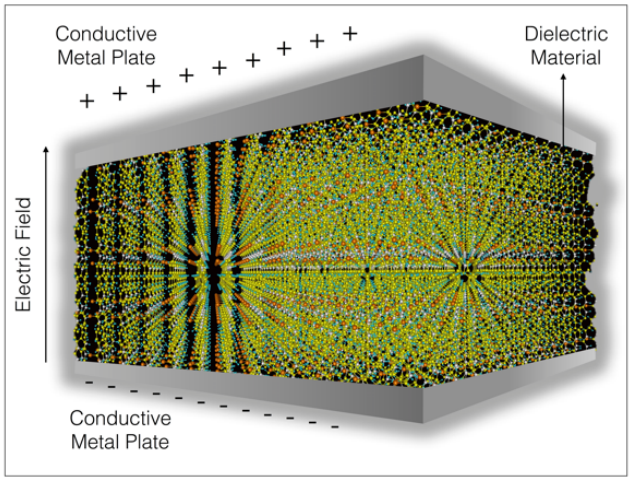


Figure 1.2. *Schematic of a capacitor with metal plates, dielectric material and applied electric field labelled.*

In any capacitor, the amount of energy that can be stored depends on the dielectric and electronic characteristics of the dielectric interface between the metal plates (**Figure 1.2**). In a linear capacitor, the charge stored (or the polarization) is directly and linearly proportional to the voltage, resulting in a constant capacitance. However, unless the dielectric interface is vacuum, there is always nonlinear dependence of the polarization on the voltage (or the electric field), as shown in **Figure 1.3**. The area under the curve between the polarization and the applied electric field, known as the D-E loop, yields the energy stored in the capacitor. This transforms into a relationship between the energy on one side and the

electric field and dielectric constant on the other, as the dielectric constant is given by the change in polarization with applied electric field. That is,

$$\frac{dP}{dE} = \varepsilon$$

$$U = \int_0^{Eb} EdP = \int_0^{Eb} \varepsilon E dE$$

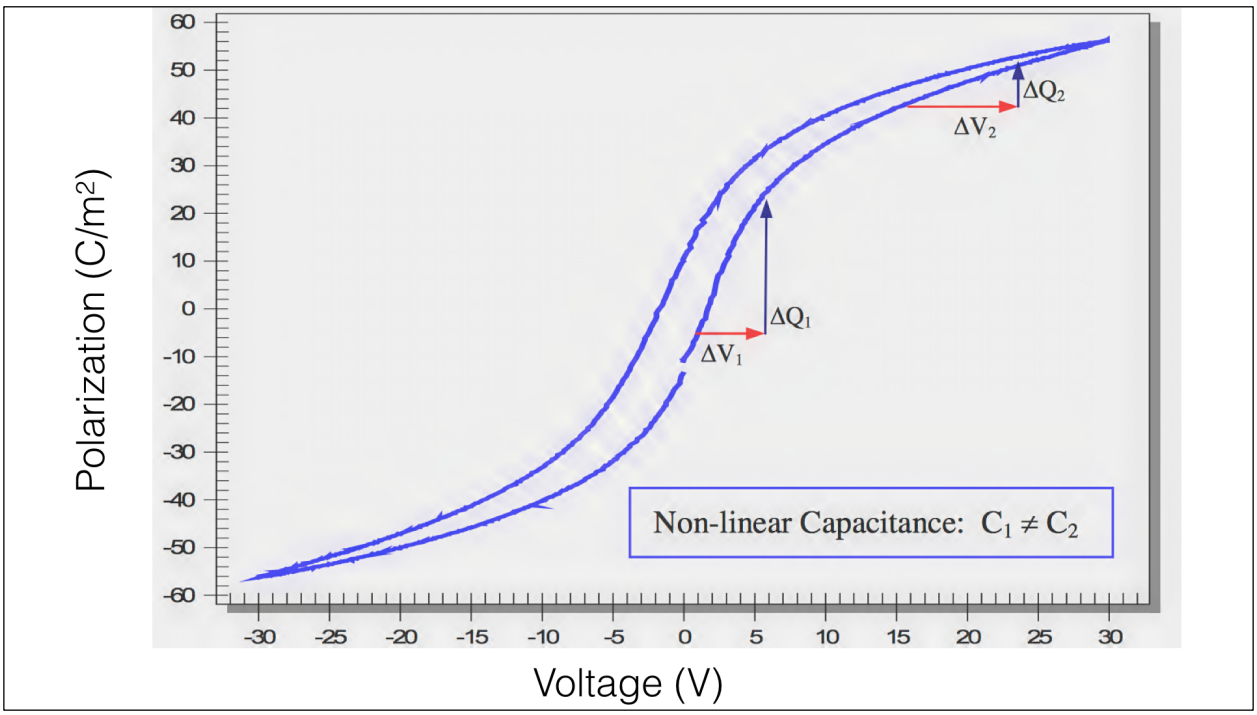


Figure 1.3. *A standard polarization curve for a nonlinear dielectric.*

where dP is the change in polarization induced by the applied electric field E, ε is the dielectric constant, E_b is the breakdown electric field and U gives the energy density of the capacitor. For a linear capacitor, this converts into the following simple equation:

$$U = \frac{1}{2} * \varepsilon * E_b^2$$

While the relationship is not quite as simple for nonlinear dielectrics, the energy storage is still dependent on ε and E_b, and increasing both quantities will help increase the energy density of the capacitor. Thus, the optimal dielectric for high energy density applications would have a high dielectric constant and would break down at high electric fields. Many kinds of ceramics have traditionally served as capacitor dielectrics, for example lead zirconate ($PbZrO_3$) and lead titanate ($PbTiO_3$). However, *polymers* are particularly attractive capacitor dielectrics for energy storage because of their easy processability, flexibility, high resistance to external chemical attacks and most importantly, propensity for graceful failure [34] [35] [36] [37] [38]. Many organic polymers such as Polyvinylidene fluoride (PVDF), Polypropylene (PP) and Polyethylene terephthalate (PET) have been used as dielectrics in a variety of energy and electronics related applications [35] [36] [39]. The key properties of consideration in dielectric polymers are not only the dielectric constant and the dielectric breakdown strength, but also the dielectric loss and mechanical properties, among others.

The current state-of-the-art polymer dielectric is Biaxially Oriented Polypropylene (BOPP), which has a modest dielectric constant (~ 2.2) and low operating temperature (85°C), but a very high dielectric breakdown strength (> 700 MV/m) and a small area (~ $1cm^2$) [40]. This leads to an energy density of ~ 6 J/cm^3. However, BOPP has quite a few limitations; while it can function at high electric fields, its low dielectric constant certainly

imposes a restriction on the energy density. BOPP also suffers from significant dielectric losses due to electronic conduction at higher temperatures [36] [37] [38]. Thus, there have been experimental as well as computational efforts in improving over BOPP as the dielectric polymer candidate.

Much of the work in this regard has taken place with Polyvinylidene fluoride (PVDF), and related modifications. BOPP and most of the other dielectric polymer candidates are nonpolar polymers; atomic and electronic polarizations alone cannot contribute sufficiently to increasing the dielectric constant. PVDF was thus pursued, given that its orientational polarization and high dipole density could be exploited for high energy densities. Defect-modified PVDF, PVDF-HFP and PVDF-CTFE, as well as PVDF with inorganic fillers added to the matrix have been studied and recommended for polymer dielectrics, with high dielectric constants of ~10 and energy densities of 30 J/cm^3 achieved [41] [42] [43] [44] [45]. However, a major problem with PVDF and its derivatives was their ferroelectric nature, which results in a hysteresis D-E loop. This causes heavy energy losses as compared to a paraelectric material, and makes the polymer unsuitable as dielectric for energy storage. Thus, PVDF and related polymers have been explored as possible dielectrics, but came up short because of their ferroelectric behaviour [41] [42] [45] [43] [44].

Polymer dielectrics for modern power electronics applications require not only high energy densities, but also high temperature capabilities and miniaturization, without

affecting the cost too much. Recently, there has been a rising demand for high energy density capacitors due to the on-going electrification of transportation, communication and military and civilian systems [34] [46] [47] [48]. Each of the current possible choices for polymer dielectric applications suffers from one shortcoming or the other. There is a pressing need to expand the pool of polymer dielectric candidates so that novel polymers with the optimal mix of relevant properties can be designed and gradually improved upon. There are significant challenges associated with this, none bigger than the vastness of the polymer chemical universe, and how little of it has been experimentally studied till date [49] [50]. The great challenge here is the difficulty of experimental consideration of a large dataset of polymers; the synthesis and property measurement of polymers in a case-by-case manner, leading (one hopes) to the eventual identification of desirable systems, is a very involved and expensive process. This makes a computation-driven treatment appropriate here: it is much faster to study many materials on a computer, and apply initial screening criteria to down-select polymers that can then be studied experimentally. Therefore, computations when combined with experiments in a rational manner can result in the quick and efficient design of new and improved polymer dielectrics [51].

1.1.3. Rational Co-Design of Polymer Dielectrics

Scientific discoveries and technological innovations have benefited enormously from seemingly "trial and error" practices, and serendipity. A classic example that is often quoted in this context is the work of Thomas Edison surrounding the discovery of suitable materials for the light-bulb filament. Although the "Edisonian Approach" has been replicated time and time again in materials science and related fields by systematically (and laboriously) experimenting on several candidate materials, recent materials discovery efforts approach this problem in a more rational manner using computations in the first screening stage (and in subsequent steps as required). The initial down-selection effort based on advanced computations, when combined with targeted additional computations, materials synthesis, testing and validation, is referred to here as "co-design". This emerging "rational materials co-design" paradigm can significantly reduce costs, provide enormous insights, and speed up the materials design process.

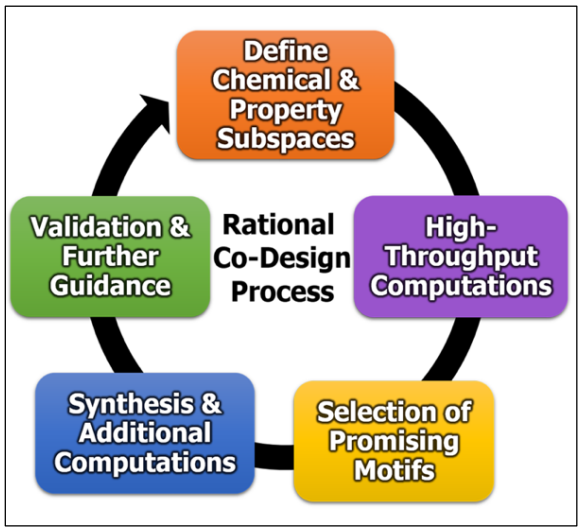


Figure 1.4. *Steps involved in a rational co-design approach.*

For the rational computation-guided co-design approach to work—at least in the manner practiced in the recent past—the problem must be amenable to rapid high-throughput computations, and it should be possible to state the (initial) screening criteria in terms of

calculable properties. If such is the case, the “domain experts” of the materials and applications subfields frame the chemical subspace to be explored defined by the atoms/structural units and the framework in which such units may be placed; this will lead to a list of combinatorial possibilities. Additionally, the domain experts specify computable properties that are relevant to the desired application. High-throughput computations are then performed on these systems, at a chosen acceptable level of theory, to determine the properties deemed important in the initial screening step, leading to a shortlist of potentially useful candidate materials. Following this, the materials synthesis specialist further reduces the shortlist by determining which cases would be amenable to synthesis, considering both the availability of starting materials and the cost of production. Only at this point are any benchtop experiments done, and attempts are made to produce the few selected materials. Those successfully synthesized undergo in-depth computations to include additional details previously ignored during the initial high-throughput screening step (such as the actual crystal structure or morphology, requirement of a higher level of theory for some properties, etc.). The computed results are then compared with measurement results for validation, and the results are analyzed. Further in-depth studies are planned that may lead to an alteration of the initial chosen chemical subspace, and the process may continue in an iterative manner. A possible workflow that captures these notions is portrayed in **Figure 1.4**.

1.2. My Thesis in a Nutshell

In this thesis, a general framework for rationally designing new polymer dielectrics using a computation-driven treatment (presented in **Figure 1.5)** was established and executed. The first step was performing high-throughput DFT computations ('high-throughput' implying the use of computational resources in an automated manner over a long period) to estimate the dielectric constants and band gaps of polymers belonging to a selected chemical subspace, followed by screening for promising candidates [49] [52]. Initial recommendations were made for synthesis, and experimental measurements of the same properties provided validation for the DFT results [49] [50]. While the two steps together constitute *rational discovery*, the design process went far beyond to include 'learning' from the DFT data: this involved looking for correlations between properties and crucial attributes of the polymers, as well as training machine learning models to facilitate property predictions for newer polymers. This learning was applied to perform chemical space expansion, i.e., to predict the properties of thousands of new polymers without the need to perform more expensive

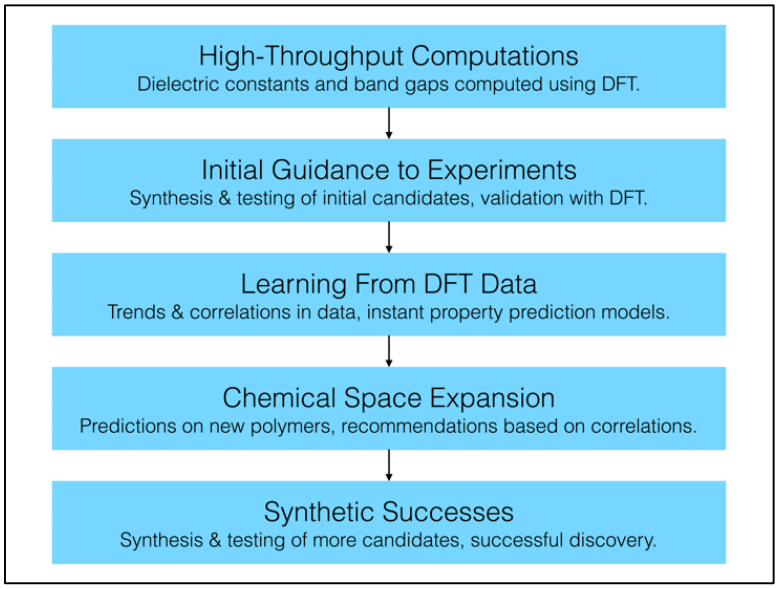


Figure 1.5. *The polymer dielectrics design strategy, involving computational guidance, targeted experiments and machine learning.*

computations [52]. These predictions provided further recommendations for experiments and fresh computations, paving the path to a successful data-driven design of polymer dielectrics. In the following sections, the computation-guided design strategy is described in detail, in the form of high-throughput computational work on organic and organometallic polymer chemical spaces, the synthetic successes that followed the initial computations, and learning from the computational data that led to useful design rules and prediction models. Further, each aspect of the rational polymer design process is presented as a different chapter in this Thesis.

The application of high-throughput DFT to a selected polymer chemical subspace first involved determining the appropriate DFT formalisms for property computation. Density functional perturbation theory (DFPT) [53] [54] [55] is a powerful technique where the dielectric constant of a material is computed by studying the system responses to external perturbations, in this case, electric fields. The band gap can be computed using the hybrid Heyd-Scuseria-Ernzerhof HSE06 electronic exchange-correlation functional [56] [57], which corrects for the band gap underestimation associated with standard DFT. Dielectric constants and band gaps computing using DFPT and the HSE06 functional respectively have been shown to match up very well with experimentally measured results for inorganic compounds as well as common polymers [49] [58]. Thus, these methods were selected for performing the high-throughput DFT computations.

While polymers are known to be either amorphous or semi-crystalline in nature, a crucial assumption made here was to consider a closely packed crystalline model. Although crystal structural information (lattice parameters and bond lengths) is available for many well-known polymers like polyethylene, PVDF and polyacetylene, there isn't sufficient diversity within the family of such common polymers to cover a large enough space for maximum payoff in terms of dielectric properties. To overcome this issue, new chemical spaces had to be devised using some of the most pervasive chemical units as polymer building blocks.

1.2.1. Organic Polymers as Dielectrics

An organic polymer chemical space consisting of seven basic building blocks—CH_2, NH, CO, C_6H_4, C_4H_2S, CS and O—was selected for initial high-throughput computations [49] [50] [52]. Any *n-block* polymer here was generated by linearly connecting *n* blocks with each of them drawn from the 7 possibilities. DFT calculations were carried out for around 300 such *4-block* polymers, which consisted of system sizes (i.e., number of atoms in the system) that were manageable for first principles computations. Crystal structure prediction for such many polymers is no trivial task, especially with scant information available in the literature given that most of these polymers would be hypothetical systems (at least at the first stage). However, recipes for computational prediction of polymeric crystal structures have been well studied in the past [59] [60]. In this work, a

structure prediction algorithm known as Minima Hopping [61] [62] was applied to determine the lowest energy relative packing arrangement of polymer chains (with all energies computed using DFT) in a unit cell, which was then taken to be the ground state crystal structure for the given polymer and used for property computation.

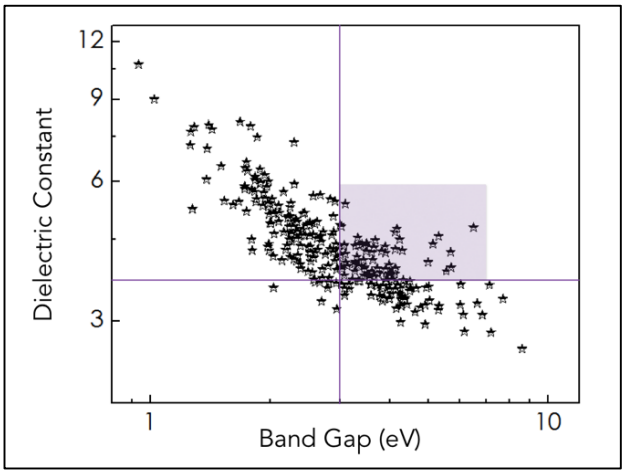


Figure 1.6. *The dielectric constants and band gaps of 4-block polymers computed using DFT.*

The DFT computed dielectric constants and band gaps for all the *4-block* polymers are plotted against each other in **Figure 1.6**. From DFPT, the dielectric constant is computed as two separate components: the electronic part, which depends on atomic polarizabilities, and the ionic part, which comes from the IR-active vibrational modes present in the system. The total dielectric constant is expressed as a sum of the electronic and the ionic parts. The shaded *high dielectric constant, large band gap* region in **Figure 1.6** provided a few promising candidates for initial experiments as well as leads on the most profitable building block combinations for simultaneously enhancing the two properties. For instance, it was observed that polymers containing urea (-NH-CO-NH-), thiourea (-NH-CS-NH-) or imide (-CO-NH-CO-) linkages alongside an aromatic ring such as $-C_6H_4-$ or $-C_4H_2S-$ were present in abundance in the shaded region [49] [50]; subsequently, a few such polymers were considered for experimental studies.

Three polymers belonging to three distinct polymer classes—polyurea, polyimide and polythiourea—were selected out of the shaded region in **Figure 1.6** and synthesized in the laboratory [50]. Appropriate monomers and reaction schemes were adopted here to yield satisfactory quantities of each polymer, following which Ultraviolet-Visible Spectroscopy (UV-Vis) was performed to estimate the band gaps and Time Domain Dielectric Spectroscopy (TDDS) to measure the dielectric constants. It was seen that the experimental results matched quite well with the computational results, providing not only a validation for the high-throughput DFT scheme, but also three novel promising polymer dielectric candidates for energy storage capacitor applications. However, it was seen that these initial polymers had solubility issues and could not be processed into thin films, which is an important capacitor dielectric requirement. To overcome these issues, newer, longer chain polymers belonging to the same and related polymer classes were pursued.

1.2.2. Moving Beyond Pure Organics: An Organometallic Polymer Chemical Space

While interesting new organic polymer motifs were identified as potential capacitor dielectrics, the low ionic dielectric constants seen throughout for the pure organics hinted at a missed opportunity. The lack of correlation between the ionic dielectric constant and the band gap suggested that the former could perhaps be enhanced without adversely affecting the latter [49]. Studies carried out for the oxides and halides of group 14

elements showed that Pb, Sn and Ge based compounds have much higher dielectric constants than their C or Si counterparts, as well as band gap values around or greater than 4 eV [63] [64]. This led to the following thought experiment: if metal based units were inserted in the backbone of an otherwise organic polymer (for instance, Polyethylene), there could potentially be an increase in the dielectric constant compared to the pure organic, while maintaining a large band gap. Metal-organic frameworks (MOFs), which are compounds containing metal clusters surrounded by organic ligands, are commonly used for gas storage, catalysis and supercapacitors [65]. Along similar lines, a metal-organic polymer framework was proposed wherein the organic polymer chain would be interrupted by a metal containing unit. For initial study, Sn was chosen as the metal atom over the poisonous Pb or expensive Ge. Polymer repeat units were generated by introducing tin difuoride (-SnF_2-), tin dichloride (-$SnCl_2$-) and dimethyltin-ester (-COO-$Sn(CH_3)_2$-COO-) units in polyethylene chains in varying amounts [66] [67] [68] [69]. DFT calculations showed that these systems indeed display superior dielectric constants compared to organics for a given band gap value; this caused much excitement in terms of prospective experiments, and the Sn-ester based polymers were duly synthesized and tested, as described in detail in the next section.

The computation-driven discovery of novel Sn-based organometallic polymers paved the path for a sweeping exploration of polymers containing different metals chosen from the periodic table. In **Figure 1.7**, DFT computed results are presented for organometallic polymers constituted of (respectively) 10 different metal atoms [70] [71]; also, shown for

comparison are all the organic polymers discussed earlier. The metal based systems clearly surpass the pure organics in terms of high dielectric constants for given values of band gap. The primary reason behind this increase is the enhanced polarity of chemical bonds in the organometallics because of bonding between electropositive metal atoms and highly electronegative atoms such as O, F and Cl. The swinging and stretching of these polar bonds at low frequencies cause fluctuations in polarization under electric fields, which means they will contribute more to ionic or dipolar parts of the dielectric constant [64] [70] [58]. As seen from **Figure 1.7**, this effect is more pronounced in some organometallics than others: it was observed that the higher the amount of metal in the system, higher is the dielectric constant. The identity of the metal atom itself and its coordination environment were other crucial factors at play here [71].

Many parallel experimental efforts were undertaken to bring the computer-modelled, potentially game-changing materials to life. These include the second generation of computation-guided organic polymers that followed from the initial recommendations, as well as an entire series of Sn-ester based polymers.

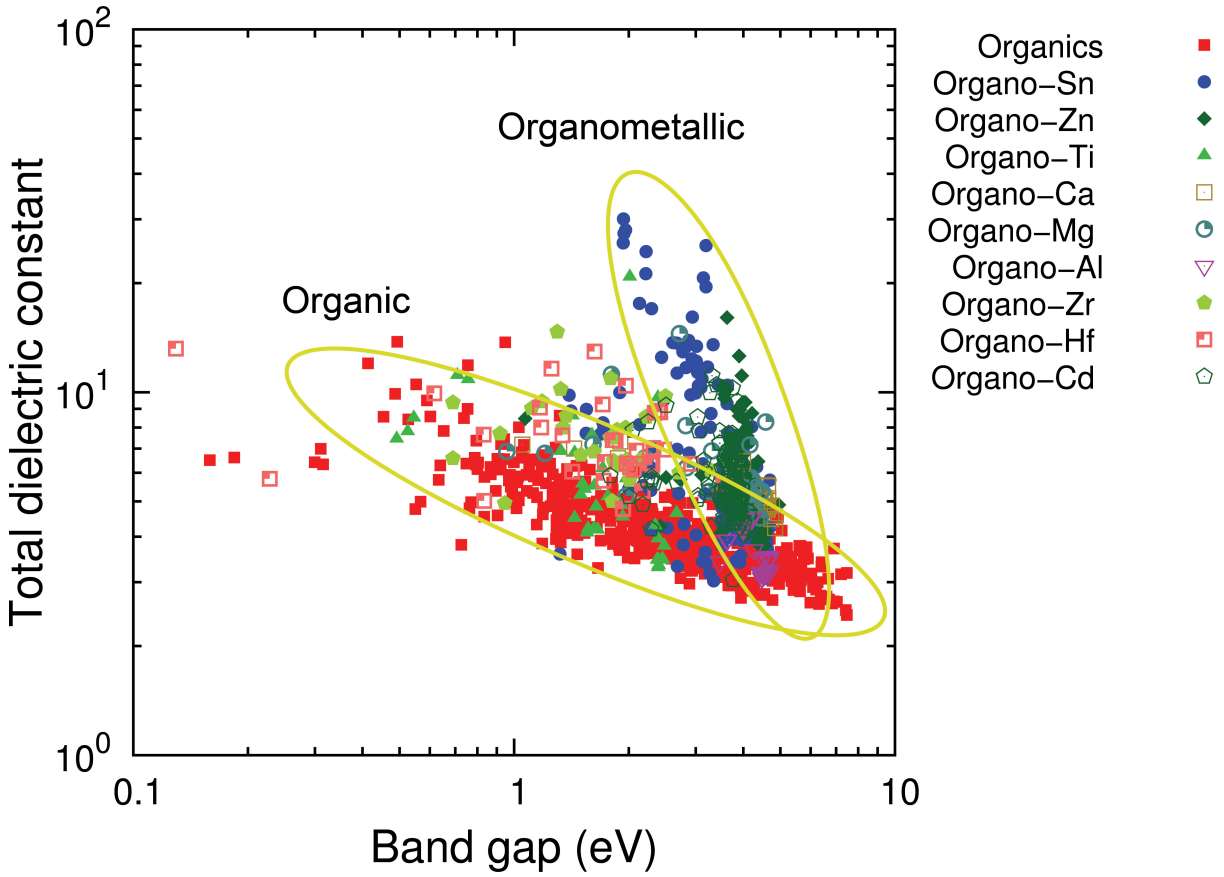


Figure 1.7. *DFT computed band gaps and dielectric constants for all organic and organometallic polymers. The organometallics show higher dielectric constants than the organics for a given large band gap.*

1.2.3. Synthetic Successes

Without the knowledge attained from modelling polymers on a computer, the polymer chemist might end up lost in a sea of possibilities, much like an explorer setting sail on a rudderless ship. The computational models may be viewed as the GPS to the

experimentalist, telling her about potentially promising directions to take. It was revealed from high-throughput computations that -NH-CO-NH-, -NH-CS-NH- or -CO-NH-CO- linkages accompanied by aromatic rings were particularly useful in boosting the dielectric constants and band gaps. While this led to initial synthesis and characterization of three new polymers, the processability and solubility concerns inspired a foray into a second generation of organic polymer motifs: several new polymers belonging to the same generic polymer classes were thus synthesized and tested [72] [73] [74] [75]. Free-standing films were made from most of these polymers, and their dielectric constants, band gaps, dielectric breakdown strengths and loss characteristics, among other properties, were experimentally measured.

Table 1.1 provides a glimpse of three newly designed organic polymers with the best characteristics, and compares their (experimentally measured) properties with the state-of-the-art polymer dielectric, BOPP. The three polymers are a polythiourea named PDTC-HDA, a polyimide named BTDA-HDA and another polyimide named BTDA-HK511, where PDTC stands for Para-phenylene Diisothiocyanate, HDA stands for Hexane Diamine, BTDA stands for Benzophenone Tetracarboxylic Dianhydride and HK511 is a jeffamine-containing ether. Apart from the properties listed earlier, the recoverable energy densities were also estimated for all the polymers using electric displacement-electric field (D-E) loop measurements. Apart from forming free-standing films, each polymer displayed an energy density 2 to 3 times higher than BOPP. In this fashion, (at least) three new organic polymers were successfully designed that can

potentially replace BOPP in capacitor applications [49]. The rationale for pursuing these kinds of polymers came from computational guidance; however, the choice of the specific polymer repeat units was determined by the polymer chemists using their experience and knowledge of chemical feasibility, solvent considerations and film formability. The experimental data thus obtained further bolsters the polymer dataset and even provides vital leads on newer chemical blocks to introduce in polymers for future computational studies.

Polymer Name	**Polymer Class**	**Dielectric Constant**	**Breakdown Field (MV m^{-1})**	**Energy Density (J cm^{-3})**
BOPP	Polypropylene	~ 2.2	~ 700	~ 5
PDTC-HDA	Polythiourea	~ 3.7	~ 685	~ 9
BTDA-HDA	Polyimide	~ 3.6	~ 812	~ 10
BTDA-HK511	Polyimide	~ 7.8	~ 676	~ 16

Table 1.1. *Measured properties for PDTC-HDA, BTDA-HDA and BTDA-HK511, three of the best novel organic polymer dielectrics designed using computational guidance and targeted experiments. Also, shown for comparison are properties for BOPP (biaxially oriented polypropylene).*

Following the fruits yielded by the computation-driven work on organic polymers, attention was diverted to the exciting new field of organometallic polymers. Synthesis of the organo-Sn polyesters proved to be challenging, but the polymer chemists were able to make 12 such polymers containing a varying number of linker -CH_2- units placed between the Sn-based units, yielding the repeat unit -[$Sn(CH_3)_2(COO)_2$-$(CH_2)_n$]-, where *n* changes from 0 to 11. Dielectric constants and band gaps were measured for all the polymers; DFT computations on these systems (this data is part of the organometallic polymers plotted in **Figure 1.7**) revealed three kinds of low energy crystal structural motifs, and properties were computed for each motif of each polymer. Computed and experimentally measured properties of the entire series of organo-Sn polyesters are shown in **Figure 1.8**.

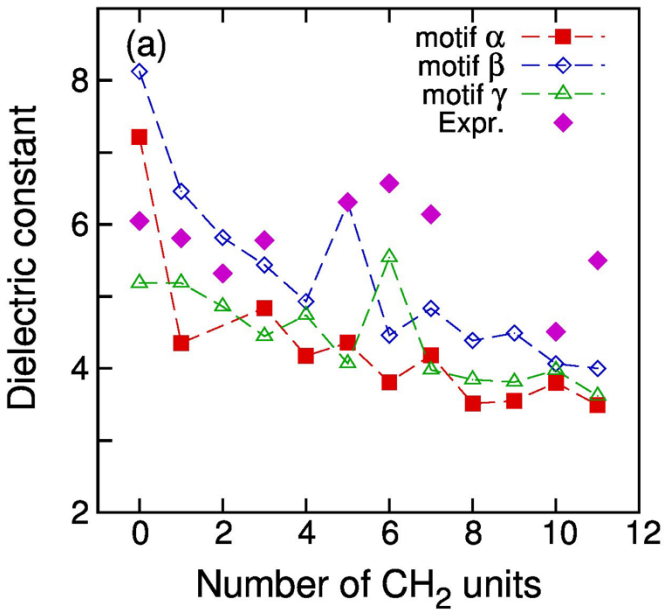


Figure 1.8. *Computational and experimental dielectric constants for a series of organo-Sn polyesters as a function of the number of linker -CH_2- units.*

While the band gaps were seen to be around 6 eV for all the polymers, the dielectric constant displayed a general decrease with increasing number of linker -CH_2- units in the polymer. Very high dielectric constants of > 6 were observed for systems with an intermediate number of linker -CH_2- units (5, 6 or 7) [67] [69]. The remarkable combination of high

dielectric constant and large band gap put the organo-Sn polymers a notch above all the organic polymers studied so far. However, issues of solubility and film formability brought them a notch down again. Co-polymerizing the Sn-ester based polymers with one another (as well as with attractive polyimide or polythiourea based units) led to cast films [68] [76], for which the measured energy densities were roughly the same as BOPP. Regardless, the work on organo-Sn polymers revealed the true promise of the organometallic chemical space, providing motivation for ongoing efforts to further optimize the polymers and obtain next generation capabilities.

1.2.4. Learning from Data

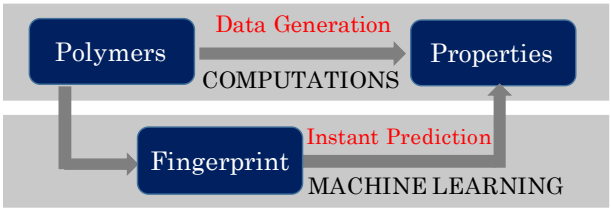


Figure 1.9. *Machine learning strategy to accelerate materials design.*

First principles computations undoubtedly accelerate the materials design process, but are quite time intensive and could benefit from statistical learning approaches. The substantial computational dataset of polymers created in this work can be mined to *learn* about how the important physical and chemical attributes of a polymer contribute to its properties, and thus make qualitative or quantitative forecasts on the behavior of newer polymers. This involves an intermediate polymer 'fingerprinting' step as shown in **Figure 1.9**, wherein every polymer would be represented as a unique string of numbers that can be mapped to the properties to yield

predictive models. In this work, we fingerprinted our polymers by quantifying the chemical composition in terms of the basic chemical units (CH_2, C_6H_4 etc., Fingerprint Type I), and in terms of the basic atomic units (4-fold C atoms, 2-fold O atoms etc., Fingerprint Type II). Within each type, three fingerprints were defined in a hierarchical manner: a) Singles, counting the number of times each unit appeared in the polymer, b) Doubles, counting the number of times each pair of units appeared in the polymer, and c) Triples, counting the number of times each triplet of units appeared in the polymer [52] [77].

ML techniques were applied on the freshly generated computational polymer data within the frameworks of fingerprint types I [52] and II [77] independently. A linear correlation analysis performed between the components of Singles & Doubles (Fingerprint Type I) respectively and 4 properties: band gap, electronic dielectric constant, ionic dielectric constant and total dielectric constant revealed that while CH_2 and O blocks, and CH_2-CH_2 and CH_2-O pairs lead to the highest band gaps, C_4H_2S and CS blocks and their pairs with each other decrease the band gap the most. The effects on the dielectric constant followed quite the opposite trend, thanks to the inverse relationship between the electronic dielectric constant and the band gap. The ionic dielectric constant, meanwhile, is positively contributed to by NH and CO groups, and NH-CO pairs. Thus, the influence of specific blocks and block pairs on the polymer properties was identified, and a similar analysis using fingerprint type II would reveal the atom types and pairs of atom types that are influential.

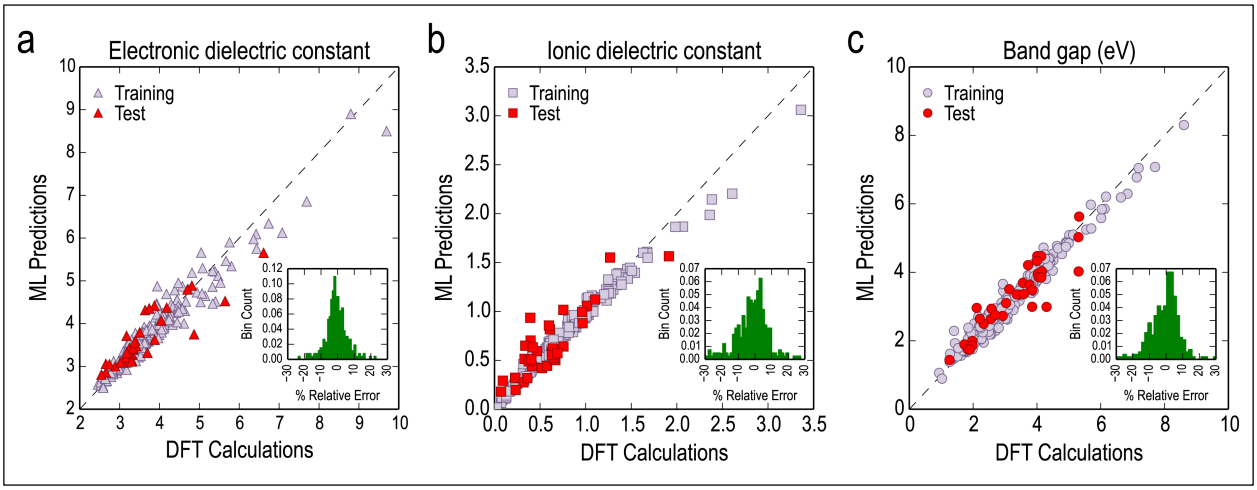


Figure 1.10. *Results of machine learning applied on the DFT data: performances of property prediction models trained using Kernel Ridge Regression.*

Next, a regression algorithm was used to train a model that converts a fingerprint input to its property, within a statistical accuracy. The benefit of having such a prediction model as opposed to mere correlations is being able to make a quantitative prediction of the dielectric constant and band gap of any new polymer, and consequently enhance the initial computational dataset to include hundreds and thousands of new polymers. Kernel Ridge Regression (KRR [78]) was applied on the dataset and predictive models were obtained whose performances are shown in **Figure 1.10**. The Triples fingerprint, given its uniqueness for *4-* to *8-block* polymers and the degree of information it contains, was used for this purpose. The necessary parameters for the KRR formalism were obtained by training the model on a subset of the dataset known as the *training set* (~90% of the total data points), while testing of the model for generality and performance evaluation is done on the *test set*. Statistically satisfactory relative errors of less than 10% were seen for

both the training and test points, as shown in **Figure 1.10**. These predictive models were then tested on nearly 40 random polymers with an arbitrary number of chemical blocks in their repeat units; impressive correspondence was seen between the ML predicted, DFT computed and experimentally measured properties.

1.2.5. Exploring the Polymer Genome

The importance of *data* in driving discovery and innovation puts the onus on scientists to catalogue their computational and experimental results, and whatever insights they may have gained from it, for the benefit of the entire scientific community. This aligns well with the goals and objectives of the *Materials Genome Initiative* [23], and efforts towards the same are evidenced by the rise of many materials databases over the last few years [79] [80] [81] [82] [83].

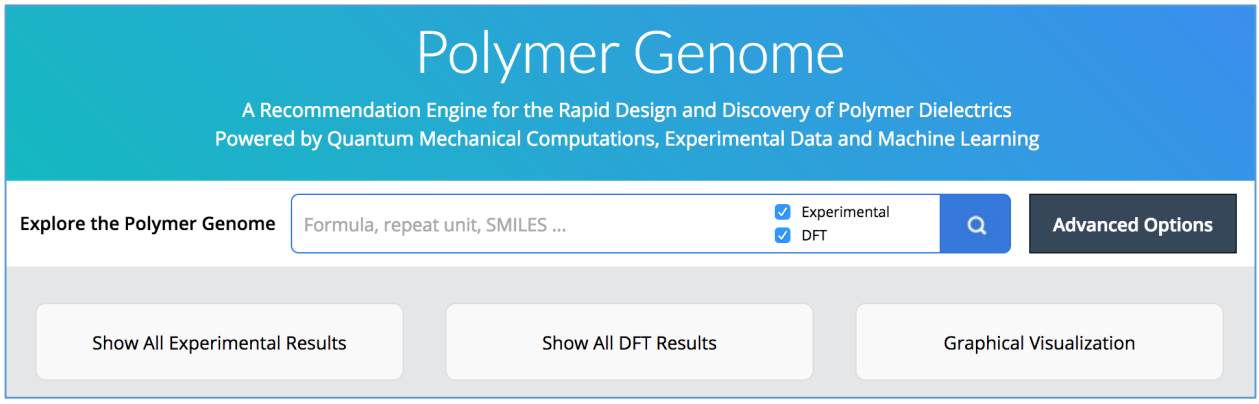


Figure 1.11. *The Polymer Genome platform.*

All the polymer data (including computationally obtained ground state structures, and the DFT estimated and experimentally measured properties) and machine learning models presented in this article may be found within the “Polymer Genome” search platform in *Khazana* [84], an online materials knowledgebase. Any user searching for a polymer by its repeat unit, chemical name or desired properties will be able to access the relevant experimental or computational data, as well as ML predicted properties, and can utilize this information to make an instant go/no-go decision on whether to pursue it for applications of interest. Fingerprinting a polymer in terms of its basic building block (the chemical unit or atom) is like tracking the polymer “genetic material” or “gene”, which is then utilized for explaining trends in the properties; hence the terminology “the polymer genome”. This knowledgebase is an attempt to unravel the polymer genome, and through the medium of past data and machine learning tools, provide ready access to meaningful spaces of the polymer chemical universe to the community.

Chapter 2

COMPUTATIONAL DATA GENERATION METHODS

2.1 Density Functional Theory (DFT)

DFT is a quantum mechanical modelling technique that is extensively used to investigate the electronic structure of any given collection of atoms. The main principle of DFT is that the energy of a system is a unique functional (a term that means the function of a function) of the charge density. The only inputs required for a DFT calculation are the electronic and ionic charges and the masses of the constituent atoms, which makes it a parameter-free method. Today, DFT-based quantum mechanical solutions are known to correctly define atomic level interactions in diverse chemical environments.

While the Schrodinger's equation can be exactly solved for a one-electron system, the DFT formalism converts a many-nuclei, many-electron problem to an effective one-electron problem. This was realized in the pioneering work of Hohenberg, Kohn and Sham in the 1960s [13] [14]. Within Kohn-Sham DFT, the following eigenvalue equation (in atomic units) is solved:

$$[-\bigtriangledown^2 + V_{eff}(r)]\Psi_i^{KS}(r) = \epsilon_i^{KS}\Psi_i^{KS}(r)$$

where $-\nabla^2$ represents the electronic kinetic energy and $V_{eff}(r)$ represents the effective potential energy as experienced by an electron. The latter contains all the electron-electron and electron-nucleus interactions, as well as the potential caused by any external electric field. In practice, the quantum mechanical part of the electron-electron interaction is approximated using a local functional within the local density approximation (LDA), a semi-local functional within the generalized gradient approximation (GGA), or nonlocal hybrid functionals. ϵ_i^{KS} and $\psi_i^{KS}(r)$ are the energy eigenvalues and wave-functions of the Kohn-Sham orbitals. For a given set of atomic positions, the above equation is solved self-consistently to result in converged charge densities (obtained from the wave functions of the occupied states), total energies (obtained from the wave functions and eigen energies of the occupied states) and atomic forces (obtained from the first derivatives of the total energy with respect to the position of any given atom). The atomic coordinates are optimized by requiring that the total energy of the system is a minimum, and the forces on each atom are close to zero.

2.2 Crystal Structure Prediction

For accurate computation of the properties of a polymer, it is essential to estimate its ground state crystalline structural arrangement. While polymers are known to be either amorphous or semi-crystalline in nature, a closely packed crystalline model was an approximation made here for the successful implementation of DFT. Crystal structural information (lattice parameters and bond lengths) is available for many well-known polymers like polyethylene, PVDF and polyacetylene. However, the polymer chemical space we considered for computational study contains several polymers that are hypothetical (at least in the current stage), and thus require crystal structure prediction.

We implemented the following strategy for obtaining ground state crystal structures: single polymer chains (that do not interact with one another) were first locally optimized to yield low energy chain configurations with correct bond lengths and bond angles. Following this, we built unit cells with two polymer chains next to each other, and estimated total energies of the system for different relative packing arrangements of the chains. The lowest energy configuration thus estimated by DFT was taken to be the ground state crystal structure for the given polymer, and used for property quantification.

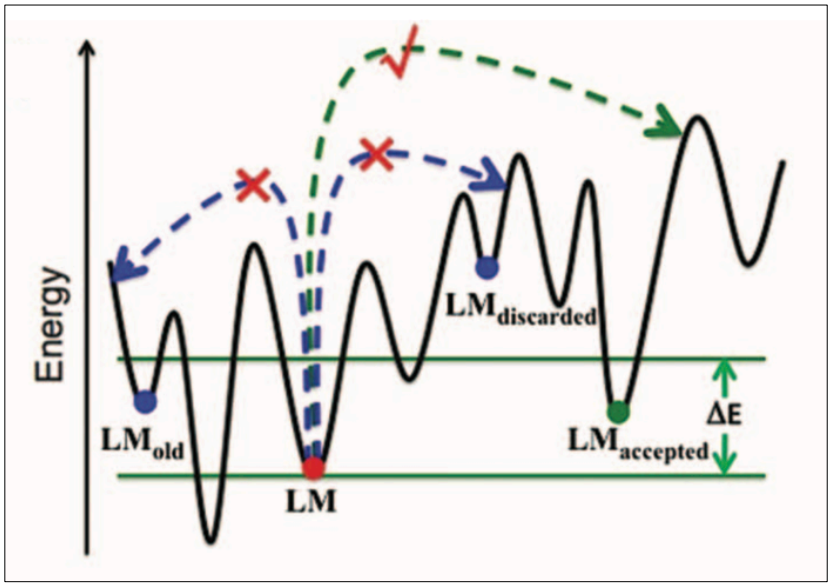


Figure 2.1. *The Minima-Hopping algorithm for crystal structure prediction. Reproduced from ref. [85]*

The exploration of the polymer configurational space was performed using the Minima Hopping algorithm, developed by Godecker et al. [61] [62] [85] and successfully applied in various problems for global optimization of the potential energy surface. Different minima on the energy landscape lead to different stable configurations. In this method, there is an inner part that deals with the jump of the system from the current minimum to a local minimum of another basin, and an outer part that concerns accepting or rejecting this new local minimum. Hopping to a new minimum is achieved using a short molecular dynamics (MD) simulation by applying kinetic energy to the atoms, causing the system to crossover a barrier (that is smaller than a pre-chosen value, $E_{kinetic}$) to a new configuration. These MD runs continue repeatedly with different values of $E_{kinetic}$ until new minima are found. Geometric relaxation into the next closest local minimum happens using standard steepest descent and conjugate gradient methods; it is accepted if the energy difference is smaller than another pre-chosen value, E_{diff}.

The Minima Hopping algorithm, as described here and pictorially depicted in **Figure 2.1**, was applied to estimate the ground state structures of all novel polymers that constituted the computational dataset in this Thesis. Using the stable structure, DFT is applied to compute two properties, the dielectric constant and band gap.

2.3 Computation of Dielectric Constant and Band Gap Using DFT

The application of high-throughput DFT to a selected polymer chemical subspace first involved determining the appropriate DFT formalisms for property computation. Density functional perturbation theory (DFPT) [54] [55] is a powerful technique where the dielectric constant of a material is computed by studying the system responses to external perturbations, in this case, electric fields. The band gap can be computed using the hybrid Heyd-Scuseria-Ernzerhof HSE06 electronic exchange-correlation functional [56] [57], which corrects for the band gap underestimation associated with standard DFT.

Using DFT as a tool for estimating the bandgap and dielectric constant, one question arises: how accurate are these estimates with respect to state-of-the-art experimental measurements? In **Figure 2.2**, we provide a comparison of the DFT computed bandgaps and dielectric constants with the corresponding experimental values for a few chosen materials. We show here some inorganic compounds that have been labeled in the plots,

a few known polymers—namely polyethylene (PE), polypropylene (PP), polystyrene (PS), poly(ethylene terephthalate) (PET), polyoxymethylene (POM), and poly(ethylene oxide) (PEO)—and, finally, some new polymers that we discuss in detail. We thus claim that an initial DFT screening step involving the computed bandgap and dielectric constant is a satisfactory approach.

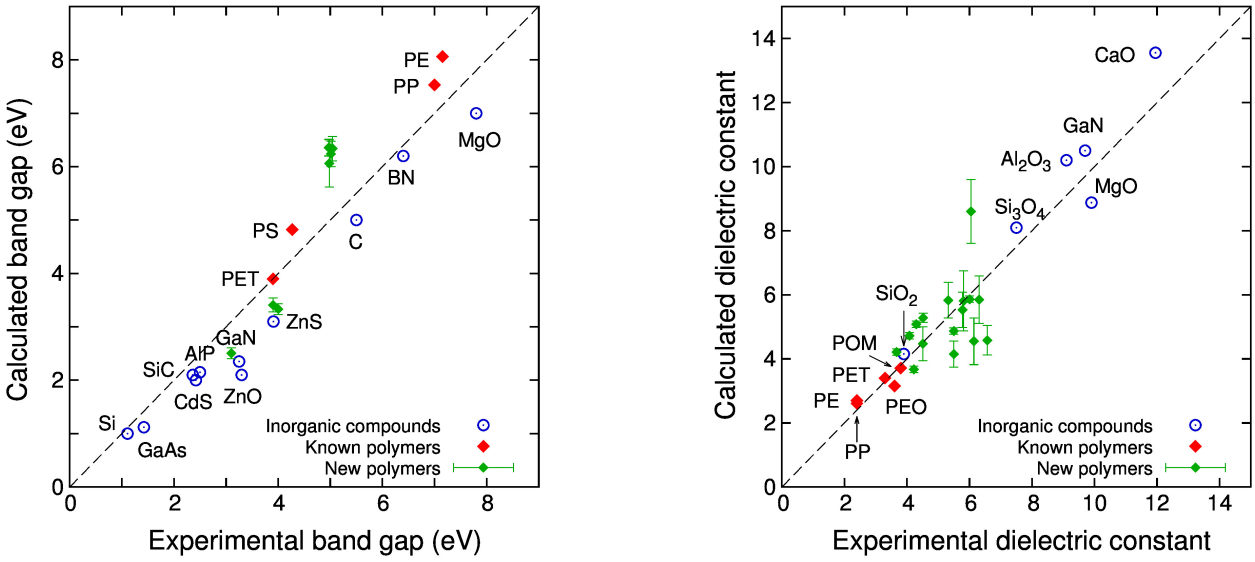


Figure 2.2. *Experimental validation of dielectric constants and band gaps computed using DFT for a few known inorganic compounds, known polymers, and new polymers studied as a part of this Thesis.*

2.4 DFT Calculation Details

All the computational data reported in this Thesis was prepared with density functional theory (DFT) calculations, using the projector augmented-wave (PAW) formalism [42] as implemented in Vienna Ab initio Simulation Package (vasp). The default accuracy level of our calculations is “Accurate”, specified by setting PREC = Accurate in all the runs with vasp. The basis set includes all the plane waves with kinetic energies up to 400 eV, as recommended by vasp manual for this level of accuracy. PAW datasets of version 5.2, which were used to describe the ion-electron interactions, are also summarized in **Table 2.1**. The van der Waals dispersion interactions, known to be important in stabilizing soft materials dominated by non-bonding interactions like polymers, were estimated with the non-local density functional vdW-DF2. The generalized gradient approximation (GGA) functional associated with vdW-DF2, i.e., refitted Perdew-Wang 86 (rPW86), was used for the exchange-correlation (XC) energies.

Element	POTCAR	Element	POTCAR	Element	POTCAR
Aluminum	Al	Bromine	Br	Carbon	C
Calcium	Ca_sv	Cadmium	Cd	Chlorine	Cl
Fluorine	F	Hydrogen	H	Hafnium	Hf_sv
Magnesium	Mg_sv	Nitrogen	N	Oxygen	O

Phosphorus	P	Lead	Pb_d	Sulfur	S
Tin	Sn_d	Titanium	Ti_sv	Zinc	Zn
Zirconium	Zr_sv				

Table 2.1. *VASP PAW potentials of the elements used for calculations in this work.*

Because the examined material structures are significantly different in terms of the cell shape, the sampling procedure of their Brillouin zones must be handled appropriately. For each structure, a Monkhorst-Pack k-point mesh of a given spacing parameter h_k in the reciprocal space was used. For the geometry optimization and dielectric constant calculations, h_k = 0.25 $Å^{-1}$ while the band gap calculations have been performed on a finer γ-centered mesh with h_k = 0.20 $Å^{-1}$. We further set the lower limit for the Monkhorst Pack mesh dimensionality, that is, the number of grid points along any reciprocal axis is no less than 3, regardless of how short the reciprocal lattice dimension along this axis is. During the relaxation step, we optimized both the cell and the atomic degrees of freedom of the materials structures until atomic forces are smaller than 0.01 eV $Å^{-1}$. Calculations for band gap E_{gap} was then carried out on top of the equilibrium structures. Because E_{gap} is typically underestimated with a GGA XC functional like rPW86, this important physical property has also been calculated with the hybrid Heyd-Scuseria-Ernzerhof (HSE06) XC functional with an expectation that the calculated result would become much closer to the

true material band gap. Both E_{gap}^{GGA} and E_{gap}^{HSE06}, the band gap calculated at the GGA-rPW86 and HSE06 levels of theory, are provided in all the entries of the dataset. Finally, the dielectric constant ε of these structures was calculated within the DFPT formalism as implemented in vasp package. Calculations of this type involve the determination of the lattice vibrational spectra at γ, the center of the Brillouin zone. This information is also used to compute the IR spectra of some structures for validation.

Chapter 3

ORGANIC POLYMERS AS DIELECTRICS

3.1 Strategy for Rational Computation-Guided Search

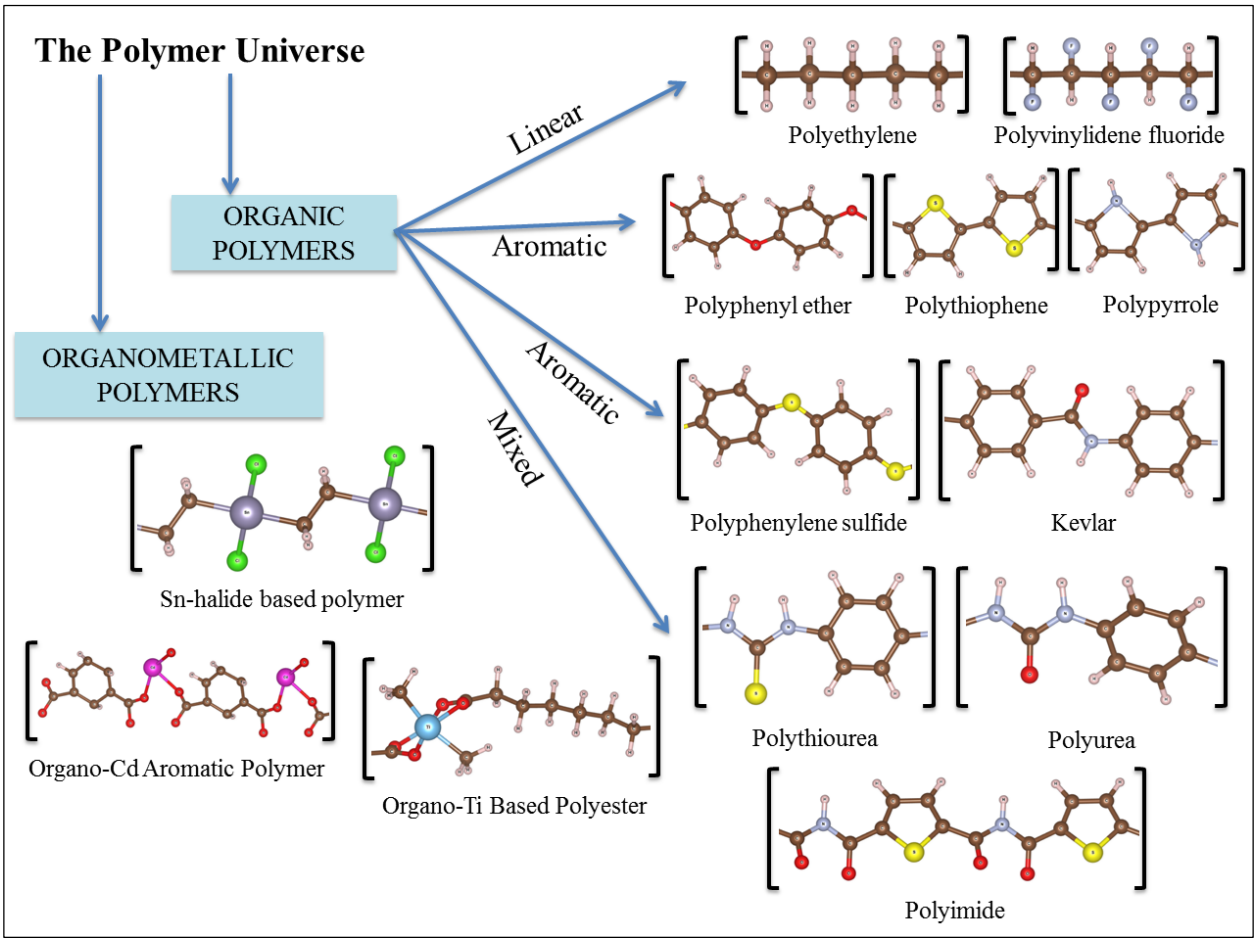


Figure 3.1. *The vast chemical space spanned by a variety of polymer building blocks.*

With respect to exploring large polymer chemical spaces for dielectric capabilities, the historical work has generally featured a few limited subclasses/families of polymers. The staggeringly large number of chemical unit possibilities, and the various kinds of possible connectivity sequences of the units giving rise to different polymer repeat units (**Figure 3.1** provides a flavor), make experimental examination of a substantial number of these systems impractical.

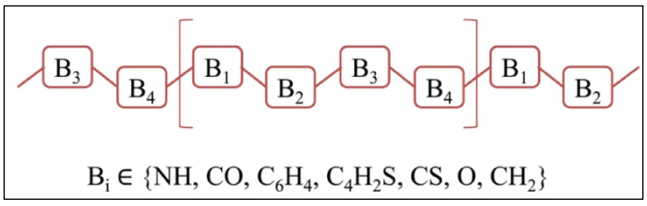


Figure 3.2. *The chemical subspace of polymers containing 7 basic building blocks.*

However, a controlled subspace selected out of this vast expanse could be a dataset of polymers that is ripe for high-throughput computational study. We thus considered a chemical subspace of polymers containing the following 7 building blocks: CH_2, CO, NH, C_6H_4, C_4H_2S, CS, and O. This set of building blocks was chosen based on how ubiquitous they are in well-known polymer systems, and was deemed to be suitable for performing a controlled computational study of organic polymers [50] [52] [71]. Several different *n-block polymers* (containing *n* blocks in the repeat unit) were generated by linearly connecting randomly chosen blocks out of the set of 7, as shown in **Figure 3.2**. This led to the possibility of hundreds of different symmetry-unique polymers; in fact, there is an exponential explosion of the total number of polymers that are possible as we go to higher values of *n*.

While there are 7^n ways to populate *n* blocks using the 7 motifs, this number reduces when we consider symmetry uniqueness in the polymers, which is determined in terms of inversion and translational invariances. The total number of symmetry unique polymers possible for different values of n are shown in **Figure 3.3** alongside the respective 7^n values for comparison. For the purposes of high-throughput computations, a dataset of purely *4-block* polymers was considered here. A total of 406 symmetry-unique *4-block* polymers can be formed using the 7 building blocks, of which only 284 were subjected to DFT computations. Chemical intuition and prior knowledge dictates that some combinations of adjoining chemical blocks make for unstable systems, leading to the elimination of all polymers consisting of O-O, CS-CS, CO-CO and NH-NH pairs; hence the reduced number [52].

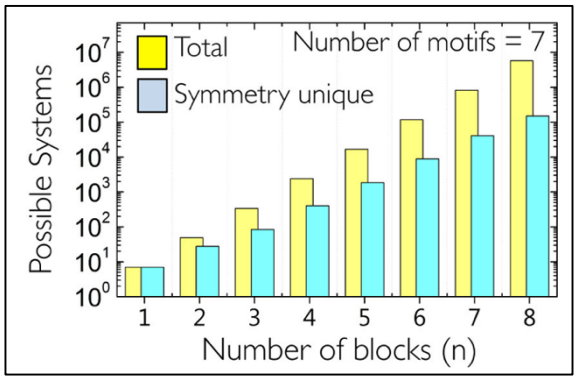


Figure 3.3. *The total number of n-block polymers for different values of n.*

First principles computations using density functional theory (DFT) were performed on all such polymers in a high-throughput manner, resulting in the bandgap [56] [57] and the electronic and ionic dielectric constants, denoted by ε_{elec} and ε_{ion}. In the terminology used here, ε_{ion} includes all non-electronic contributions to the dielectric response, including bond stretching and bond (dipole) rotations allowed within a crystalline lattice. The sum of these two quantities, namely, ε_{elec} and ε_{ion}, typically computed within the perturbation formalism of DFT [54] [55], is the total dielectric constant ε_{tot}, which is relevant for

comparison with measurements. For dielectric polymers to maximize the amount of energy stored in a capacitor, the dielectric constant, as well as the dielectric breakdown field, should be high (and the dielectric loss should be low). Given the difficulty in computing the breakdown field (especially the true engineering breakdown field) and the dielectric loss at the low frequencies (kHz) of interest from first principles, the bandgap (known to be correlated with the breakdown field and dielectric loss [44]) was used as a proxy instead. Thus, an initial screening criterion of "high dielectric constant" and "large bandgap" was used to down-select suitable polymers.

3.2 High-throughput Computations

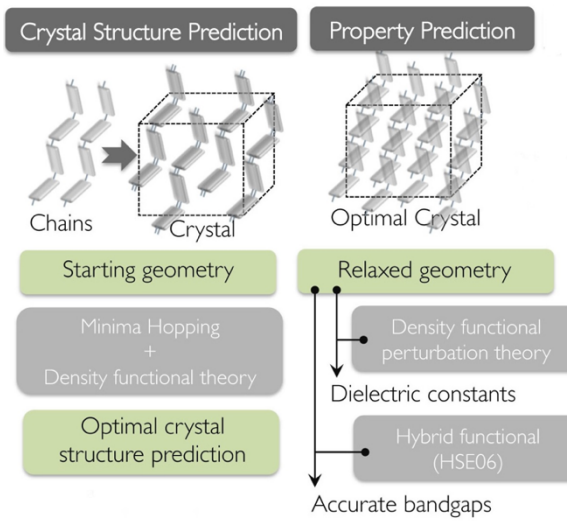


Figure 3.4. *Computational data generation framework.*

For the creation of the initial dataset via first principles computations, we restricted ourselves to *4-block* polymers, that is, polymers built with 4 blocks in the repeating unit (with each of these drawn from the pool of 7 building blocks shown in **Figure 3.2**). The crystal structures of all 284 4-block polymers were determined using the minima hopping method, with the necessary total potential

energies and atomic forces computed using density functional theory (DFT). Details of the structure prediction and property estimation using DFT are provided in **Chapter 2**. With the stable 3-dimensional arrangements of polymer chains determined for all 284 polymers, their relevant properties were calculated: the bandgap (E_{gap}), computed using hybrid electron exchange-correlation functional, and the electronic (ε_{elec}), ionic (ε_{ion}) and total (ε_{tot}) dielectric constant, computed using density functional perturbation theory. A computational workflow is depicted in **Figure 3.4**.

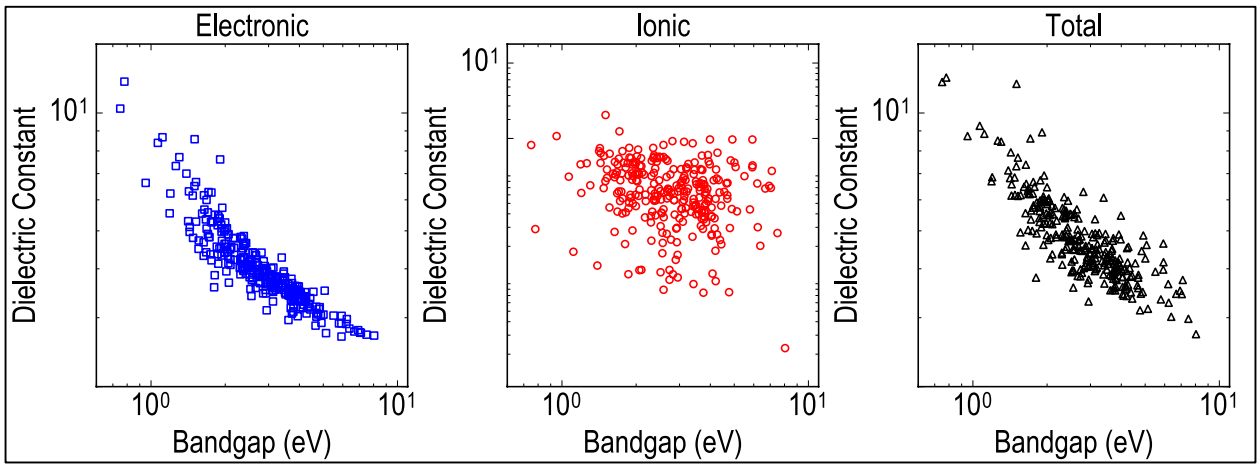


Figure 3.5. *The dielectric constants (divided into electronic and ionic parts) and band gaps of 284 polymers computed using DFT.*

With the idea of selecting promising polymer units for high-energy density applications, the dielectric constants (electronic, ionic and total) were plotted against the electronic bandgaps, as shown in **Figure 3.5**. Several important insights emerged from this analysis (as well as actual guidance for synthesis). These include:

i) The electronic part of the dielectric constant is inversely correlated to the bandgap; hence, a large electronic part of the dielectric constant, although desirable (owing to the short timescales of this response) is not safe, as it will lead to poor insulators;

ii) The ionic part of the dielectric constant is immune to the above trend, i.e., it is uncorrelated, or only weakly correlated, to the bandgap; this contribution to the dielectric constant should thus be exploited;

iii) The best polymers for energy-density applications are those with the best tradeoff between the total dielectric constant and the bandgap. Although a wide spectrum of dielectric constant values (~2 to ~12) and band gap values (~1 eV to ~9 eV) were covered by this chemical subspace of polymers, only about 10% of the total points populated the *high dielectric constant, large band gap* region. To limit the large amount of data to a sample size that could be considered for experiments, the region was defined by setting a threshold for the dielectric constant at 4 and the lower limit for the bandgap at 3 eV. The systems in this region were seen to be predominantly composed of at least one of the polar units, namely NH, CO, and O, and at least one of the aromatic rings, namely C6H4 and C4H2S. NH, CO, and O tend to enhance the ionic part of the dielectric constant, whereas the aromatic groups boost the electronic part. This immediately provided experimentalists their first vital leads.

3.3 Initial Computational Guidance and Synthetic Validation

Based on insights from the high-throughput computational study, while a number of polymers (and a number of chemical groups/combinations of groups) were seen to be promising avenues to pursue, three polymers were recommended for synthesis and characterization: –NH–CO–NH–C_6H_4–, –CO–NH–CO– C_6H_4–, and –NH–CS–NH– C_6H_4– [49] [50]. As a synthetic starting point, these three polymers were ideal, as they represented three different polymer classes, namely polyureas, polyimides, and polythioureas, while maintaining the same aromatic unit, C_6H_4.

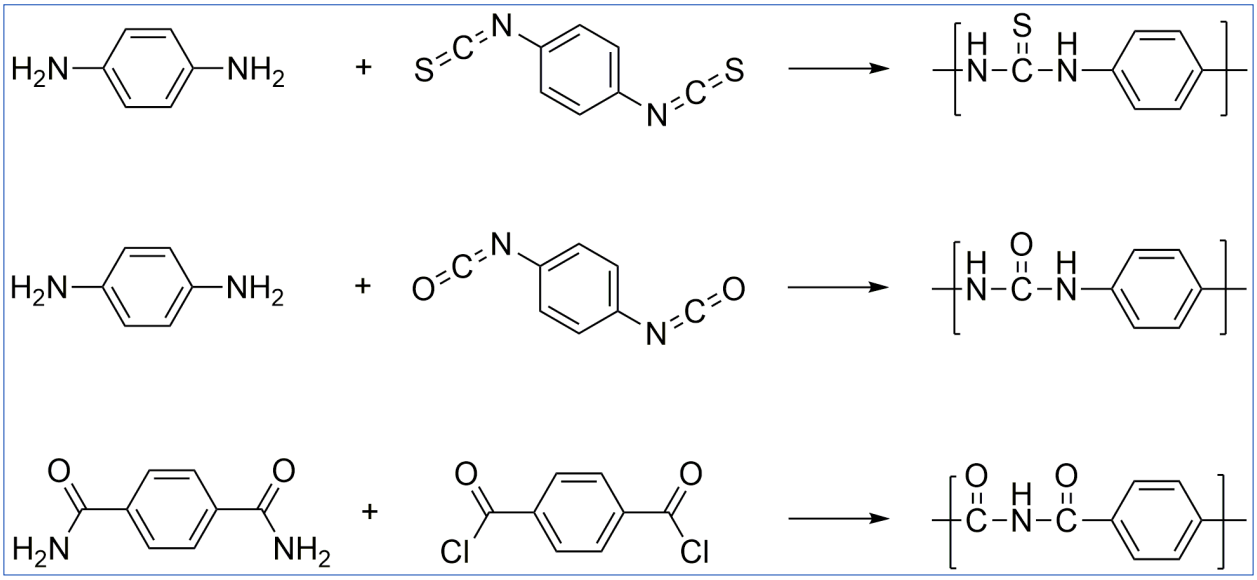


Figure 3.6. *Synthetic scheme for the three organic polymers down-selected from high-throughput DFT.*

Several synthetic routes had to be considered here, as it was not possible to synthesize all the predicted structures by the chemistry known at the time. To fit the criteria of using only 4 structural units, monomers were chosen to obtain repeat structure units as shown in **Figure 3.6**. Polymerizations to create all three down-selected candidates proceeded in a stepwise mechanism leading to condensation polymers, which limited the quantity of byproducts and side reactions that would ultimately become impurities and interfere with dielectric spectroscopy measurements.

Traditional polymer characterization techniques were employed to study the synthesized polymers. Fourier transform infrared (FTIR) spectroscopy was used to confirm the chemical structure, X-ray diffraction (XRD) to determine the crystalline structure, and UV-vis spectroscopy to estimate the bandgap. Furthermore, time-domain dielectric spectroscopy (TDDS) was employed to study the dependence of the dielectric properties on the frequency. Meanwhile, the crystal structures, morphologies, and relevant properties were studied in greater detail and with more accuracy using computations. The dielectric constants and band gaps computed from DFT and measured experimentally are presented in **Table 3.1**. While the dielectric constants were found to range from 4 to 6, which is double that of Polyethylene or Polypropylene, the bandgaps were seen to be greater than 3 eV.

Polymer		DFT E_{gap} (eV)	Expt. E_{gap} (eV)		DFT ε_{tot}	Expt. ε_{tot}
-[NH-CO-NH-C_6H_4]$_n$-		~ 3.5	~ 3.9		~ 4.9	~ 5.6
-[CO-NH-CO-C_6H_4]$_n$-		~ 4.1	~ 4.0		~ 5.7	~ 4.5
-[NH-CS-NH-C_6H_4]$_n$-		~ 2.7	~ 3.1		~ 5.8	~ 6.2

Table 3.1. *Experimentally measured properties for initial recommendations (listed using the polymer repeat units) from high-throughput DFT, and a comparison with DFT computed values.*

Even though all first principles computations are performed for purely crystalline structures and the synthesized polymers are largely semi-crystalline or amorphous, the measured dielectric constant ranges matched up very well with predictions for the three polymers. The closeness of the computed properties and the experimental measurements, as evident from **Table 3.1**, provided validation for the high-throughput DFT scheme, and three novel promising polymer dielectric candidates emerged for capacitive energy storage applications. Thus, a "computations → experiments → computations" synergistic loop was successfully pursued in the design of new organic polymer dielectrics.

The success of the initial mating between DFT calculations and synthetic efforts for dielectric studies gave way to other possible systems to be developed using the same

initial computational data. The synthetic efforts branched into three different studies, each involving a different polymer class to further understand the theoretical and experimental properties of proposed organic dielectrics. As shown in **Figure 3.7**, the three polymers studied at first and discussed above gave way to the study of a number of: i) polythioureas, ii) polyureas, and urethanes, and iii) polyimides.

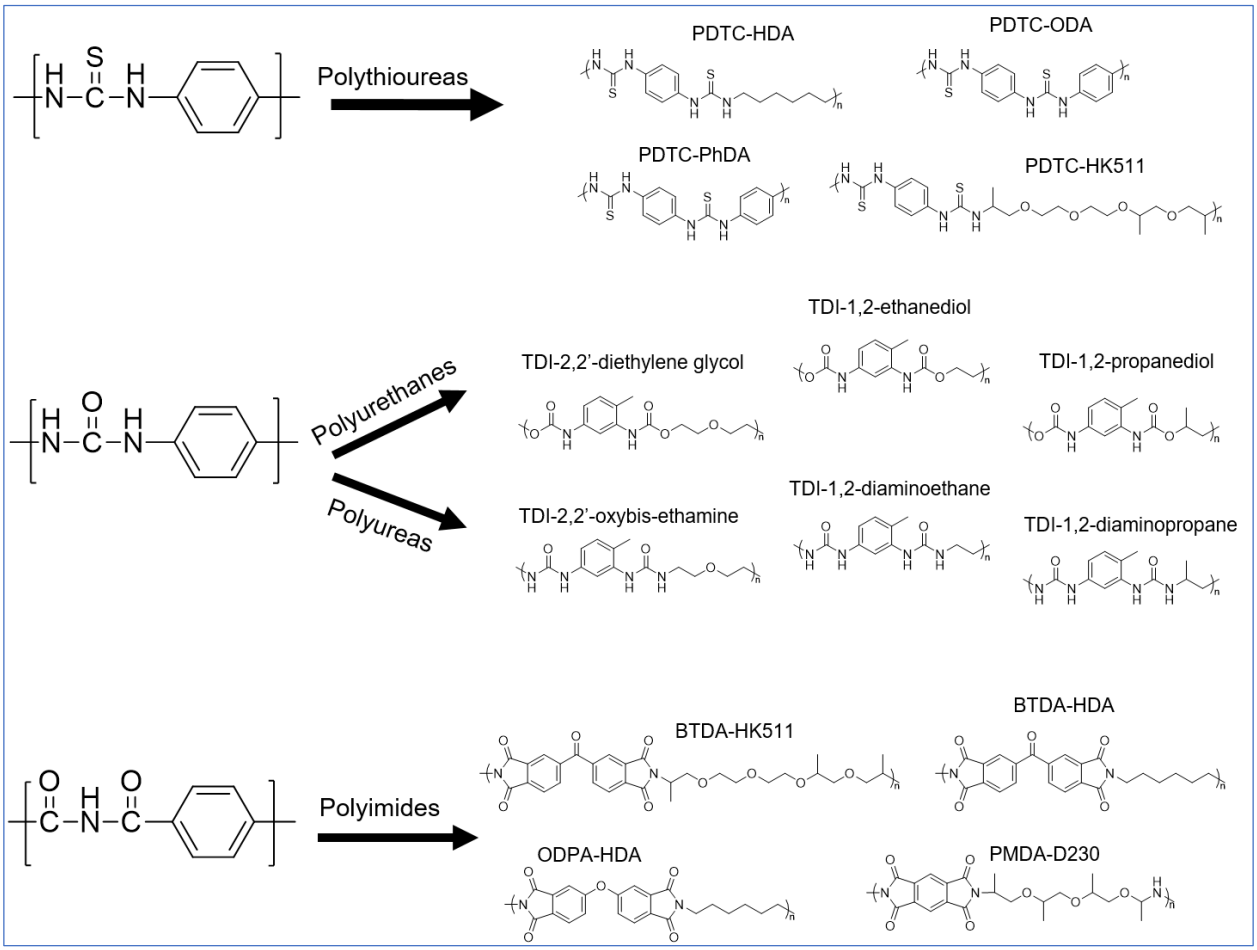


Figure 3.7. *Extensions to new polythioureas, polyureas, polyurethanes and polyimides.*

3.4 Extensions to New Polymers

3.4.1 Polythioureas

The first polythiourea synthesized, -NH-CS-NH-C_6H_4-, provided a simple structure for calculations and synthesis; however, the actual synthesized material proved to be insoluble and was seen to melt at or above its degradation temperature, and was therefore not processable into thin films. Polythioureas with longer and more flexible chains were thus studied [75], with the idea of improving the processability. By keeping one monomer constant, para-phenylene diisothiocyanate (PDTC), and by varying the diamines between different aromatic, aliphatic, and polyether monomers, structure-property relationships were derived for the dielectric constant of this family of polythioureas. Five different diamines were chosen: 4,4'-oxydianiline (ODA), Bis(4-aminophenyl)methane (MDA), 1,4-Diaminobenzene (PhDA), hexane-1,6-diamine (HDA) and Jeffamine HK511. Further, a thiophosgene reaction was employed to mimic an industrial scale reaction for the polymerization of a related thiourea compound reported by Wu *et al.* [86] The measured (and some computed) properties, namely ε_{elec}, ε_{tot}, dielectric loss (tan δ) and E_{gap}, of the six polymers thus studied are listed in **Table 3.2**, and the structures of some of them are shown in **Figure 3.7**.

Polymers	ε_{elec}	ε_{tot}	tan δ	E_{gap} (eV)
PDTC-ODA	3.20 (3.86)	4.52 (5.42)	0.0233	3.22 (3.27)
PDTC-MDA	3.28 (3.69)	4.08 (4.59)	0.0348	3.16 (3.41)
PDTC-PhDA	N/A	4.89	0.0144	3.07
PDTC-HDA	2.92 (3.29)	3.67 (4.01)	0.0267	3.53 (3.75)
PDTC-HK511	2.69	6.09	0.0115	3.51
Thiophosgene-MDA	3.03	3.84	0.0226	3.3

Table 3.2. *Experimental and computational (shown in brackets) data for polythioureas. Measured ε_{tot} and tan δ correspond to room temperature (r.t.) and a frequency of 1 kHz; ε_{elec} is reported as the squared value of the measured refractive index.*

Extensive polymer characterization was carried out on these polythioureas; this included obtaining their respective FTIR, XRD and solution NMR spectra to determine the structure. ε_{elec} was estimated experimentally by measuring the refractive index using ellipsometry—the square of the refractive index is equal to the electronic component of the dielectric constant. Further computations were performed on these specific polythioureas to obtain more accurate property estimates, and were compared with experimental results. The low energy polymer configurations were subjected to band gap and dielectric constant calculations. The experimentally determined ε_{elec} and ε_{tot} are comparable to but lower than the calculated DFT values, which is due to calculations being done solely on the crystalline polymer state while experimental measurements are

averages over the crystalline and amorphous regions. This was further confirmed by comparison of calculated and experimental IR and XRD spectra of both fiber precipitates and solution cast films.

The dielectric spectrum obtained using TDDS for one of these polythioureas (that showed very attractive properties), namely PDTC-HDA (the polymer made from 1,4-Diisothiocyanatobenzene and hexane-1,6-diamine), are shown in **Figure 3.8 (a)** (this polymer is shown at the top of **Figure 3.7**). While ε_{tot} increased with operational temperature due to the chains becoming more mobile and thus enhancing dipole alignment, this effect diminished at high frequencies, as the dipoles are unable to align as quickly. E_{gap} provides a good theoretical substitute for the dielectric breakdown field strength since a higher band gap would imply a higher threshold for impact ionization; however, access to the breakdown field (E_b) is possible through either direct breakdown measurements, or electric displacement-electric field (D-E) loop measurements. The latter measurements also provide a pathway to assess linearity and energy recovery, and to obtain energy density estimates. Such D-E measurements were done for PDTC-HDA, and this is shown in **Figure 3.8 (b)**. Further, the recoverable energy density as a function of the applied electric field is shown in **Figure 3.8 (c)**. The ability to operate at high electric fields would lead to a significant increase in energy density. For PDTC-HDA, an energy density of 9.3 J/cm^3 was achieved at a maximum applied field of 685 MV/m, which is a substantial improvement over BOPP (almost double its value). The maximum energy density is expected to further improve with better processing conditions to remove

contaminants such as dust impurities and residual solvent as this will lead to higher values of the breakdown field. An important point worth noting is that although the initial computational screening was based just on the dielectric constant and band gap, the directions identified, in terms of materials subclasses to pursue, have led to polymers with acceptable dielectric loss and high breakdown field (and hence, energy density).

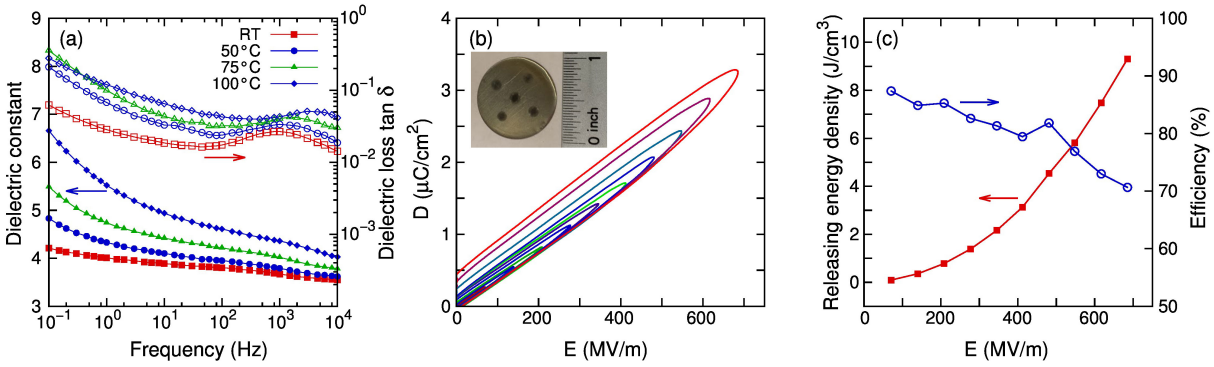


Figure 3.8. *(a) Dielectric constant and dielectric loss tan δ measured at room temperature (RT), 50°C, 75°C, and 100°C, (b) D-E loops, and (c) the releasing energy density and efficiency of PDTC-HDA. A film of PDTC-HDA is also shown in (b) as an inset. Figures adapted from Ref. [75].*

3.4.2 Polyureas and Polyurethanes

After having a successful correspondence between expected values from DFT and experimental results for the 4-block polyurea originally recommended, namely, -NH-CO-NH-C_6H_4- (with ε_{tot} greater than 5 and E_{gap} above 3 eV), synthetic efforts returned to the

urea structure. To increase experimental variation in the tested systems, two sets of nearly identical polymers were studied: one comprising of polyureas and the other of a related class, polyurethanes [73]. Polyurethanes were attractive for this comparison as the reaction is very similar, substituting out diamines for dialcohols and adding in a small amount of dibutyltin diluarate catalyst. All other reaction conditions were held the same in both cases to give comparable results.

In this case, 5 diamines and their respective diols (which act as polar segments) were polymerized along with toluene diisocyanates (TDI); select polyureas and polyurethanes are shown in **Figure 3.7**. The polymers were purified and dried before being characterized structurally, thermally and electronically; the results for a series of polyureas (labeled 1A-5A; here, different alkyl and aromatic groups are flanked by -NH-CO-NH- units) and the corresponding polyureathanes (labeled 1B-5B; the same groups are flanked by -NH-CO-O- units) are tabulated in **Table 3.2**.

Polymer	1A	2A	3A	4A	5A	1B	2B	3B	4B	5B
$\varepsilon_{1\mathrm{kHz}}$	5.18	4.29	3.47	2.08	6.19	6.35	6.74	5.81	4.09	10.5
$\tan\delta_{1\mathrm{kHz}}$ (%)	0.758	0.889	1.73	3.12	4.29	1.26	1.54	1.39	1.56	1.88

Table 3.3. *Measured ε_{tot} and tan δ values for the polyureas and polyurethanes.*

In general, the polyurethanes showed a higher ε_{tot} than their corresponding polyurea cousins, which can be explained by the higher electronegativity of the urethane group compared to the urea group. Also following the same electronegativity argument is the fact that more carbon atoms in the backbone decreased ε_{tot} across the board, as carbon has a diluting effect on the urea and urethane linkages as shown in previous studies [70]. The increase in ε_{tot} seen in polymers 5A and 5B shows the beneficial effect of adding ethers into the backbone of polymers on dielectric constant, and agree with previously reported values for polyether urethanes [87]. In summary, this study confirmed that the best ways to increase the dielectric constant in aromatic polyurea and polyurethanes involves maximizing polarizability through electronegative atoms such as oxygen, and decreasing carbon in the backbone to maximize the contribution from the functional groups. Further extensive studies and optimization are required to realize practically useful polyureas and polyureathanes. Such work is in progress.

3.4.3 Polyimides

Polyimides are attractive for dielectric applications due to their high thermal stability, which allows them to have a higher operational temperature than traditional polymers such as BOPP. Inspired by the identification of polyimides as an attractive subclass in the initial computation-based screening, ten polyimides were synthesized [74] by choosing four different rigid aromatic dianhydrides, namely Pyromellitic dianhydride (PMDA), 3,3',4,4'-Benzophenone tetracarboxylic dianhydride (BTDA), 4,4'-Oxydiphthalic

anhydride (OPDA) and 4,4'-Hexafluoroisopropylidene diphthalic anhydride (6-FDA), along with two flexible diamines with aliphatic chains of different lengths, propane-1,3-diamine (DAP) and hexane-1,6-diamine (HDA). Also chosen were two different ethers containing Jeffamines (D230 and HK511), based on previous positive results. In this fashion, 10 polyimides were studied, some of which are shown in **Figure 3.7**. The measured band gap, dielectric constant and dielectric loss of these polymers are shown in **Figure 3.9**.

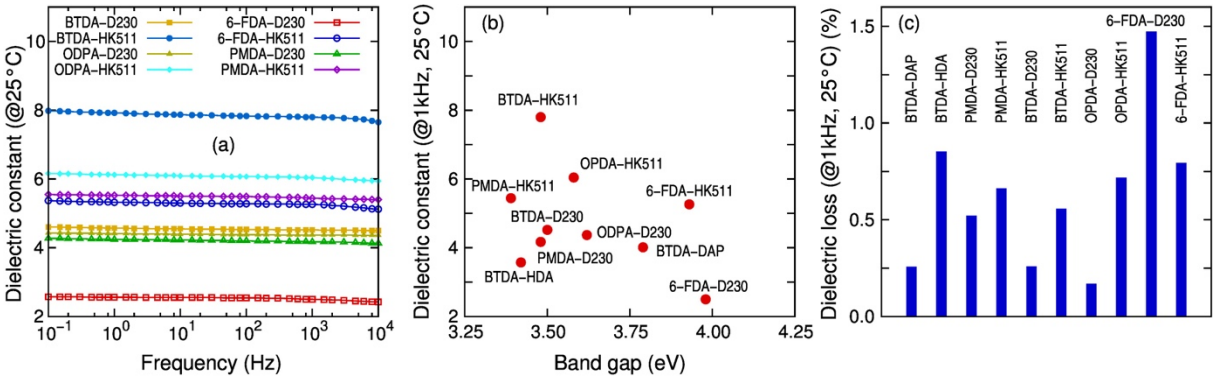


Figure 3.9. *(a) The dielectric constant measured for all the polyimides at room temperature (25°C) given as function of frequency, (b) dielectric constants measured at 1kHz plotted against the band gaps, and (c) dielectric losses measured at 1kHz. Figures adapted from Ref. [74].*

The dipolar polarizability of the imide functional group leads to all the polyimides having higher dielectric constants than BOPP. It can be seen from **Figure 3.9 (b)** that the polyimide BTDA-HK511 showed the highest ε_{tot} of 7.8, which is in large part due to orientational polarization imparted by the polyether section. BTDA-HK511 was also seen

to have one of the lowest dielectric loss values of all the polyimides shown in **Figure 3.9 (c)**, around 0.5%, while being able to operate at temperatures up to 75°C. Large scale free standing films could be made from this polymer, as shown in **Figure 3.10 (a)**. Results of TDDS measurements performed at increasing temperatures on BTDA-HK511 are plotted in **Figure 3.10 (b)**, and a Weibull plot of the breakdown measurements for BTDA-HK511 is shown in **Figure 3.10 (c)**.

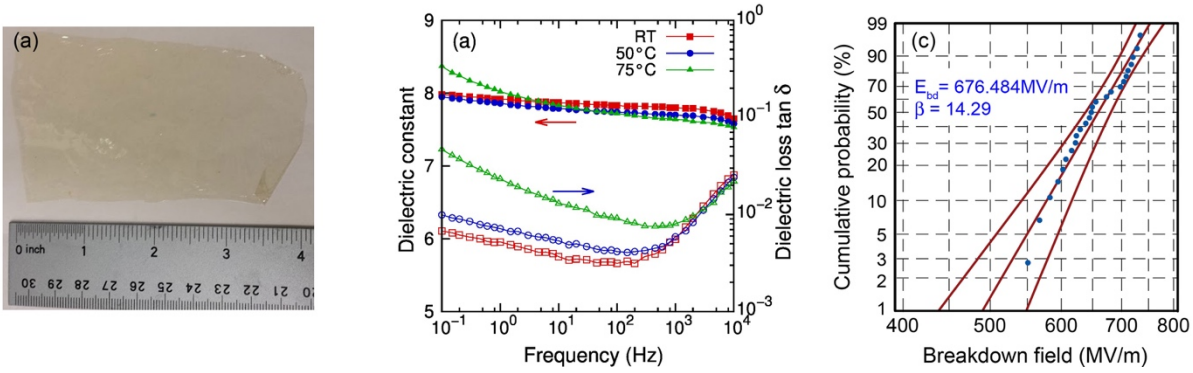


Figure 3.10. *(a) A solvent cast free standing film of BTDA-HK511 with a thickness of 12 microns, (b) the dielectric constant and loss at the room temperature (RT), 50°C, and 75°C, (c) Weibull plot of dielectric breakdown, with the characteristic breakdown field and the slope parameter indicated. Figures adapted from Ref. [74].*

It was concluded that while the dielectric constant decreases with frequency due to slower orientation of the dipoles with alternating electric fields, dielectric loss increases because of chain relaxations. The Weibull analysis was used to determine a characteristic breakdown field of 676 MV/m, as shown in **Figure 3.10 (c)**. For a straight comparison, the same exercise was performed for the polyimide that formed the best films, namely,

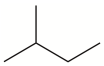
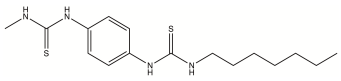
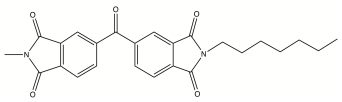
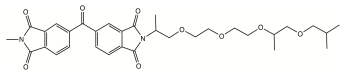
BTDA-HDA, which displayed the highest breakdown field (among all polyimides considered here) of 812 MV/m, although with a modest dielectric constant of less than 4. The respectable breakdown field of BTDA-HK511 along with its high dielectric constant of 7.8 corresponds to a potential energy density of 15.77 J/cm^3. *This is nearly three times that of BOPP.* The co-design approach has thus lead to quite a few polymer dielectrics that could potentially surpass BOPP in actual applications.

3.5 Major Synthetic Successes

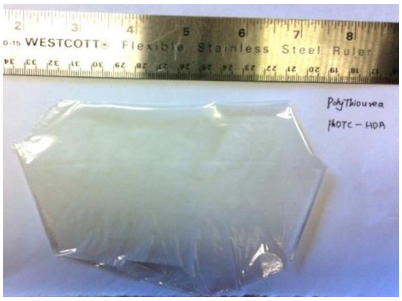
In summary, the efforts on rational co-design of purely organic polymer dielectrics for capacitive energy storage led to successful synthetic studies of several novel polymers belonging to generic polymer classes such as thiourea, urea, urethane and imide, showing attractive dielectric and electronic characteristics [49]. Processability issues with the initial computation-driven candidates inspired the foray into longer chain polymers, many of which were studied computationally as well. As a result, this entire design process can be said to happen in a cyclical "computations → experiments → computations" manner, truly embodying the philosophy described by **Figure 1.4** in **Chapter 1**. Three polymers thus designed that showed the most attractive combination of properties (i.e., high dielectric constant, breakdown at large electric fields, low dielectric losses, satisfactory thermal stability and good film formability, among other features) out

of all the polythioureas, polyureas, polyurethanes and polyimides are shown in **Figure 3.11**.

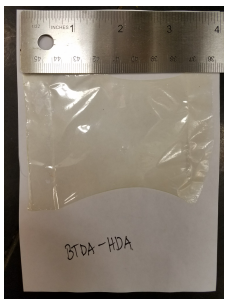
a

Polymer Name	Polymer Repeat Unit	Dielectric Constant	Breakdown Strength (MV/m)	Energy Density (J/cm^3)
BOPP		2.2	700	~ 5
PDTC-HDA (Polythiourea)		3.7	685	~ 9
BTDA-HDA (Polyimide)		3.6	812	~ 10
BTDA-HK511 (Polyimide)		7.8	676	~ 16

b



c



d

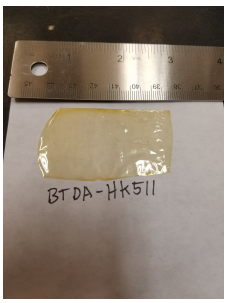


Figure 3.11. *Three of the best novel organic polymers synthesized and characterized as part of the rational co-design process, and their properties as compared with BOPP, the current state-of-the-art capacitor dielectric. Also shown are the free-standing films of each polymer.*

One of the key things to note here is that the ε_{tot} of these polymers are much higher than BOPP while each of them shows a comparable or higher E_b, because of which the energy densities observed are 2 to 3 times as high as the 5 J/cm^3 that is characteristic of BOPP. This is a significant result, and immediately raises the prospects of commercializing three novel metallized polymer film dielectrics for energy storage capacitors. While free-standing films were made from these polymers (the three respective films are also pictures in **Figure 3.11**), efforts are currently underway to synthesize blends and copolymers of the most impressive polythiourea and polyimide homopolymers. It is hoped that this will lead to further improved film formation, higher thermal stability and possibly, even increased energy densities. The success of the rational co-design approach in the discovery of several novel organic polymer dielectrics as discussed here was recently reviewed by us within the context of capacitor dielectrics design [49].

Chapter 4

BEYOND PURE ORGANICS: ORGANOMETALLIC POLYMERS

4.1 Compounds of Group 14 elements: building blocks for advanced insulator dielectrics design

Being in the group with the most diverse set of properties among all in the periodic table, the Group 14 elements (C, Si, Ge, Sn, and Pb) are particularly interesting candidates for structure–property investigation. Motivated by the need to create new insulators for energy storage and electronics applications, we studied a few compounds based on Group 14 elements, namely their dihydrides, dichlorides, and difluorides. Using density functional theory (DFT) calculations, we established patterns in their properties, including favored coordination chemistry, stability, electronic structure, and dielectric behavior. While a coordination number (CN) of 4 is commonly associated with Group 14 elements, there is a significant deviation from it down the group, with CNs as high as 7 and 8 common in Pb. Further, there is an increase in the relative stability of the +2 oxidation state as opposed to +4 when we go from C to Pb, a direct consequence of which is the existence of the di-compounds of C and Si as polymers, whereas the compounds of Ge,

Sn, and Pb are strictly 3D crystalline solids. The coordination chemistries are further linked with the band gaps (E_{gap}) and dielectric constants (ε_{tot}, divided into the electronic component ε_{elec} and the ionic component ε_{ion}) of these compounds. We also see that the more stable difluorides and dichlorides have large E_{gap} and a modest ε_{elec}, and most of the Ge and Sn compounds have remarkably large ε_{ion} because of the presence of polar and more flexible bonds. The staggering variation in properties displayed by these parent compounds offers opportunities for designing derivative materials with a desired combination of electronic and dielectric properties.

As part of this work, structure-property investigations were carried out for several compounds with the formula unit XY_2, where X is one of the Group 14 elements and Y is either H or one of two halogen atoms, Cl and F. It should be noted that Group 14 elements form XY_4 type molecules as well with varying stabilities, but the di-Y formula unit gives us the opportunity to consider solids and thus look at properties originating from the crystalline nature of the material. We considered all the XY_2 compounds in five previously known prototypical structures to obtain the stable structural conformations, the XY_2 formation and cohesive energies, the band gaps, and the dielectric constants, for a total of 15 systems.

Density functional theory (DFT) [13] [14], as implemented in the Vienna ab initio software package (VASP) [88] [89], was applied to determine the electronic structure and properties of the 15 XY_2 compounds. The Perdew, Burke and Ernzerhof (PBE) [90]

functional was used with projector augmented wave (PAW) [91] pseudopotentials. All calculations were carried out with a tight convergence criterion of 10^{-8} eV and an energy cut-off of 600 eV. Since the traditional PBE functionals are unable to capture van der Waals (vdW) interactions correctly, we incorporated the DFT-DF vdW correction [92] [93] into our calculations. Further, it is known that PBE calculations underestimate band gap values, and this deficiency is overcome (to a large extent) using the Heyd-Scuseria-Ernzerhof (HSE) [56] functional instead. The relaxed geometries of the structures that we obtained were used as input for the density functional perturbation theory (DFPT) [54] [55] calculations, which provided us with the dielectric constant tensors that included the electronic components as well as the ionic components. The reported dielectric constant values are obtained by averaging the diagonal elements of the tensors.

4.1.1. Structures and Coordination Chemistries

Each XY_2 compound is known to exist in one of five different crystal structural prototypes shown in **Figure 4.1**, referred to as Types A to E. The specific case of CH_2 is nothing but polyethylene (PE) that occurs in structure Type-A, in which individual PE chains can be discerned characteristic of typical polymers. It was observed that all the dihydrides (of which CH_2 is the only one experimentally known [94]), as well as the difluorides and dichlorides of C and Si, adopt structure Type-A, and are thus polymers isostructural with

PE. The difluorides and dichlorides of Ge, Sn and Pb are not polymers but closely packed 3D crystals, and are found to exist in the other four structures (Type-B to Type-E).

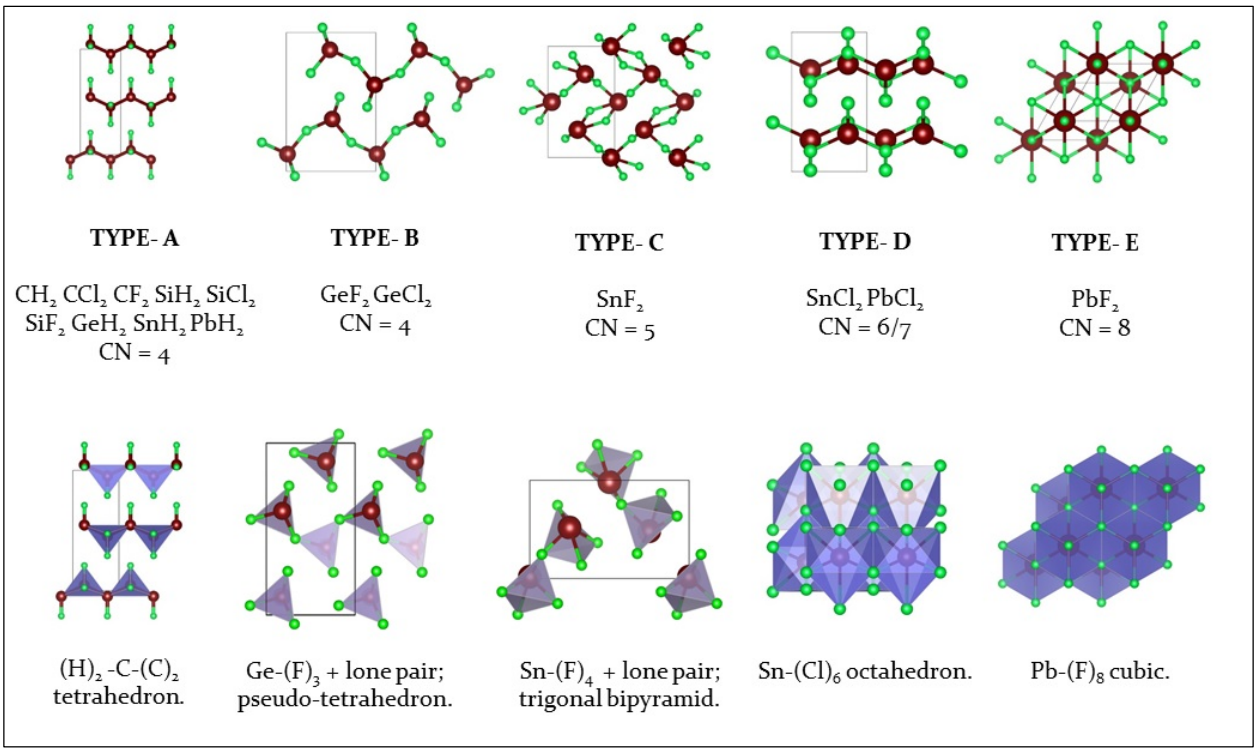


Figure 4.1. *Structure Types A to E along with the respective CNs and the example systems. Note that as many as 9 systems adopt structure Type-A: all C and Si-based systems, and dihydrides of Ge, Sn and Pb.*

The PE structure [94] [95] consists of an orthorhombic unit cell with every C atom at the center of a tetrahedron whose 4 corners are formed by the 2 H atoms and 2 C atoms it is connected to. C is in a 4-fold coordination and we see long chains of connected CH_2 units that are arranged in a crystal with weak interactions between adjoining chains [96]. This is a strictly polymeric structure, and the stability of the long chains lends PE some of its

most vital properties. Replacing the H in PE by F or Cl does not change the CN or crystal structure (Type-A) at all, as we can expect from C being a stubborn CN 4 element [97]. SiH_2, SiF_2 and $SiCl_2$ are also seen to crystallize as polymers in the same Type-A structure with Si in a clear 4-fold coordination. It is interesting that all the C and Si based systems here are polymers; this comes from both elements favoring the tetravalent (IV) state to the divalent (II) state, as well as from the stability of C-C and Si-Si bonds.

Down the group, the divalent state increases in stability, meaning Ge(II), Sn(II) and Pb(II) are very common. That said, GeH_2, SnH_2 and PbH_2 all prefer structure Type-A again. These hydrides are not experimentally known and it stands to reason that if the Group 14 elements are forming dihydrides at all, they are going to be polymers isostructural to PE. It is when we go to the respective difluorides and dichlorides that we finally see something different. Although Ge, Sn and Pb are in an overall +2 oxidation state, it is known from the works of Cotton [97], Trotter et al. [98] and Denes et al. [99] that more electrons than the 4 present in the valence band are involved in bonding, which leads to a higher CN and to 3D crystal structures with little or no 'polymeric' behavior.

Both GeF_2 and $GeCl_2$ crystallize in orthorhombic structure Type-B [98] [100], where GeY_2 units (Y = F, Cl) are successively linked to each other via a bridge Y atom. Ge is strongly bonded to 3 Y atoms, but has one other Y atom as its neighbor that it weakly bonds to, which means that although we see stacking of chains, there is stronger bonding between them than seen in PE. As previously explained by Doll et al. [100], the 3 F atoms around

Ge along with the lone pair of electrons on it result in a pseudo-tetrahedral geometry (a CN of 4). It should be noted that structure Type-B is a metastable state for SiF_2 and $SiCl_2$, and Si(II) may indeed be the preferred state at higher temperatures.

The difluorides and dichlorides of Sn and Pb deviate a fair bit from the structures seen so far. Unlike structure Types A and B, clear 3D networks are seen here which results in an increase in coordination around the X atom. SnF_2 adopts the Type-C structure, with a tetragonal unit cell where each SnF_2 unit is linked to two other SnF_2 units via terminal F atoms [99] [101]. Every Sn atom makes strong bonds with four F atoms and there is a lone pair of electrons on it, resulting in a trigonal bi-pyramidal geometry and thus, a CN of 5. The Type-C structure is in excellent agreement with the stable γ-SnF_2 polymorph that has been experimentally studied [102]. The dichlorides of Sn and Pb both adopt crystal structure Type-D, where an orthogonal unit cell contains chains of XCl_2 units (X = Sn, Pb) linked to each other through a bridging Cl atom. These chains are interconnected by means of 3 weaker X-Cl bonds [103] (4 in case of Pb [104]), resulting in little or no polymeric behavior, and a CN of 6 and 7 for $SnCl_2$ and $PbCl_2$ respectively. The last remaining compound is PbF_2, which adopts structure Type-E that is isostructural with the Fluorite structure [105] (seen in compounds like CaF_2 and ZrO_2). There is a conspicuous absence of any kinds of 1D chains of connected XY_2 units here, and the system is a pure crystalline solid. A cubic unit cell contains the Pb atoms in FCC positions and the F atoms in tetrahedral voids [106] [107]. This has been shown with the rhombohedral primitive cell

in **Figure 4.1** to better exhibit the 8-fold coordination adopted by Pb here, which is the highest we have encountered among the XY_2 systems.

Shown in **Figure 4.2 (a)** are the CNs for all the XY_2 compounds in their most stable structures, clearly seen to be increasing from C to Pb. The lattice parameters of the 15 compounds we studied are listed in Table I. The experimentally reported values are shown as well and generally seen to be in good agreement with the DFT-DF results. Table II lists the X-Y bond length comparisons of DFT values with experimental values for a few of the XY_2 compounds, with encouraging agreement once again.

While both X-X and X-Y bonds are present in structure Type-A, only X-Y bonds are there in the other four structure types. This is because of the reduced tendency for catenation as we go from C to Pb. While C-C and Si-Si have very high bond strengths, Ge, Sn and Pb are more likely to form bonds with other electronegative elements than with themselves. All the Type-A structures have X in an oxidation state of +4, whereas in the other four structures, X is in a +2 state owing to the lack of polymeric chain linkages. It is noted that there is an increased contribution of inner shell electrons to the bonding as we go from C-based compounds to Pb-based compounds, and the increase in CN down the group makes sense.

4.1.2. Energetics

Next, we explored the energetics of the XY_2 crystal structures relative to the elements in their standard states, as well as relative to isolated XY_2 chains. The latter is an attempt to quantify the tendency of these systems to exist as polymers with distinct 1D chains. We thus estimated two kinds of energies: the formation energy E_{form}, and the cohesive energy E_{coh}, defined as

$$E_{form} = E(XY_2 \text{ crystal}) - (E(X) + E(Y_2)) \quad (1)$$

$$E_{coh} = E(XY_2 \text{ crystal}) - E(XY_2 \text{ chain}) \quad (2)$$

$E(XY_2$ crystal) and $E(XY_2$ chain) are the respective DFT energies (per XY_2 unit) of the XY_2 crystal and the XY_2 chain [64], while E(X) is the DFT energy of X in its elemental standard state and $E(Y_2)$ is the DFT energy of a Y_2 molecule [64]. For elemental standard states, the diamond structure was considered for C, Si, Ge and Sn while Pb was considered in an FCC structure [108]. The XY_2 chain being considered here for each system consists of isolated chains of repeating XY_2 units, much like a PE chain, in the possible assumption that this is how a hypothetical polymer chain of said XY_2 system would exist.

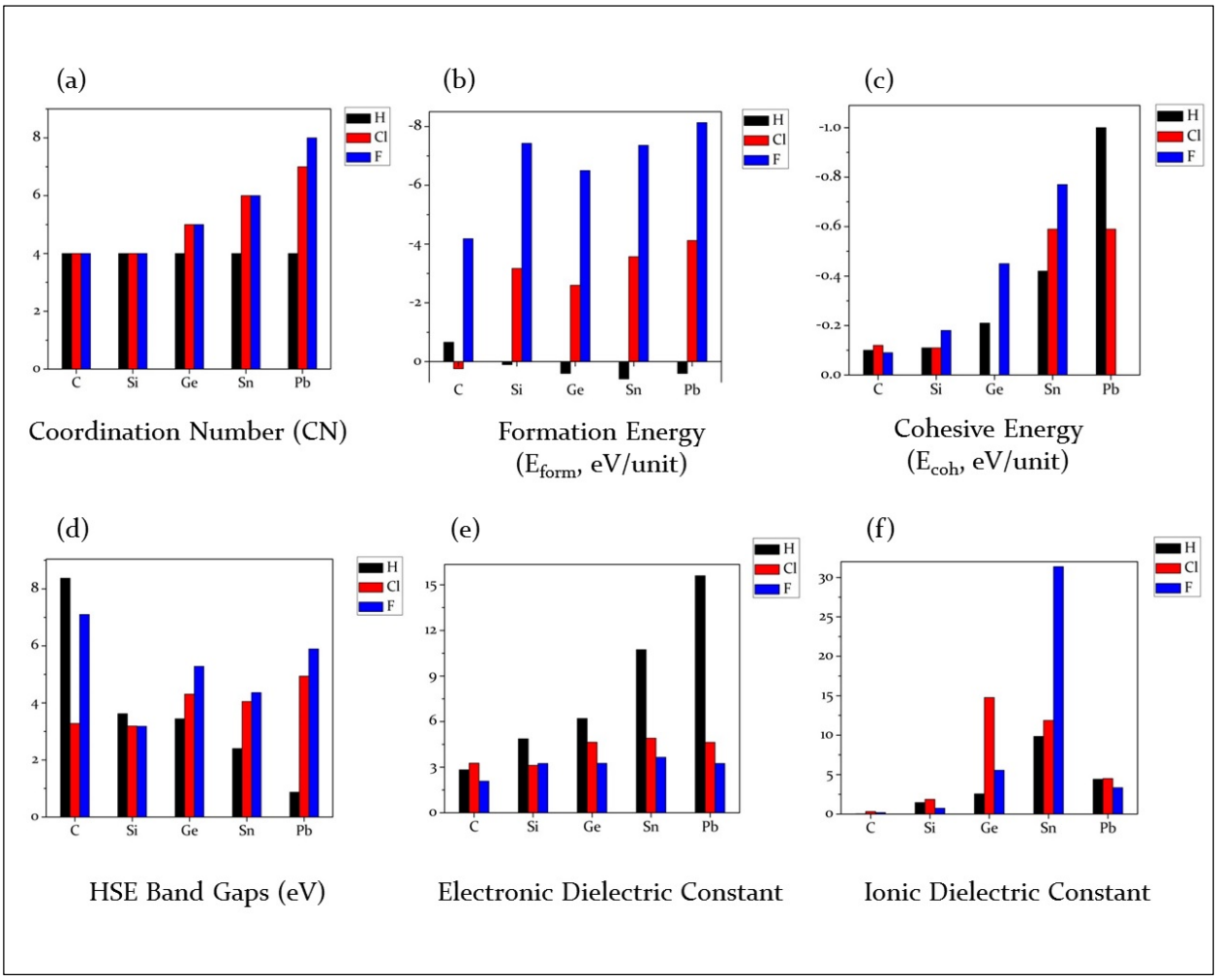


Figure 4.2. *Plots showing the following features of the 15 compounds: (a) coordination number (CNs), (b) formation energy (E_{form}), (c) cohesive energy (E_{coh}), (d) electronic dielectric constant (ε_{elec}), (e) ionic dielectric constant (ε_{ion}), and (f) Band Gaps (E_{gap}).*

Figure 4.2 (b) and **(c)** capture the calculated values of E_{form} and E_{coh} respectively. The formation energies are mostly large negative numbers, which means X and Y_2 would rather form the XY_2 crystal than not, except, as can be seen from the small positive values, the dihydrides of Si, Ge, Sn and Pb. As mentioned earlier, these dihydride polymers are not experimentally known, which could explain their possible instability. E_{form}

is seen to become less negative as we go from the F containing to Cl containing to H containing compounds, which means that relative to elemental states, the difluorides are more stable than the dichlorides, which are more stable than the dihydrides.

Meanwhile, the values of E_{coh} seem to become more negative on going from the C-based systems to the Pb-based systems. This follows from the previous discussion: the systems that adopt Type-A structure essentially contain 1D motifs arranged together through weak interactions in a crystal, and will thus have stabilities close to that of isolated chains. The 3D crystalline structures of the difluorides and dichlorides of Ge, Sn and Pb will have much lower stabilities for 1D motifs and thus, much lower cohesive energies. The Type-A crystal structures are all polymeric and thus close in energy with the individual chain energies, while the other structure types are not. Note that E_{coh} has not been mentioned for $GeCl_2$ and PbF_2, as the chains for these systems could not be isolated at all.

4.1.3. Electronic Properties

The electronic band gap values, shown in **Figure 4.2 (d)**, are seen to go from the lows of around 2 eV to the highs of 8 eV, except for PbH_2 that shows an unusually low E_{gap} of less than 1 eV. Most of the compounds lie approximately in the 3 eV to 5.5 eV range, which puts them somewhere in the semiconductor to insulator category. The C-based compounds have the highest E_{gap}, which comes from the low polarizability of C-Y bonds. Increased polar nature of the bonds causes more dispersion of charge and thus lower

E_{gap} in the remaining compounds; however, the band gaps are generally both high and relatively low in Si, Ge, Sn and Pb based systems.

The dihydrides follow the simplest of trends here: E_{gap} successively goes down from CH_2 to PbH_2. An explanation for this can be drawn from the stability arguments in the previous section: the stability of the compound goes down from CH_2 to PbH_2, resulting in decreased X-H bond strength [97] and more free electrons, which reduces E_{gap}. Further, for any given X, all the XF_2 compounds show a higher E_{gap} than the corresponding XCl_2 compounds, which again relates to the higher stabilities of the difluorides as compared to the dichlorides. The bond strengths are higher for the X-F bonds than for the X-Cl bonds [97], leading to less free electrons in XF_2 compounds than in XCl_2. It can also be seen among the difluorides and dichlorides that moving towards a higher CN and increased 3D nature of the structure seems to cause an increase in E_{gap}. Of course, the more rigid a crystal structure is, the more immobile would be the electrons of the constituent atoms and thus, E_{gap} would be higher. The difluorides and dichlorides of Ge, Sn and Pb all have high E_{gap}, because of being very stable crystalline compounds.

4.1.4. Dielectric Properties

To further understand the implications of the bonding and chemical coordination, and keeping in mind possible applications, dielectric constants were determined for the 15

XY_2 compounds. DFPT calculations give as output the total dielectric constant (ε_{tot}) tensor divided into two component tensors: the electronic part (ε_{elec}) and the ionic part (ε_{ion}). While ε_{elec} is a function of the polarizabilities and vdW volumes of all the atoms, the ionic part ε_{ionic} depends on the strength and flexibility of the X-Y dipoles [64] [66].

It can be seen from **Figure 4.2 (e)** that ε_{elec} is increasing for the dihydrides from C to Pb, while for any given Group 14 element, it generally decreases from the dihydrides to the dichlorides to the difluorides. These observations can be deconstructed in the following manner: in the presence of an external electric field, there will be an induced dipole moment in the system, the strength of which depends on the ease of distortion of the electron cloud around an atom. This distortion becomes successively easier on moving to lower stability systems, leading to higher values of ε_{elec}; thus, the gradual decrease in stability of the dihydrides going from C to Pb and the increased stabilities of the difluorides and dichlorides explain the trends. Further, we could easily correlate the pattern of variation in ε_{elec} with what we saw with the E_{gap} values, which in general were decreasing for the dihydrides down the group and increasing for the difluorides and dichlorides. Higher polarizabilities lead to lower E_{gap} [109], which may be understood by realizing that the polarizability of a bulk system can be written as a sum over electronic transitions from the valence to conduction band manifolds with the corresponding transition energies appearing in the denominator [95]. Thus, there can be said to exist an inverse relationship between ε_{elec} and E_{gap}.

ε_{ion} follows quite a different kind of trend. The values for C and Si-based compounds are low, whereas Ge and Sn-based compounds are much higher, and Pb-based compounds, surprisingly, are low. We can try to understand this with some bond strength and dipole moment arguments. The contribution to ε_{ion} comes from the presence of structural units having high dipole moments, and how easily the realignment of these dipoles takes place in the presence of external electric field [66]. For any X-Y bond, the dipole moment is known to increase with the bond length and the electronegativity difference between X and Y. Since the Group 14 elements (bar C) all have similar electronegativity (~ 2), the respective X-Y electronegativity differences can be said to be equivalent for any Y (Y = H, Cl, F), meaning we should look at the bond lengths to determine which dipole moments are higher. It can be seen from **Figure 4.2 (f)** that for any given Y, ε_{ion} is increasing from C to Si to Ge to Sn, and then decreasing again for Pb. The X-Y bond lengths increase as X goes from C to Pb, resulting in increased dipole moments and thus higher ε_{ion} values. It can be argued that Ge-Y and Sn-Y bonds are more susceptible to stretching/wagging than the bonds in the heavier Pb compounds, which leads to the drop in ε_{ion} in PbY_2.

4.1.5. Observations from the Study of the Compounds of Group 14 elements

From this study of the different XY_2 compounds, it is quite clear the oxidation states and electronegativities of X, the strength of the X-Y bonds, as well as the role of lone pair electrons have a major influence on the coordination geometries and stable

conformations, and subsequently, the properties. The trends make no secret of the intriguing nature of Group 14 that we talked about in the Introduction: from nonmetallic C to metallic Pb, we see what is very much a logical transition in the structures and properties. The electronic and dielectric properties, especially, reveal interesting trends that can be utilized in many ways. For instance, structural units containing Group 14 elements can be introduced into existing polymeric structures to tune the overall band gaps or dielectric constants, useful in applications like capacitors, organic electronics, photonics and photvoltaics. Units like GeF_2, SnF_2 and $SnCl_2$ could be useful in applications requiring high dielectric constant and band gap. Based on these observations, the study of polymers incorporating such metal-based units in their backbones was undertaken, and their structures and properties were studied.

4.2 Organo-Sn Polyesters

4.2.1 Rationale for Exploring Chemical Spaces Beyond Purely Organic Systems

Organometallic polymers, i.e., those containing metal atoms covalently bonded within their backbones, are outside the chemical subspace of the organic polymers. The development of such polymers for energy storage [67] [68] [69] was guided by several rational considerations, aiming specifically at boosting the ionic dielectric constant ε_{ion},

given a certain large E_{gap} (the importance of the ionic contribution was already pointed out in the context of organic polymers in **Chapter 3**, and the insights that emerged from the analysis of **Figure 3.5**). Two of the primary considerations in favor of incorporating non-carbon species (e.g., metals or even semiconducting systems in their bulk forms) in polymers are as follows. First, metal-containing bonds may be highly polar, depending on the nature of the metal. Second, the lattice vibrations involving these bonds are generally low in frequency. Both factors are crucial for an improvement of ε_{ion} at the low-frequency limit [58], while according to Ref. [70], the electronic dielectric constant ε_{elec} of polymers in this class is also confined by the same theoretical limit shown in **Figure 3.5** and discussed in **Chapter 3**. It should be noted that ε_{ion} includes all non-electronic contributions to the dielectric response, including bond stretching and bond (dipole) rotations allowed within a crystalline lattice.

The expected improvement of ε_{ion} was soon confirmed in a high-throughput screening work based on DFT computations [66]. By considering several single polymeric chains containing different blocks based on C, Si, Ge and Sn, E_{gap}, ε_{elec} and ε_{ion} were computed. A summary of the resulting data is shown in **Figure 4.3**; an overall inverse relationship can be seen between the two properties, like the trends observed for the purely organic polymers in **Figure 3.5**. There is a bound on one property when increasing the other; the polymer chains containing polar units such as SnF_2, $SnCl_2$, GeF_2 and $GeCl_2$ dominate the upper left portion of the plot, where dielectric constants are as high as 30 while band gaps are around 3 - 4 eV. Most importantly, this study also revealed key correlations between

ε_{ion} and the dipole moments and rotational barriers to the dipoles from adjoining groups. This interesting observation aligns well with the systematic study of the compounds of Group 14 elements presented in the previous section in this chapter. The DFT computed ε_{elec} and ε_{ion} for the hydrides, fluorides and chlorides based on these elements are shown in **Figure 4.3**. Compared to C and Si-based compounds, the ionic dielectric constant of the Ge, Pb, and especially, Sn-based materials, is extremely high. These initial studies provided the rationale for the recommendation that incorporation of Sn (Ge and Pb were avoided at this stage because of their expense and poisonous nature, respectively) into typical organic polymer backbones, potentially bonded with highly electronegative atoms such as F, Cl or O, may be beneficial.

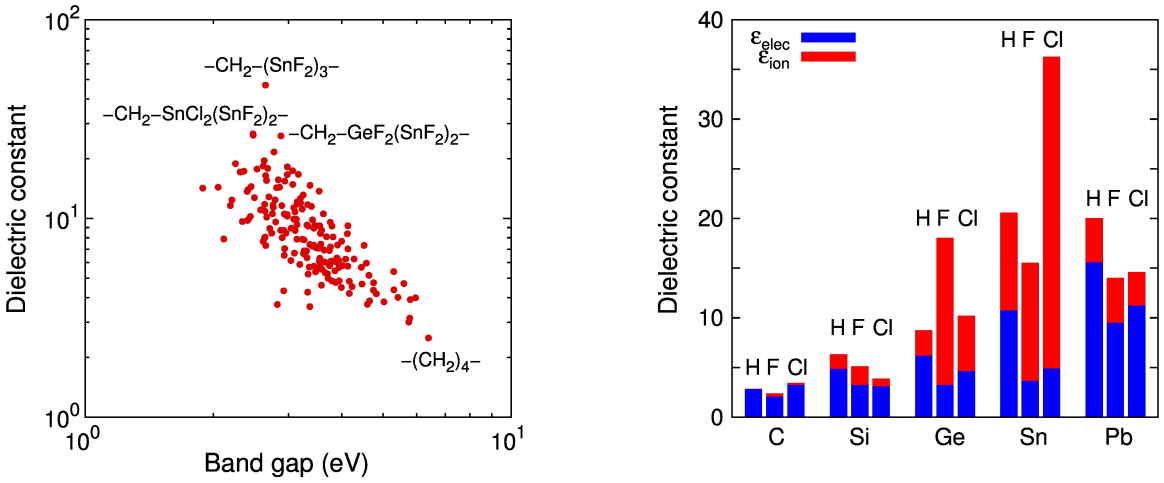


Figure 4.3. *(Left) Computed dielectric constants shown vs band gaps of single chain polymers formed from C, Si, Ge and Sn based units [66], and (right) the electronic and ionic dielectric constants of compounds of Group 14 elements [63].*

4.2.2 Poly(dimethyltin glutarate) and Poly(dimethyletin esters)

Based on the computational guidance, Sn was selected for developing new organometallic polymers [67] [69]. An organotin functional group, Sn-ester, was identified as the starting point due to the large electronegativity difference between Sn and oxygen (O). The actual synthesis was performed using dimethyltin dichloride and glutaric acid, resulting in poly(dimethyltin glutarate), or p(DMTGlu), a new polymer in which the repeat unit contains a dimethyltin group flanked on either side by a carboxylate group, with a linear chain of 3 methylene (CH_2) units acting as the linker [69]. The synthesis scheme, as shown in **Figure 4.4**, was altered from that proposed by Carraher, Jr. [110] by using tetrahydrofuran (THF) instead of hexane. This change allows for a traditional condensation polymerization instead of an interfacial polymerization, while the added polarity of THF helped to solubilize the growing chain and to produce higher molecular weight polymers. The resulting polymer, p(DMTGlu), is thermally stable at temperatures up to 235˚C while showing a high dielectric constant larger than 6.

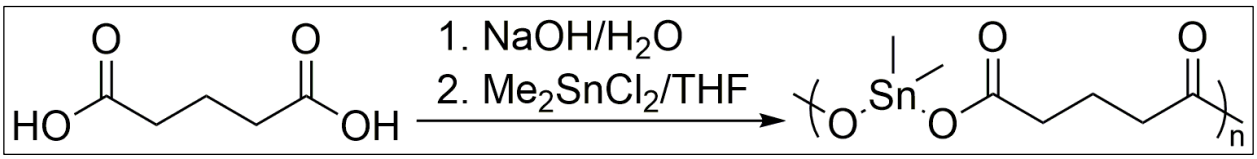


Figure 4.4. *Synthetic route towards p(DMTGlu). The repeat unit of the resulting polymer contains a dimethyltin group and a linker of 3 methylene (CH_2) groups* [69].

In terms of high temperature capability and dielectric constant, p(DMTGlu) is superior to most of the currently used organic dielectric polymers, e.g., BOPP, which has a dielectric constant of 2.2 and works at temperatures below 105˚C [111]. Although the thermal stability of p(DMTGlu) is still below that of some recently developed polyimides [72] [74], there is a clear indication that tin-based organometallic polymers can combine both high temperature capability and a high dielectric constant. Given the success in developing p(DMTGlu), a complete series of related poly(dimethyltin esters) was then synthesized [67]. All the polymers in this family are based on the dimethyltin-ester group, differing from p(DMTGlu) and from each other by the number of methylene units in the linker, which ranges from 0 to 11. The whole family of polymers has since been processed and characterized [67] [69]. As not all of them are soluble, desired measurements had to be performed either on cast films or on pellets made of these polymers.

In parallel with the experimental efforts, detailed first-principles calculations were performed for this family of organotin polymers [67] [69]. In brief, the structures of these polymers were determined by the minima-hopping structure prediction method [61, 62], starting from the polymeric chains of the predetermined repeat units. Because Sn can adopt a variety of coordination environments, the geometry of the Sn-containing units is not well defined. Thus, the structure prediction step had to be done without constraints. Further calculations were then performed on the most stable structures at suitable levels of DFT, determining the dielectric constant with DFPT [53] and band gap using either the Perdew-Burke-Ernzerhof (PBE) [90] or HSE electronic exchange-correlation functional.

[56] Calculations at the PBE level of DFT is fast but the result is typically underestimated by 30% or more [57] while HSE offers more reliable results at sufficiently higher computational cost. For methodology validation, IR spectra and XRD patterns were obtained from simulations as well.

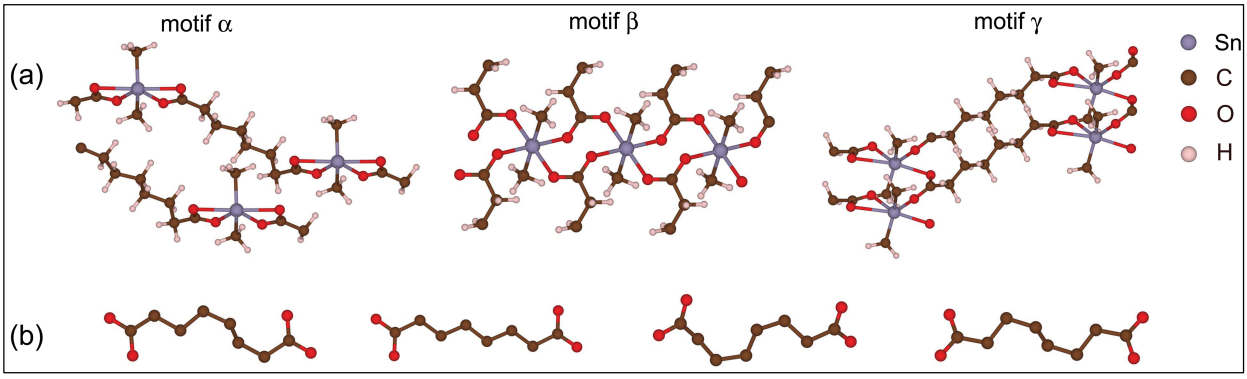


Figure 4.5. *(a) Three basic structural motifs (α, β, and γ) computationally predicted for poly(dimethyltin esters) and (b) some folding patterns of the methylene linker. Figure reprinted from Ref. [67] with permission from ACS.*

In the predicted structures of poly(dimethyltin esters), all the Sn atoms are six-fold coordinated. The Sn-O bonds, on the other hand, can link different polymeric chains in various ways [67] [69]. This feature distinguishes these organometallic polymers from their organic counterparts, in which the polymeric chains are essentially isolated. Depending on the arrangement of the Sn-O bonds, three basic motifs, namely intra-chain (α), inter-chain (β), and hybrid (γ, which combines some features of α and β), were identified. In the α motif, four Sn-O bonds link the central Sn atom with four O atoms from two carboxylate groups in the same repeating unit, thus realizing the hypothesized intra-

chain motif. Motif β, on the other hand, contains two Sn-O bonds that link the chains together. Out simulations predicted that hybridization between the inter-chain and intra-chain motifs is also possible, and this was realized in the form of motif γ, where the first carboxylate group is connected to the Sn atom by two Sn-O bonds, and of the two other Sn-O bonds, one links the central Sn atom with the second carboxylate group while the other interlinks the chains. These motifs, which are shown in **Figure 4.5**, can exist simultaneously in the synthesized samples because they only differ from each other by a meV/atom. Of these, the intra-chain and inter-chain motifs have been documented in the literature for some organotin materials [112]. The existence of these motifs in the synthesized samples was confirmed by comparison of the computed and measured IR spectra and XRD patterns [67] [69].

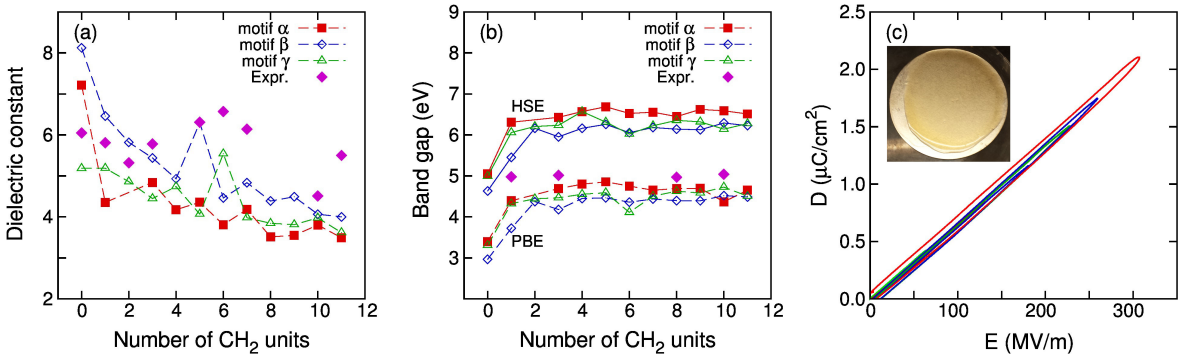


Figure 4.6. *Computed and measured data for (a) dielectric constants and (b) band gaps (calculated at PBE and HSE levels of theory) of the poly(dimethyltin esters) in different motifs (α, β, and γ) with different linker length, ranging from 0 to 11 methylene (CH_2) units. In (c), the DE loops measured on the 20/80 (wt/wt) blend of pDMTSub/pDMTDMG are*

shown together with a film cast for this polymer in the inset. Figure plotted from the data reported in Ref. [67].

The dielectric constant and the band gap of the poly(dimethyltin esters) depend on the length of the linker (or the number of the methylene units) in certain ways. **Figure 4.6** compiles the computed and measured data for these essential quantities, revealing that, in general, the dielectric constant is decreased and the band gap is increased when the linker is longer. There are, however, some "optimal" lengths of the linker (4-7 methylene units) at which both the band gap and the dielectric constants are high. For the particular case of poly(dimethyltin suberate) (or pDMTSub, the poly(dimethyltin ester) that contains a linker of 6 methylene groups), the dielectric constant can be as high as 7, and at the same time, the band gap reaches a value of nearly 7 eV. The refractive index, the square of which is the electronic dielectric constant ε_{elec}, was measured for some cases, leading to a good agreement with the calculated data [67] [69].

To access the dielectric breakdown strength of the poly(dimethyltin esters), their charge-discharge behavior was determined through the D-E hysteresis loop. Because these polymers form large crystalline phases upon drying whose size depends on the length of the methylene linker, they were blended with a second homopolymer, which is poly(dimethyltin 3,3,-dimethylglutarate), or pDMTDMG, to produce an amorphous morphology. Films can then be casted for the desired measurements. The hysteresis loops obtained for a blend consisting of a 20/80 (wt/wt) pDMTDMG/pDMTSub are shown in **Figure 4.6 (c)** while those of the pDMTDMG/pDMTGlu are reported in Ref. [69]. The

measured data suggests that pDMTGlu and pDMTSub are linear dielectrics with breakdown strengths of roughly 400 MV/m and 300 MV/m, respectively, leading to an energy density of roughly 4 J/cm^3. Although this parameter is still below that of BOPP (5 J/cm^3), the new chemical subspace of the organometallic polymers looks promising given that we are at the very initial stages of optimization of this entirely new polymer subclass. More importantly, the pathway leading to the development of these organotin polymers can be used for further exploration into this subspace (and for expanding the search space beyond just organotin polymers, as briefly discussed later).

4.2.3 Experimental Validation of Computations

Bond	α-Motif 1	Complex 5	α-Motif 2	Complex 4
Sn-O1	2.174	2.113	2.114	2.140
Sn-O2	2.517	2.511	2.432	2.552
Sn-O3	2.171	2.113	3.098	> 3
Sn-O4	2.553	2.511	2.105	2.136
Sn-C1	2.127	2.109	2.125	2.119
Sn-C2	2.126	2.109	2.132	2.130

Table 4.1. *Sn-O and Sn-C bond lengths (in Å) of the two α motifs given in a comparison with those of Complex 4 and Complex 5 reported in ref.* [112].

To test the validity of the DFT obtained structures and properties of the poly(dimethyltin) esters, we compared the results for p(DMTGlu) with appropriate data from the available literature. Four low energy structures were computationally estimated for p(DMTGlu), with two structures adopting the intra-chain (α) motif, one adopting the inter-chain (β) motif and one adopting the hybrid (γ) motif (with stability going up from β to γ to α).

The theoretically predicted structural models are strongly supported by a very recent experiment. In particular, the two α motifs shown by p(DMTGlu) were observed in organotin carboxylates, the related crystals of which are also based on the -COO-Sn-OOC- unit [112]. The Sn-O and Sn-C bond lengths of these motifs, as shown in **Table 4.1**, agree well with the corresponding bond lengths of Complex 5 and Complex 4, two experimentally determined structures of the organotin carboxylates [112]. The formation of the Sn-O bond was confirmed by the shift of the carbonyl group of glutaric acid to a lower energy as seen in the Infrared (IR) spectrum, shown in **Figure 4.7**. Carraher reports two absorption ranges for the carbonyl group in poly(tin carboxylates), 1635–1660 and 1550–1580 cm^{-1} with pDMTGlu carbonyl absorptions at 1673 and 1563 cm^{-1}. Also indicative of the tin oxygen bond formation is the combination of skeletal C-CO-O with Sn-O stretching at 645 cm^{-1} [113]. It is well documented that tin mono- and di-carboxylates form coordination complexes [114]. Peruzzo et al. hypothesized that both intra- and inter-chain complexes described above are present, with the bridging and non-bridging symmetric carbonyl bands exhibiting different absorptions, 1410–1430 and

1350–1370 cm^{-1}, respectively [115]. The IR of pDMTGlu shows the formation of both complexes with absorptions at 1406 and 1378 cm^{-1}.

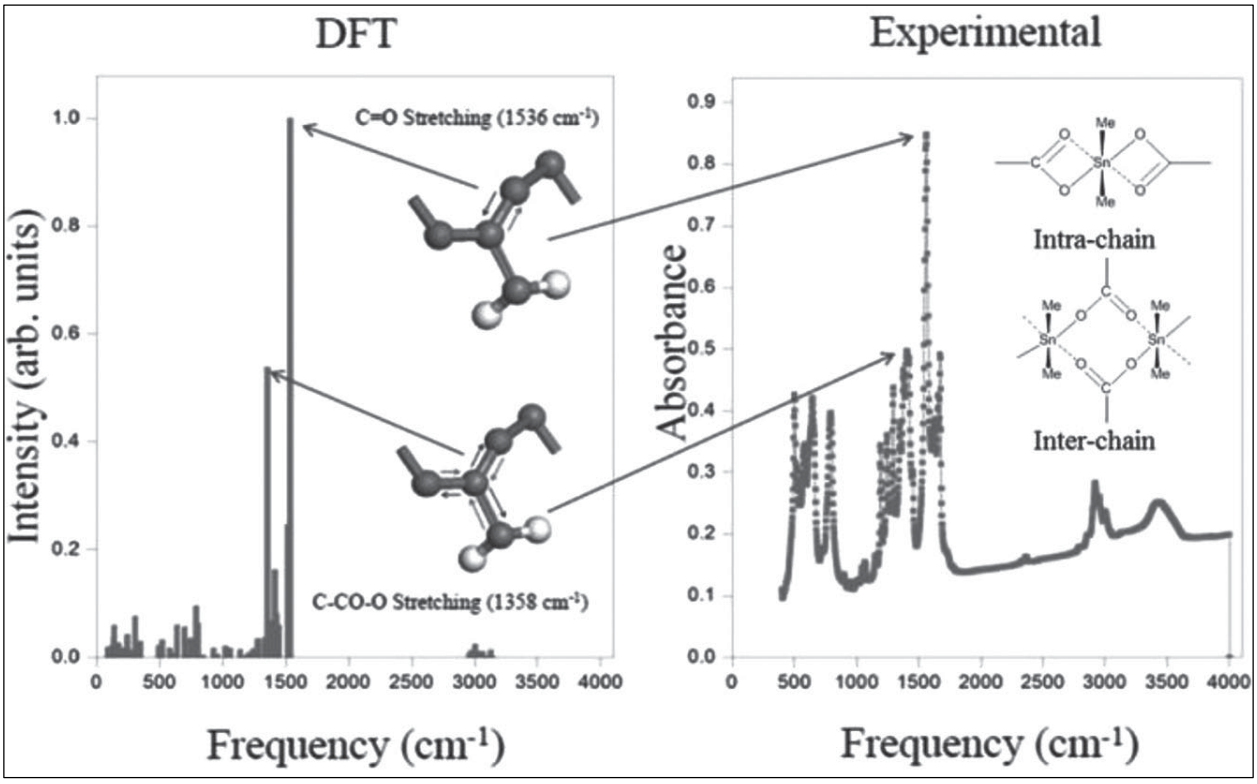


Figure 4.7. *The C-O and C-CO-O stretching modes for the lowest energy* α *motif, shown on the IR plots obtained from DFT results as well as from experimental results* [69].

Using the same perturbation theory formalism within DFT that yielded dielectric constants, the IR spectrum of the stabilized polymers was calculated and the characteristic vibrational modes of the lowest energy α motif of pDMTGlu were identified. It is observed that the peaks in the IR intensity versus frequency plot obtained computationally match with the experimental IR peaks seen in the transmittance versus frequency plot. Furthermore, the following observed stretching mode frequency matched with the

frequency reported in Carraher's work; the C-CO-O stretching at 1294 cm^{-1} (1250-1290 cm^{-1}). **Figure 4.7** further illustrates the comparison of the experimental and computational IR spectra.

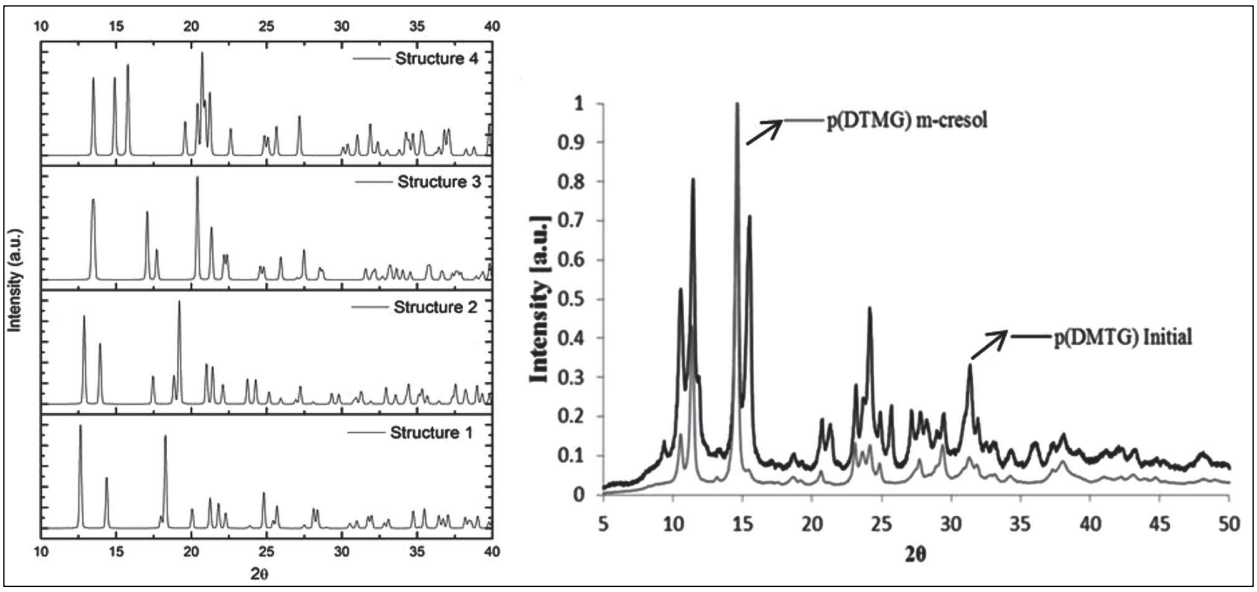


Figure 4.8. *Predicted X-ray diffraction pattern (XRD) of the four stable motifs of p(DMTGlu) (Structures 1 to 4, in order, the β motif, the γ motif and the two α motifs) and experimental XRD pattern of the precipitated polymer and after solubilizing and recovering from m-cresol* [69].

To confirm our hypothesis that the four stable predicted motifs are present in the sample in unequal amounts, the computationally predicted X-ray diffraction (XRD) pattern of the motifs and experimentally collected XRD of the polymer after precipitation from the reaction as well as after solubilizing and recovery from m-cresol was compared (**Figure 4.8**). The initial polymer powder shows a conglomeration of all possible stable structures

with rearrangement to one predominant crystal structure after dissolution in m-cresol which is signified by the disappearance of peaks in the XRD at 2θ values of 11.50 and 15.52. Comparison of the calculated and experimental XRD shows that the rearrangement of the crystal structure has the propensity to stabilize itself in the lowest energy α motif configuration (Structure 4 in **Figure 4.8**), since this structure as well the polymer exhibit a peak at approximately a 2θ value of 15.

4.2.4 Effects of aromatic and chiral groups on the dielectric properties of poly(dimethyltin esters)

While the initial rationale leading to the development of poly(dimethyltin esters) focuses on the tin-containing groups, the linker does play an important role. When the linker contains a given number methylene units only, the effect of the linker length on the dielectric constant and the band gap is shown in **Figure 4.8**. If other building blocks like aromatic and chiral groups are introduced in the linker (see **Figure 4.9**), the dielectric properties can be further manipulated.

A study on organo-Sn polymers containing the aromatic groups shown in **Figure 4.9** led to the conclusion that the size and the electron density/polarity of the aromatic rings have certain effects on the dielectric constant. As the size of the ring is increased, the resulting dielectric constant would be lower. The nature of the aromatic rings is also relevant. In

particular, polymers containing the (electron neutral) benzene rings would have higher average dielectric constants compared to those having (electron donating) thiophene rings or (electron withdrawing) pyridine rings. Chiral groups can be used to control the crystallinity of the polymers, which, in turn, affects the averaged dielectric constant. It is clear that substantial room is available for optimizing the dielectric properties of the poly(dimethyltin esters).

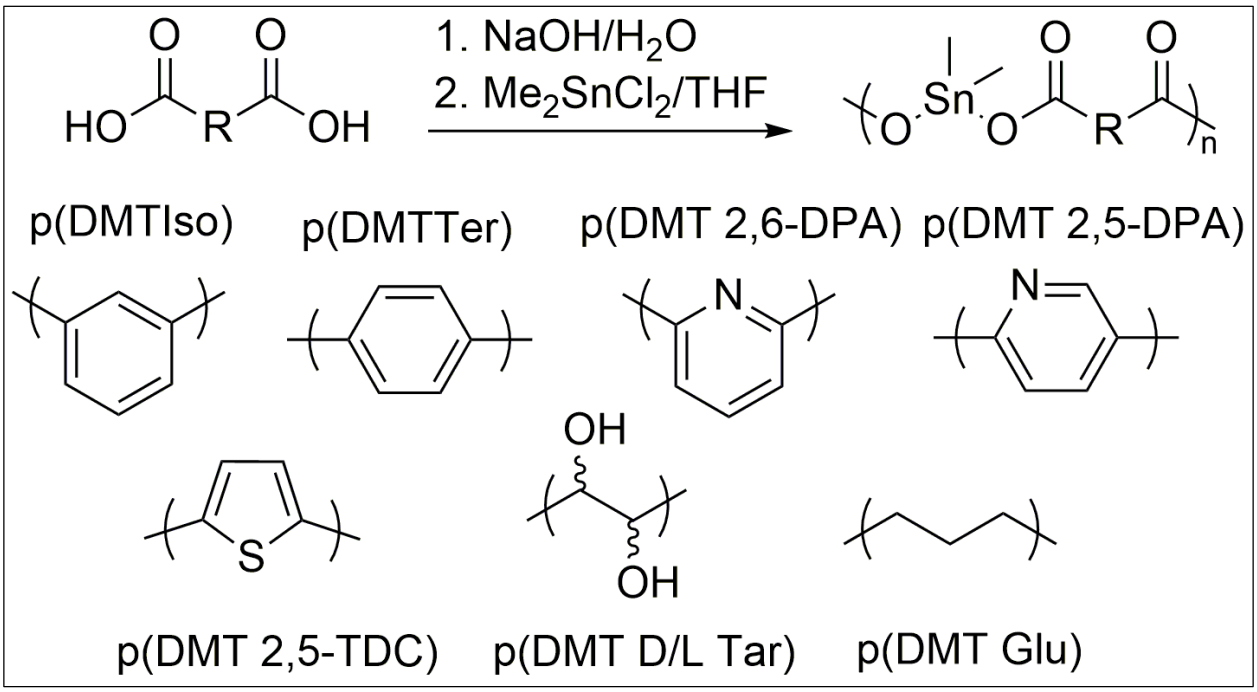


Figure 4.9. *Synthetic scheme of poly(dimethyltin esters) with aromatic (pyridine, benzene, and thiophene) and chiral (tartaric acid) groups* [68].

4.2.5 Optimization of Organo-Sn polymers via Blending and Copolymerization

Computational and experimental studies on aromatic and chiral poly(dimethyltin esters) showed that there is a benefit from the incorporation of metal tin atoms in the backbone of a polymer chain through a bond between tin and oxygen. Whereas nanoparticles can drop out of a polymer, the tin atoms are dispersed throughout the polymer and unable to aggregate together, which alleviates dispersion difficulties. The ester linkage to the tin atom is an ideal way to bind the metal into the polymer backbone as it provides an increased atomic polarization and can further increase dipole interactions. By varying the length of methylene groups between diacid monomers, aliphatic poly(dimethyltin esters) were able to be produced with different weight percentages of tin [67] [69]. During the study on aliphatic poly(dimethyltin esters), a suggestion was made that film morphology based on methylene spacer length was responsible for variations in dielectric constant and band gap. The regularly repeating and polar nature of these poly(dimethyltin esters) allowed for high degrees of crystallinity in the polymer powders and films.

Based on these insights, blends and copolymers of poly(dimethyltin suberate) p(DMTSub) (the organo-Sn polymer showing the highest E_{gap} of ~ 7 eV and a high ε_{tot} of 5.5) and p(DMTDMG) (which showed E_{gap} of ~ 6 eV and ε_{tot} of > 5) were made [76]. Increasing amounts of p(DMTSub) were used, from 10% to 50%, to find a balance between electronic properties and film morphology. Both blend and copolymer systems showed improved results over the homopolymers with the films having ε_{tot} of 6.8 and 6.7

at 10 kHz with losses of 1% and 2% for the blend and copolymer system, respectively. The energy density of the film measured as a D-E hysteresis loop was 6 J/cc for the copolymer, showing an improvement compared to 4 J/cc for the blend. This improvement is hypothesized to come from a more uniform distribution of di-acid repeat units in a copolymer compared to a blend, leading towards improved film quality and subsequently higher energy density. The measured ε_{tot}, dielectric loss and Egap values for 5 different polymer blends and 5 different copolymers are shown in Table 4.2, and the corresponding values for the pure p(DMTSub) and p(DMTDMG) polymers are also shown for comparison.

Polymer	p(DMTSub)	p(DMTDMG)	BP10	BP20	BP30	BP40	BP50	CP10	CP20	CP30	CP40	CP50
Dielectric Constant (10 kHz)	5.6 (5.5)	6.3 (5.3)	6.8	6.6	6.2	5.7	5.1	6.7	6.6	6.4	6.0	5.9
Loss (10 kHz)	1.7	1.2	0.9	0.8	0.7	0.9	0.5	1.8	1.0	1.0	1.4	1.8
Band gap (eV)	6.7 (6.2)	4.8 (5.9)	4.8	4.8	4.8	4.8	4.8	4.8	4.9	4.9	4.9	4.9

Table 4 2. *Dielectric constant and loss values taken at 10 kHz for the Sn polyester blends and copolymers, as well as their measured band gaps* [76].

4.3 Extensions to Other Organometallic Polymers

4.3.1 Organo-Zn and Organo-Cd Polyesters

Incorporating metal-oxygen bonds into the main chain of polymers helps maintain the large band gaps characteristic of polymers, while increasing the dielectric constants to approach those of the metal oxides. Zinc oxide (ZnO) and cadmium oxide (CdO) show dielectric constants of 8.5 and 6.2 respectively. In similar fashion to the Sn polyesters studied and described in the previous section, Zn- and Cd-based units could be introduced in the backbone of polymers to create novel organometallics with improved dielectric properties. We thus considered a series of Zn ($3d^{10}$) and Cd ($4d^{10}$) aliphatic coordination complex polyesters with varying numbers of methylene spacer(s), changing from 1 to 8. DFT calculations (involving structure prediction and property estimation, as described earlier) were performed to compute ε_{elec}, ε_{ion} and E_{gap} for the 8 organo-Zn and the 8 organo-Cd polymers. Further, all these polymers were synthesized and characterized for their dielectric properties [116].

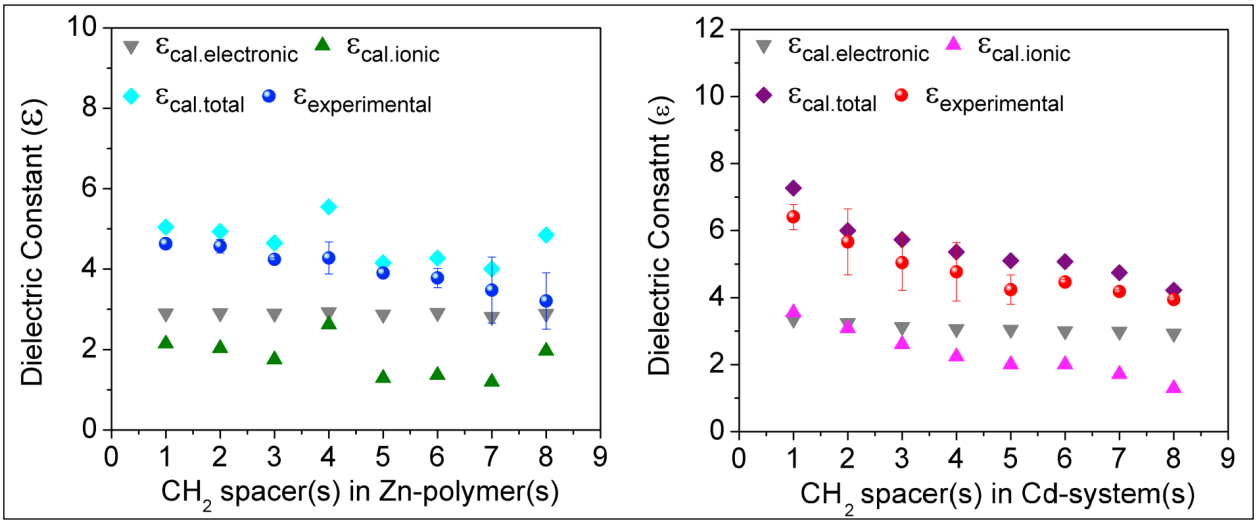


Figure 4.10. *Calculated dielectric constants (electronic, ionic and total) of all the Zn and Cd-based polymers, compared with the experimental measurements as a function of number of CH_2 spacers [116].*

The computed and measured dielectric constants for all the Zn and Cd systems are shown in **Figure 4.10**. The Cd-systems were seen to have higher ε_{tot} (ca. 6.7 to 4.0) compared to their Zn counterparts (ca. 4.6 to 3.6), as was predicted by DFT. Significant improvements were seen in the dielectric loss values, which were generally less than 1%. Both the measured and computed E_{gap} values were observed to lie between 5 and 6 eV for all sixteen systems, which exceeds the typical optical band gap energies of ZnO and CdO. This close agreement between calculated and experimental results provides the impetus for further research to be directed towards the expansion of the organometallic polymer chemical space.

4.3.2 All Organometallic Polymers Dataset

The computation-driven discovery of novel Sn-based, Zn-based and Cd-based organometallic polymers paved the path for a sweeping exploration of polymers containing different metals chosen from the periodic table. In **Figure 4.11**, DFT computed results are presented for organometallic polymers constituted of any of 10 different metal atoms [70] [71]; also, shown for comparison are all the pure organics that were discussed in **Chapter 3**. The metal based systems clearly surpass the pure organics in terms of high dielectric constants for given values of band gap. The primary reason behind this increase is the enhanced polarity of chemical bonds in the organometallics because of bonding between electropositive metal atoms and highly electronegative atoms such as O, F and Cl. The swinging and stretching of these polar bonds at low frequencies cause fluctuations in polarization under electric fields, which means they will contribute more to ionic or dipolar parts of the dielectric constant [64] [70] [58]. As seen from **Figure 4.11**, this effect is more pronounced in some organometallics than others: it was observed that the higher the amount of metal in the system, higher is the dielectric constant. The identity of the metal atom itself and its coordination environment were other crucial factors at play here [71]. The great promise shown by the computed dielectric properties of organometallic polymers provides motivation to experimentally pursue a lot more cases than polyesters of Sn, Zn and Cd studied so far. In **Chapter 6**, results of 'learning' from this comprehensive polymer dataset will be discussed, and efforts will be made to uncover the crucial features of the polymers that control their properties. The success of the

rational co-design approach in the discovery of several novel organometallic polymer dielectrics as discussed here was recently reviewed by us within the context of capacitor dielectrics design [49].

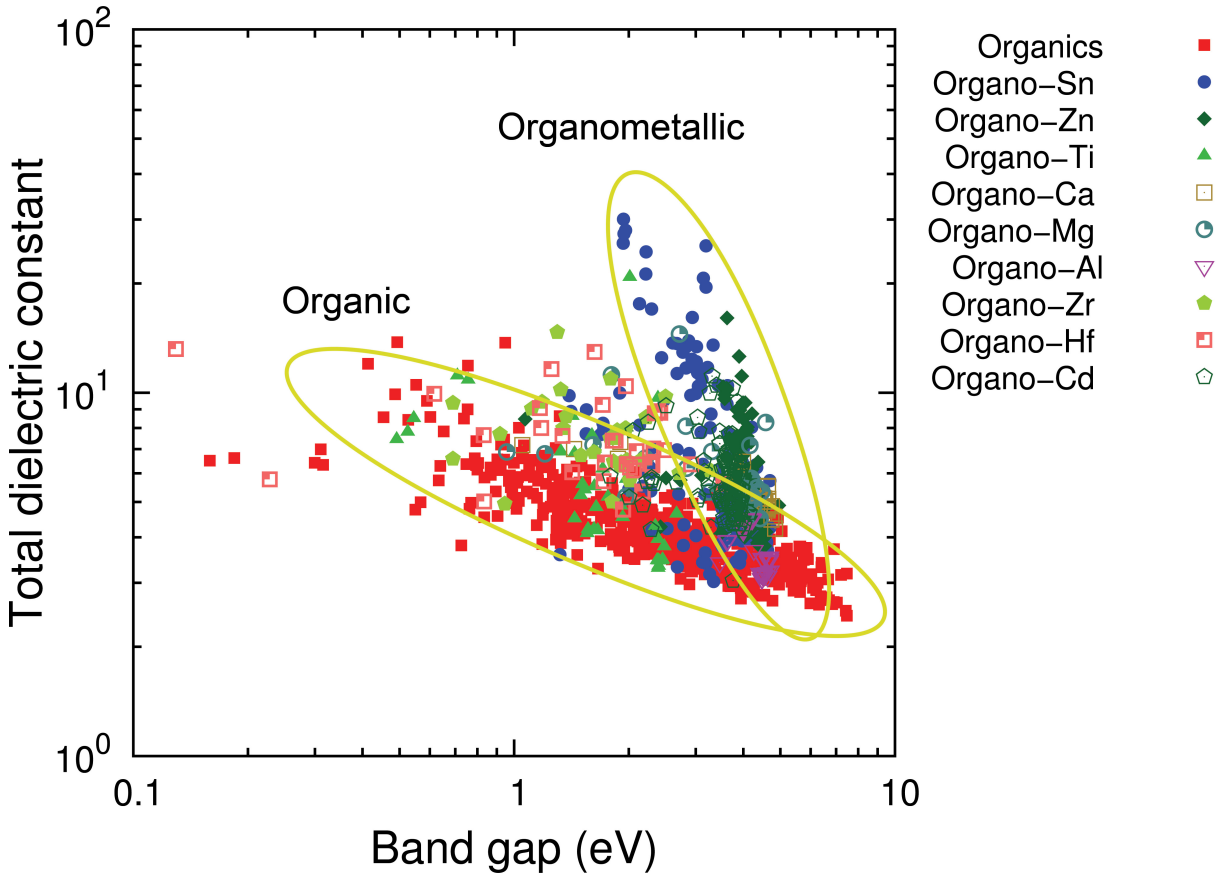


Figure 4.11. *Computed properties of more than a 1000 polymers. The organometallics include polymers containing any of 10 different metal atoms; clearly, the organometallic polymers out-perform the pure organic polymers in terms of high dielectric constants [70].*

Chapter 5

MACHINE LEARNING STRATEGY FOR POLYMER DIELECTRICS DESIGN

5.1 Introduction

The materials design process requires the identification of materials that meet a desired application or property need. The traditional routes adopted thus far to meet such design goals involve the determination of the relevant properties of several potential candidates, via high-throughput experiments or computations, and choosing the best cases for further studies and optimization. While powerful and successful, this strategy suffers from two primary drawbacks. First, the consideration of each material in a case-by-case manner is laborious and time consuming, especially if one were to ignore the availability of past data on the same or similar candidate materials. Second, the prevalent strategy addresses the materials design problem in an 'inverted' manner, i.e., instead of approaching the "desired properties → suitable materials" design problem (previously referred to as inverse design), the "materials → properties" problem is tackled, and the former design aspect is addressed indirectly through enumeration, i.e., explicit consideration of an enormous number of candidate materials. Confronting both these hurdles is critical to accelerate, streamline and focus the materials design process [52].

While the rational co-design approach described in **Chapter 3** and **Chapter 4** led to the successful identification of several promising polymer candidates for electrostatic and energy storage applications, the shear enormity of the polymer chemical space means it is extremely likely that significant untapped opportunities remain hidden. A more diverse spectrum (than currently available) of new, better and more suitable candidates will constantly be needed to meet growing future needs mandated by performance measures, amenability to synthesis and compatibility with other parts of devices. Rational and accelerated polymer design strategies and solutions that further improve upon the speed-up obtained from computational screening would thus be extremely useful. This is where machine learning strategies come into the picture.

The field of materials science that deals with using machine learning (ML) to accelerate materials design is often referred to as *materials informatics* [52] [77] [117] [118] [119] [120] [121]. In recent years, informatics approaches have been used for the prediction and classification of crystal structure types [122] [123] [124], stability of phases [125] [126], band gaps [118] [127] [128], elastic moduli [129], dielectric breakdown [130] [131] and instant atomic forces [132] [133] [134]. The most crucial aspect of materials informatics is *fingerprinting*, or the numerical representation of a material in terms of its most important attributes [124] [135] [136]. For instance, if one were to fingerprint the organic polymers belonging to the chemical subspace described in **Chapter 3** using their band gap values, one could qualitatively predict a new polymer's dielectric constant based on the magnitude of its band gap. However, the purpose of fingerprinting materials is to

have easily attainable, general and unique vectors that can be mapped to the properties of interest [77] [124]. Materials scientists have used elemental properties such as electronegativity and ionization energy [118] [130] [137], oxidation states [138], HOMO-LUMO levels [118] [139], shape and structural parameters [27] [140], chemical composition [137] [141], radial distribution functions [133] [134] [142] and Coulomb matrices [143] [144] for fingerprinting materials.

For implementing a machine learning based approach, our starting point was the generation of reference property data (using first principles computations) for a benchmark set of polymers spanning a selected chemical subspace. This was already achieved in **Chapter 3** with the high-throughput computations performed on nearly 300 organic polymers; these properties were presented in **Figure 3.5**. Given the uniform, well-defined (in terms of chemical building blocks) nature of this chemical space, we selected this dataset of polymers for the machine learning study. We applied commonly used interpolative statistical (or machine) learning concepts on this data to train an on-demand instant property prediction model, via an intermediate (and critical) 'fingerprinting' step that converts every polymer to a numerical string (c.f. **Figure 5.1**). The prediction scheme produces accurate results for cases not used in the training phase (but falling within the same chemical subspace), as demonstrated by comparisons with more laborious first principles computations and experimental measurements. Such a model via an enumeration scheme can be used to predict the properties of a plethora of new candidate materials to search for cases that fulfill a desired property need. Furthermore, one can

make rapid go/no-go decisions on whether a new synthesizable polymer is worth pursuing or not.

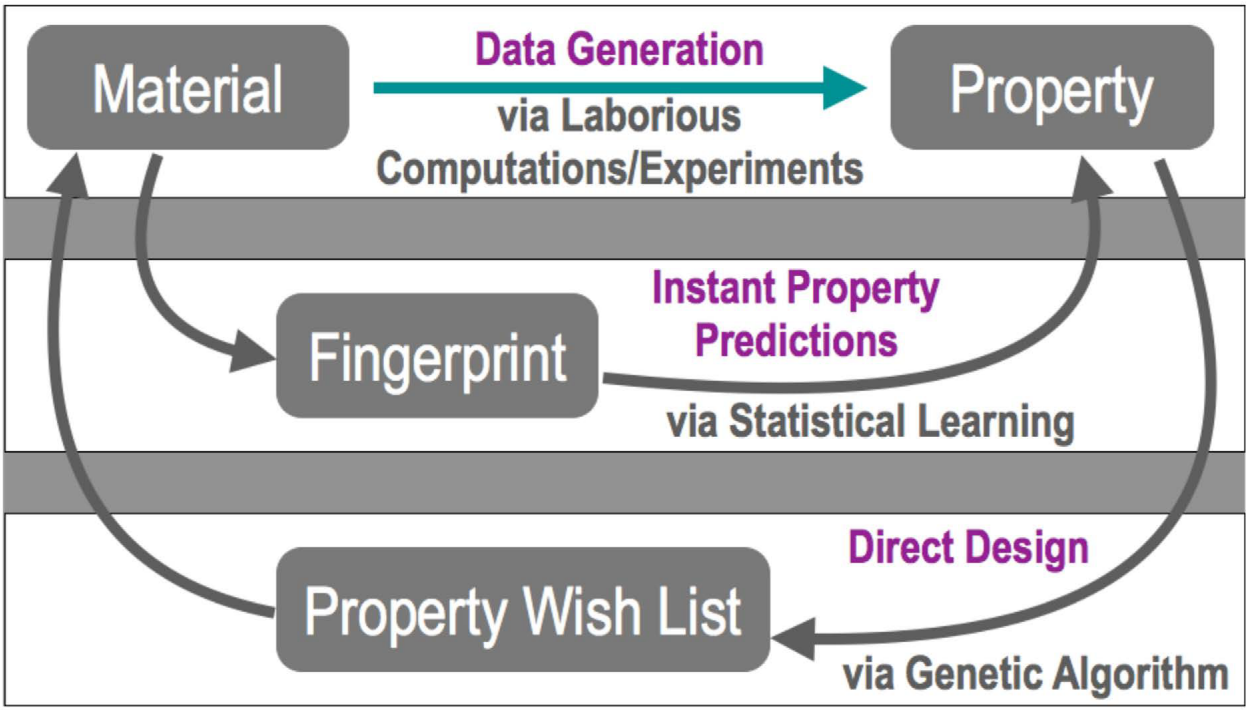


Figure 5.1. *Accelerated materials design using statistical (or machine) learning and genetic algorithm.*

The enumeration approach to materials design is not the most efficient one, as it involves consideration of an enormous number of cases, most of which will not be viable in the end (thus leading to low success rates). A better approach is to use the on-demand property prediction scheme within a genetic algorithm, to directly tackle the "desired properties → suitable materials" design problem. Several polymers that meet a property requirement criterion are designed directly here using such a strategy at a minuscule

fraction of the time required for enumeration. The predicted property results of the designed polymers are validated by explicit first principles computations.

The suite of tools and strategies that emerge from this effort take us a step closer to rational, accelerated and direct design of materials in general, and polymer dielectrics in particular. These strategies can also be extended to larger polymer chemical and property subspaces. The essential ingredients of this effort are illustrated in **Figure 5.1**, and described in detail in the following.

5.2 Polymer Fingerprinting

While high-throughput data generation efforts can provide useful 'lead candidates' with desired properties, the natural question that arises is whether one can understand the origins of the attractive behavior, and harness this understanding to search for other suitable options. Within the context of polymeric materials under investigation here, the origins should be traceable to the identities of their basic chemical building blocks. This comes from the theory that electronic and dielectric properties of organic polymers can be effectively expressed in terms of a sum of contributions from different constituent groups [145] [146]. These contributions are in the form of polarizabilities and dipole-dipole interactions from the groups, with different weights attached to different groups. In the

case of our polymers, different building blocks or combinations of blocks are expected to have different influence on the properties being studied.

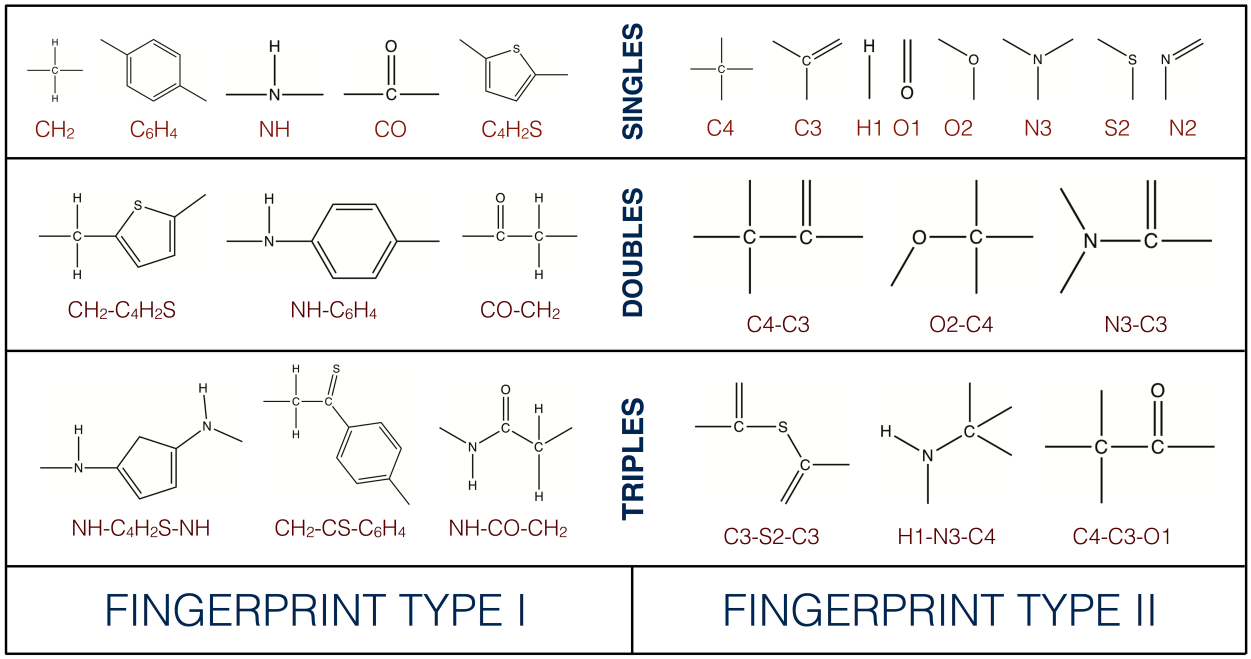


Figure 5.2. *Examples of the basic building blocks, building block pairs and building block triplets that help define fingerprint types I (where chemical units like CH_2 and C_6H_4 are building blocks) and II (where atoms like 4-fold C (C4) and 2-fold O (O2) are building blocks).*

Thus, if we can numerically represent—or fingerprint—our polymers based on their building block identities, correlations can potentially be established between the fingerprints (or parts of it) and properties. Indeed, numerically representing materials is emerging as an active topic of inquiry within materials science, physics and chemistry in

recent years. Descriptors such as this have historically been used in cheminformatics and related fields like medicinal chemistry and drug discovery. Key requirements of such representations are that the fingerprints should be intuitive, easily computable, invariant with respect to translations and rotations of the material, invariant to permutations of like atoms or motifs, and generalizable to all cases within the same chemical subspace. With the idea that the polymer properties are dictated by “group contributions” from its basic building elements, we proposed a chemo-structural fingerprinting scheme that quantifies the chemical build-up of the polymer in terms of its constituent basic chemical units, like the building blocks shown in **Figure 3.2** or types of atoms like C, H or O.

A simple polymer fingerprint could therefore be a count of the number of different types of building blocks (e.g., the number of CH_2 blocks, the number of C_6H_4 blocks, etc.), normalized by the total number of blocks in the repeat unit. This would give rise to a 7-dimensional vector, each component of which corresponds to one of the blocks and is related to the number of times it appears in the given polymer repeat unit. We call this fingerprint M_I. While it is a simple and elegant way of representing a polymer, M_I does not take the effects of neighboring blocks into account. Thus, we go a step higher in complexity and propose fingerprint M_{II}, which is a count of the number of different types of pairs of building blocks in the polymer, normalized again by the total number of blocks in the repeat unit. M_{II} is defined as a 7 x 7 matrix, every component of which corresponds to any one pair of two neighboring blocks (eg. CH_2-NH pairs, CS-O pairs, etc.). Similarly, a fingerprint M_{III} can be defined which would be a 7 x 7 x 7 matrix, each component of

which refers to any triplet of blocks (CH_2-NH-CO triplets, C_4H_2S-C_6H_4-CS triplets, etc.). In this fashion, we could go to higher dimensional fingerprints with more information added at every step; in the limit that we consider n-tuple block combinations, we can uniquely represent any polymer out of an n-block polymer repeat unit chemical space.

We refer to the scheme of fingerprinting in terms of chemical building blocks (such as CH_2 and NH) as *Fingerprint Type I*, and it is depicted pictorially in **Figure 5.2**. In the same fashion, constituent atom types (such C, H and O, and what kind of coordination environment they adopt) could be considered instead of blocks, leading to *Fingerprint Type II* as also depicted in **Figure 5.2**. A hierarchy of fingerprints can be defined within this type as well, like M_I, M_{II} and M_{III}. Fingerprint Type II will be described in further detail and utilized for learning purposes in **Chapter 6**, while Fingerprint Type I was used in the results presented in this chapter.

The fingerprints M_I, M_{II} and M_{III} are characterized by a few key mathematical constraints which have been listed below:

i) The sum of all the elements in any fingerprint should be equal to the total number of blocks in the polymer (N). Thus:

$$\sum_{i=1}^{7} M_I(i) = N$$

$$\sum_{i,j=1}^{7} M_{II}(ij) = N$$

$$\sum_{i,j,k=1}^{7} M_{III}(ijk) = N$$

ii) The sum of elements in any row or column of M_{II} should be equal to the total number of blocks of that kind in the polymer. This can be written as:

$$\sum_{j=1}^{7} M_{II}(ij) = M_{I}(i)$$

Similarly, the sum of elements in any given 7x7 matrix plane in M_{III} should be equal to the total number of blocks of that kind in the polymer, which can be written as:

$$\sum_{j,k=1}^{7} M_{III}(ijk) = M_{I}(i)$$

iii) The periodic symmetry in the polymer dictates that the fingerprint matrix diagonal acts as a mirror; the corresponding elements on either side of it should be equal. That is, $M_{II}^{ij} = M_{II}^{ji}$ and $M_{III}^{ijk} = M_{III}^{kji}$.

iv) The diagonal elements in any fingerprint matrix should be integer values, that is, M_{II}^{ii} and $M_{III}^{iii} \in$ the set of non-negative integers.

With the present prescription, the fingerprint for any given n-block polymer is populated by assigning a certain score to every block or pair of blocks or triplet of blocks that is encountered, with the counting done from either end of the polymer repeat unit to take

periodicity and inversion into account. The scores are always averaged and normalized by the total number of blocks in the repeat unit. The averaging step ensures that sum rules are satisfied, and normalization assures that the fingerprints are generalizable to repeat units of arbitrary length. It should be noted that this polymer fingerprint does not consider spatial degrees of freedom or other structural factors, and would thus not distinguish between two polymers with the same repeat unit but different crystal structural arrangements.

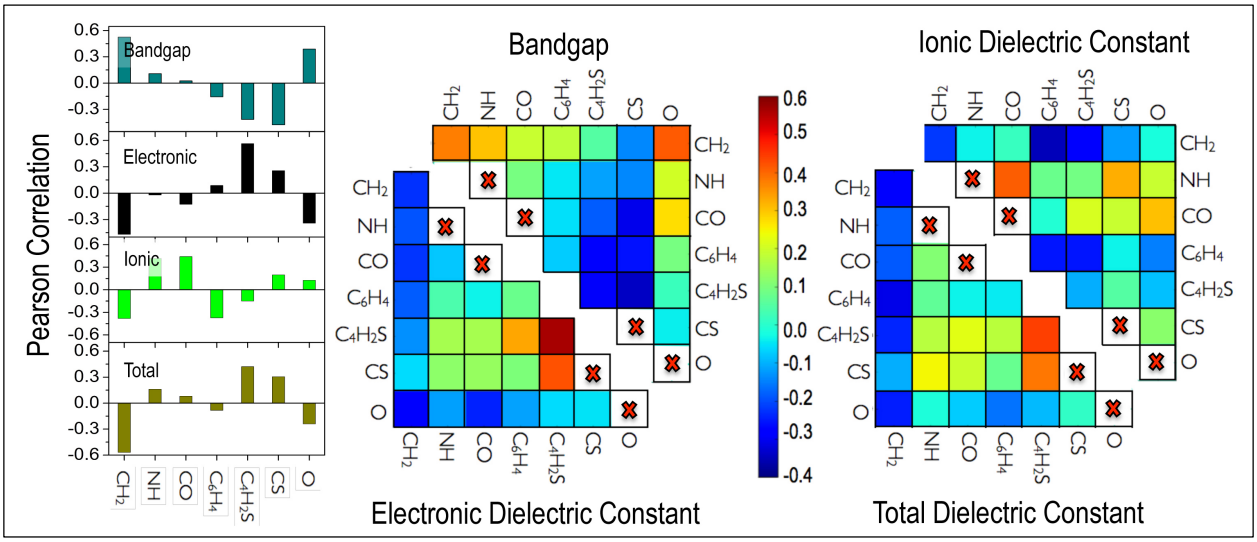


Figure 5 3. *Correlations between different components of M_I and M_{II} with the different properties.*

For ease of initial discussion, we considered the fingerprints M_I and M_{II}. Correlations between the different components of fingerprint MI and 4 properties (ε_{elec}, ε_{ion}, ε_{tot} and E_{gap}) are shown in **Figure 5.3**. The coefficients plotted on the y-axes were obtained using the Pearson correlation analysis, which gives us values between –1 and +1 showing the

degree of negative or positive correlation between any property and any component of the fingerprint vector [147]. The opposite behavior of ε_{elec} and E_{gap} can be ascertained by observing their respective plots: CH_2 and O blocks make notable positive contributions to E_{gap} and negative contributions to ε_{elec}, whereas C_4H_2S and CS contribute positively to ε_{elec} and negatively to E_{gap}. The same effects largely translate to ε_{tot} as well; CO and NH blocks were seen to contribute the most to ε_{ion}.

Results for a similar Pearson correlation analysis between M_{II} and the 4 properties are shown in **Figure 5.3** in the form of half-matrix heat maps. The shade of the color in any matrix component (based on the adjoining color scale) shows how positively or negatively that pair of blocks is correlated with the given property. Once again, it can be seen how the heat map for E_{gap} is opposite to that of ε_{elec} or ε_{tot} in terms of the spectrum of colors (dark blue to dark red). While C_6H_4-C_4H_2S, C_4H_2S-C_4H_2S and C_4H_2S-CS pairs make the most positive contributions to ε_{elec} and CH_2-O and CO-O pairs make the most negative contributions, the roles of these pairs are just reversed when considering their contributions to E_{gap}. In case of ε_{ion}, NH-CO, NH-CS and CO-O pairs contribute to its increase while CH_2-C_6H_4 and CH_2-C_4H_2S pairs have the opposite effect. It is now possible for us to come up with educated combinations of different kinds of pairs of building blocks targeted towards increasing the dielectric constant or the bandgap or indeed, both. Considering these insights, it is not surprising that polymers with [-NH-CO-NH-C_6H_4-], [-NH-CS-NH-C_6H_4-], and [-NH-CO-NH-C_6H_4-] repeat units were singled out in past work as promising dielectrics for energy storage applications. Thus, the influence of specific

blocks and block pairs on the polymer properties was identified, and a similar analysis using fingerprint type II would reveal the atom types and pairs of atom types that are influential.

5.3 Machine Learning Applied on a Polymer Dataset

5.3.1. On-Demand Property Prediction

While qualitative notions such as discussed above are useful, a quantitative property prediction model that is fast (because it by-passes the DFT route to property predictions) would satisfy several practical needs. Following previous work, we used kernel ridge regression (KRR) to establish a quantitative mapping between the polymer fingerprints on the one hand and the relevant properties (namely, E_{gap}, ε_{elec}, and ε_{ion}) on the other. KRR is a statistical or machine learning algorithm capable of handling nonlinear relationships [77] [117]. By comparing the fingerprint, say M_{III}, of a new polymer with those of a set of reference cases for which property values are known, an interpolative prediction of the property of the new polymer may be obtained.

In practice, the machine learning prediction model is developed for a subset of the available dataset, referred to as the training set, and the performance of the model is

tested on the remainder of the dataset, referred to as the test set. Model development based on the training set also included internal cross-validation to minimize over-fitting and ensure model generality. In the present work, about 90% of the 284 4-block polymer dataset was taken to be the training set, and the remaining 10% was placed in the test set. The optimal training set size was determined by studying the ML model performances for different training set sizes.

The plots in **Figure 5.4**, **Figure 5.5** and **Figure 5.6** show E_{gap}, ε_{elec} and ε_{ion} as predicted using the KRR-based machine learning (ML) model versus the respective DFT values, using fingerprint M_I, M_{II} and M_{III}, respectively. It was seen that the prediction errors (both training and test) steadily decreased from M_I to M_{II} to M_{III}, indicating that the higher-dimensional fingerprint M_{III} is required within the KRR formalism to obtain predictive models with satisfactory accuracies. In **Figure 5.6**, the insets show the relative error distribution for each property prediction, indicating that the average error for all three properties is of the order of 10% or less. We thus have a model in our hands that will convert a fingerprint (M_{III}, in the present illustration) to property values with errors that are reasonable (given the efficiency of the prediction process relative to DFT).

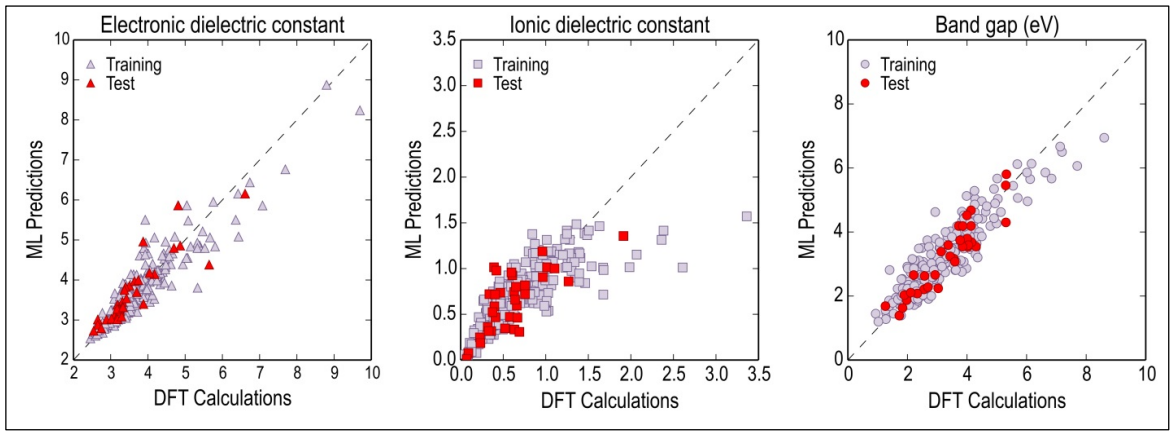


Figure 5.4. *ML-DFT parity plots for models trained with M_I for the three properties.*

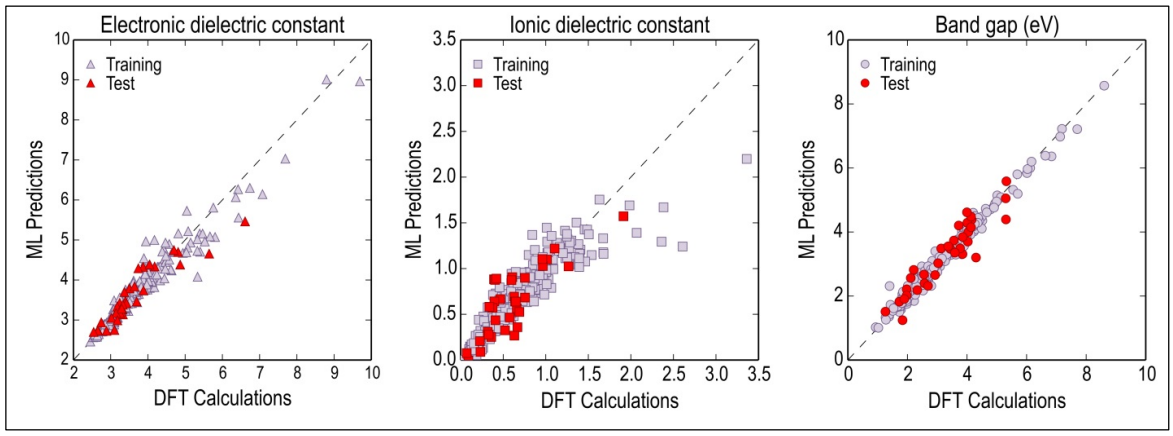


Figure 5.5. *ML-DFT parity plots for models trained with M_{II} for the three properties.*

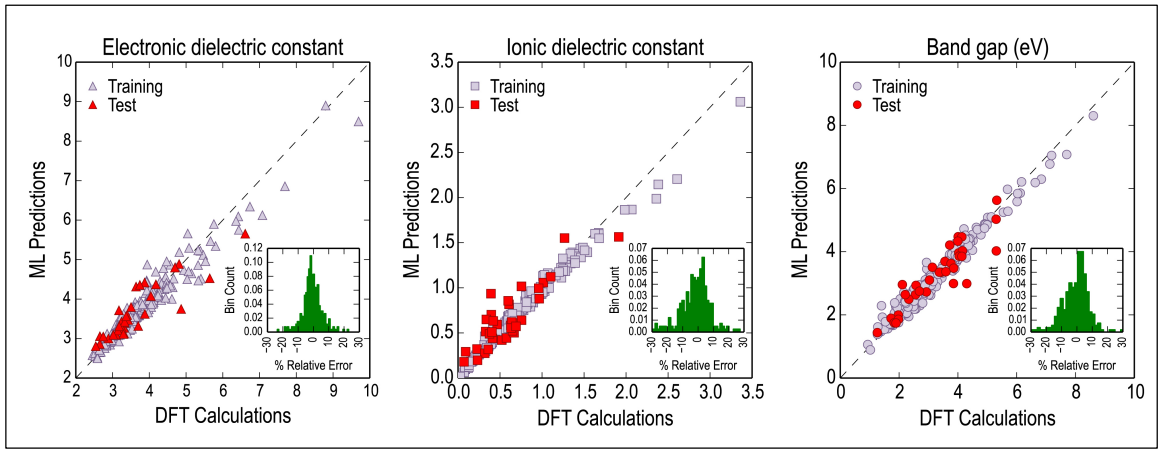


Figure 5.6. *ML-DFT parity plots for models trained with M_{III} for the three properties.*

The true power of such a property prediction model is its ability to instantly predict E_{gap}, ε_{elec} and ε_{ion} for a polymer with arbitrarily long repeat unit (but with the building blocks drawn from the same pool of 7), without needing to pursue the cumbersome approach of structure prediction and DFT. The workflow involved in predicting the properties of new *n-block* polymers is depicted in **Figure 5.7 (a)**.

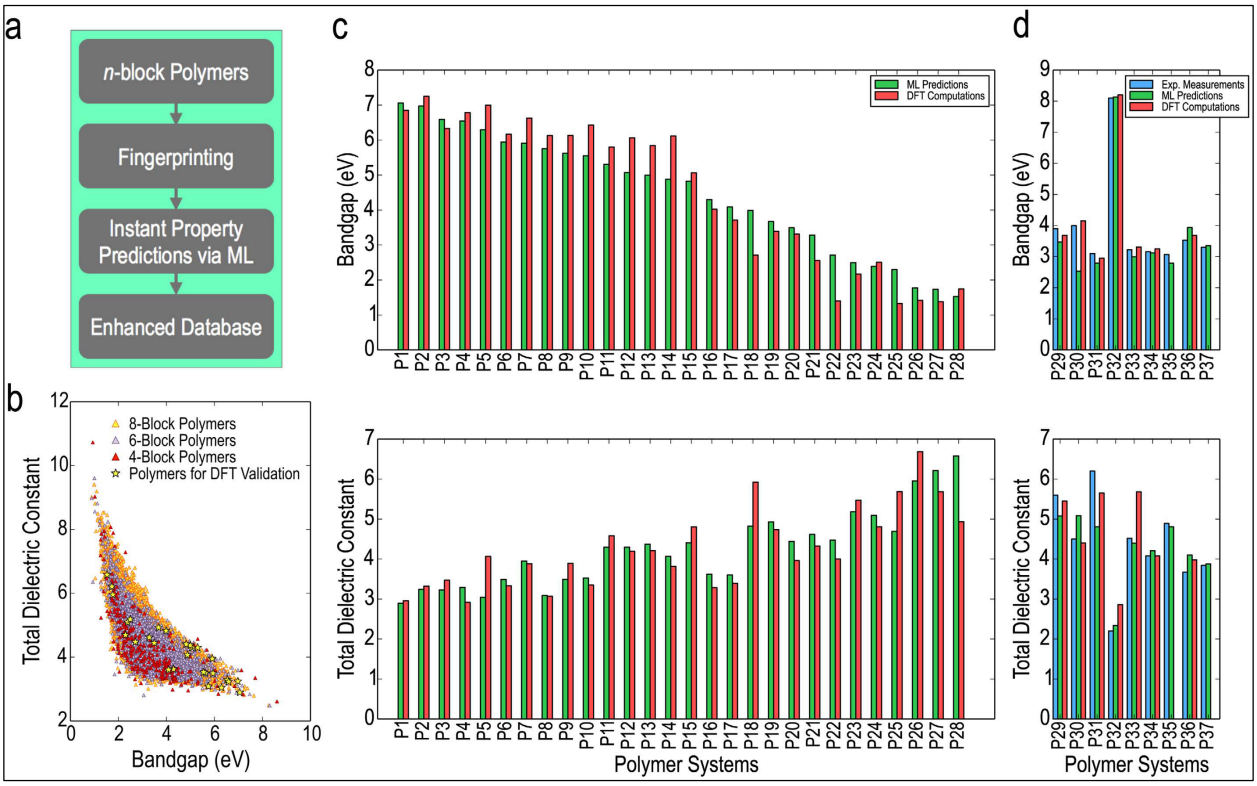


Figure 5.7. *On-demand property prediction of polymers. (a) The steps involved in predicting properties of any given n-block polymer using the instant prediction models. (b) Dielectric constants and bandgaps from the prediction models plotted against each other for all 6-block polymers and 8-block polymers, with the computational data for 4-block polymers also shown for reference. (c) Machine learning predicted and DFT computed*

properties of 28 polymers obtained by applying the direct design scheme to different ranges of dielectric constants and bandgaps. (d) The machine learning predicted, DFT computed and experimentally measured properties of some previously synthesized polymers.

If one were to pursue the enumeration approach, it is straightforward to list all possible n-block polymers for any given n, if n is a small enough number. To illustrate this, we came up with all the possible symmetry-unique 6-block polymers (~ 6,000 in number) and 8-block polymers (~ 200,000 in number), determined their respective fingerprints, and estimated their properties using our ML model. **Figure 5.7 (b)** shows the predicted ε_{tot} (= $\varepsilon_{elec} + \varepsilon_{ion}$) plotted against the predicted E_{gap} values for all the 6-block polymers and 8-block polymers, as well as for the considerably smaller number of the 4-block polymers. **Figure 5.7 (b)** is a demonstration of how one may use interpolative statistical learning methods to densify the population within a chemical subspace. We thus have thousands of more options to choose from than we did in **Figure 3.5**.

The predictive performance of our model can be put to test in two ways: by comparing our predictions with actual DFT calculations, and by comparing them with available laboratory measurements. First, we validate our ML model against DFT calculations. A selection of 8-block polymers ranging from low (high) to high (low) values of ε_{tot} (E_{gap}) was chosen out of **Figure 5.7 (b)** (shown by stars in figure; incidentally, these were also the cases identified by our genetic algorithm, discussed in the next section, but the same

examples serve the present purpose of ML model validation). The stable crystal structures of these 8-block polymers were determined using Minima Hopping as before, following which their dielectric constants and band gaps were calculated using DFT. **Figure 5.7 (c)** compares the ML prediction with the corresponding DFT results. As can be seen, the agreement is impressive indicating that the prediction model trained on 4-block cases is transferable to polymers with repeat units of arbitrary size.

Next, in **Figure 5.7 (d)**, we compare the on-demand predictions with experimental values for polymers synthesized and tested in the recent past [50] [75], as well as the corresponding DFT results, for completeness. These polymers were synthesized following the earlier work on high-throughput computational data generation using the isolated polymer chains model; this means we have available experimental as well as computational quantification of ε_{tot} and E_{gap} for several polymers which are predictable with our prediction models. These polymers are, of course, the same synthesized polythioureas, polyureas, polyimides, etc. that were explained in detail in **Chapter 3**. Clearly, again, the performance of the ML model is impressive. The closeness of our predictions with first principles as well as with actual experiments allows us to state with some confidence that we have the means to instantly, and with reasonable accuracy, predict the properties of any *n-block* polymer belonging to the chemical subspace under consideration. All the polymers plotted in **Figure 5.7 (c)** and **(d)** are denoted by labels P1 to P37, and the polymer repeat unit corresponding to each label is shown in **Figure 5.8**.

The ML predictions are always close to the experimental values, validating our claim of accelerating property prediction for arbitrarily long polymer chains.

Label	Polymer Repeat Unit	Label	Polymer Repeat Unit
P1	CH_2-O-CH_2-O-CH_2-CH_2-CH_2-CH_2	P20	O-C_6H_4-CO-C_4H_2S-CO-NH-O-CO
P2	CH_2-O-CH_2-O-CH_2-CH_2-CH_2-O	P21	CH_2-CH_2-O-CS-NH-CS-C_6H_4-NH
P3	CH_2-NH-CH_2-CH_2-CH_2-O-CH_2-O	P22	C_6H_4-C_6H_4-CH_2-CS-C_4H_2S-CS-CH_2-O
P4	CH_2-CH_2-O-CO-O-CH_2-CH_2-O	P23	C_6H_4-NH-C_6H_4-CS-NH-C_4H_2S-CO-NH
P5	CO-O-CH_2-CH_2-CH_2-CH_2-CH_2-O	P24	CO-C_4H_2S-NH-CS-O-C_4H_2S-NH-C_4H_2S
P6	CH_2-CH_2-O-CO-NH-CH_2-CH_2-O	P25	CS-CO-CH_2-CH_2-NH-C_6H_4-CS-C_6H_4
P7	CH_2-NH-CO-NH-CH_2-O-CH_2-O	P26	C_6H_4-NH-C_4H_2S-C_4H_2S-CS-C_4H_2S-C_4H_2S-NH
P8	CH_2-CH_2-CH_2-CH_2-NH-CO-CH_2-CH_2	P27	C_4H_2S-C_4H_2S-C_4H_2S-CS-C_4H_2S-NH-CS-NH
P9	CO-NH-O-CH_2-CH_2-CH_2-CH_2-O	P28	C_4H_2S-CS-C_4H_2S-CS-CO-NH-C_6H_4-C_4H_2S
P10	CH_2-O-CO-NH-CH_2-CH_2-NH-CH_2	P29	NH-CO-NH-C_6H_4
P11	CH_2-NH-CH_2-NH-CO-NH-CO-NH	P30	CO-NH-CO-C_6H_4
P12	CH_2-NH-CO-O-NH-CO-NH-O	P31	NH-CS-NH-C_6H_4
P13	CO-NH-CO-O-CO-NH-CH_2-NH	P32	CH_2-CH_2-CH_2-CH_2
P14	CO-NH-CO-NH-CH_2-CH_2-CH_2-NH	P33	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4-O-C_6H_4
P15	CO-NH-CO-CH_2-NH-CO-O-NH	P34	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4-CH2-C_6H_4
P16	C_6H_4-O-CO-CH_2-CO-CH_2-CH_2-O	P35	NH-CS-NH-C_6H_4-NH-CS-NH-C_6H_4
P17	CH_2-CH_2-CO-O-CO-CH_2-C_6H_4-C_6H_4	P36	NH-CS-NH-C_6H_4-NH-CS-NH-$[CH_2]_6$
P18	CO-NH-O-NH-CO-NH-C_4H_2S-NH	P37	NH-CS-NH-C_6H_4-CH_2-C_6H_4
P19	CO-NH-CO-NH-CO-NH-C_4H_2S-NH		

Figure 5.8. *Polymer repeat units denoted by the labels P1 to P37 in* ***Figure 5.7****.*

5.3.2. On-Demand Direct Design

Although the entire expanse of the chemical space can be covered using enumeration, it is essentially a brute-force search for suitable polymers, and as such not the best possible design strategy. For instance, enumerating for 8-block, 10-block and 12-block polymers will lead to ~ 200,000, ~ 5,000,000 and ~ 50,000,000 systems respectively (this

exponential explosion in the chemical subspace was captured earlier in **Figure 3.3** in **Chapter 3**), which are unreasonably large numbers considering the property domain of interest may restrict us to a small fraction of that. We thus attempted to find an efficient way of obtaining specific *n-block* polymers that simultaneously show a certain desirable dielectric constant and a desirable bandgap, without having to individually consider every possible polymer. Such a model would make the "desired properties → suitable materials'" route an instant, on-demand reality [77] [141] [148] [149].

We applied a genetic algorithm (GA) approach as the means to optimize the polymers given the target properties. It has been shown that GA is a very efficient approach in searching for materials with desired properties when compared to other approaches like random search and even chemical-rules based search [150]. The idea here is to start with a random initial population of *n-block* polymers (for any given *n*) and let them undergo evolution (in terms of constituent blocks and their neighbors) based on the principles of GA, finally yielding a set of polymers with properties closest to the provided targets. At any step, the properties of the polymers are computed instantly using the on-demand prediction ML model we developed and explained in the previous section. The series of steps followed in this method are shown in **Figure 5.9 (a)**. The same philosophy was implemented by us in an earlier work as well [77], but using a simulated annealing approach instead of GA for designing organic molecules with specific target properties.

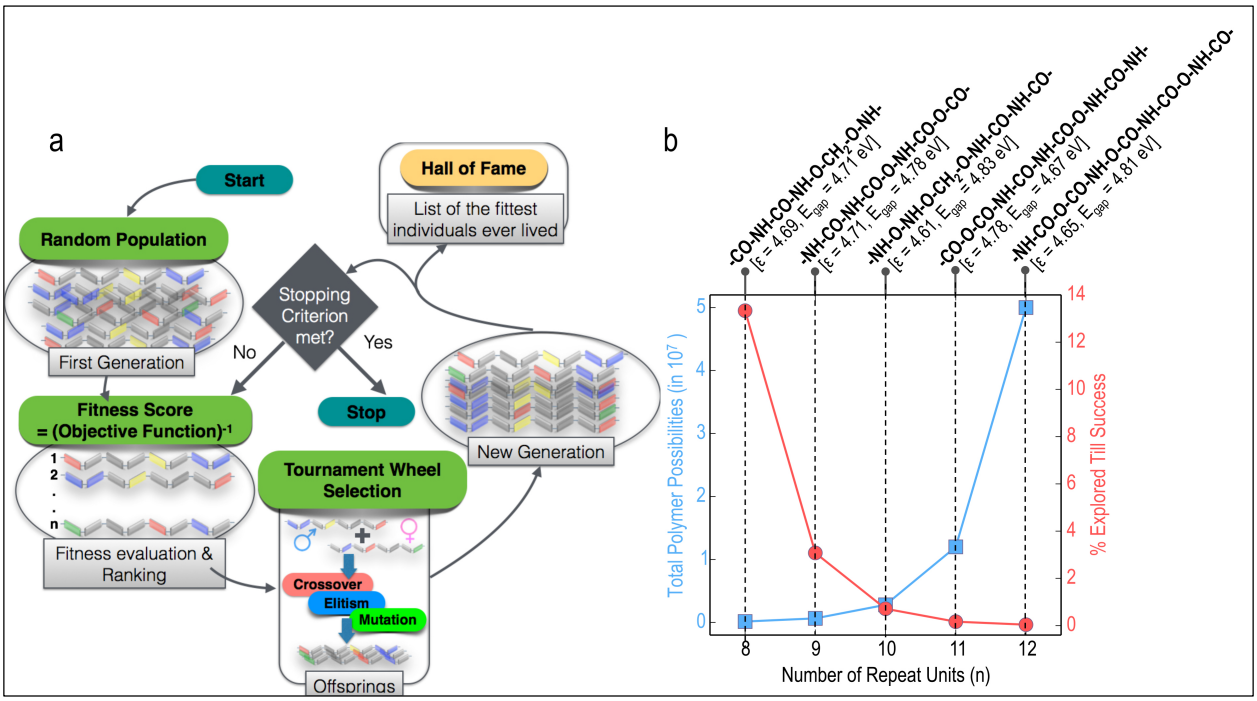


Figure 5.9. *(a) The steps involved in the genetic algorithm (GA) approach leading to direct design of polymers. (b) The exponential increase in total polymer possibilities for increasing number of repeating blocks, and the simultaneous decrease in the percentage of points to be explored till success. Also shown are one optimal polymer each for each case for a target dielectric constant and bandgap of 5 and 5eV respectively.*

Given the target ε_{tot} and E_{gap}, and the number of blocks in the polymer repeat unit (the value of *n*), the algorithm generates a list of 300 *n-block* polymers which serve as the first generation. Based on the predicted property values, a fitness score is assigned to every polymer and all the polymers are ranked according to this score. While polymers with satisfactory fitness scores survive (this is called elitism), the rest undergo different kinds of evolution, namely crossover and mutation [150]. New generations of polymers are produced in this manner; a stopping criterion is provided based on the fitness score, and

once polymers with suitable fitness scores are obtained, the algorithm stops. From every generation, the polymers with fitness scores above a certain threshold are compiled as the list of best solutions. At the end of the algorithm, this list contains the final set of optimal polymers showing the desired ε_{tot} and E_{gap}.

Based on the target dielectric constant and band gap, an objective function was defined as the following:

$$W = [\varepsilon_{tot} - \varepsilon_{tot}^{target}]^2 + [E_{gap} - E_{gap}^{target}]^2$$

where $\varepsilon_{tot}^{target}$ and E_{gap}^{target} are the target dielectric constant and bandgap values respectively, while ε_{tot} and E_{gap} are the dielectric constant and bandgap of the polymer undergoing optimization. This function would be minimized when the difference between either property of the polymer and the respective target property is the least. Further, a fitness score (mentioned in the previous paragraph) was defined as the inverse of the objective functional value, and acted as the measure of suitability of any system. We devised a polymer encoding system that converted any *n-block* polymer into an *n*-component vector, assigning a number between 0 and 6 to each of the 7 motifs respectively. Using completely random values for this vector, an initial population of 300 polymers was generated. Properties were instantly calculated for all these polymers using the on-demand prediction models, and the fittest polymers (showing the highest fitness scores) were selected. Mating is performed between these individuals using a

combination of crossovers, elitism and mutation, giving rise to the 'offspring' polymers that then go forth to the next generation of polymers. In crossover, some of the vector components of the parent polymers were simply exchanged to generate the children. Elitism refers to preserving a few of the fittest parent polymers in the next iteration, whereas with mutation, we changed some of the vector components of the parents randomly to obtain the children. Thus, generation after generation of polymers was studied and those with the highest fitness scores at every generation went into the list of best solutions. In the end, this list would contain the best individuals that ever lived (that is, the polymers with properties closest to the target values $\varepsilon_{tot}^{target}$ and E_{gap}^{target}), and these would be our solutions.

For a demonstration and validation of this approach, we restricted our initial search to 8-block polymers only, as this provides us with a substantial population of systems to explore while ensuring the system size does not become so large as to render subsequent first principles validation extremely expensive. We took 6 different ($\varepsilon_{tot}^{target}$, E_{gap}^{target}) combinations as the targets, and allowed the algorithm to search for suitable 8-block polymers showing the best combination of properties. **Figure 5.7 (c)** gives a glimpse of the results: we show a few polymers each obtained for the different targets we provided. The ML model predicted property values for these polymers are always close but not the same as the target values; we further show here the DFT computed values obtained after performing crystal structure prediction for these polymers. As mentioned earlier, there is excellent agreement between the ML predictions and the DFT results.

To understand exactly how valuable the direct design scheme is, we need to quantify the speed of the GA approach when compared to enumeration. Taking the example of *8-block* polymers, while there are a total possible ~ 200,000 such systems, GA is able to traverse a small percentage of the points in determining the required polymer(s). Upon going to higher block systems, like *9-block* or *10-block* polymers, the total possibilities are exponentially higher but the percentage of points the algorithm needs to explore is even smaller. **Figure 5.9 (b)** shows that despite the exponential increase in total polymer possibilities, as the number of repeating units *n* increases, a smaller and smaller percentage of points need to be considered by the algorithm in order to obtain the optimal polymer(s). Also shown in **Figure 5.9** (b) are certain *n-block* polymers obtained for different values of *n* for $\varepsilon_{tot}^{target} = 5$ and $E_{gap}^{target} = 5$ eV (a very desirable combination of properties for energy storage capacitor dielectrics). Thus, with actual polymer outputs (with arbitrarily long chains) as well as a quantification of the speed-up, we have in our hands an efficient polymer design model that negates the need for enumeration followed by down-selection of desired systems.

5.4 Critical Assessment of regression-based machine learning methods

Apart from the availability of robust, uniformly generated data, there are a few other essential factors in the machine learning process that need to be taken care of for optimal learning. These include defining a suitable fingerprint, choosing a learning algorithm, and determining the necessary subset of the data that is needed for training the learning model. The fingerprints we chose and tested in ref. [52] were chemo-structural in nature, that is, they quantified the types and combinations of different constituent blocks in the polymer. Three fingerprints were used: a count of the different types of building blocks in the polymer, called fingerprint M_I, a count of the types of block pairs (fingerprint M_{II}), and a count of the types of block triplets (fingerprint M_{III}). The fingerprints were normalized and generalized for any number of blocks in the polymer repeat unit, and used to train a regression model for the three properties of interest.

Whereas all three fingerprints were tested in ref. [52], the learning algorithm used was Kernel Ridge Regression (KRR), a nonlinear regression technique that works on the principle of similarity. Euclidean distances between fingerprints were used to quantify the similarity. A distance kernel goes into the definition of the property here, for which a Gaussian kernel was used. Around 90% of the entire polymer dataset was used to train the KRR model, and predictions were made on the remaining points as a test of the performances. Mean absolute errors (MAE) in prediction of less than 10% with respect to the DFT values were seen, which is satisfactory for a statistical model and the best

performance that could be obtained using the current optimal learning parameters. The optimal fingerprint used here was M_{III}, with M_{II} and M_{I} discarded owing to larger prediction errors.

Machine Learning Parameters	**Choices used so far**	**Choices explored here**
Fingerprint	M_{I}, M_{II}, M_{III}	M_{III}
Regression Algorithm	KRR	KRR, SVR, AdaBoost
Type of Kernel	Gaussian	Gaussian, Laplacian, Linear, Polynomial
Training Set Size	90% of Data	Learning Curves
Error Definition	MAE	RMSE, $1 - R^2$

Table 5.1. *A comparison of various choices of machine learning parameters used in Ref.* [52] *and explored here. The acronyms used stand for: Kernel Ridge Regression (KRR), Support Vector Regression (SVR), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and goodness of fit (R^2).*

Although we obtained learning models as described above to predict polymer properties with reasonable accuracies, a detailed study of all the different possible machine learning (or regression) parameters is due. Such a study can be very valuable in terms of truly testing the capabilities of our machine learning philosophy for the given polymer dataset, and indeed, improving the performances. In **Table 5.1**, we try to capture all these different parameters, mentioning the specific choices that we used in ref. [52] as well as the other possible options explored here. Although the fingerprint choices were already rigorously

tested, each of the other parameters provide room for further testing, and thus possible performance improvement.

We took the same polymer dataset and analyzed the machine learning prediction performances for different regression algorithms, different distance kernel choices, different training set sizes and different error definitions [128]. Possible alternative algorithms to KRR include, but are not limited to: Linear Regression (LR), Support Vector Regression (SVR), Gaussian Process Regression (GPR) and SVR with AdaBoost. Whereas we used KRR with a Gaussian kernel in ref. [52], Linear, Laplacian or Polynomial kernels can be used as alternatives in any kernel-based regression algorithm. Further, the training set size can be varied systematically to study the prediction errors. The prediction errors can be quantified in different ways, such as mean absolute error (MAE), root mean square error (RMSE) and error based on the coefficient of determination ($1\text{-}R^2$).

5.4.1 Kernel Ridge Regression (KRR)

In this section, we delve deeper into KRR, the algorithm that formed the basis of all machine learning prediction models in ref. [52]. KRR is a similarity based regression algorithm [78] that inherently takes the nonlinearity of the system into account. The ‘similarity’ between any two data points is defined using some standard mathematical

measure of distance, such as a Euclidean distance. For any two polymers *i* and *j* having fingerprints $\vec{x}_i$ and $\vec{x}_j$ respectively (where $\vec{x}_i$ is an *m* dimensional vector with components x_i^1, x_i^2, x_i^3... x_i^m), the Euclidean distance between them will be defined as:

$$d(\vec{x}_i, \vec{x}_j) = \| \vec{x}_i - \vec{x}_j \|_2 = \sqrt{(x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \cdots + (x_i^m - x_j^m)^2} \quad (1)$$

The smaller (larger) is this distance, the more similar (dissimilar) the two polymers are. Now, KRR involves defining the property of interest (the output) as a function of such a distance measure, so that the property of any polymer can be estimated by taking its distances from all the other polymers. Mathematically, the predicted property of polymer *j*, denoted by *P(j)*, will be defined as follows:

$$P_{pred}(j) = \sum_{i=1}^{n} \alpha_i K(\vec{x}_i, \vec{x}_j) \quad (2)$$

The summation is performed over the entire training set size *n*, and $K(\vec{x}_i, \vec{x}_j)$ is the *kernel function* that is defined in terms of $d(\vec{x}_i, \vec{x}_j)$, the distance between polymer *i* (in the training set) and polymer *j*. The purpose of the kernel function is to transform the points (the polymers) from the fingerprint space to a higher dimensional space, thus making nonlinear mapping possible [151]. The two crucial parameters that need to be optimized here are the kernel coefficients α_i and the parameters that go into the kernel definition, such as the Gaussian width for a Gaussian kernel. Training of a KRR model essentially

involves an iterative minimization of prediction errors leading to the optimal parameter choices.

In practice, as mentioned in the Introduction, the total available dataset is divided into two parts: the training dataset and the test dataset. When training the model using the former, an important step that must be carried out is *cross-validation*, wherein the training set itself is divided into subsets. One of the subsets is used as a temporary test set while training is performed on the remaining subsets, and this procedure is repeated for each of the subsets. The optimal regression parameters are obtained corresponding to minimum average prediction errors on the temporary test sets; subsequently, the error computed over the entire training set with these parameters is referred to as the `cross-validation error', or sometimes the cross-validated `training error'. The purpose of cross-validation is to avoid overfitting in the data and to make the model more generalizable, that is, to ensure that the model predictions would work reasonably for points outside the training dataset.

Mathematically, the training process involves a minimization of the following expression:

$$\arg\min \sum_{i=1}^{n}\left(P_{pred}(i) - P_{actual}(i)\right)2 + \lambda \sum_{i=1}^{n} ||\alpha_i||_2^{\,2} \qquad (3)$$

where $P_{pred}(i)$ is the KRR model predicted property value of polymer *i* as defined in **Equation 2** and $P_{actual}(i)$ is its actual property value; ($P_{pred}(i)$ - $P_{actual}(i)$) is thus a measure

of the prediction error. However, the second term in the expression involves the regularization parameter λ. Regularization [117] is an important step that is again aimed at preventing overfitting, and involves adding extra information to the expression being minimized. The solution to **Equation 3** is given by:

$$\vec{\alpha} = \left(\vec{K}_{train} + \lambda I\right)\text{-}1\ \vec{P}_{actual} \qquad (4)$$

where $\vec{\alpha}$ is the vector of all α_i values, $\vec{K}_{train}$ represents the kernel matrix for the entire training set, and $\vec{P}_{actual}$ represents the vector of actual property values for all points in the training set. Based on the above discussion, it would appear that the two important parameters that need to be optimized during the training process are the following: regularization parameter λ, and the relevant kernel parameters. A set of values for these parameters are tested here towards the minimization of the expression in **Equation 2**, thus yielding the final form of **Equation 3** that can be used for predictions on the test set.

5.4.1.1. Learning with different kernels

When applied in **Equation 3**, any kernel function will be expressed in terms of the distance between two polymers as defined by **Equation 1**. Whereas a given polymer *i* exists as $\vec{x}_i$ in the fingerprint space, implementing a kernel function is simply a way of projecting the polymer to the kernel space, which is what makes the application of a

technique such as kernel ridge regression possible. Many different types of kernel definitions can be applied in **Equation 2**, as shown in **Table 5.1**, such as a linear kernel, polynomial kernel, Gaussian kernel and Laplacian kernel. Here, we consider three different kinds of kernels and compare the KRR prediction performances with each, with prediction errors given using two error definitions: the root mean square error (RMSE) and ($1-R^2$), where R^2 is known as the coefficient of determination and represents goodness of the fit. We should thus be able to determine the best performing kernel with respect to one regression algorithm: KRR.

Gaussian Kernel

A Gaussian kernel (an example of a radial basis function kernel) is defined for any two polymers *i* and *j* as:

$$K\left(\vec{x}_i, \vec{x}_j\right) = \exp\left(\frac{-||\,\vec{x}i - \vec{x}j\,||2}{2\sigma 2}\right) \qquad (5)$$

Here, the numerator inside the exponential term contains the Euclidean distance measure, or the L_2-norm, and the denominator contains σ, a kernel parameter known as the Gaussian width. One of the most important things to note here is that σ is an adjustable parameter that affects the kernel performance in a major way. Given the square scaling relationship, even a slight overestimation of σ can cause the exponential to start acting linearly, which leads to a loss in nonlinearity of the kernel projection and thus, the KRR algorithm. On the other hand, an underestimation of σ can lead to overfitting in the training data and consequently, poor prediction performances on the test

set. Estimating the optimal σ value is thus of utmost importance, and the two parameters that need to be optimized while performing KRR with a Gaussian kernel are λ and σ.

Laplacian Kernel

A Laplacian kernel is also a radial basis function kernel, and can be expressed mathematically as:

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(\frac{-||\vec{x}i - \vec{x}j||1}{2\sigma 2}\right) \qquad (6)$$

The distance measure in the numerator of the exponential term here is the Manhattan distance, or the L_1-norm. The observations made about σ in the discussion of Gaussian kernels are applicable here as well.

Polynomial Kernel

Whereas the two kernels described above are exponential functions, yet another choice for a kernel could be a polynomial function. Such a kernel can be expressed as follows

$$K(\vec{x}_i, \vec{x}_j) = (\gamma(\vec{x}_i, \vec{x}_j) + c)\text{^}d \qquad (6)$$

where (.,.) denotes the dot product in the space of the input feature vectors and the adjustable parameters are the constant term *c* and the degree of the polynomial, *d*. Here, we take $c = 0$ and $\gamma = 1$ for simplicity, which leaves *d* as the one important kernel parameter to be optimized.

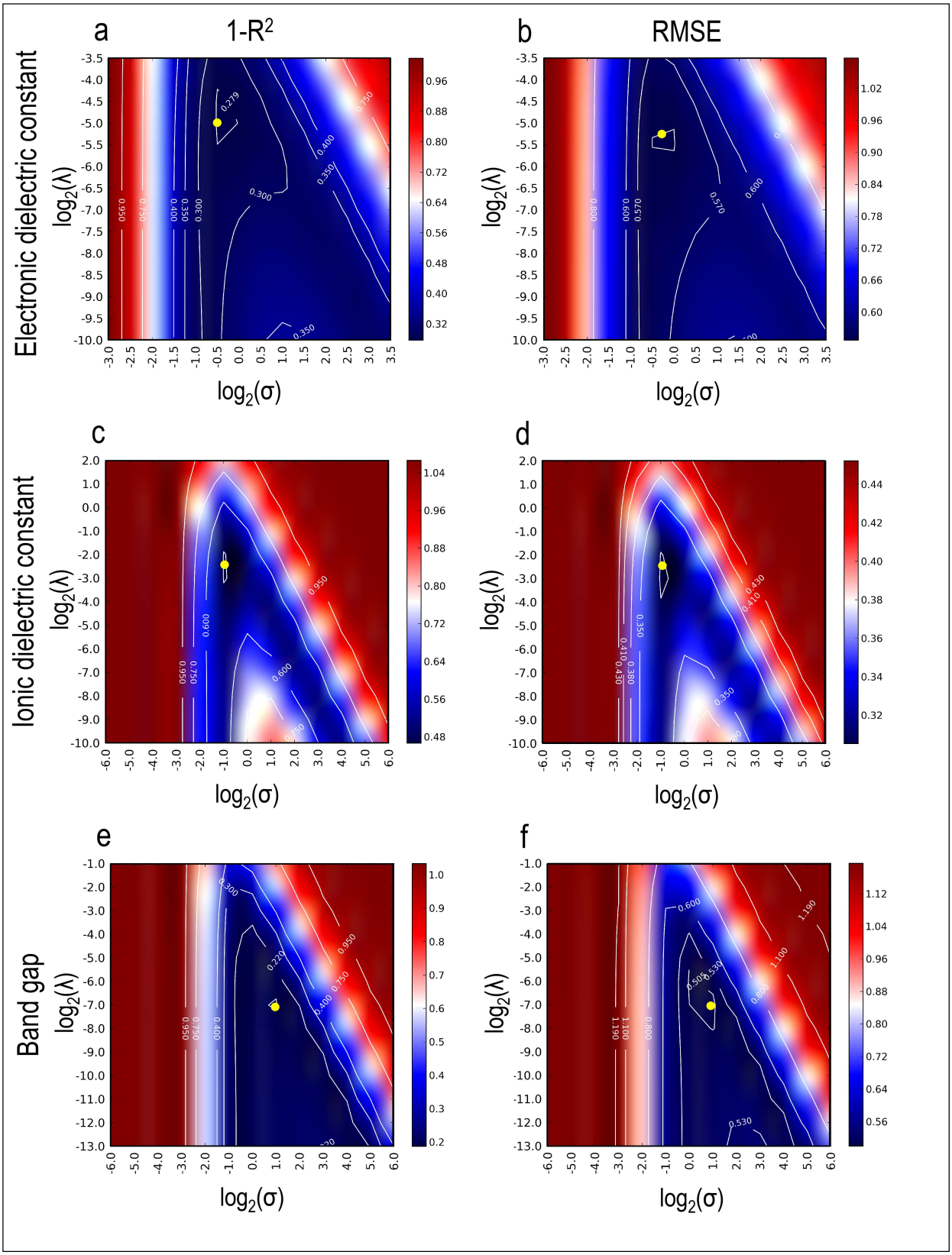


Figure 5.10. *Optimal parameter selection for KRR models with Gaussian kernels.*

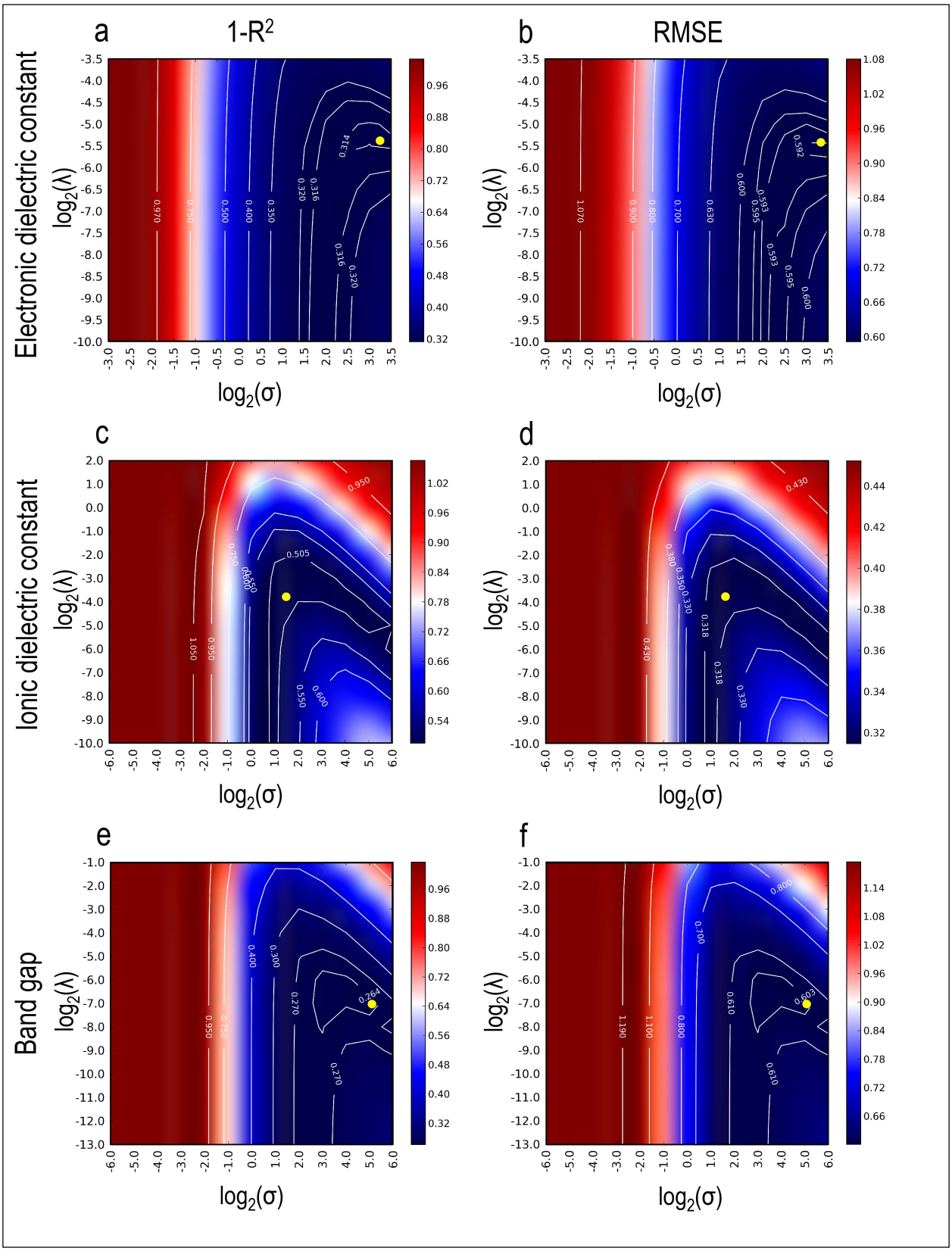


Figure 5.11. *Optimal parameter selection for KRR models with Laplacian kernels.*

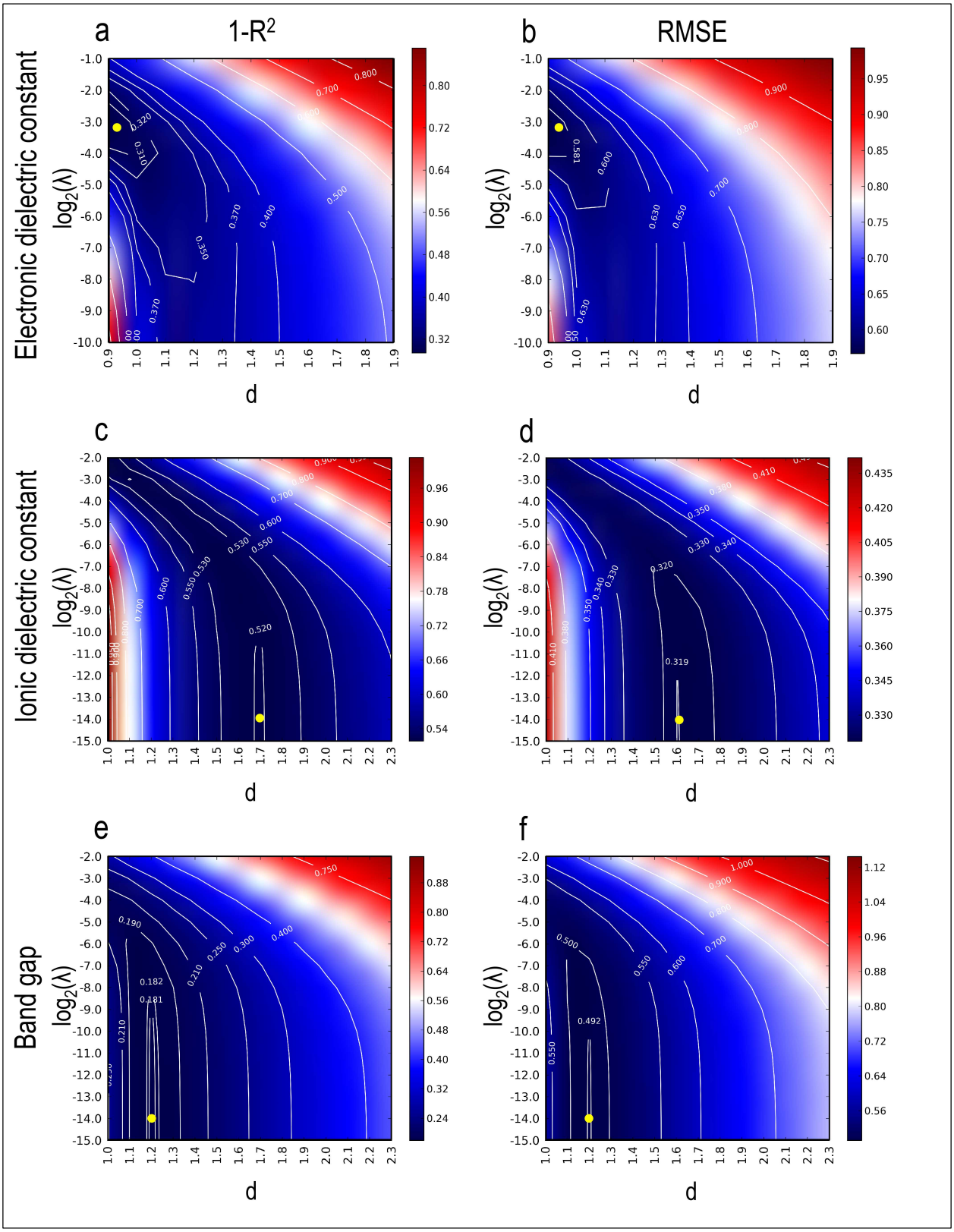


Figure 5.12. *Optimal parameter selection for KRR models with Polynomial kernels.*

Figure 5.10, **Figure 5.11** and **Figure 5.12** show plots between two vital KRR parameters when using the Gaussian, Laplacian and polynomial kernel respectively. Whereas the plot is between λ and σ in **Figure 5.10** and **Figure 5.11** (on a logarithmic scale), the plot in **Figure 5.12** is between parameter *d* and λ. Shown in different colors (according to the adjoining color bar) in each of the plots, for the three properties and using two different error definitions, are the respective prediction errors corresponding to any combination of the two parameters. The prediction errors are estimated for the training set points and the test set points respectively, as the averaged RMSE or ($1\text{-}R^2$) errors over all the points, and the test errors are depicted.

The plots in **Figures 5.10**, **5.11** and **5.12** enable us to determine regions of unfavorable parameter values as well as the region where the optima will be found. For example, a combination of $\sigma = 4$ and $\lambda = 2^{-7}$ appears to provide the minimum ($1\text{-}R^2$) and RMSE errors for E_{gap} predictions. The optimal [σ, λ] or the optimal [*d*, λ] values can similarly be obtained for KRR models for each property, using each kind of kernel. The lowest training and test prediction errors thus observed for the optimal parameter choices with the three different kernels are listed in **Table 5.2**. It should be noted that the optimal λ values obtained using the polynomial kernel, especially for ε_{ion} and E_{gap}, are many orders of magnitude smaller than those with the two exponential kernels. This has important consequences, as we explain below.

The plots of most interest following this study are the ones presented in **Figure 5.13**. KRR performances using the three kernels (based on the optimal parameter choices for each) are shown here for the three properties—ε_{elec}, ε_{ion} or E_{gap}—in the form of parity plots between the KRR predicted values and the actual DFT values. **Figure 5.13 (a)** shows the KRR performances using a Gaussian kernel, which is the same as the machine learning models that were presented in ref. [52]. It can be seen that there is no clear improvement in the prediction performances on the test set points upon going from the Gaussian to the Laplacian (**Figure 5.13 (b)**) and the polynomial (**Figure 5.13 (c)**) kernels, which vindicates the prior usage of the Gaussian kernel.

For ε_{elec}, the performance worsens with the polynomial kernel when compared to the exponential kernels. For the two other properties, whereas the test prediction errors are more or less the same with every kernel, there is a problem of overfitting in the data to some extent with the Laplacian kernel, but to a large extent with the polynomial kernel. This is because of the smaller values of the regularization parameter λ as pointed out earlier, which leads to a shrinkage of the second term in **Eq. 3**. Given that our model selection is based on the lowest cross-validation errors that can be obtained from the training set, the scatter in the test set points seen in **Figure 5.13 (c)** shows the inadequacy of the polynomial kernel in representing the properties as a function of the fingerprint. This further points towards the exponential kernels, and specifically the Gaussian kernel, being the best choice for KRR among the kernels and applications considered here.

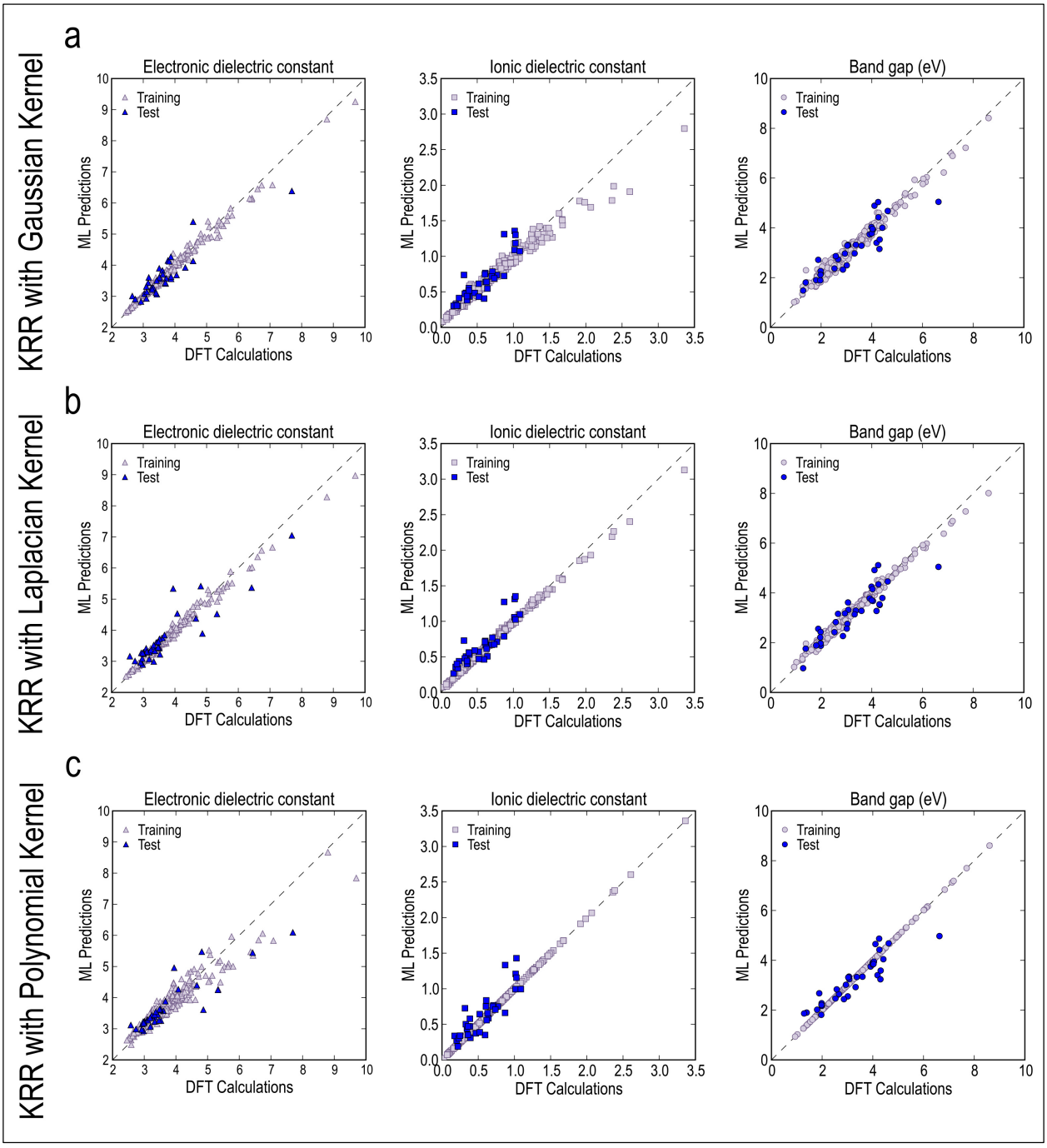


Figure 5.13. *Prediction performances of the KRR models using different kernels.*

5.4.1.2. Optimal Training Set Size: Learning Curves

While we have considered a training set size of 250 (approximately 90% of the entire dataset) in all the analyses presented so far, a rigorous demonstration of how we obtained this number is missing. In any statistical learning treatment, determining the minimum number of data points necessary for training a satisfactory model is of utmost importance. One may not possess sufficient data to train a respectable model, or one may possess excess data, in which case some points can safely be put aside for model testing purposes. Here, we present a systematic study of the adequacy of the training data set with respect to obtaining acceptable statistical prediction errors, using KRR with the three different kinds of kernels as before.

Shown in **Figure 5.14** for the three properties, for KRR with each kernel, are plots between the prediction errors ($1 - R^2$) and the training set sizes, referred to in machine learning practices as *learning curves* [121]. We increase the training set size from 50 (~ 20% of the dataset of 284) in steps of 5% of the entire dataset, all the way to 250 (~ 90% of the dataset); the test set is, of course, all the remaining points in the dataset. In each of the 9 cases, we consider 50 different randomly chosen training set populations for a given training set size, and measure the prediction errors (on the test set points) using the respective trained models. What we have plotted in **Figure 5.14** are the averaged test set prediction errors as well as the standard deviation in errors, for different training sizes. As one would expect, the general trend exhibited in each of the plots is a gradual

decrease in the average error as the training set size increases, which is simply owing to the improvement of the prediction model with a higher number of points trained upon. Whereas the standard deviations do not necessarily decrease the same way, the maximum and minimum errors generally follow the same trend as the average errors. In fact, the standard deviations seem to be higher in many of the cases for a large training set size, which happens because while some prediction models are excellent (reflected in the low minimum errors), there could be data overfitting in some others given the few remaining points that constitute the test set may not be well predictable (reflected in the high maximum errors). The average errors steadily decrease all the way till a training set size of 250, which justifies our optimal training size selection in ref. [52], except for band gap predictions with Gaussian kernel where the error minimum occurs around 220 points.

The learning curves are, for most parts, very smooth in nature and follow the average decreasing trend we expect. The standard deviations are a consequence of the dataset at hand, where selection of the appropriate `number' as well as `nature' of training set points has a strong effect on the prediction model. For instance, the absence of certain combinations of constituent polymer chemical blocks in the training set would make predictions on the test set containing such polymers quite poor, despite perhaps many data points being present in the training set. This is what leads to a high standard deviation in prediction errors in some cases; nevertheless, the learning curves do tell us that a large enough training set size would enable us to train regression models with sufficiently low prediction errors.

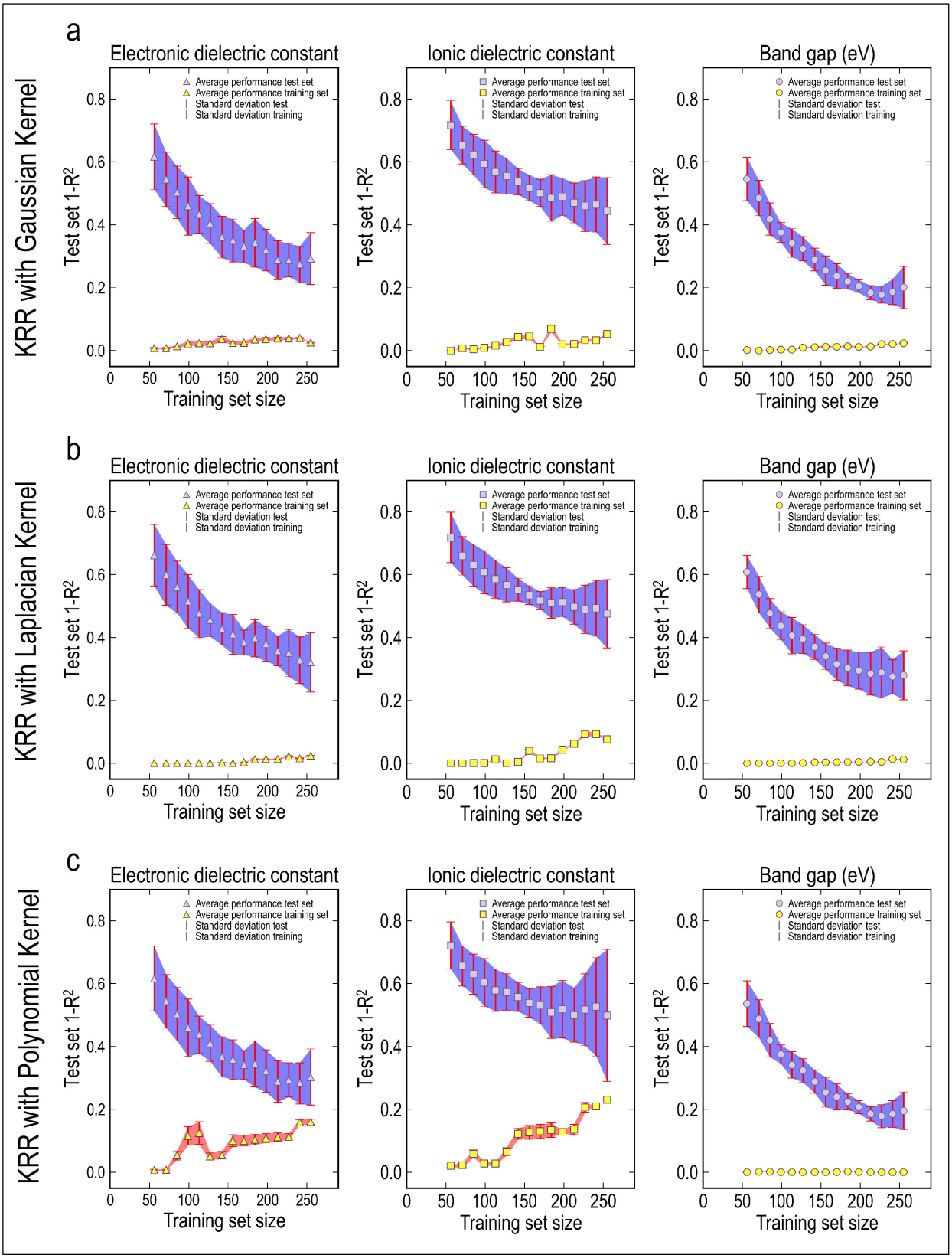


Figure 5.14. *Learning curves for KRR models with different kernels.*

Learning Algorithm	Kernel Used	Error in ε_{elec}	Error in ε_{ion}	Error in E_{gap}
Test Set				
KRR	Gaussian	0.193	0.368	0.203
KRR	Laplacian	0.187	0.361	0.214
KRR	Polynomial	0.233	0.407	0.194
SVR	Gaussian	0.250	0.259	0.183
SVR + Boosting	Gaussian	0.310	0.667	0.155
Training Set				
KRR	Gaussian	0.013	0.051	0.026
KRR	Laplacian	0.016	0.006	0.013
KRR	Polynomial	0.097	0.001	0.002
SVR	Gaussian	0.098	0.171	0.006
SVR + Boosting	Gaussian	0.070	0.160	0.003

Table 5.2. *Training and test prediction errors (1 - R^2) with all the regression algorithms.*

5.4.2 Support Vector Regression (SVR)

While Kernel Ridge Regression has provided reasonable prediction accuracies so far, the machine learning community has been known to use many other learning algorithms with varying degrees of success. One such algorithm is Support Vector Machines (SVM), supervised learning techniques developed at AT&T Bell Laboratories by Vapnik and co-workers [152] [153], and widely used in classification problems. When applied to regression and function estimation problems, SVMs are called Support Vector

Regression (SVR) and constitute a very popular regression algorithm which is implemented in most of the standard machine learning packages. SVMs are efficient tools for going beyond linear classification or regression owing to the implementation of the 'kernel trick', which as explained earlier, simply involves transforming data points to a higher dimensional kernel induced feature space to incorporate nonlinearity.

Given the input variables (the polymer fingerprint) and the response variable (the polymer property), in the form of training data $\{(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)\} \subset \chi \times \mathbb{R}$, where χ denoted a d-dimensional feature space, an ε-SVR algorithm tries to find a function *f(x)* that has at most ε deviation from the targeted property values y_i, and at the same time is as flat as possible. Any deviation larger than ε is not acceptable.

For linear regression, the function *f(x)* can take the following form:

$$f(x) = < w, x > + b, \qquad (8)$$

where $w \subset \chi$, b $\subset \mathbb{R}$. $Flatness$ of the function *f(x)* in this case means that we seek a vector w with a small norm, i.e., $\| w \|^2 = < w, w >$. The convex optimization problem can then be written as:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i - \langle w, \vec{\mathbf{x}}_i \rangle - b \leqslant \epsilon \\ & \langle w, \vec{\mathbf{x}}_i \rangle + b - y_i \leqslant \epsilon. \end{aligned} \tag{9}$$

In writing the above expression, we tacitly assume that the convex optimization problem is feasible, or in other words, there exists a function *f* that approximates all training pairs ($\vec{x}_i$,y_i) with at least ε precision. However, in practice, this may not be the case many a times, and we have to allow for some errors. This is done by incorporating a "soft margin" loss function through slack variables ξi and ξi^* in the otherwise infeasible optimization problem. Introduction of the slack variables in **Equation 9** leads to the following formulation:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \langle w, \vec{\mathbf{x}}_i \rangle - b \leqslant \epsilon + \xi_i \\ & \langle w, \vec{\mathbf{x}}_i \rangle + b - y_i \leqslant \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geqslant 0. \end{aligned} \tag{10}$$

where the positive constant C determines the trade-off between the flatness of *f* and the amount up to which deviations larger than ε are tolerated through the slack variables. The objective presented in the above minimization problem (**Equation 10**) is also referred to as the *primal* objective function, which is solved by constructing a Lagrange function $\mathcal{L}$, and accounting for the constraints through the positively constrained Lagrange multipliers α_i, α_i^*, β_i and β_i^*, as follows:

$$\mathcal{L}:$$
$$= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i\left(\epsilon + \xi_i - y_i + \langle w, \vec{\mathbf{x}}_i\rangle + b\right)$$
$$- \sum_{i=1}^{n}\alpha_i^*\left(\epsilon + \xi_i^* + y_i - \langle w, \vec{\mathbf{x}}_i\rangle - b\right) - \sum_{i=1}^{n}(\beta_i\xi_i + \beta_i^*\xi_i^*). \qquad (11)$$

In accordance with the saddle point condition, the partial derivatives of $\mathcal{L}$ with respect to the variables *w*, *b*, ξi and ξi^* lead to linear equations which when substituted back into **Equation 11** lead to the so called dual optimization problem of SVR, as given below,

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\langle\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j\rangle \\ & - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*) \qquad (12) \\ \text{subject to} \quad & \sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]. \end{aligned}$$

By solving the above dual optimization problem, α_i, α_i* and *b* can be determined, which can then be used to make predictions on new systems with a given input *x* as:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle\vec{\mathbf{x}}_i, x\rangle + b. \qquad (13)$$

Note that once the α_is and α_i*s have been determined, *w* can be written as $\sum_{i=1}^{n}(\alpha_i - \alpha_i *)$. This is known as *Support Vector expansion*, which describes w as a linear combination of the training data points. At this point it is important to note that the only data points for which the Lagrange multipliers (i.e., either α_i or $\alpha_i *$; both cannot be

simultaneously non-zero) are non-zero play a role in determining *w* and therefore enter **Equation 13**. From Karush-Kuhn-Tucker (KKT) conditions [154], it also follows that only for training data points for which the prediction error (i.e., $|f(\vec{x}_i) - y_i|$) is greater than ε, the Lagrange multipliers may be nonzero. Therefore, we have a sparse expansion of *w* in terms of $\vec{x}_i$ (i.e. not using all $\vec{x}_i$ to describe *w*). These non-vanishing coefficients are called Support Vectors.

Thus far, we have only considered a linear SVR problem. However, moving to a non-linear case from here is relatively straightforward and can be done by defining a kernel function $\phi(x)$ that takes a point *x* in the feature space and transforms it non-linearly in the kernel space. Furthermore, since the SVR algorithm's dual optimization in **Equation 12** only depends on the dot products between patterns x_i, for the non-linear case, it should suffice to know the analogous dot product $K(x,x')$ in the kernel space given by $<\phi(x), \phi(x')>$. This allows us to restate the non-linear SVR optimization problem as:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\mathcal{K}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) \\ & -\epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \qquad (14) \\ \text{subject to} \quad & \sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]. \end{aligned}$$

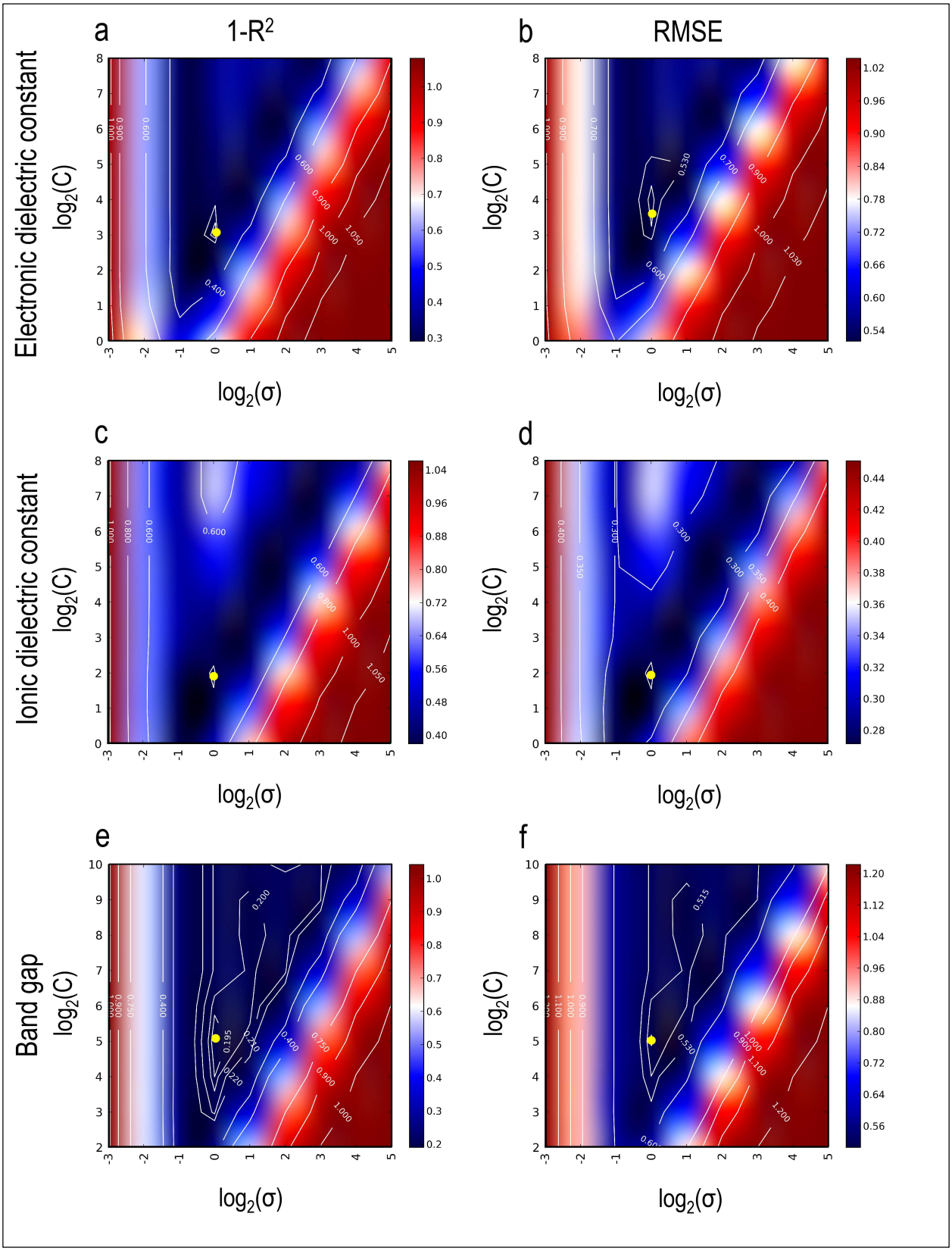


Figure 5.15. *Optimal parameter selection for the SVR models with Gaussian kernels.*

Note that in the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space. Furthermore, to be an admissible kernel, $K(x,x')$ is required to satisfy Mercer's condition [155]. Finally, following an analogous expression to the **Equation 13**, predictions on new systems for the non-linear case can be made as:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(\vec{\mathbf{x}}_i, x) + b. \qquad (15)$$

We take the bias term *b* to be zero, which leaves the Kernel parameter and the tradeoff constant C as the parameters that need to be optimized. In our case, after analyzing initial test performance on several kernels (such as linear, polynomial and Gaussian; all of which are admissible SVR kernels), we decided to go forward with the Gaussian kernel. Our results for this kernel are presented in **Figure 5.15** and **Figure 5.16**. Given that the parameters being optimized are σ and C, we measure model prediction errors for each combination of the two parameters and make plots showing the errors in **Figure 5.15**, similar to **Figures 5.11**, **5.12** and **5.13**. The optimal σ values are always around 1, whereas C takes different optimal values from 4 to 2^5. These respective pairs of optimal parameters were taken for the final SVR prediction models, whose performances are shown in **Figure 5.15**.

It can be seen from the parity plots in **Figure 5.16** that the regression performances are slightly worse than with KRR using a Gaussian kernel (shown in **Figure 5.14 (a)**). The

training and test set prediction errors for the three properties have been listed in **Table 5.2**. Whereas the training performances are worse for the electronic and ionic dielectric constants (than with Gaussian KRR), there is a clear problem of overfitting in the data for the band gap, which is again owing to the tradeoff constant C being higher (similar to the explanation for KRR with a polynomial kernel). Conventional SVR thus appears to not improve upon Gaussian kernel based ridge regression, and we attempt to rectify this in the following section with a technique known as 'AdaBoost'.

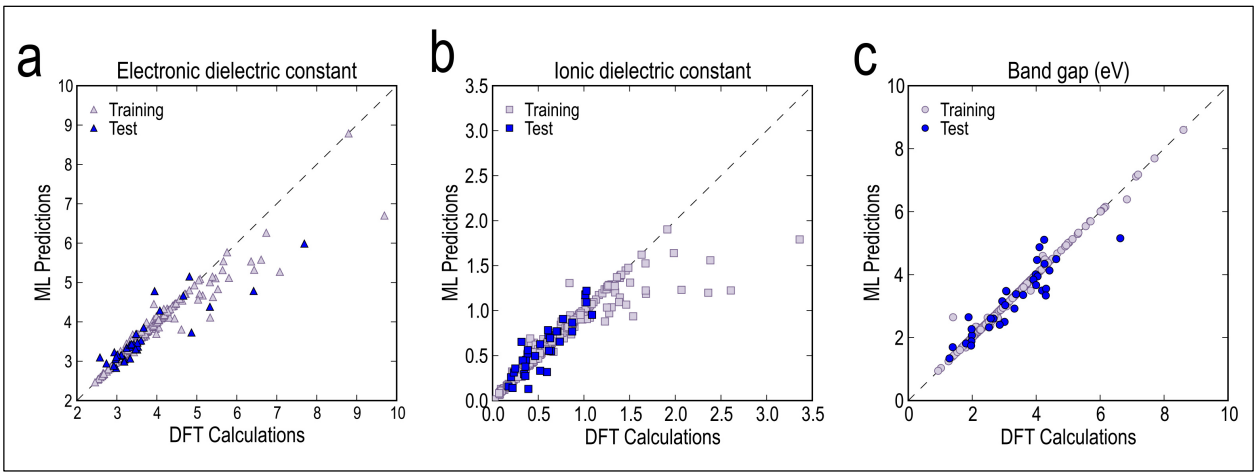


Figure 5.16. *Prediction performances of the SVR models with a Gaussian kernel.*

5.4.3 Adaboost

Boosting refers to the general problem of coming up with an accurate prediction algorithm by optimally combining different weak learners. Belonging to this family, AdaBoost, short for Adaptive Boosting, is a Godel Prize winning machine learning technique that has commonly been applied in conjunction with regular regression algorithms (such as SVR,

as we considered here) for improving their performances [156]. Boosting involves focusing on the points that have not been predicted well with SVR, that is, the difficult data points. If certain parameters could be modified to improve predictions on those points without affecting the predictions on all other points, we would have a better model than with regular SVR.

The AdaBoost algorithm is conceptually very simple. It is an iterative process where during each iteration, a new regressor is trained on the training set, with weights that are associated with each data point in the training set. These weights are modified at each iteration according to how successfully that data point has been predicted in the past iterations. The data points in the training set with larger prediction errors (i.e., those that are difficult to predict) are assigned larger weights. In practice, for a regressor such as SVR the boosting procedure involves training the model a number of times by changing the parameters σ and C as explained above, such that we will have different models with different accuracies of prediction on the poorly predicted points. There may be some models where predictions are better than the others, and these models deserve special attention. The overall prediction model is reported as a weighted median of all these models (as discussed below), with higher weights given to the specific models where predictions on the difficult data points show low errors. This means that the result of boosting is an optimal prediction model that is a weighted combination of all the different models.

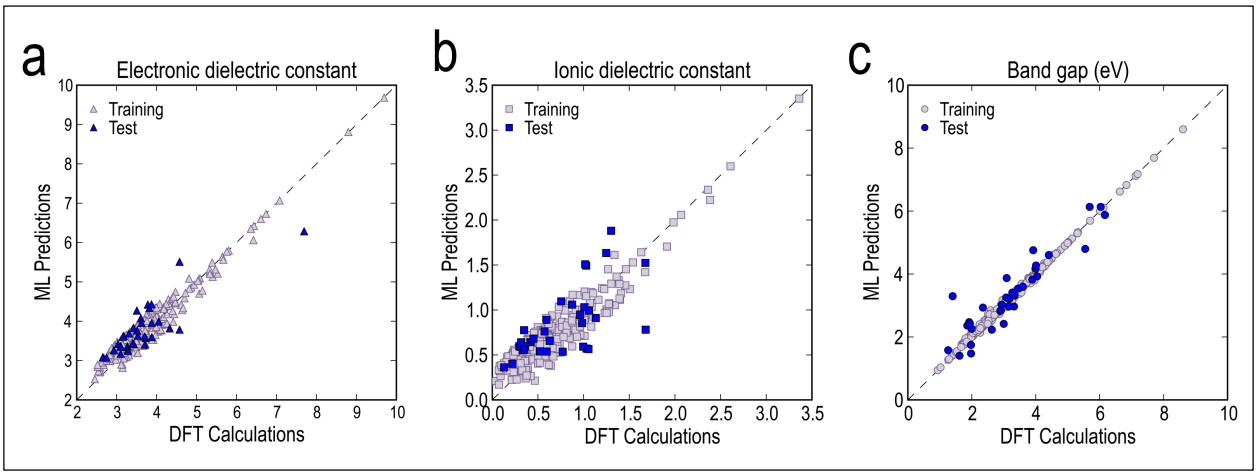


Figure 5.17. *Prediction performances of the SVR models with adaboost.*

Based on the algorithm as applied to our data, we obtain the cumulative predictions for every point, and **Figure 5.17** shows parity plots similar to **Figure 5.14**. The respective training and test set errors obtained here are again listed in **Table 5.2**. It can be seen that while the training performances (as compared with regular SVR) have definitely improved with Boosting for the three properties, the test performance is only slightly better for the electronic dielectric constant and worse for the ionic dielectric constant and the band gap. This means that while boosting can possibly improve upon regular SVR, there are some points that are quite poorly predicted with SVR, especially for the ionic dielectric constant. Further, the test errors with SVR + Adaboost are still higher than the errors with KRR using a Gaussian kernel.

The performances with SVR and AdaBoost can be explained as a consequence of the nature of the data we have. Whereas typical materials science data mining problems would include large amounts of data [133] [157], our dataset of 284 polymers and their

properties constitutes a `small dataset'. The regression performances with both regular SVR and AdaBoost could be improved for a larger, more diverse set of polymers, where the fraction of poorly predicted points could perhaps be minimized. As such, KRR with Gaussian kernel is the algorithm that performs better on average than these techniques, as captured in **Table 5.2**, thus bringing a measure of redemption to the practices followed by us and others using materials science data in the recent past.

5.4.4 Observations from this Study

In conclusion, we applied different kinds of machine learning treatments on a dataset of organic polymers and gained some insight on the appropriate choices for learning parameters. Given our objective to develop accurate, robust prediction models by mapping polymer fingerprints (the input) to the properties (the output) using regression, we explored a number of different kinds of regression strategies. Regression performed using different learning algorithms, different distance kernel definitions and different training set sizes revealed that Kernel Ridge Regression with a Gaussian kernel and a sufficient training set size resulted in the best prediction performances. While KRR with a Laplacian kernel performed almost as well, the polynomial kernel appeared to be unsuitable. Another major algorithm, Support Vector Regression, was also used and it was seen that SVR, even when used with AdaBoost, did not improve upon the best KRR performance.

The prediction accuracies are limited by the size and nature of the computational data, as well as by the quality of the fingerprints used. Since the polymer fingerprint only takes the population of different combinations of chemical building blocks into account, factors such as the conformation and the planarity of polymer chains---which could have varying effects on the properties---are ignored here. While KRR with a Gaussian kernel appears to be the best regression algorithm to use on the given data, performances can further be improved with larger datasets and an improved fingerprint that contains more information than currently used.

5.5. Uncertainty Quantification

We have developed various types of property prediction models for the dielectric constants and band gaps of organic polymers so far. However, something very crucial that is still missing is an estimate of 'uncertainty' in the predictions made using the KRR, SVR or any other model. The regression algorithms covered till now, while valuable, do not have ready remedies for quantifying the error bars in predicted values. A solution for this problem is imperative, as our statistical models are only around 90% accurate at best, and appropriate uncertainties attached to property predictions could affect the utility of the material for the application of interest.

The predictions are only as good as the polymer training data, which means if certain chemical components (such as blocks, block pairs or triplets, following from the fingerprinting definition) are not present or represented infrequently in the set of training points, predictions made on new polymers containing those components would need to be taken with a pinch of salt. On the other hand, polymers with components that are very well represented in training (for example, the CH_2 group) can safely be assumed to possess properties very close to what is predicted. Thus, we could come up with some measure of confidence in terms of error bars for any prediction that is made. Here, two different techniques to quantify uncertainties in regression predictions are discussed.

5.5.1. Gaussian Process Regression (GPR)

GPR is another widely used non-linear regression algorithm [158] along the lines of KRR and SVR. A Gaussian process can be thought of as defining a distribution over functions, leading to the predictions taking the form of a full predictive distribution. Based on the spread in predictions, it is possible to estimate the uncertainty associated with any value. The GPR algorithm is presented in **Figure 5.18**, and is similar to the KRR algorithm described in the previous section of this chapter, with the added advantage of getting error bars along with the prediction. The important parameters here are the coefficient σ_f, the Gaussian width σ_l and the noise σ_n. While the noise must be provided while training based on the known errors in data (for example, the chemical accuracy that can be

attained with DFT today is around 0.1%), σ_f and σ_l will be optimized via grid sampling in a similar fashion as with KRR.

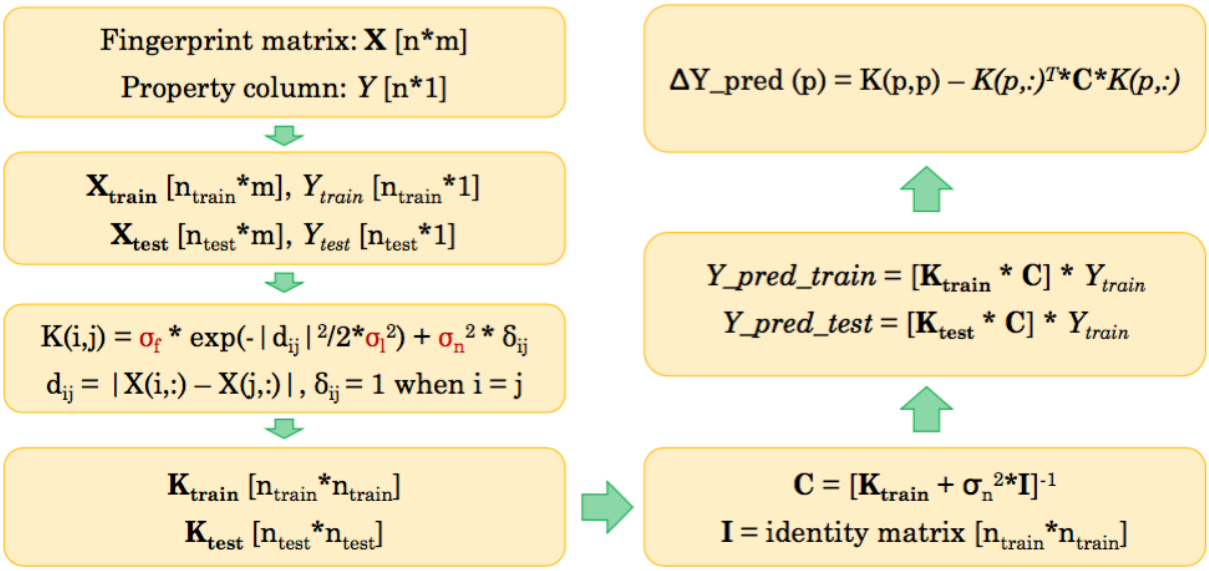


Figure 5.18. *The Gaussian Process Regression algorithm.*

The predicted property Y_pred and the uncertainty in the prediction ΔY_pred are expressed as a function of the Kernel matrix (each component of which corresponds to a (test point, training point) combination) and σ_n. For the same dataset and fingerprints described in the previous sections, GPR was used to train the prediction models shown in **Figure 5.19**, where error bars have been shown for each prediction. It is seen that these errors are greatly influenced by the noise parameter, which should be determined astutely to get meaningful uncertainties.

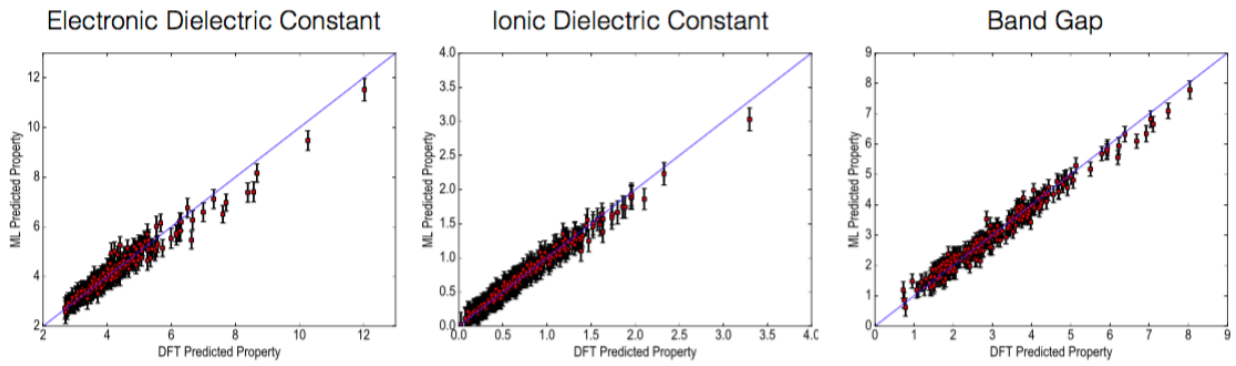


Figure 5.19. *Prediction performances using GPR, along with uncertainties for every prediction.*

5.5.2. Bootstrapping

Another useful way of obtaining uncertainties is by inducing disturbances in the training data and measuring the resulting deviations in property predictions; this method is known as Bootstrapping [25] [26]. The algorithm is presented in **Figure 5.20**, and KRR is regression algorithm of choice used here. If the best possible KRR model is trained with *N* unique points, it is kept aside and *N* points are sampled multiple times with repetition, such that there will be less than N unique points in every combination. Different models are trained every time and used to make predictions on all the points; the spread in the predicted values for each point provides a measure of how uncertain the overall model is about it. The eventual prediction can then be expressed as P ΔP, where P is the prediction from the best model and ΔP is the standard deviation from bootstrapping. Regression

models similar to **Figure 5.19** are shown in **Figure 5.21** with the uncertainties estimated using bootstrapping this time.

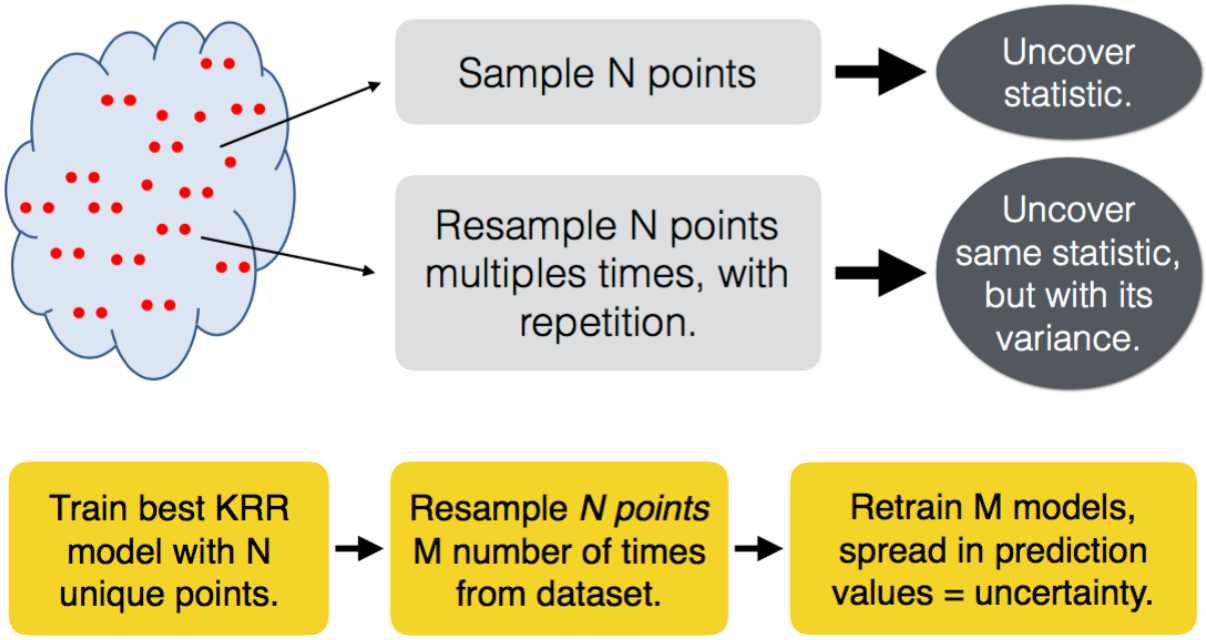


Figure 5.20. *Bootstrapping technique to induce disturbance into a data distribution and probe for the uncertainty in property estimation.*

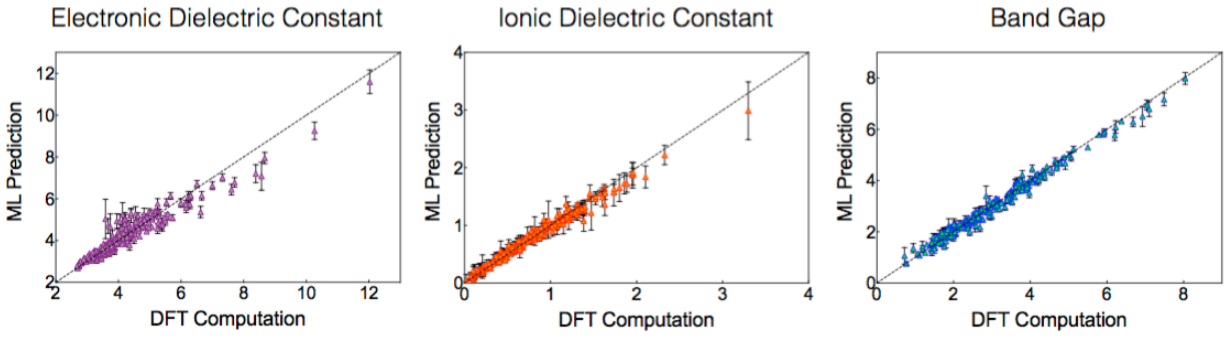


Figure 5.21. *Prediction performances and uncertainties using a combination of KRR and Bootstrapping.*

Chapter 6

DESIGN OF ADVANCED POLYMER DIELECTRICS: LEARNING FROM DATA

6.1 Introduction

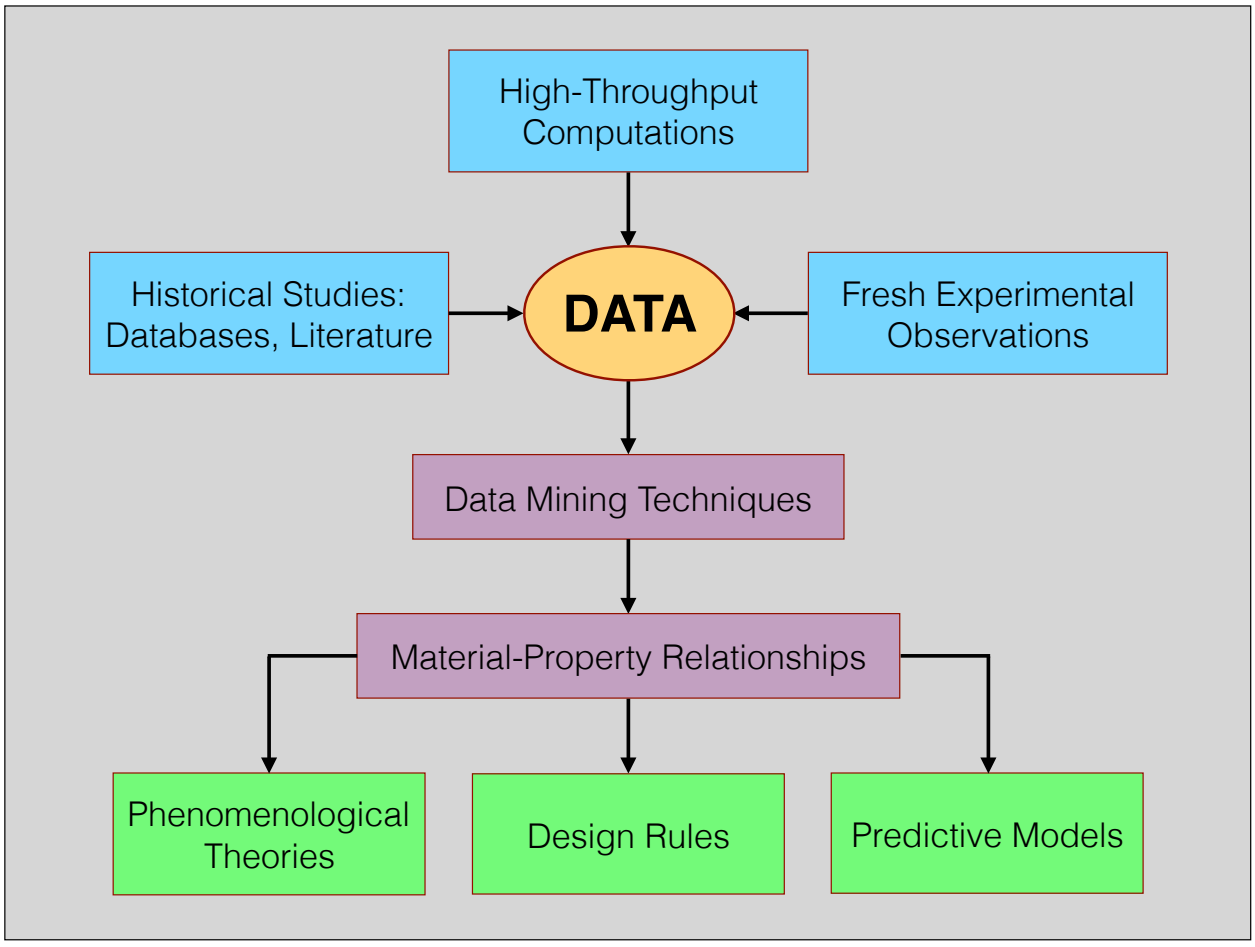


Figure 6.1. *A data-driven materials design philosophy.*

For a long time, empirical data has helped build chemical intuition and scientific insights, and supported the formulation of chemical and physical laws. Some classic examples cited in this regard are Hume-Rothery's set of rules for miscibility in a solid solution [4], and the Hall-Petch relationship between material strength and grain size [5]. Analysis of data procured from meticulous experimentation was key to developing these rules, showing that data-driven approaches have sometimes been a great ally of the materials scientist. Experimental data, while invaluable, could suffer from being time-intensive, non-uniform and possibly irreproducible; on the other hand, computational methods provide the means to generate data much faster at a uniform level of theory. Computational data provides a quick and efficient way of identifying promising candidates for applications of interest, as well as unearthing the role of chemistry, structure and other crucial material characteristics in determining the properties.

Figure 6.1 tries to capture the mechanism of converting materials data into learning in the form of laws, rules and models. The ideas presented here were implemented as a computation-guided, data-driven strategy for the rational design of new and advanced polymer dielectrics for capacitive energy storage applications, which has been the subject of **Chapters 3, 4** and **5** in this Thesis. The dielectric constant and band gap provided useful initial screening criteria for dielectrics for capacitive energy storage. Density functional theory (DFT) was used to compute the two properties for several organic and organometallic polymers, leading to the synthesis and characterization of several candidates that could potentially replace the current standard capacitor dielectric, biaxially

oriented polypropylene (BOPP). Machine learning techniques were also applied on the computational data to yield correlations between polymer building blocks and its properties, as well as to develop predictive models.

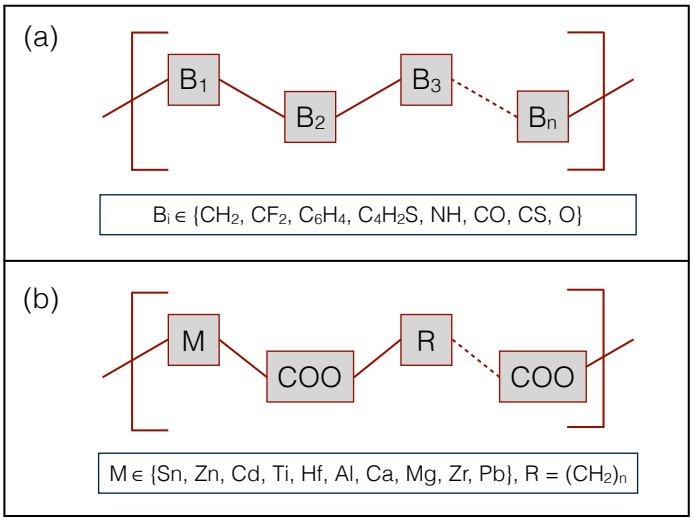


Figure 6.2. *The chemical space of (a) organic polymers and (b) organometallic polymers that constitute the computational dataset.*

As part of the search for new and advanced polymer dielectrics, a comprehensive first principles dataset of more than 1100 polymers and related materials was generated. Specifically, this dataset comprised of computed ground state crystal structures, electronic band gaps and dielectric constants for commonly known polymers, novel organic polymers, newly proposed metal containing polymers and several molecular crystals; a glimpse of these materials is provided in **Figure 6.2**, showing the chemical building units in organic and organometallic polymers respectively. While the organic polymers were built from linear combinations of a few selected chemical blocks, the organometallic polymers contain metal-ester units flanked by linker CH_2 units, with the metal atom chosen from a set of a few selected metals. This combined dataset of organic and organometallic polymers, described in **Chapter 3** and **Chapter 4** respectively, was supplemented by addition of molecular crystals containing the same atoms, providing more data to learn from.

We attempted to mine the substantial computational polymer dataset to obtain a critical understanding of how factors such as the chemical composition and coordination environment affect the dielectric constants and band gaps. This involves converting the materials in our dataset to unique representative fingerprints, and mapping the fingerprints to the properties. Fingerprinting of materials is typically performed using easily attainable physical and chemical characteristics such as composition, elemental properties, easily calculable properties, etc., in a way that is unique, general and reproducible. Mapping the fingerprint to the property helps reveal the correlation between any fingerprint component and the property in question. Further, regression algorithms, which are the staple of modern statistical learning approaches, can be applied to train models that yield properties of any material given its fingerprint.

Our results showed that while chemical bonds between 4-fold C atoms and H atoms or 2-fold O atoms enhance the band gap, bonds between 3-fold C atoms and S atoms increase the electronic component of the dielectric constant. However, perhaps the most important observation was that 6-fold metal atoms such as Sn, Zn and Cd bonded to electronegative atoms like O and F improve the ionic dielectric constant (and consequently, the total dielectric constant) significantly. In the following sections, the constituents of the dataset, the fingerprinting scheme and correlations between the fingerprints and the properties are explained in further detail. We are thus able to formulate some guidelines for property optimization in the present class of polymers for dielectric applications, the subject of discussion in [71].

6.2 Computational Data Visualization

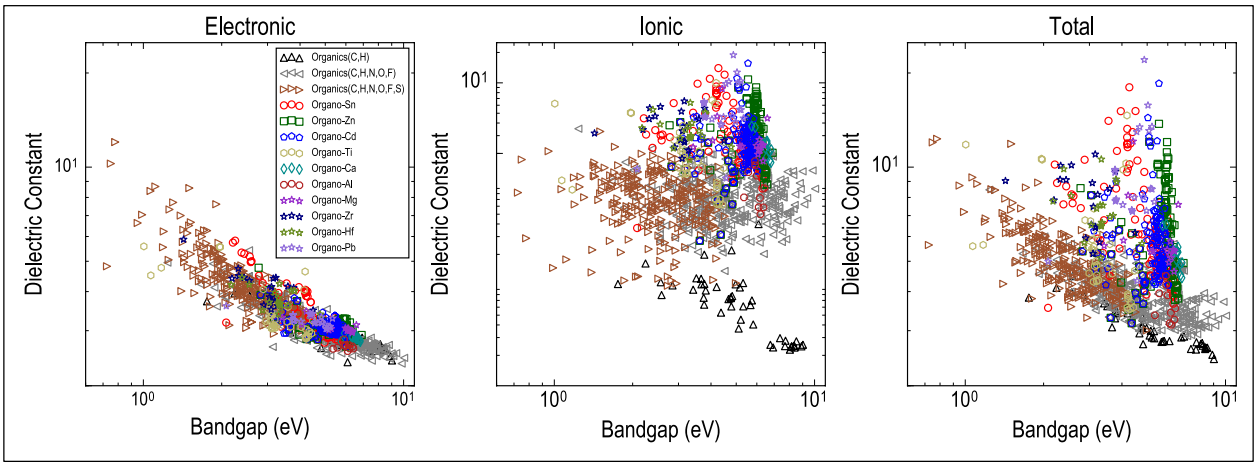


Figure 6.3. *The electronic, ionic and total dielectric constants plotted against the band gaps for the entire computational polymer dataset.*

Figure 6.3 shows all the computational data in the form of plots between the band gaps (E_{gap}) and the dielectric constants (electronic (ε_{elec}), ionic (ε_{ion}) and total (ε_{tot})). The dataset can broadly be divided into the organics (the purely organic polymers and molecular crystals, containing the following atoms: C, H, O, N, S and F) and the organometallics (the metal based polymers and metal-organic frameworks, each containing any one of the following atoms aside from the organic atoms: Sn, Zn, Cd, Pb, Mg, Ca, Al, Ti, Zr and Hf). We created further subdivisions in the data as shown in **Figure 6.3** to see correlations between the presence of specific atoms and the corresponding properties. The organics were divided into systems containing only C and H (Organics-1), systems containing C, H, O, N and F (Organics-2), and systems containing C, H, O,

N and S (Organics-3), and the organometallics were divided into ten subsets based on the identity of the constituent metal atom.

A visual examination of the plots in **Figure 6.3** reveals that while ε_{elec} correlates inversely with E_{gap}, ε_{ion} does not, and can thus lead to a ε_{tot} (= ε_{elec} + ε_{ion}) that fails to correlate with E_{gap}, especially for points possessing a high ionic. The organics, which span the entire expanse of the ε_{elec} spectrum but generally show very low ε_{ion}, have ε_{tot} values that inversely correlate with E_{gap}. All the organometallics, on the other hand, show a similar trend in ε_{elec} but clearly surpass the organics in ionic and thus total.

Atom	**Presence in Dataset**	**Polarizability (Cm^2/V)**	**Electronegativity (Pauling)**
C	All Organics, All Organometallics	11.0	2.55
H	All Organics, All Organometallics	4.5	2.1
O	Organics-2, Organics-3, All Organometallics	6.04	3.44
N	Organics-2, Organics-3, All Organometallics	7.43	3.04
F	Organics-2, Organo-Sn	3.76	3.98
S	Organics-3	19.6	2.58
Cl	Organo-Sn	14.7	3.16
Sn	Organo-Sn	52	1.96

Zn	Organo-Zn	39.2	1.65
Cd	Organo-Cd	46.3	1.69
Pb	Organo-Pb	46	2.33
Hf	Organo-Hf	109	1.3
Zr	Organo-Zr	121	1.33
Mg	Organo-Mg	71.7	1.31
Ca	Organo-Ca	160	1.0
Al	Organo-Al	56.3	1.61
Ti	Organo-Ti	99	1.54

Table 6.1. *All the constituent atoms across the entire dataset, the respective subsets that contain them, their polarizability and electronegativity.*

It has been shown in the past that ε_{elec} is determined by the atomic polarizabilities of the constituent atoms, whereas ε_{ion} depends on the dipoles in the system, the ease with which they can swing and stretch, and their characteristic vibrational modes [52] [58] [66]. **Table 6.1** lists all the different constituent atoms in our dataset along with their measured polarizabilities [159] and well documented electronegativities [160]. From **Figure 6.3**, it can be seen that while Organics-1 (containing only C and H atoms) show the lowest dielectric constants with nearly negligible ε_{ion} owing to the closeness of electronegativities of C and H, Organics-2 (C, H, O, N and F atoms) show higher ε_{ion} values due to the presence of dipoles formed by highly electronegative atoms like O, N and F bonding with

C and H. Organics-3 (S atoms included) show the highest ε_{elec} values among the organics because of the higher relative atomic polarizability of S as listed in **Table 6.1**, which also leads to their lower E_{gap}, especially when the concentration of S atoms is higher.

The addition of polar atoms like O or N and polarizable atoms like S clearly have a marked effect on the properties of systems containing C and H atoms. **Table 6.1** also shows that all the metal atoms are much more polarizable compared with the organic atoms, leading to a high ε_{elec} which is brought down by the presence of C and H atoms throughout the organometallics. The electropositive nature of the metal atoms and the high electronegativities of O and F lead to the presence of large dipoles in the organometallics. Aside from being polar, these bonds also display stretching and wagging vibrational modes that are soft in nature, leading to higher IR intensities at low frequencies [58] [63] [64] and thus, the highest ε_{ion} values. The organometallics contain no clear demarcations between the different subsets, with high ε_{ion} as well as medium to large E_{gap} shown by all. The actual concentration of the metal atom and the coordination environment it adopts play a crucial role here, as discussed later.

These observations provide an example of how we can extract patterns out of data simply by visual analysis; however, the problem today is that the rate of data generation far surpasses our intrinsic ability to process the data. Consequently, advanced machine learning and data mining techniques are needed. This involves, as stated earlier, converting materials to numerically representative fingerprints and developing models by

mapping them to the properties. Following our observations from the current dataset that certain kinds of atoms have positive or negative effects on the properties, a logical way to fingerprint the materials is in terms of the constituent atom types and the surrounding chemical environment. In subsequent sections, we explain this procedure and demonstrate the utility of fingerprinting with useful and meaningful models.

6.3 Fingerprinting

The fingerprinting scheme used here follows from past work on purely organic polymers and organic molecular crystals in the past [52] [77]; the former was applied and described in detail in **Chapter 5**. A chemo-structural fingerprint is required to take the contributions to the polarizability and dipole-dipole interactions from different chemical constituents into account. Thus, we used a fingerprint that encodes compositional and configurational information by quantifying the fraction of different types of atoms in the system and the different types of chemical bonds they form.

Atom types are defined by their chemical identities (such as C, O, Sn etc.) along with their coordination number (like 3-fold, 4-fold, 6-fold etc.). For instance, C4 refers to a C atom forming 4 different bonds with neighboring atoms, and Sn6 refers to an Sn atom forming 6 bonds, each based on previously known typical bond length ranges. It should be noted

that while characteristic bond lengths are well documented in the case of purely organic compounds (for instance, a C4-H1 bond length will be 1Å and a C4-C4 bond length will be 1.5 Å [77]), the distribution of possible bond lengths is more diverse for metal atoms in the organometallic systems. As an example, the Sn-O bond length (known experimentally to be 2 Å in tin oxide [114]) is seen to range from 2 Å to 3 Å in the organo-Sn systems in our dataset. Defining appropriate bond length minimum and maximum cut-offs is thus of utmost importance for obtaining meaningful fingerprints. The bond length cut-offs used for fingerprint definition in this work are shown in **Figure 6.4**.

Bond	Maximum Bond Length (Å)		Bond	Maximum Bond Length (Å)
C—C	1.7		Zn—C	2.2
C—H	1.3		Cd—O	2.5
C—O	1.6		Cd—C	2.2
C—F	1.6		Al—O	2.5
O—H	1.4		Al—C	2.2
Sn—C	2.3		Mg—O	2.5
Sn—O	2.6		Mg—C	2.2
Sn—F	3.0		Ca—O	2.5
Sn—Cl	3.0		Ca—C	2.2
C—N	2.0		Hf—O	2.5
N—H	1.4		Hf—C	2.2
N—O	2.0		Pb—O	3.0
C—S	2.0		Pb—C	2.2
Ti—O	2.5		Zr—O	3.0
Ti—C	2.2		Zr—C	2.2
Zn—O	2.5		Al—Cl	2.5

Figure 6.4. *Bond length cut-offs defined for fingerprinting purposes.*

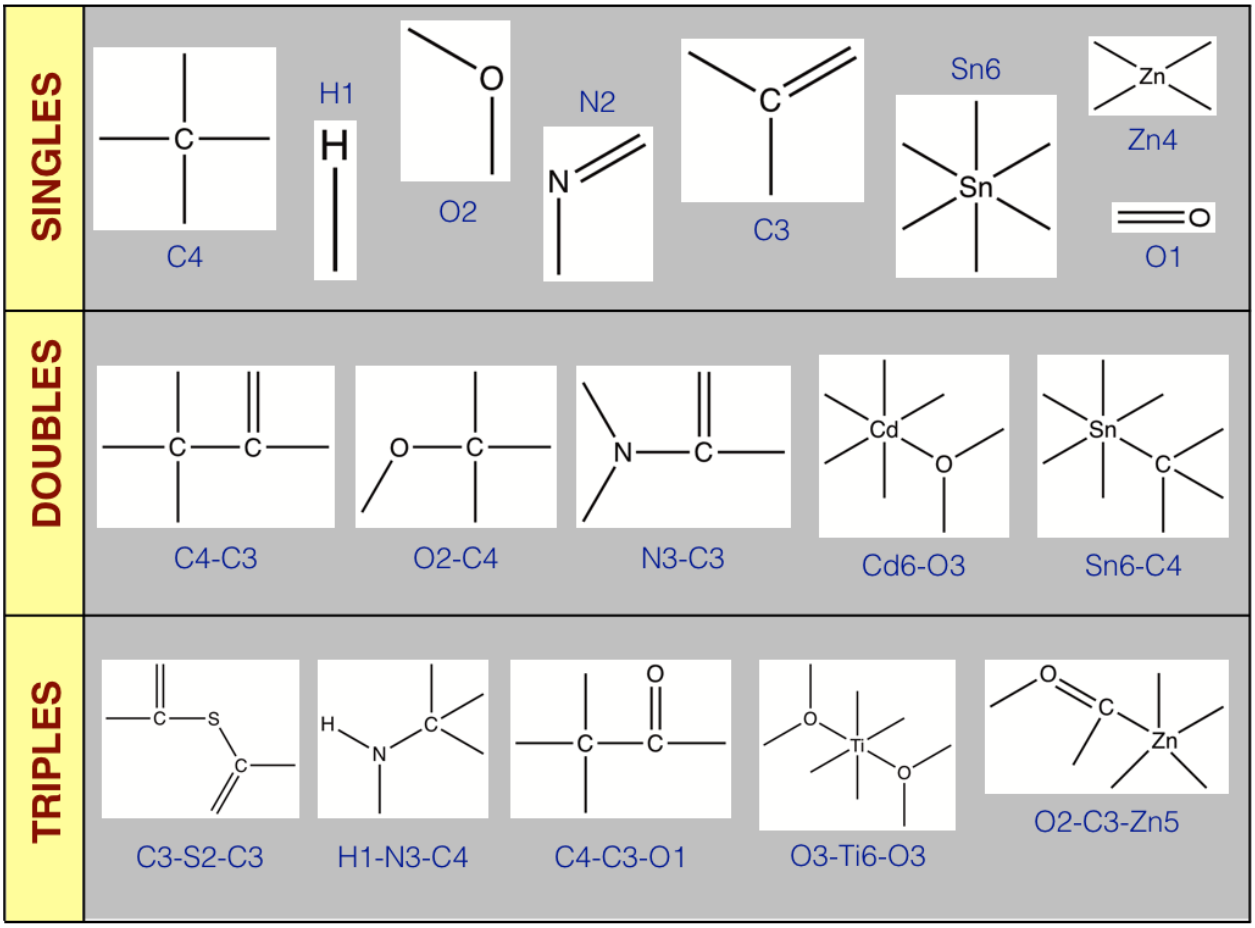


Figure 6.5. *Fingerprinting technique, showing examples of various types of singles, doubles and triples components found in our polymer dataset.*

In **Chapter 5**, a hierarchy of fingerprints was described, going from singles to doubles to triples: these refer to the complexity and dimensionality of the fingerprint. Here, we consider three kinds of fingerprints again with increasing amounts of information and increasing complexity:

1. When the count of each atom type (C4, O2, Sn4 etc.) is considered, the fingerprint is called ‘singles’. The fingerprint dimensionality is equal to the distinct types of

atoms, m, present across the dataset (m = 54 for our dataset). The singles help take into account the atomic polarizabilities in the system.

2. When the count of each pair of atom types (C4-C3, Sn6-O3, Zn4-O2 etc.) is considered, it is called 'doubles'. Fingerprint dimensionality would typically be m^2, but it is ~ 150 for our dataset once the components that are zero throughout (that is, the pairs that do not exist in any of the materials) are eliminated. Note here that the chemical bonds are being included along with the atoms, thus taking into account the (non-zero) dipole moments present in the system.

3. When the count of each triplet of atom types (C4-C3-H1, Sn6-O3-C3, Zn4-O2-C3 etc.) is considered, it is called 'triples'. Fingerprint dimensionality would typically be m^3, but it is ~ 500 for our dataset once the components that are zero throughout (that is, the triplets that do not exist in any of the materials) are eliminated. Note that here, we are taking atoms, chemical bonds as well as chemical conjugation into account.

Figure 6.5 shows a few examples of the different types of singles, doubles and triples components that exist in the materials that constitute our dataset. It should be noted that in each type of fingerprint, the count is normalized with respect to the total number of atoms in the system, which means every fingerprint component is a fraction. Periodicity

is accounted for, which means that a system if doubled or tripled in size would have the same fingerprint.

6.4 Fingerprint-Property Relationships

Figure 6.6 shows the linear correlation coefficients [147] between all the components of the singles and the three properties (in the form of bar charts), as well as between some selected components of the doubles and the properties (in the form of heat maps). The singles components correlating the most positively and the most negatively with the properties have been highlighted; the most relevant doubles components are shown in the middle covering up the singles that show little or no correlation. Any component of the heat map refers to a specific pair of atom types and the color shows the amount of positive or negative (or no) correlation.

As seen from **Figure 6.6 (a)**, the most positive correlations to ε_{elec} come from 3-fold C and 1- or 2-fold S atoms, while the negative correlations are provided by 4-fold C, H, and 2- and 3-fold O atoms. The correlations to E_{gap} follow exactly the opposite trend, as shown in **Figure 6.6 (c)**. These observations can be explained with the help of the atomic polarizabilities listed in Table 1 as well as the frequencies of occurrence of different types of atoms. While the metal atoms have by far the largest polarizabilities, they do not

contribute as much to ε_{elec} because their relative concentrations compared to C, H and O atoms are very small. Meanwhile, S atoms are present in comparatively higher concentrations in Organics-3, and their effect in increasing ε_{elec} and decreasing E_{gap} is considerable.

The bonds that contribute to large ε_{elec} are C3-C3, C3-S2, C3-S1 and C3-H1. The presence of S atoms in the systems in Organics-3 is in the form of thiophene (C_4H_2S) groups and thiol (CS) groups, both of which contain 3-fold C atoms singly bonded and doubly bonded respectively to S atoms, thus explaining the results in **Figure 6.6 (a)**. It can be seen from the heat map in **Figure 6.6 (c)** that the same chemical bonds decrease E_{gap}. C4-C4, C4-H1 and C4-O2 bonds, on the other hand, increase E_{gap} and decrease ε_{elec}. This is owing to all the data subsets other than Organics-3, in which C, H as well as O atoms exist in abundance and S atoms do not.

Figure 6.6 (b) shows that the largest positive correlations to ε_{ion} are dominated by high coordination number (CN) metal atoms like Sn6, Zn6, Pb6 and Cd6, and by atoms with the high electronegativity like O, F and Cl, as evident from **Table 6.1**. The heat map shows that it is indeed chemical bonds between high CN metals and O2, O3, F2 or Cl2 atoms that, owing to the dipole moments they introduce in the system, as touched upon earlier, contribute the highest to ε_{ion}. Bonds between C3 and O atoms as well as bonds between C4 and Sn6 atoms also show positive correlations, because of the high abundance of these bonds in organometallic systems.

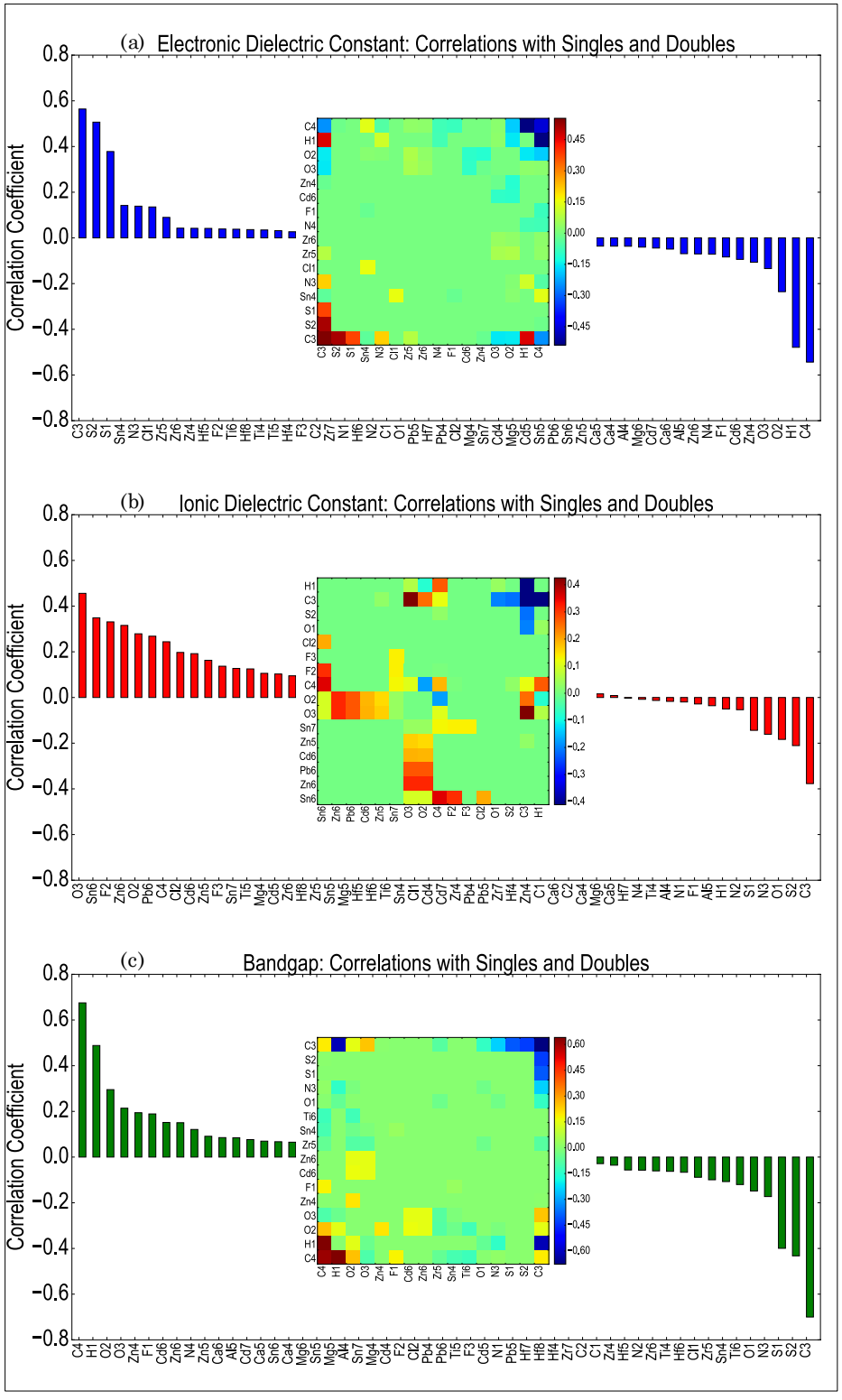


Figure 6.6. *Correlations between singles / doubles components and three properties: (a)* ε_{elec}*, (b)*ε_{ion}*, and (c)* E_{gap}*.*

Negative correlations with ε_{ion} are shown by C3, O1 and the S atoms, owing to relatively lower polarities in systems where 3-fold C atoms are singly bonded to H, doubly bonded to O and singly or doubly bonded to S. It is interesting to note that while O2 and O3 increase ε_{ion} greatly, O1 has the opposite effect; this is because of the abundance of 2- and 3-fold O atoms bonded to C or metal atoms in the organometallic systems (a glimpse of this can be had from the fragments shown in **Figure 6.5**), which show much higher ε_{ion} values than the Organics-2 where double bonds between C and O1 atoms are common. Moreover, the C3-O1 double bond stretching mode of vibration occurs at a much higher frequency than the C4/C3-O2/O3 single bond mode, and the former also shows a slightly lower dipole moment owing to a shorter bond length; this leads to the C4/C3-O2/O3 bonds contributing to higher ε_{ion}.

6.5 Guidelines for Property Optimization

The data analysis presented in the previous section can be utilized to engineer novel polymers that are likely to show desirable properties for capacitor applications, that is, high dielectric constants and large band gaps. It was seen that building systems with only C, H, O or N will lead to very high band gaps but low dielectric constants, despite the presence of dipoles like C-O and N-H that could potentially boost the ionic contribution. In our dataset, these systems (Organics-1 and Organics-2) consist of chemical blocks

such as CH_2, CO, NH and aromatic rings like C_6H_4. Upon addition of S-containing groups like C_4H_2S and CS (Organics-3), the electronic contribution to the dielectric constant sees a significant increase while the band gap drops dramatically. A middle ground exists in the pure organics with the right mix of chemical blocks leading to polymers with both high dielectric constants and large band gaps: this was pursued in the past [49] [50] to yield all new organic polymers (such as the polyurea -[NH-CO-NH-C_6H_4]$_n$- and the polythiourea -[NH-CS-NH-C_6H_4]$_n$-) that were synthesized, tested and shown to possess dielectric constants as high as 5 and 6 with band gaps always above 3 eV.

However, the most exciting insights obtained here pertain to the increase in the ionic dielectric constant caused by the low frequency stretching, wagging and other vibrational modes of the highly polar bonds in organometallic polymers. Metal atoms like Sn, Zn, Cd and Pb (introduced in the polymer backbone to approximately mimic the chemical environments they show in their well-known oxides or other compounds) are seen to adopt 4-, 5- or 6-fold coordination (occasionally even 7- and 8-fold), with the octahedral coordination the most preferred. This leads to the metal atoms forming 4 to 6 polar bonds with electronegative atoms like O, F and Cl, and even when the concentration of these bonds is diluted by the presence of organic linkers, the dielectric constant is seen to be enhanced in comparison to their purely organic counterparts.

From **Figure 6.3 (b)**, we saw that although the organometallic systems in the dataset display the highest ionic, they cover a wide range of values. This comes from the varying

concentrations of metal atom in the system, an effect we have explored in previous studies [67] [116]. For instance, it was shown that a series of Sn-polyesters with a varying number of linker methylene (CH_2) groups showed, in general, a decrease in ionic (and consequently, total) as the length of the linker chain increased and metal concentration went down. The same trend was observed for organometallics containing any of the other metals. This allowed us to say that, generally, higher the metal content, higher is the dielectric constant of the polymer. In the limiting case that the system contains the maximum possible amount of metal, the metal based compound (oxide or fluoride, for instance) is reproduced and dielectric constant is likely to be the maximum. However, a certain number of organic linkers, whether methylene(-CH_2-), carbonyl (-CO-), phenyl (-C_6H_4-) or thiophene (-C_4H_2S-) groups, is necessary in an organometallic polymer to ensure easy synthesis and processability into free-standing thin films for dielectric applications.

The fingerprint-property correlations presented in the previous section show that when it comes to the dielectric constant, the metal coordination number and the identity of its surrounding atoms are just as important as the amount of metal present in the system. To probe the effect of various factors on the properties of organometallic polymers, we plotted the dielectric constant as a function of the identity of metal atom in **Figure 6.7**. While different colors correspond to different metal coordination numbers (which varies from 4 to 8), the size of any data point correlates with the volume fraction of the metal in the polymer. The latter quantity is the fraction of the (previously documented) metal

covalent volume [161] to the total crystalline volume of the polymer, as estimated from its computationally obtained crystal structure. The metal volume fractions range from the lower limit of ~ 2 % to a high of ~ 20 %.

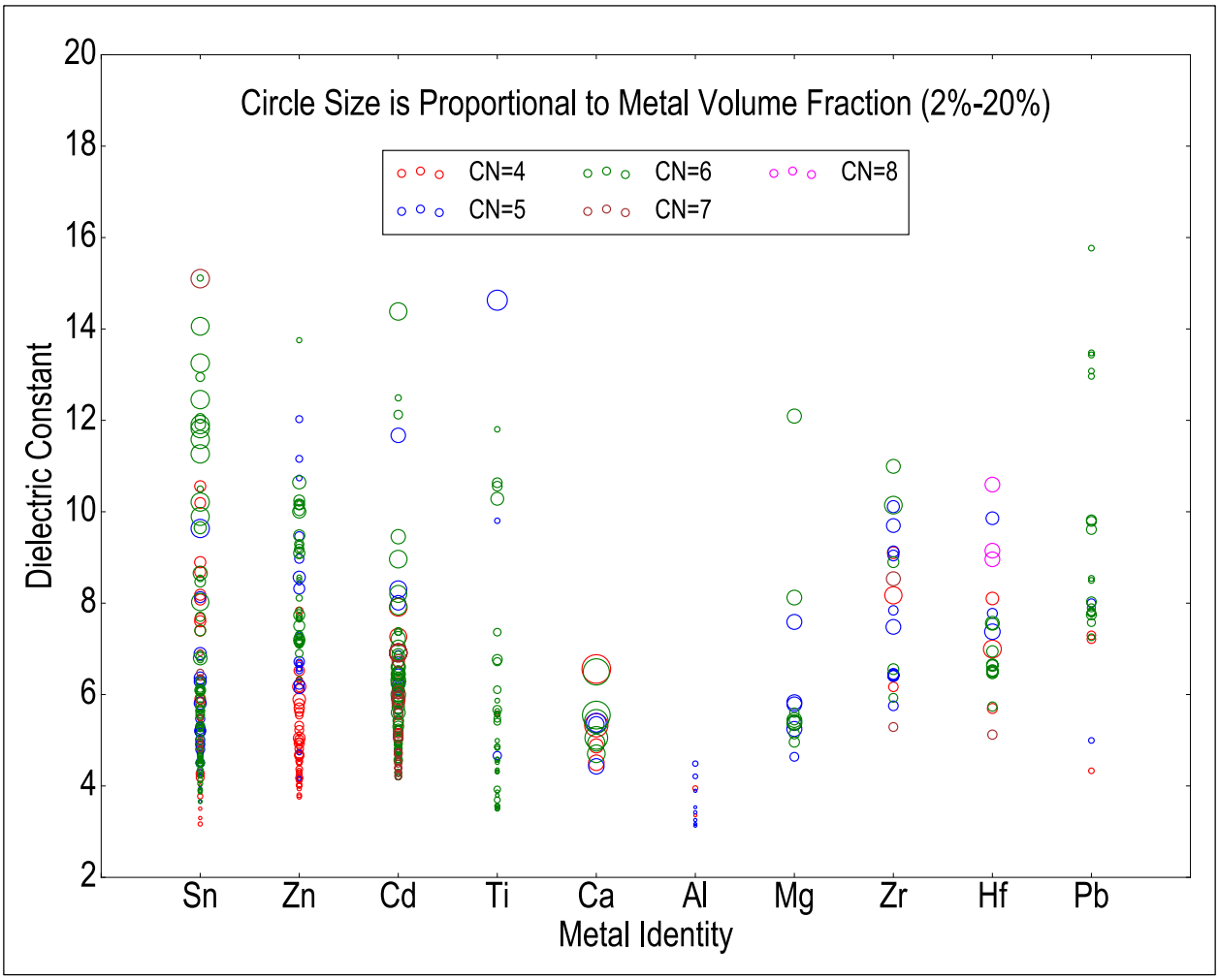


Figure 6.7. *Dependence of dielectric constant on metal identity, volume fraction and coordination number.*

It can be seen from **Figure 6.7** that circle sizes do not always go up upon increasing the dielectric constant, implying that a direct relationship does not exist between the property and the metal content. In general, the highest dielectric constants are shown by systems where the metal exists in a 6-fold coordination. Further, it is observed that certain metals like Zn, Cd, Ti and Pb exist in cases that display a dielectric constant > 10, but with very little metal volume fractions of 2 % - 5 %. This is an important insight, as it implies that it is possible to boost the dielectric constant of organometallic polymers without the need to insert a very large amount of metal. As discussed earlier, organic portions help improve the polymer film processability, and as seen here, high dielectric constants can be achieved for the same, theoretically. Motivated by these ideas, blends and copolymers of organometallic and organic polymers are currently being synthetically pursued, with the hopes of achieving polymer(s) with good processability and significantly improved dielectric behavior.

All the important results obtained here can be boiled down to a number of useful design rules as listed in **Table 6.2**. This includes all the atom types and bonds that have a serious positive or negative impact on the dielectric constant and band gap.

Atom	Effect on Dielectric Constant	Effect on Band Gap
C4	Decreases ε_{elec}, increases ε_{ionic} via bonds with high CN metals.	Largest possible increase in E_{gap}.
C3	Increases ε_{elec} via bonds with S, increases ε_{ionic} via bonds with O.	Decreases E_{gap} via bonds with S.
H1	Effect only via bonding with C/O.	Large increase in E_{gap} when bonded to C.
O2	Decreases ε_{elec}, increases ε_{ionic} via bonds with high CN metals and C.	Increases E_{gap} when bonded to C.
O3	Increases ε_{ion} via bonding with high CN metals and C.	Increases E_{gap}.
F2	Increases ε_{ion} via bonding with high CN Sn.	Increases E_{gap}.
S1	Increases ε_{elec} via bonding with C.	Decreases E_{gap} via bonding with C.
S2	Increases ε_{elec} via bonding with C.	Decreases E_{gap} via bonding with C.
Sn6	Increases ε_{ion} via bonding with O/F.	Slightly increases E_{gap}.
Zn6	Increases ε_{ion} via bonding with O.	Slightly increases E_{gap}.
Pb6	Increases ε_{ion} via bonding with O.	Little or no effect.
Cd6	Increases ε_{ion} via bonding with O.	Slightly increases E_{gap}.

Table 6.2. *A list of the atom types in the database that effect the dielectric constants and band gaps the most. Appropriate combinations of atoms can be chosen in the system to increase or decrease one or both properties.*

6.6 Property Prediction Models Using Regression

While the qualitative understanding obtained from the computational data is very valuable, a quantitative predictive model is missing. Ideally, one would wish to reduce the desired property to an explicit function of the fingerprint, meaning the property of any new material can be instantly estimated by fingerprinting it. However, very often in materials science, relationships exist but cannot be written down in the form of straightforward equations. Using a regression algorithm is a means to train a model that converts the fingerprint input to the property output with a statistical accuracy, based on the available data. In this work, we use Kernel Ridge Regression (KRR) [152] to develop prediction models for the dielectric constants and band gaps of polymers.

KRR is a method that has been widely used in materials science problems to obtain nonlinear regression models. It involves transforming the materials from the input (fingerprint) space to Kernel space and defining the output (property) as a function of the Kernel, a covariance matrix and the input set of property values. In practice, the input set (also referred to as the training set) is taken to be a subset of the entire dataset and the model is trained with these points, following which testing is done on the remaining points. The best prediction models for three properties—ε_{elec}, ε_{ion} and E_{gap}—are reported in the form of parity plots.

To apply KRR on our data, we selected the triples as the fingerprint to serve as input to the model. This comes from past work, explained in detail in **Chapter 5**, showing that the level of complexity provided by the triples (which contains information about composition, bonding environment and chemical conjugation) is necessary to train predictive models with high accuracy and generalizability. It was observed that when the entire dataset of organic and organometallic polymers was considered for model development, poor performances were obtained because of the lack of sufficient data when it comes to systems containing the metal atoms. Thus, the only models presented here were trained on a restricted dataset of purely organic polymers, i.e., systems containing the following set of atoms: C, H, O, N, F and S. This is an improvement over our previous attempt at training predictive models for organic polymers that is documented in **Chapter 5**, wherein a lesser dataset was used along a fingerprint based on chemical building blocks (such as CH2, C6H4, etc.) as opposed to the atom types.

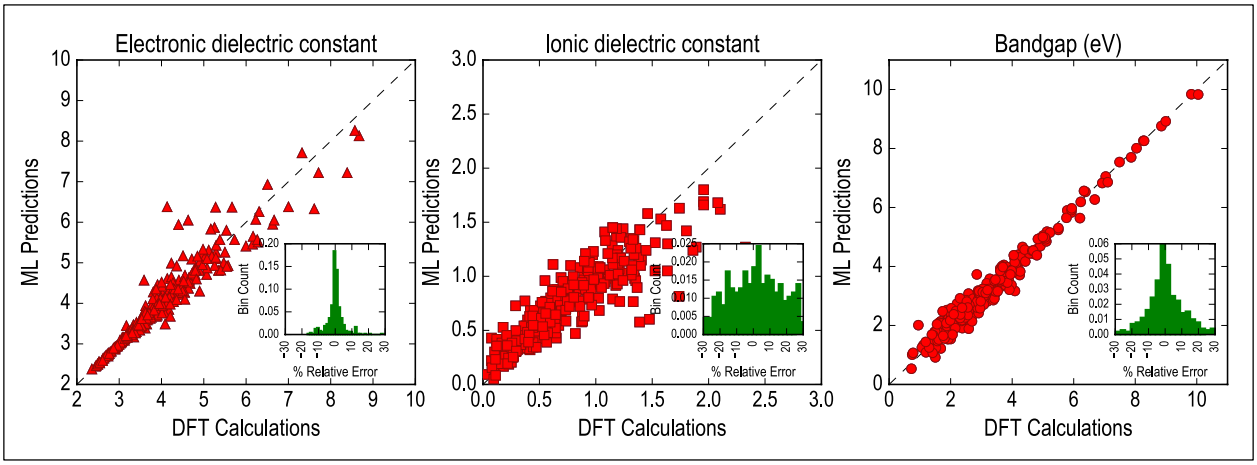


Figure 6.8. *KRR parity plots for the 3 properties.*

The best prediction models thus obtained are shown in **Figure 6.8**. The prediction performances are the best for E_{gap}, followed by ε_{elec} and ε_{ion}, which follows from the observations from **Figure 6.6** and the discussion in the previous section: the magnitude of positive and negative correlations with the fingerprints are highest for E_{gap}, followed by ε_{elec} and ε_{ion}. This means our fingerprint space is best representative of the trends in band gaps and not as good for the dielectric constants. A consequence is the failure of the model in certain cases, especially for higher values of ε_{elec} and ε_{ion}, where the data is scarce. The insets in the plots in **Figure 6.8** show the distribution of relative errors across the dataset, which is normal for ε_{elec} and E_{gap}, but much higher errors are seen for ε_{ion}.

The poor predictions for the dielectric constants could be attributed to an incompleteness in the fingerprint as much as it is down to the lack of sufficient data. With its atomic, bond and conjugation information, the triples fingerprint captures almost all the necessary details required for E_{gap} and thus leads to a well correlated model. However, other crucial factors such as the crystal structure and the spatial conformation of polymer chains are known to affect the dielectric constant, especially ε_{ion}, which could explain partly the inadequacy of the current models. Model performance for ε_{elec} is good for values < 8; only a handful of organic polymers with high S content show exceptionally high ε_{elec} between 8 and 12 (an example being polythiophene, $-[C_4H_2S]_n-$), and the scarcity of points in this property regime leads to the poorly predicted systems for values > 8. With infusion of newer data in the high dielectric constant regions, predictions performances can be

improved. Further, an increase in the population of each type of organometallic polymer can help extend the KRR models to systems containing metal atoms as well.

6.7 Conclusions: Learning from Data

It was demonstrated here that a large dataset of materials belonging to related classes can be mined to extract trends, design rules and prediction models. While the quantitative models trained using regression currently have limited capabilities, the qualitative insights obtained from the computational polymer dataset are undeniably valuable. Fingerprinting in terms of atom types, chemical bonds and chemical conjugation in the system enabled us to take almost all the contributing factors into effect. By understanding the role of specific atoms and their coordination environments in relation to the dielectric constants and band gaps, we can determine the chemical build-up of polymers to optimize them for dielectric applications. **Table 6.2** gives a glimpse of the essential rules learned in this work, which can only be reinforced and improved with the addition of more data. The next computation and/or experiment on a potentially promising polymer will be guided by **Table 6.2**; a surprising result would indicate a tweak in our learning models, whereas the expected result would spark a new discovery.

Some of the most important observations made here can be summarized as follows:

- Organic polymers containing a majority of 4-fold C atoms bonded to H atoms show very large band gaps but low electronic dielectric constants and almost negligible ionic dielectric constants.
- Organic polymers containing S atoms bonded to 3-fold C atoms show very high electronic dielectric constants and low band gaps.
- In organic polymers, O and N atoms lead to slightly high ionic dielectric constants via bonds with C while maintaining large band gaps.
- Organometallic polymers far outperform pure organics in terms of simultaneously enhancing the dielectric constant and the band gap. The metal atom, the metal concentration and its coordination chemistry are the important factors that control the polymer's dielectric constant.
- 6-fold Sn, Zn, Cd and Pb atoms bonded with O or F lead to the highest ionic dielectric constants observed in this work, with band gaps in the moderate to high regions. For many of these polymers, modest metal fractions (by volume) of 2 % - 5 % were sufficient to obtain high dielectric constants.

The results presented in **Figure 6.6** and **Table 6.2** reveal precisely the atoms and bonds that are required to manipulate any property. In practice, choosing the right mix of chemical constituents to optimize different properties simultaneously is no trivial task, and is further complicated by the issues of experimental viability, processability and stability. However, this task is simplified to some extent, and made more rational because of all

the insights obtained here. The prediction models shown in **Figure 6.8** have the potential to be extremely valuable in guiding quick, targeted experiments: one simply needs to input any new, hypothetical, possibly unrealistic polymer repeat unit, and the model would return its properties instantaneously. We demonstrated the power of this with highly accurate predictive models previously developed for purely organic polymers and with the development of online polymer design tools. With higher amounts of data (that is being generated on a regular basis by the community) and perhaps an improved fingerprint, the parity plots in **Figure 6.8** can be made better and the on-demand design of new dielectric polymers can be extended far beyond the pure organics.

Chapter 7

THE POLYMER GENOME PROJECT

7.1 A Computational Database of Polymers

The importance of *data* in driving discovery and innovation puts the onus on scientists to catalogue their computational and experimental results, and whatever insights they may have gained from it, for the benefit of the entire scientific community. This aligns well with the goals and objectives of the *Materials Genome Initiative* [23], and efforts towards the same are evidenced by the rise of many materials databases over the last few years [79] [80] [81] [82] [83]. All the polymer data (including computationally obtained ground state structures, and the DFT estimated and experimentally measured properties) and machine learning models presented in this article may be found within the *Polymer Genome* platform in *Khazana* [84], an online materials knowledgebase. Any user searching for a polymer by its repeat unit, chemical name or desired properties will be able to access the relevant experimental or computational data, as well as ML predicted properties, and can utilize this information to make an instant go/no-go decision on whether to pursue it for applications of interest. Fingerprinting a polymer in terms of its basic building block (the chemical unit or atom) is like tracking the polymer "genetic material" or "gene", which is

then utilized for explaining trends in the properties; hence the terminology “the polymer genome”. This knowledgebase is an attempt to unravel the polymer genome, and through the medium of past data and machine learning tools, provide ready access to information in meaningful spaces of the polymer chemical universe.

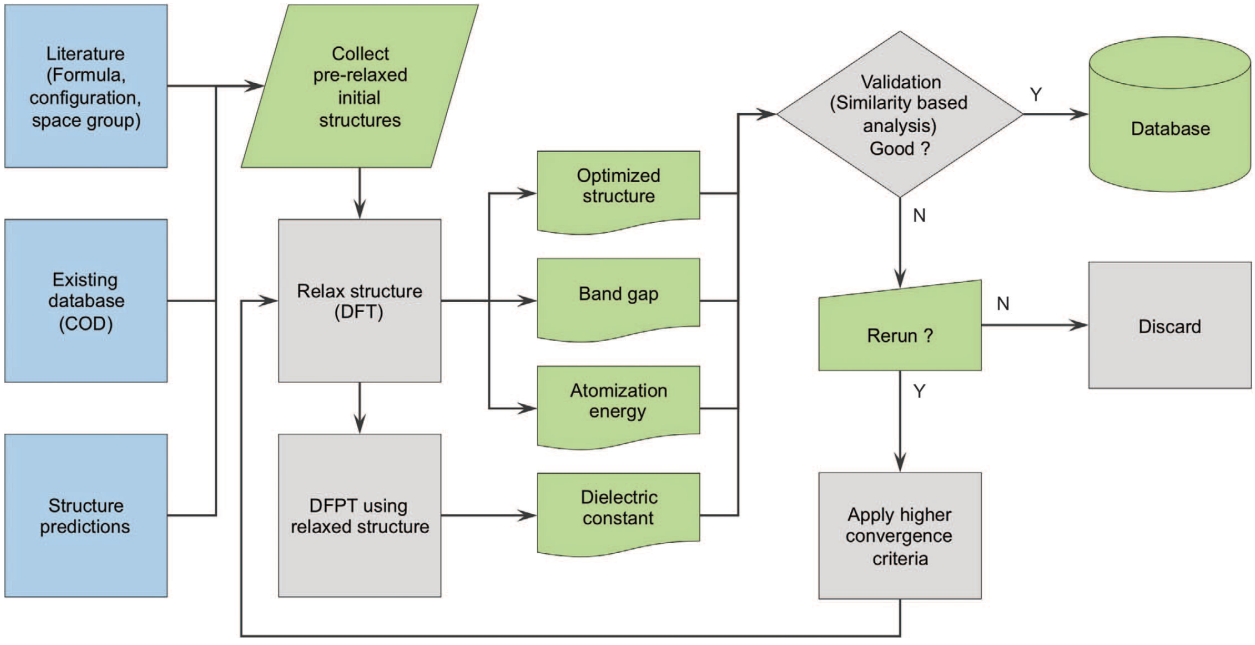


Figure 7.1. *Scheme for preparing a database of polymers and related materials* [70].

The organic and organometallic polymers discussed and presented in **Chapter 3** and **Chapter 4** respectively together constitute a polymer dataset of nearly a thousand systems. We fortified this dataset further with several known polymers obtained from the experimental literature (like Polyoxymethylene, PVDF, Polyacrylonitrile, etc.), as well as a few hundred molecular crystals collected from Crystallography Open Database (COD) [162], containing such atoms as C, H, O and Sn. Band gap and dielectric constant

calculations were performed on all these systems within the DFT formalism as before. This process helped diversify the types of chemical environments present in the database and thus enhance *learning* (leading to all the results presented in **Chapters 5** and **6**). The workflow shown in **Figure 7.1** summarizes the preparation of this database of polymers and related materials. While the necessary computational details have been discussed in **Chapters 2**, **3** and **4**, a workflow such as this is essential for ensuring the uniformity and reliability of data. The composition of the database in terms of different material classes (organic crystals, novel polymers, etc.) is shown in **Figure 7.2**.

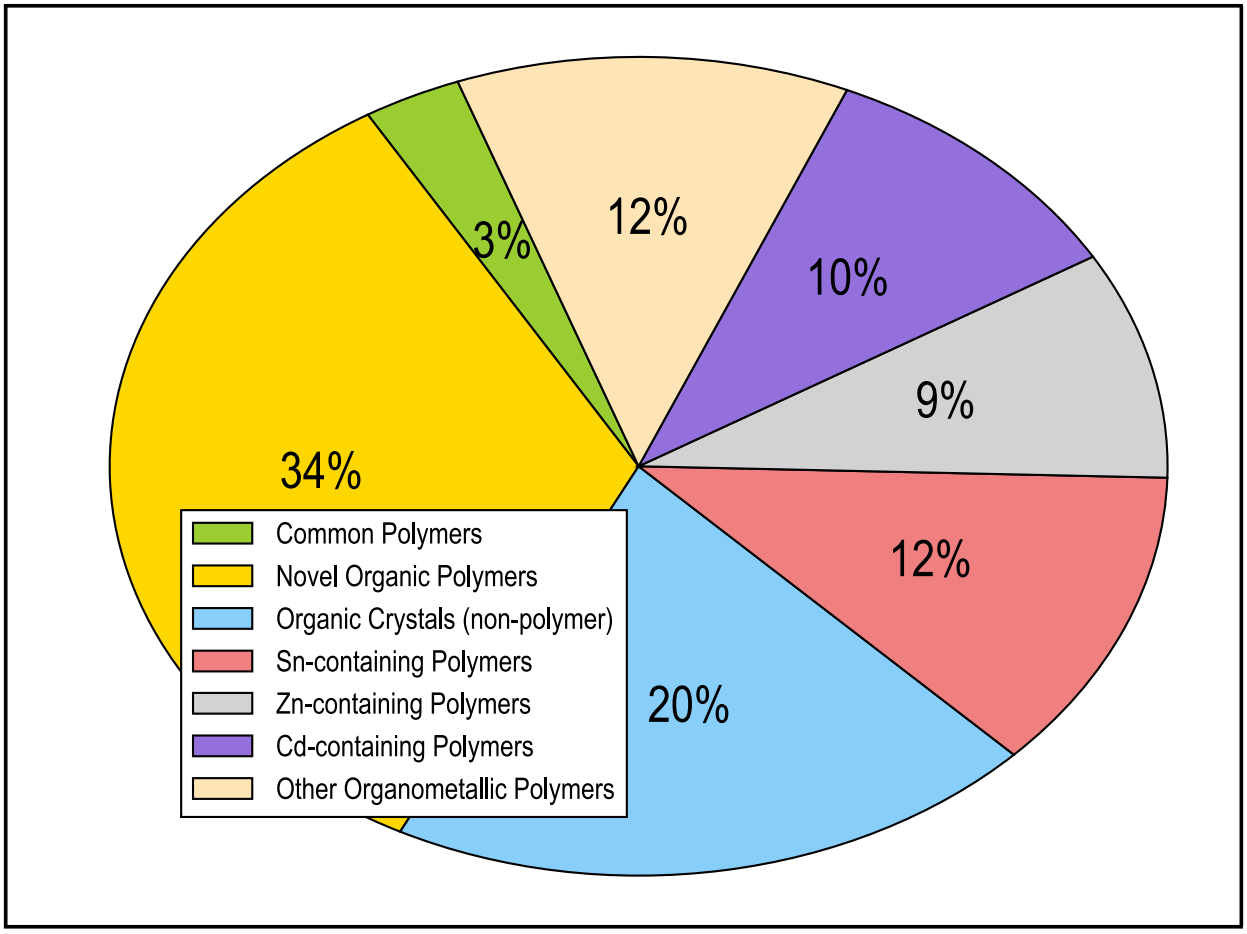


Figure 7.2. *The compositions of different types of materials present in the database [71].*

For materials that required structure prediction, namely the organic and organometallic polymers from **Chapter 3** and **4**, the calculations were subjected to a preliminary filter for removing any obvious redundancy of identical structures. Then, the selected structures were optimized by DFT calculations, yielding the equilibrium structures and their atomization energies. E_{gap} was then calculated on a dense grid of k-points while ε_{tot} (composed of ε_{elec} and ε_{ion}), was computed within the framework of density functional perturbation theory (DFPT).

In the next step, the computational scheme and the calculated results were validated with available measured data, including the measured band gap, the dielectric constant and/or the infrared spectroscopy (IR) or X-ray diffraction (XRD) measurements. In both **Chapter 3** and **Chapter 4**, for several newly synthesized polythioureas, polyureas, polyimides and Sn-polyesters, the measured properties, and IR and XRD patterns were seen to match up very well with the DFT results. However, for those cases which did not agree with the available experimental data, the materials were subjected to further calculations at tighter convergence criteria of residual atomic force (compared to the calculation details discussed in **Chapter 2**), and if better agreement was not achieved, these points were removed from the dataset. A post-filtering step was finally performed on the whole dataset, keeping only distinct data points. Thus, it was ensured that all the computational data that went into the *Polymer Genome* platform was well-converged and meaningful.

7.2 Computation of Relevant Properties

As part of this Thesis, several polymers have been computationally studied for capacitive energy storage, which mainly involved the calculation of two properties, the band gap and the dielectric constant. The successes so far could be broadly divided into the following categories: (a) computational polymer database of structural, electronic and dielectric properties, (b) rational co-design of novel polymer dielectrics, and (c) accelerated design and discovery using machine learning. What this means is that similar strategies when applied for other properties appropriate for various applications can lead to similar successes. For instance, even for capacitor dielectrics, properties such as the dielectric loss and the breakdown field are important, but are not amenable to first principles computational studies with present methodologies. However, the very same polymers designed here could have applicability for various other applications.

One such application is in organic semiconductors, which are built from thin-film organic polymers containing atoms such as C, H, O, N and S. These polymers typically require an electron affinity between 2 eV and 4 eV and an ionization energy between 4.5 eV and 6.5 eV, aside from a semiconductor appropriate band gap [163] [164]. Another example is of organic photovoltaic cells, which contain organic polymer layers between metallic conductors. The properties are determined by whether the polymer consists of electron donating or accepting groups, which in turn, depends on the ionization energy and the electron affinity. Yet another application where the band gap, dielectric constant,

ionization energy and mechanical properties like the bulk and shear modulus are important is optoelectronics.

Property	Computation Using DFT
Structural Parameters (Lattice Constants, Bond Lengths)	Optimization of crystalline structure computed using MHM, or obtained from the literature.
Density (g/cm^3)	$\frac{\text{Molecular Weight of Repeat Unit}}{\text{Volume of Optimized Structure}}$
Atomization Energy (eV)	Energy of optimized structure relative to the energies of stable elemental forms of the constituent atoms.
PBE Band Gap (eV)	Band gap computed at the PBE level of theory.
HSE Band Gap (eV)	Band gap computed using the hybrid HSE06 functional.
Dielectric Constant	Electronic component ε_{elec} and ionic component ε_{ion} computed using density functional perturbation theory (DFPT).
Refractive Index	Calculated as the square root of ε_{elec}.
Ionization Energy (eV)	Negative of the HOMO energy level of the optimized polymer single chain structure.
Electron Affinity (eV)	Negative of the LUMO energy level of the optimized polymer single chain structure.
Cohesive Energy (eV)	Energy of optimized polymer crystal structure relative to the energy of the optimized polymer single chain structure.

Table 7.1. *Various relevant properties calculated for materials in the database using DFT.*

Several applications and properties can be identified in this fashion, given that DFT can be applied to compute them. Keeping this in mind, we established a list of relevant properties, which consist of those that have been computed so far and a few newer properties. **Table 7.1** presents these properties and shows how to compute each of them using DFT. The computational database discussed in **Section 7.1** contains the optimized, ground state, polymer crystalline structures, accounting for structural parameters such as lattice constants, bond lengths as well as space groups. The optimized structure further provides the crystalline density of the polymer, based on a ratio of the polymer molecular weight (given that the crystalline structure consists of two polymer chains, this would be twice the molecular weight of the polymer repeat unit) and the volume of its unit cell. The atomization energy for every material can be computed based on the difference between the DFT energy of the optimized structure and the DFT energies of the elemental states of its constituent atoms. The band gaps at the PBE level and the HSE level of theories have both been computed. DFPT was used compute ε_{elec} and ε_{ion}, the sum of which is the total dielectric constant. The refractive index is defined as the square root of ε_{elec}, and thus can be computed easily.

For every polymer, the isolated single chain has been optimized using DFT, which yields not only the local polymer chain geometry but also accurate HOMO and LUMO energy levels (at the HSE level of theory as opposed to the PBE level). The ionization energy and electron affinity of the polymer is defined as the difference between the vacuum energy level and the HOMO and LUMO levels, respectively [165] [166] [167] [168] [169].

While inorganic semiconductors contain charges trapped in dangling surface-gap states, an organic polymer semiconductor surface would contain flat bands as they do not exhibit dangling bonds. This makes a calculation of the molecular orbital energy levels of single chain polymers appropriate and representative of the polymer. The difference between the DFT energies of the polymer crystalline form and the single chain form also yields the cohesive energy, which is a measure of how likely the polymer is to exist in a 3D structural arrangement.

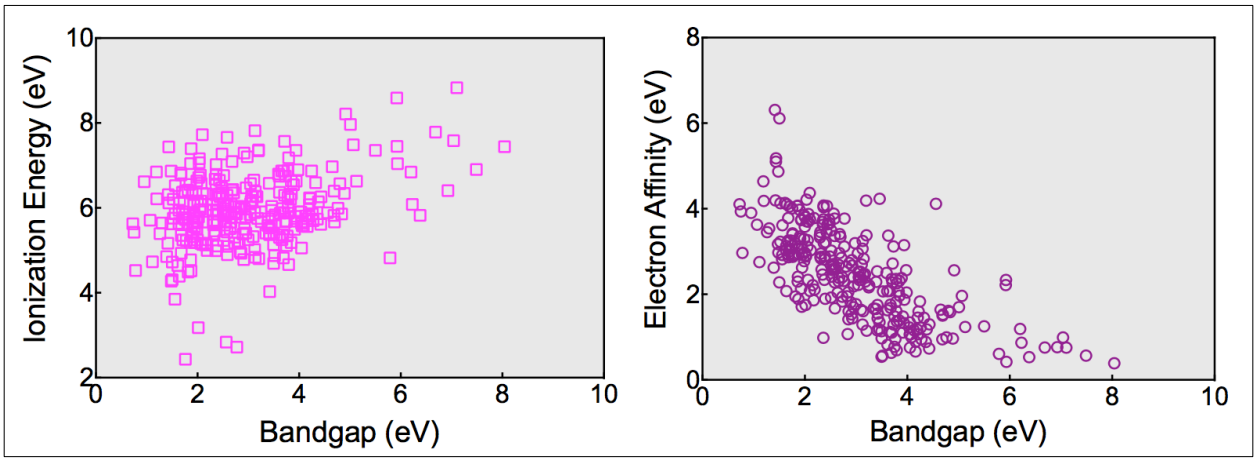


Figure 7.3. *The ionization energies and electron affinities computed for all the organic polymers in the database, plotted as a function of the band gap.*

All the properties listed in **Table 7.1** were computed for the dataset of organic polymers presented in **Section 7.1**, with calculations on organometallic polymers currently underway. **Figure 7.3** shows the ionization energy and electron affinity plotted against the band gap. The computed electron affinity of polyethylene is around 8 eV, which is in good agreement with reported experiments [170]. Given these new properties, regression

models can be trained like before, and the scope of applicability of machine learning within currently studied chemical spaces can be broadened.

7.3 The Polymer Genome Platform

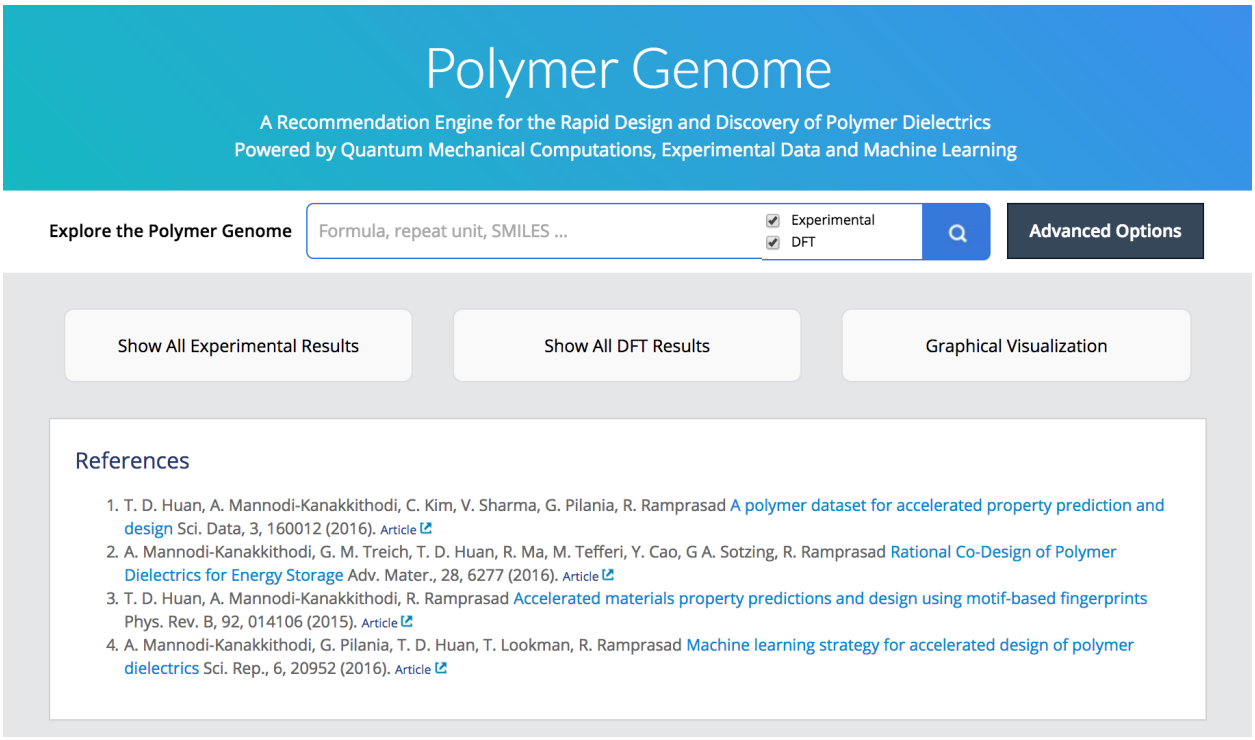
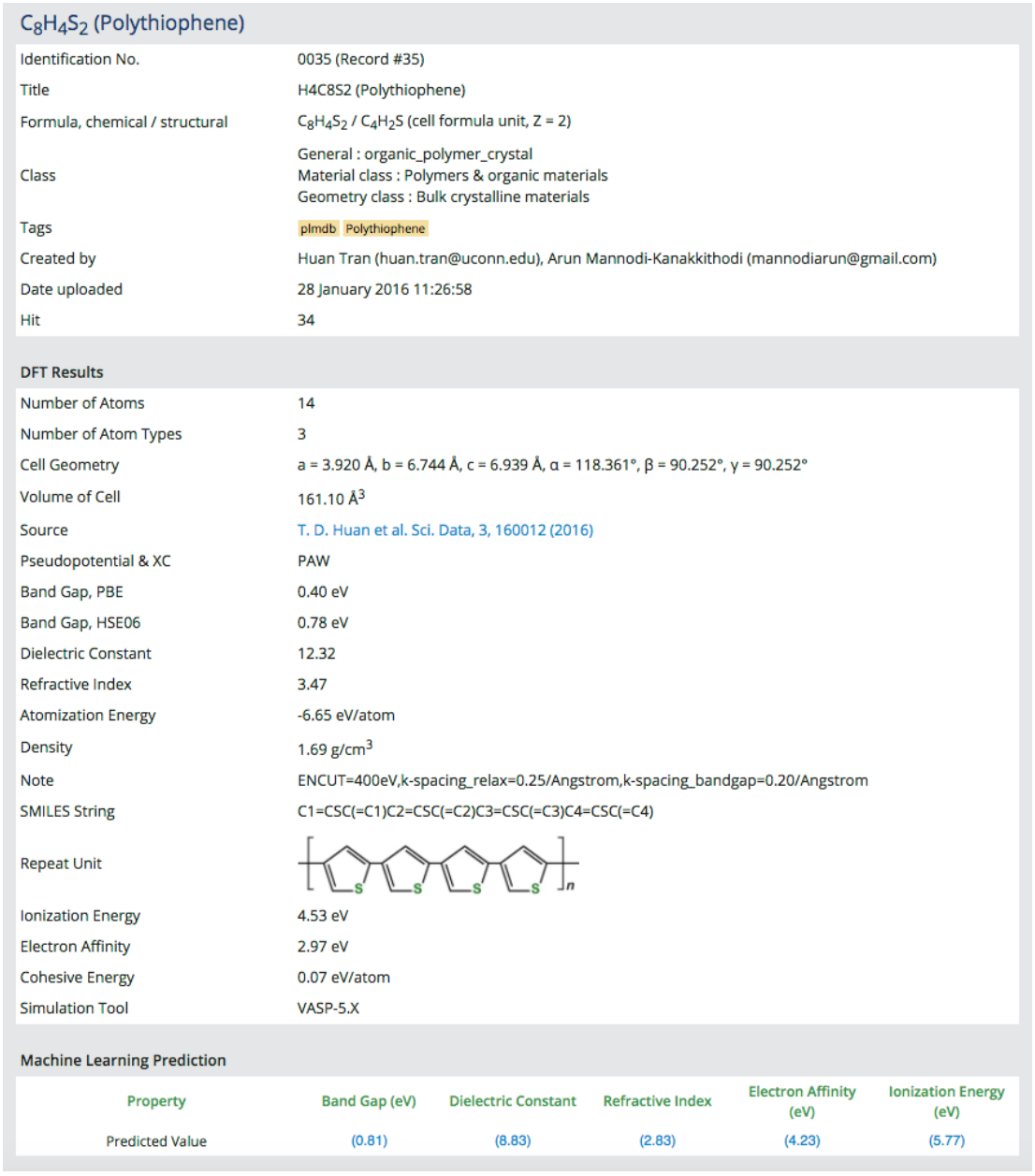


Figure 7.4. *The Polymer Genome platform where any organic polymer can be searched for in terms of its chemical building blocks, SMILES notation or name, and its properties can be accessed via documented experimental or DFT data or via ML predictive models [84].*

The Polymer Genome platform, pictured in **Figure 7.4**, has been developed to function as a recommendation engine for the design of advanced polymer dielectrics. All the DFT computed properties, specifically the ones listed in **Table 7.1**, are available for over 1100 polymers and related materials, while experimentally measured properties (such as dielectric constant, dielectric breakdown field, dielectric loss, etc.) are available for around 70 polymers that were synthesized and tested and discussed in **Chapters 3** and **4**. If any of these polymers are queried in terms of their chemical names, chemical constituents or SMILES, their properties and chemical structure can immediately be pulled up. As an example, shown in **Figure 7.5** are the results of searching for the polymer Polythiophene (-$[C_4H_2S]_n$-); the computed lattice parameters, band gap, dielectric constant, ionization energy, etc. have been listed. Another example is shown in **Figure 7.6**, where searching for the polymer Poly(ether ether ketone) (also known as PEEK) yields its experimentally measured dielectric constant, glass transition temperature and other properties.

Figures 7.5 and **7.6** also show the machine learning predicted properties for the respective polymers, namely the band gap, dielectric constant, refractive index, ionization energy and electron affinity. Each of these models was trained using the methodology described in **Chapter 5**. To make the predictions general in terms of atomic constituents (and not restricted to certain chemical blocks), the motif-based fingerprint explained and used in **Chapter 6** was used here as well. Fresh models were developed for the five properties by training on a dataset of nearly 400 organic polymers using Kernel Ridge Regression and the triples type of fingerprint; parity plots for the same are shown in

Figure 7.7, along with the uncertainties in prediction estimated using Bootstrapping. These models were then incorporated in the Polymer Genome platform in the form of user-friendly design and prediction tools, and are currently employed in making on-demand predictions for any desired polymer.



$C_8H_4S_2$ (Polythiophene)

Identification No.	0035 (Record #35)
Title	H4C8S2 (Polythiophene)
Formula, chemical / structural	$C_8H_4S_2$ / C_4H_2S (cell formula unit, Z = 2)
Class	General : organic_polymer_crystal Material class : Polymers & organic materials Geometry class : Bulk crystalline materials
Tags	plmdb Polythiophene
Created by	Huan Tran (huan.tran@uconn.edu), Arun Mannodi-Kanakkithodi (mannodiarun@gmail.com)
Date uploaded	28 January 2016 11:26:58
Hit	34

DFT Results

Number of Atoms	14
Number of Atom Types	3
Cell Geometry	a = 3.920 Å, b = 6.744 Å, c = 6.939 Å, α = 118.361°, β = 90.252°, γ = 90.252°
Volume of Cell	161.10 $Å^3$
Source	T. D. Huan et al. Sci. Data, 3, 160012 (2016)
Pseudopotential & XC	PAW
Band Gap, PBE	0.40 eV
Band Gap, HSE06	0.78 eV
Dielectric Constant	12.32
Refractive Index	3.47
Atomization Energy	-6.65 eV/atom
Density	1.69 g/cm^3
Note	ENCUT=400eV,k-spacing_relax=0.25/Angstrom,k-spacing_bandgap=0.20/Angstrom
SMILES String	C1=CSC(=C1)C2=CSC(=C2)C3=CSC(=C3)C4=CSC(=C4)
Repeat Unit	
Ionization Energy	4.53 eV
Electron Affinity	2.97 eV
Cohesive Energy	0.07 eV/atom
Simulation Tool	VASP-5.X

Machine Learning Prediction

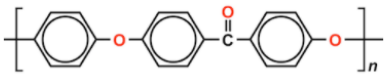
Property	Band Gap (eV)	Dielectric Constant	Refractive Index	Electron Affinity (eV)	Ionization Energy (eV)
Predicted Value	(0.81)	(8.83)	(2.83)	(4.23)	(5.77)

Figure 7.5. *Search results for 'Polythiophene' on the Polymer Genome platform.*

$C_{19}H_{12}O_3$ (Poly(ether ether ketone))

Identification No.	E0015 (Record #E91)
Title	C19H12O3 (Poly(ether ether ketone))
Formula, chemical / structural	$C_{19}H_{12}O_3$ /
Tags	plmdb Polyether, Polyketone Poly(ether ether ketone) PEEK
Hit	8

Experimental Results

Repeat Unit	
Reference	ARL-TR-4880 (REPORT)
Dielectric Constant	3.10
Dielectric Loss, RT 1KHz	0.00 tan(δ)
Dielectric Breakdown Field	322.00 MV/m
Energy Density	1.40 J/cc
Glass Transition Temperature	149.00 °C
Melting Temperature	342.00 °C
State	32.4% Crystallinity
SMILES string	C(C=C1)=CC=C1OC(C=C2)=CC=C2C(=O)C(C=C3)=CC=C3O

Machine Learning Prediction

Property	Band Gap (eV)	Dielectric Constant	Refractive Index	Electron Affinity (eV)	Ionization Energy (eV)
Predicted Value	(3.52)	(4)	(1.85)	(2.6)	(6.14)

Figure 7.6. *Search results for 'PEEK' on the Polymer Genome platform.*

The input to the polymer genome search feature can happen via the chemical formula, the polymer repeat unit, the SMILES notation [171], the polymer chemical name and the polymer class, among other identifying features. The power of a platform such as this is that searching for any polymer will return its properties—computational or experimental—if they exist, but more importantly, the properties of any new polymer that is real or hypothetical can be instantly predicted. Without resorting to expensive computations or measurements, one can make a judgement on the possible utility of a material with the help of such prediction tools. Going forward, *Polymer Genome* will be reinforced with a

host of fresh data, fresh properties, and fresh (more accurate, more general) models, making it truly a “live” database of polymers and a very useful recommendation tool for various properties and applications of interest.

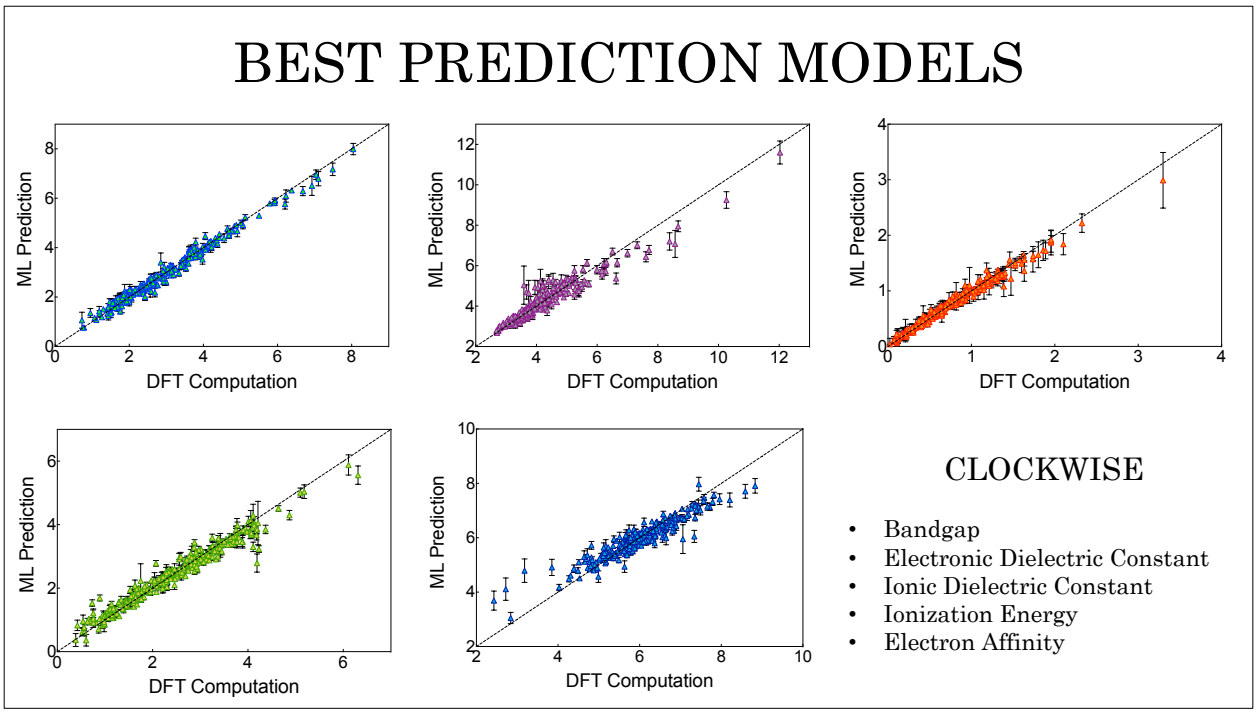


Figure 7.7. *KRR parity plots for 5 properties trained on the organic polymer dataset, and incorporated in the prediction tools on Polymer Genome.*

Chapter 8

SUMMARY AND FUTURE WORK

8.1 Summary

In this Thesis, the importance of computation-guided and data-driven strategies for the rational design of materials was highlighted with the example of advanced polymer dielectrics for energy storage capacitor applications. A design strategy involving high-throughput density functional theory (DFT), guided experiments and machine learning (ML) based insights was executed here, culminating in the successful discovery of several novel organic and organometallic polymer dielectric candidates. DFT was used to compute two crucial properties—dielectric constant and band gap—for a few hundred organic polymers, followed by a few hundred organometallic polymers. After a first stage of screening yielded promising candidates that were synthesized and tested to provide validation for the DFT computations, subsequent generations of polymers were experimentally studied to overcome the issues posed by the initial polymers. Further, ML methods were applied on the DFT data to obtain design rules based on correlation analysis and regression-based prediction models, which facilitate quick and easy estimation of the properties of new polymers. All the computational and experimental data

generated as part of this work, along with the ML models, were collected in the form of an “online materials knowledgebase” within a platform we call *Polymer Genome*. Such data repositories and design tools are critical to the future of materials design, providing ready guidance to future experiments and computations, consequently leading to faster, more efficient design and discovery.

The synergistic use of computations and experiments in a *rational co-design* formulation enabled the design of new polymer dielectrics much faster than implementing standalone experiments. Any computations are incomplete without accompanying experiments, which provide validation as well as realization of modelled materials; experiments, on the other hand, suffer from a lack of direction without computational insights. A marriage between the two is truly a recipe for success in the modern materials research environment. Further, the ability to learn from uniform, curated (experimental or computational) data, and apply this learning on new materials, is truly transformative in terms of accelerating materials design. This is the rapidly progressing field of *materials informatics*, which deals with developing phenomenological theories, design rules and predictive models based on learning from data, as well as logically determining next computations or experiments that should be performed to improve the models and expand the pool of promising materials. Regular improvements in computing power and the increasing use of machine learning based approaches presents endless possibilities in materials research in the coming years.

8.2 Limitations of Current Approach

Although it has been demonstrated that high-throughput computational screening can lead the way for fruitful experimentation and successful discovery, there are certain limitations in the computational approach that one needs to be mindful of. Each computation involves an accurate and extensive crystal structure search, followed by separate DFT calculations for dielectric constant, band gap and other properties. A complete set of calculations for any average polymer in our dataset could take up to 48 hours of computation time on a 32-core computational cluster, which amounts to several months to a few years of total computation time for > 1000 polymers. Besides, the DFT computational expense scales as a square of the number of atoms in the system. These factors make it a tough task to run computations on an enormous number of materials, limiting the growth of the computational dataset and thus the guidance being provided to experiments. However, several massive supercomputers do exist today, such as the majestic *Titan* that resides in all its glory in Oak Ridge National Lab, assisting thousands of scientists in their computational study of materials galore.

Further, all computations employed in this Thesis were performed on purely crystalline renditions of polymers, with the results further restricted by the 0K approximation in DFT. It is known that real polymers may contain significant amorphous portions and would seldom (or never) be purely crystalline. With current state-of-the-art in first principles based computational techniques, the consideration of amorphous nature of materials is

not a problem with a trivial solution. Thus, it is entirely possible that computationally estimated structures and properties are not reproduced experimentally. That said, the close correspondence of many of the computationally obtained XRD patterns and IR spectra with those seen experimentally (as discussed in parts of **Chapter 3** and **Chapter 4**) is an encouraging sign that what is being modeled is not as far away from reality as one might fear.

One of the most crucial aspects of this work is the application of ML techniques for the development of statistical models. An important limitation of this approach is that the trained models are always only as good as the training data used. For instance, using the ML models presented in **Chapter 5, 6** or **7**, predictions on polymers containing side chains, fresh chemical units (i.e., units that never appear in any of the training set polymers) or newer coordination environments may not stand up to a stricter quantum mechanical test. Thus, there would be a requirement of constant data infusion and model retraining to obtain systematic and progressive improvement.

With respect to the ML approach, another vital limitation is in the fingerprint used to represent the materials. While using the 'triplets' of atoms types to train the models has been fruitful here, it should be noted that the fingerprint does not consider factors such as the crystal structure, the morphology, the inter-chain distances, etc., each of which could have mild to major influences on the dielectric constant and band gap. Moreover, the presence of long side chains and/or aromatic rings are known to affect the free volume

and the nature of inter-chain interactions in the polymer, and the current fingerprint may not explicitly account for these factors. The consideration of various such descriptors would undoubtedly improve the prediction performances of regression models. However, it must be stated that the purpose of using machine learning is first and foremost to have a relatively easier, less cumbersome access to the properties of materials than permitted by experiments or computations. The use of a simple, intuitive fingerprint such as what is used here enables that simplicity, combined with satisfactory learning.

8.3 Going Forward

8.3.1 Expansion of Chemical and Property Spaces

The polymers studied as a part of this Thesis contained certain selected chemical blocks built from C, H, O, N, S and F atoms in the case of pure organics, and the added metal-based units (with metals such as Sn, Ti, Zn, Cd, etc.) in the case of organometallics. In terms of expanding the chemical space of polymers, there are numerous options possible for chemical blocks that populate many of the known polymers today. Some of these blocks are pictured in **Figure 8.1**. Fresh polymer repeat units can be formed combining the blocks already studied with these new blocks, and the computational procedure established in previous chapters can lead to a substantially enhanced dataset of

polymers. The addition of certain new blocks, such as the pyridine group shown in **Figure 8.1**, could be beneficial in improving the ionic part of the dielectric constant owing to the introduction of fresh dipoles in the system. Many such groups could have favourable effects on the polymer properties, aside from adding to the diversity and quantity of the data itself. The latter would in turn lead to fresh and more accurate machine learning models that would be widely applicable.

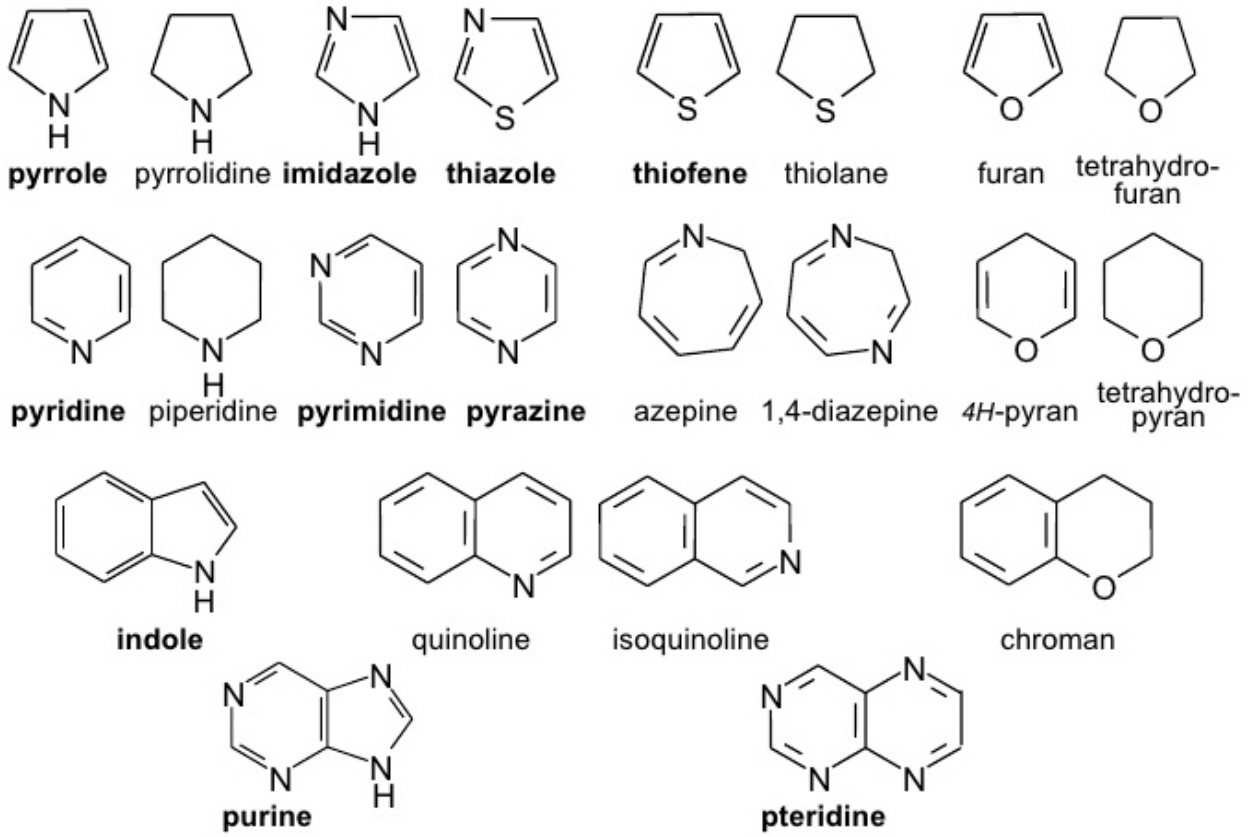


Figure 8.1. *Several new chemical building blocks that could be incorporated in polymer repeat units for fresh computations [172].*

It is also extremely important to expand our computations to other properties which are relevant not only to dielectric applications, but others like organic electronics,

photovoltaics, batteries etc. Apart from the many properties listed in **Table 7.1** in **Chapter 7** (which have been computed in this Thesis), there are properties like elastic constants, breakdown strength and electron mobility that are possible to be computed using DFT related approaches. However, the consideration of these properties and more in a long-term picture requires the application of methodologies beyond DFT, like force-field based simulations as well as meticulous experimentation. Properties like glass transition temperature, melting temperature, solubility parameter, etc. are of great importance to polymers for any application, and have been well documented experimentally over the years [146]. This data can be collected and used to train fresh prediction models for these vital properties which can only be obtained experimentally, thus expanding the power of the machine learning approach way beyond the current capabilities.

8.3.2 An Adaptive Learning Approach

The consideration of new materials for computations and the improvement of the accuracy and applicability of the machine learning model can be achieved in a symbiotic manner. This involves establishing an *adaptive* learning approach, wherein fresh computational data would be added when available and the ML models would be retrained to make fresh, more accurate predictions on newer regions of the chemical space. This process could be repeated in an iterative manner as follows: “ML Model → Predictions → Fresh Computations → Retrained ML Model”, thus establishing a strategy

of slowly but surely pushing the boundaries of the polymer chemical space and progressively expanding the predictive regions via an adaptive learning framework. A strategy for obtaining fresh computational data could be as follows: the current ML model can be used to makes predictions on thousands of new materials, and the specific cases with desirable anticipated properties (implying promise for new materials design) as well as large uncertainties (implying systems sufficiently distinct from present data, i.e., polymers containing new chemical units and environments) could be selected for new computations. These results fortify the computational dataset and enable possible improvements in machine learning, which in turn would facilitate more accurate predictions for similar materials.

REFERENCES

[1] M. F. Ashby and D. R. H. Jones, Engineering Materials 2, Oxford: Pergamon Press, 1992.

[2] E. Starke and J. Staley, "Application of modern aluminum alloys to aircraft," *Progress in Aerospace Sciences,* vol. 32, no. 2, pp. 131-172, 1996.

[3] W. J. Buehler, J. W. Gilfrich and R. C. Wiley, "Effects of Low-Temperature Phase Changes on the Mechanical Properties of Alloys Near Composition TiNi," *Journal of Applied Physics,* vol. 34, no. 5, p. 1475–1477, 1963.

[4] W. Hume-Rothery, Atomic Theory for Students of Metallurgy, London: The Institute of Metals, 1969.

[5] E. Hall, "The Deformation and Ageing of Mild Steel: III Discussion of Results," *Proc. Phys. Soc.,* vol. 64, p. 747–753, 1951.

[6] R. Dugas, A History of Mechanics, Dover Publications, 1988.

[7] F. J. Ragep, "Tusi and Copernicus: The Earth's Motion in Context," in *Science in Context*, vol. 14 , Cambridge University Press., 2001, p. 145–163.

[8] J. Mehra and H. Rechenberg, The historical development of quantum theory, 1 ed., vol. 4, New York: Springer-Verlag., 1982.

[9] R. Feynman, R. Leighton and M. Sands, The Feynman Lectures on Physics, vol. 3, Reading, MA: Addison-Wesley Pub. Co., 1964.

[10] T. Dauxois, M. Peyrard and S. Ruffo, "The Fermi-Pasta-Ulam "numerical experiment": history and pedagogical perspectives," *European Journal of Physics,* vol. 26, no. S3, 2005.

[11] B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method," *J. Chem. Phys.,* vol. 31, no. 2, 1959.

[12] A. Rahman, "Correlations in the Motion of Atoms in Liquid Argon," *Physical Review,* vol. 136, no. 2A, pp. A405-A411, 1964.

[13] P. Hohenberg and W. Kohn, "Inhomogenous electron gas," *Phys. Rev.,* vol. 136, no. 3B, p. 864–871, 1964.

[14] W. Kohn and L. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.,* vol. 140, no. 4A, p. 1133–1138, 1965.

[15] G. Ceder and K. Persson, "The Stuff of Dreams," *Sci. Am.,* vol. 309, pp. 36-40, 2013.

[16] K. S. Ágnes Nagy, "Special Issue "50th Anniversary of the Kohn–Sham Theory—Advances in Density Functional Theory"," *Computation,* vol. 4, no. 45, 2016.

[17] T. H. Jorg Neugebauer, "Density functional theory in materials science," *Comput. Mol. Sci.,* vol. 3, p. 438–448, 2013.

[18] K. Lejaeghere, V. V. Speybroeck, G. V. Oost and S. Cottenier, "Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals," in *Critical Reviews in Solid State and Materials Sciences*, vol. 39, Taylor & Francis, 2014, pp. 1-24.

[19] A. D. Becke, "Perspective: Fifty years of density-functional theory in chemical physics," *Journal of Chemical Physics,* vol. 140, no. 18A301, 2014.

[20] H. S. Yu, S. L. Li and D. G. Truhlar, "Perspective: Kohn-Sham density functional theory descending a staircase," *Journal of Chemical Physics,* vol. 145, no. 130901, 2016.

[21] L. M. Ghiringhelli, "Application of (Kohn–Sham) Density-Functional Theory to Real Materials," in *Many-Electron Approaches in Physics, Chemistry and Mathematics: A Multidisciplinary View*, 2014, pp. 191-206.

[22] J. Hafner, C. Wolverton and G. Ceder, "Toward Computational Materials Design: The Impact of Density Functional Theory on Materials Research," *MRS Bulletin,* vol. 31, no. 9, p. 659–668, 2006.

[23] "Materials Genome Initiative (MGI)," 2011. [Online]. Available: https://www.whitehouse.gov/mgi.

[24] G. Ceder, Y.-M. Chiang, D. R. Sadoway, M. K. Aydinol, Y.-I. Jang and B. Huang, "Identification of cathode materials for lithium batteries guided by first-principles calculations," *Nature,* vol. 392, pp. 694-696, 1998.

[25] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," *Nature Communications,* vol. 7, no. 11241, 2016.

[26] T. Lookman, P. V. Balachandran, D. Xue, J. Hogden and J. Theiler, "Statistical inference and adaptive design for materials discovery," *Current Opinion in Solid State and Materials Science,* 2016.

[27] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier and A. Zunger, "Prediction and accelerated laboratory discovery of previously unknown 18-electron ABX compounds," *Nat. Chem.,* vol. 7, pp. 308-316, 2015.

[28] A. Mannodi-Kanakkithodi and R. Ramprasad, "Rational Design of Polymer Dielectrics: An Application of Density Functional Theory and Machine Learning," in *Computational Materials Discovery*, (Under Review).

[29] A. R. Leach and V. J. Gillet, An Introduction To Chemoinformatics, Springer, 2007.

[30] A. M. Lesk, Introduction to Bioinformatics, Oxford University Press, 2008.

[31] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials,* vol. 4, no. 5, 2016.

[32] K. Rajan, "Materials Informatics: The Materials "Gene" and Big Data," *Annual Review of Materials Research,* vol. 45, pp. 153-169, 2015.

[33] H. Chial, "DNA Sequencing Technologies Key to the Human Genome Project," *Nature Education,* vol. 1, no. 219, 2008.

[34] H. S. Nalwa, Ed., Handbook of Low and High Dielectric Constant Materials and Their Applications, Academic Press , 1999.

[35] J. Ho, R. Ramprasad and S. Boggs, "Effect of Alteration of Antioxidant by UV Treatment on the Dielectric Strength of BOPP Capacitor Film," *IEEE Trans. Dielectr. Electr. Insul.,* vol. 14, no. 5, 2007.

[36] M. Rabuffi and G. Picci, "Status Quo and Future Prospects for Metallized Polypropylene Energy Storage Capacitors," *IEEE Trans. Plasma Sci.,* vol. 30, no. 1939, 2002.

[37] E. J. Barshaw, J. White, M. J. Chait, J. B. Cornette, J. Bustamante, F. Folli, D. Biltchick, G. Borelli, G. Picci and M. Rabuffi, "High Energy Density (HED) Biaxially-Oriented Poly-Propylene (BOPP) Capacitors For Pulse Power Applications," *IEEE Transactions on Magnetics,* vol. 43, no. 1, pp. 223-225, 2007.

[38] J. Tortai, N. Bonifaci and A. Denat, "Diagnostic of the self-healing of metallized polypropylene film by modeling of the broadening emission lines of aluminum emitted by plasma discharge," *Journal of Applied Physics,* vol. 97, no. 053304, 2005.

[39] N. Tu and K. Kao, *J. Appl. Phys.,* vol. 85, no. 7267, 1997.

[40] J. Ho and T. R. Jow, "Characterization of High Temperature Polymer Thin Films for Power Conditioning Capacitors," Army Research Laboratory, Adelphi, MD, 2009.

[41] L. Yang, J. Ho, E. Allahyarov, R. Mu and L. Zhu, "Semi-crystalline Structure−Dielectric Property Relationship and Electrical Conduction in a Biaxially Oriented Poly(vinylidene fluoride) Film under High Electric Fields and High Temperatures," *ACS Appl. Mater. Interfaces,* vol. 7, pp. 19894-905, 2015.

[42] Q. M. Zhang, V. Bharti and X. Zhao, "Giant Electrostriction and Relaxor Ferroelectric Behavior in Electron-Irradiated Poly(vinylidene fluoride-trifluoroethylene) Copolymer," *Science,* vol. Science, pp. 2101-4, 1998.

[43] W. Li, L. Jiang, X. Zhang, Y. Shen and C. W. Nan, "High-energy-density dielectric films based on polyvinylidene fluoride and aromatic polythiourea for capacitors," *J. Mater. Chem. A,* vol. 2, pp. 15803-7, 2014.

[44] L. Jiang, W. Li, J. Zhu, X. Huo, L. Luo and Y. Zhu, "Great reduction of loss at high electric field in the polyvinylidene fluoride/aromatic polythiourea blend films along with an irreversible phase transition," *Appl. Phys. Lett.,* vol. 106, p. 052901 , 2015.

[45] S. Zhang, C. Zou, D. Kushner, X. Zhou, R. J. Orchard, N. Zhang and Q. Zhang, "Semicrystalline polymers with high dielectric constant, melting temperature, and charge-discharge efficiency," *IEEE Trans. Dielectr. Electr. Insul.,* vol. 19, pp. 1158-66, 2012.

[46] J. Ennis, F. MacDougall, X. Yang, R. Cooper, K. Seal, C. Naruo, B. Spinks, P. Kroessler and J. Bates, "Recent advances in high voltage, high energy capacitor technology," in *16th IEEE International Pulsed Power Conference*, Albuquerque, NM, USA, 2007.

[47] F. E. J. MacDougall, X. H. Yang, R. Cooper, J. Gilbert, J. Bates, C. Naruo, M. Schneider, N. Keller, S. Joshi, T. Jow, J. Ho, C. Scozzie and S. Yen, "High energy density capacitors for pulsed power applications," in *IEEE Pulsed Power Conference*, Washington, DC , 2009.

[48] H. Bluhm, Pulsed Power Systems: Principles and Applications, Springer, 2006.

[49] A. Mannodi-Kanakkithodi, G. M. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing and R. Ramprasad, "Rational Co-Design of Polymer Dielectrics for Energy Storage," *Advanced Materials,* vol. 28, no. 30, 2016.

[50] V. Sharma, C. 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," *Nat. Commun.,* vol. 5, p. 4845, 2014.

[51] C. C. Wang, G. Pilanina, S. Boggs, S. Kumar, C. Breneman and R. Ramprasad, "Computational strategies for polymer dielectrics design," *Polymer,* vol. 55, no. 4, p. 979, 2014.

[52] 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, no. 20952, 2016.

[53] S. Baroni, S. de Gironcoli, A. Dal Corso and P. Giannozzi, "Phonons and related crystal properties from density-functional perturbation theory," *Rev. Mod. Phys.,* vol. 73, no. 2, p. 515, 2001.

[54] X. Gonze and C. Lee, "Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory.," *Phys. Rev. B,* vol. 55, p. 10355–10368, 1997.

[55] X. Gonze, "First-principles responses of solids to atomic displacements and homogeneous electric fields: Implementation of a conjugate-gradient algorithm.," *Phys. Rev. B,* vol. 55, p. 10337–10354, 1997.

[56] J. Heyd, G. E. Scuseria and M. Ernzerhof, "Hybrid functionals based on a screened Coulomb potential," *J. Chem. Phys.,* vol. 118, no. 18, p. 8207 , 2003.

[57] J. P. Perdew, "Density functional theory and the band gap problem," *Int. J. Quant. Chem.,* vol. 28, pp. 497-523, 1985.

[58] T. Huan, S. Boggs, G. Teyssedre, C. Laurent, M. Cakmak, S. Kumar and R. Ramprasad, "Advanced Polymeric Dielectrics for High Energy Density Applications," *Prog. Matter. Sci.,* 2016.

[59] Q. Zhu, A. R. Oganov, C. W. Glass and H. T. Stokes, "Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications," *Acta Cryst. ,* vol. B68, pp. 215-226, 2012.

[60] Q. Zhu, V. Sharma, A. R. Oganov and R. Ramprasad, "Predicting polymeric crystal structures by evolutionary algorithms," *J. Chem. Phys.,* vol. 141, no. 15, p. 154102, 2014.

[61] S. Goedecker, "Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems," *J. Chem. Phys.,* vol. 120, no. 21, p. 9911, 2004.

[62] M. Amsler and S. Goedecker, "Crystal structure prediction using the minima hopping method," *J. Chem. Phys.,* vol. 133, p. 224104, 2010.

[63] A. Mannodi-Kanakkithodi, C. C. Wang and R. Ramprasad, " Compounds based on Group 14 elements: building blocks for advanced insulator dielectrics design," *J. Mater. Sci.,* vol. 50, no. 2, p. 801, 2015.

[64] C. Wang, G. Pilania and R. Ramprasad, "Dielectric properties of carbon-, silicon-, and germanium-based polymers: A first-principles study," *Phys. Rev. B,* vol. 87, p. 035103, 2013.

[65] H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, "The Chemistry and Applications of Metal-Organic Frameworks," *Science,* vol. 341, no. 6149, 2013.

[66] G. Pilania, C. C. Wang, K. Wu, N. Sukumar, C. Breneman, G. Sotzing and R. Ramprasad, "New Group IV Chemical Motifs for Improved Dielectric Permittivity of Polyethylene," *J. Chem. Inf. Model.,* vol. 53, no. 4 , p. 879–886, 2013.

[67] A. F. Baldwin, T. D. Huan, R. Ma, A. Mannodi-Kanakkithodi, M. Tefferi, N. Katz, Y. Cao, R. Ramprasad and G. A. Sotzing, "Rational Design of Organotin Polyesters," *Macromolecules,* vol. 48, pp. 2422-2428, 2015.

[68] A. F. Baldwin, R. Ma, T. D. Huan, Y. Cao, R. Ramprasad and G. A. Sotzing, "Effect of Incorporating Aromatic and Chiral Groups on the Dielectric Properties of Poly(dimethyltin esters)," *Macromol. Rapid Commun.,* vol. 35, no. 24, p. 2082, 2014.

[69] A. F. Baldwin, R. Ma, A. Mannodi-Kanakkithodi, T. D. Huan, C. C. Wang, M. Tefferi, J. E. Marszalek, M. Cakmak, Y. Cao, R. Ramprasad and G. A. Sotzing, "Poly(dimethyltin glutarate) as a Prospective Material for High Dielectric Applications," *Adv. Mater.,* vol. 27, p. 346, 2015.

[70] T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania and R. Ramprasad, "A Polymer Dataset for Accelerated Property Prediction and Design," *Sci. Data.,* vol. 3, no. 160012, 2016.

[71] A. Mannodi-Kanakkithodi, T. Huan and R. Ramprasad, "Mining Materials Design Rules from Data: The Example of Polymer Dielectrics".

[72] A. F. Baldwin, R. Ma, C. Wang, R. Ramprasad and G. Sotzing, "Structure–property relationship of polyimides based on pyromellitic dianhydride and short-chain aliphatic diamines for dielectric material applications," *J. Appl. Polym. Sci.,* vol. 130, pp. 1276-80, 2013.

[73] R. Lorenzini, W. Kline, C. Wang, R. Ramprasad and G. Sotzing, "The rational design of polyurea & polyurethane dielectric materials," *Polymer,* vol. 54, no. 14, p. 3529, 2013.

[74] R. Ma, A. F. Baldwin, C. C. Wang, I. Offenbach, M. Cakmak, R. Ramprasad and G. A. Sotzing, "Rationally Designed Polyimides for High-Energy Density Capacitor Applications," *ACS Appl. Mater. Interfaces,* vol. 6, no. 13, p. 10445, 2014.

[75] R. Ma, V. Sharma, A. F. Baldwin, M. Tefferi, I. Offenbach, M. Cakmak, R. Weiss, Y. Cao, R. Ramprasad and G. A. Sotzing, "Rational design and synthesis of polythioureas as capacitor dielectrics," *J. Mater. Chem. A,* vol. 3, p. 14845, 2015.

[76] G. Treich, S. Nasreen, A. Mannodi-Kanakkithodi, R. Ma, M. Tefferi, J. Flynn, Y. Cao, R. Ramprasad and G. Sotzing, "Optimization of Organotin Polymers for Dielectric Applications," *Appl. Mater. Interfaces,* vol. 8, no. 33, p. 21270– 21277, 2016.

[77] T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, "Accelerated materials property predictions and design using motif-based fingerprints," *Phys. Rev. B,* vol. 92, p. 014106, 2015.

[78] K. Vu, J. Snyder, L. Li, M. Rupp, B. Chen, T. Khelif, K. Muller and K. Burke, "Understanding kernel ridge regression: common behaviors from simple functions to density functionals," *Int. J. Quant. Chem.,* vol. 115, no. 16, p. 1115– 1128, 2015.

[79] S. Curtarolo, W. Setyawan, S. Wanga, J. Xue, K. Yang, R. Taylor, L. Nelson, G. Hart, S. Sanvito, M. Nardelli, N. Mingo and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations," *Computational Materials Science,* vol. 58, p. 227–235, 2012.

[80] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, "AiiDA: automated interactive infrastructure and database for computational science," *Computational Materials Science,* vol. 111, p. 218–230, 2016.

[81] A. Jain, S. Ong, G. Hautier, W. Chen, W. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. Persson, "Commentary: The Materials Project: A materials

genome approach to accelerating materials innovation," *APL Materials,* vol. 1, no. 011002, 2013.

[82] T. Bligaard, M. Dulak, J. Greeley, S. Nestorov, J. Hummelshoj, D. Landis, J. Norskov and K. Jacobsen, "The Computational Materials Repository," *Computing in Science & Engineering,* vol. 14, pp. 51-57, 2012.

[83] J. Saal, S. Kirklin, A. Muratahan, B. Meredig and C. Wolverton, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," *JOM,* vol. 65, no. 11, p. 1501–1509, 2013.

[84] [Online]. Available: http://khazana.uconn.edu/.

[85] Y. Wang and Y. Maa, "Perspective: Crystal structure prediction at high pressures," *J. Chem. Phys.,* vol. 140, no. 040901, 2014.

[86] S. Wu, W. Li, M. Lin, Q. Burlingame, Q. Chen, A. Payzant, K. Xiao and Q. M. Zhang, "Aromatic Polythiourea Dielectrics with Ultrahigh Breakdown Field Strength, Low Dielectric Loss, and High Electric Energy Density," *Adv. Matter.,* vol. 25, p. 1734–1738, 2013.

[87] A. North and J. C. Reid, "Dielectric relaxation in a series of heterophase polyether polyurethanes," *Eur. Polym. J.,* vol. 8, pp. 1129-38, 1972.

[88] G. Kresse and J. Hafner, *Phys. Rev. B.,* vol. 47, no. 558, 1993.

[89] G. Kresse and J. Furthmuller, *Phys. Rev. B.,* vol. 54, no. 11169, 1996.

[90] J. P. Perdew, K. Burke and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Phys. Rev. Lett.,* vol. 77, pp. 3865-8, 1996.

[91] P. E. Blochl, *Phys. Rev. B.,* vol. 50, no. 17953, 1994.

[92] M. Dion, H. Rydberg, E. Schröder, D. Langreth and B. Lundqvist, *Phys. Rev. Lett.,* vol. 92, no. 246401, 2004.

[93] J. Klimeš, D. R. Bowler and A. Michaelides, *Phys. Rev. B.,* vol. 83, no. 195131, 2011.

[94] G. Avitabile, R. Napolitano, B. Pirozzi, K. D. Rouse, M. W. Thomas and B. T. M. Willis, *J. Polymer Sc.,* vol. 13, no. 6, 1975.

[95] C. Nakafuku and T. Takemura, *J. Appl. Phys.,* vol. 14, no. 599, 1975.

[96] J. Klimes, D. Bowler and A. Michaelides, "Van der Waals density functionals applied to solids," *Phys Rev B.,* vol. 83, p. 195131–195144, 2011.

[97] F. A. Cotton and G. Wilkinson, Advanced Inorganic Chemistry, New York: Wiley, 1993.

[98] J. Trotter, M. Akhtar and N. Bartlett, "The crystal structure of germanium difluoride.," *J Chem Soc A.,* p. 30–33, 1966.

[99] G. Denes, J. Pannetier and J. Lucas, "About SnF2 (stannous fluoride II): Crystal structure of SnF2.," *J Solid State Chem,* vol. 33, pp. 1-11, 1980.

[100] K. Doll and M. Jansen, "Ab initio energy landscape of GeF2: a system featuring lone pair structure candidates," *Angew Chem,* vol. 50, p. 4627–4632, 2011.

[101] G. Denes, "Phase transitions and structural relationships between Ge5F12, GeF2, SnF2, and TeO2," *J Solid State Chem,* vol. 78, p. 52–65, 1989.

[102] J. Pannetier, G. Denes, M. Durand and J. Buevoz, "SnF2 phase transition: neutron diffraction and NMR study," *J Phys. France,* vol. 41, p. 1019–1024, 1980.

[103] J. van den Berg, "The crystal structure of SnCl2," *Acta Cryst.,* vol. 14, p. 1002–1003, 1961.

[104] F. Coleman, G. Feng, R. Murphy, P. Nockemann, K. Seddon and M. Swadzba-Kwasny, "Lead (II) chloride ionic liquids and organic/inorganic hybrid materials: a study of chloroplumbate (II) speciation.," *Dalton Trans,* vol. 42, p. 5025–5035, 2013.

[105] J. Kudrnovsky, N. Christensen and J. Maek, "Electronic structure of fluorite-type compounds and mixed crystals," *Phys Rev B.,* vol. 43, p. 12597–12606, 1991.

[106] C. Erk, L. Hammerschmidt, D. Andrae, B. Paulus and S. Schlecht, "Low-temperature formation of cubic b-PbF2: precursorbased synthesis and first-principles phase stability study," *J. Chem. Phys.,* vol. 13, p. 6029–6035, 2011.

[107] S. Hull, P. Berastegui, S. Eriksson and N. Gardner, "Crystal structure and superionic conductivity of PbF2 doped with KF," *J Phys. Condens. Matter.,* vol. 10, p. 8429–8446, 1998.

[108] X. Wen, T. Cahill and R. Hoffmann, "Exploring Group 14 Structures: 1D to 2D to 3D.," *Chem. Eur. J.,* vol. 16, p. 6555–6566, 2010.

[109] W. Harrison, "Bond-orbital model and the properties of tetrahedrally coordinated solids," *Phys. Rev. B.,* vol. 8, p. 4487–4498, 1973.

[110] C. E. Carraher Jr. and M. R. Roner, "Organotin polymers as anticancer and antiviral agents," *J. Organomet. Chem. ,* vol. 751, pp. 67-82, 2014.

[111] Q. Li, L. Chen, M. R. Gadinski, S. Zhang, G. Zhang, H. Li, A. Haque, L. Q. Chen, T. Jackson and Q. Wang, "Flexible high-temperature dielectric materials from polymer nanocomposites," *Nature,* vol. 523, pp. 576-9, 2015.

[112] X. Xiao, X. Han, Z. Mei, D. Zhu, K. Shao, J. Liang, M. Tian and L. Xu, "Organotin(IV) carboxylates based on amide carboxylic acids: Syntheses, crystal structures and characterizations," *J. Organomet. Chem.,* vol. 729, pp. 28-39, 2013.

[113] C. J. Carraher, *Die Angew. Macromol. Chem.,* vol. 31, no. 115, 1973.

[114] A. Davies, Organotin Chemistry, Wiley-VCH, 2003.

[115] V. Peruzzo, G. Plazzogna and G. Tagliavini, "The preparation and properties of some allytin carboxylates," *J. Organomet. Chem.,* vol. 40, no. 1, pp. 129-133, 1972.

[116] S. Nasreen, G. Treich, M. Baczkowski, A. Mannodi-Kanakkithodi, A. Baldwin, S. Scheirey, Y. Cao, R. Ramprasad and G. Sotzing, "A Materials Genome Approach to Dielectrics Design through incorporating Zinc and Cadmium in Main Chain Organic Polymers," *(Submitted),* 2017.

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

[118] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis and T. Lookman, "Machine learning bandgaps of double perovskites," *Sci. Rep.,* vol. 6, no. 19375, 2016.

[119] T. Mueller, A. G. Kusne and R. Ramprasad, "Machine Learning in Materials Science: Recent Progress and Emerging Applications," in *Reviews in Computational Chemistry*, vol. 29, John Wiley & Sons, 2016.

[120] K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. Müller and E. Gross, "How to represent crystal structures for machine learning: towards fast prediction of electronic properties,," *Phys. Rev. B.,* vol. 89, no. 205118, 2014.

[121] F. Faber, A. Lindmaa, O. v. Lilienfeld and R. Armiento, "Crystal structure representations for machine learning models of formation energies," *Int. J. Quant. Chem.,* vol. 115, no. 16, p. 1094–1101, 2015.

[122] G. Pilania, J. Gubernatis and T. Lookman, "Classification of octet AB-type binary compounds using dynamical charges: A materials informatics perspective," *Sci. Rep.,* vol. 5, no. 17504, 2015.

[123] G. Pilania, J. Gubernatis and T. Lookman, "Structure classification and melting temperature prediction in octet AB solids via machine learning," *Phys. Rev. B.,* vol. 91, no. 214302, 2015.

[124] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl and M. Scheffler, "Big data of materials science: critical role of the descriptor.," *Phys. Rev. Lett.,* vol. 114, no. 105503, 2015.

[125] J. Hattrick-Simpers, J. Gregoire and G. Kusne, "Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge," *APL Materials,* vol. 4, no. 053211, 2016.

[126] A. Kusne, D. Keller, A. Anderson, A. Zaban and I. Takeuchi, "High-throughput determination of structural phase diagram and constituent phases using GRENDEL," *Nanotechnology,* vol. 26, no. 444002, 2015.

[127] A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, T. Lookman and J. Gubernatis, "Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers," *Comput. Mater. Sci.,* vol. 125, no. 92, 2016.

[128] A. Mannodi-Kanakkithodi, G. Pilania and R. Ramprasad, "Critical assessment of regression-based machine learning methods for polymer dielectrics," *Comput. Mater. Sci.,* vol. 125, no. 123, 2016.

[129] M. De Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta and A. Gamst, "A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds," *Sci Rep.,* vol. 6, no. 34256, 2016.

[130] C. Kim, G. Pilania and R. Ramprasad, "From Organized High-throughput Data to Phenomenological Theory: The Example of Dielectric Breakdown," *Chem. Mater.,* vol. 28, no. 1304, 2016.

[131] C. Kim, G. Pilania and R. Ramprasad, ""Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites," *J. Phys. Chem. C,* vol. 120, no. 14575, 2016.

[132] Z. Li, J. Kermode and A. De Vita, "Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces," *Phys. Rev. Lett.,* vol. 114, no. 096405, 2015.

[133] V. Botu and R. Ramprasad, "Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics," *International Journal of Quantum Chemistry,* vol. 115, no. 16, p. 1074–1083, 2015.

[134] V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," *Phys. Rev. B.,* vol. 92, no. 094306, 2015.

[135] M. Rupp, A. Tkatchenko, K. Muller and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning.," *Phys. Rev. Lett.,* vol. 108, no. 058301, 2012.

[136] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning.," *Phys. Rev. B,* vol. 89, no. 094104, 2014.

[137] A. Seko, T. Maekawa, K. Tsuda and I. Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," *Phys. Rev. B.,* vol. 89, no. 054303, 2014.

[138] V. Sharma, G. Pilania, G. Rossetti, K. Slenes and R. Ramprasad, "Comprehensive examination of dopants and defects in BaTiO3 from first principles," *Phys. Rev. B,* vol. 87, no. 134109, 2013.

[139] E. I. L. Roman M. Balabin, "Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies," *Journal of Chemical Physics,* vol. 131, no. 074104, 2009.

[140] C. L. Phillips and G. A. Vothab, "Discovering crystals using shape matching and machine learning," *Soft Matter,* vol. 9, no. 8552, 2013.

[141] G. Hautier, C. C. Fischer, A. Jain, T. Mueller and G. Ceder, "Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory," *Chem. Mater.,* vol. 22, p. 3762–3767, 2010.

[142] O. A. v. Lilienfeld, R. Ramakrishnan, M. Rupp and A. Knoll, "Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties," *International Journal of Quantum Chemistry,* vol. 115, p. 1084–1093, 2015.

[143] M. Rupp, M. Bauer, R. Wilcken, A. Lange, M. Reutlinger, F. Boeckler and G. Schneider, "Machine Learning Estimates of Natural Product Conformational Energies," *PLOS Computational Biology,* vol. 10, no. 1, 2014.

[144] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Muller and A. v. Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," *New Journal of Physics,* vol. 15, no. 095003, 2013.

[145] R. L. Miller, Crystallographic Data and Melting Points for Various Polymers, John Wiley and Sons Inc., 2003.

[146] D. Van Krevelen and K. Nijenhuis, Properties of Polymers, Elsevier, 2009.

[147] S. M. Stigler, "Francis Galton's Account of the Invention of Correlation," *Statist. Sci.,* vol. 4, no. 2, pp. 73-79, 1989.

[148] L. Yu, R. Kokenyesi, D. Keszler and A. Zunger, "Inverse Design of Hihg Absorption Thin-Film Photovoltaic Materials," *Adv. En. Mat.,* vol. 3, no. 4348, 2013.

[149] M. d'Avezac, J. Luo, T. Chanier and A. Zunger, "Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors," *Phys. Rev. Lett.,* vol. 108, no. 027401, 2012.

[150] A. Jain, I. Castelli, G. Hautier, D. Bailey and K. Jacobsen, "Performance of genetic algorithms in search for water splitting perovskites," *J. Mat. Sc.,* vol. 48, p. 6519–6534, 2013.

[151] B. Schölkopf, "The Kernel Trick for Distances," in *Advances in Neural Information Processing Systems 13*, T. K. L. a. T. G. D. a. V. Tresp, Ed., Cambridge, MA, MIT Press, 2001, p. 301–307..

[152] V. Vapnik, The Nature of Statistical Learning Theory, New York: Springer-Verlag, 2000.

[153] V. Vapnik, Statistical Learning Theory, New York: John Wiley and Sons, 1998.

[154] H. Kuhn and A. Tucker, "Nonlinear programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1951.

[155] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," vol. 209, p. 415–446, 1909.

[156] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996.

[157] F. Faber, A. Lindmaa, O. Von Lilienfeld and R. Armiento, "Machine learning energies of 2 M elpasolite (ABC2D6) crystals.," *Phys. Rev. Lett.,* vol. 117, no. 135502, 2016.

[158] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning, 2006.

[159] P. Schwerdtfeger, "Table of experimental and calculated static dipole polarizabilities for the electronic ground states of the neutral elements," 2014. [Online].

[160] R. Barrett, "J. Chem. Education," vol. 39, no. 251, 1962.

[161] "The Periodic Table by WebElements," [Online]. Available: https://www.webelements.com/.

[162] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. S. N. Quirós, P. Moeck, R. Downs and A. Bail, "Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration," *Nucleic Acids Res.,* vol. 40, no. D1, pp. D420-D427, 2012.

[163] A. Bakhshi, "Theoretical tailoring of electrically conducting polymers: some new results," *Materials Science and Engineering,* vol. C, no. 3, pp. 249-255, 1995.

[164] A. Bakshi, "Molecular engineering of small band gap polymers: heteroaromatic bicyclic polymers," *J. Mol. Struct. (Theochem),* vol. 361, pp. 259-268, 1996.

[165] S. Bouzzine, G. Salgado-Morán, M. Hamidi, M. Bouachrine, A. Pacheco and D. Glossman-Mitnik, "DFT Study of Polythiophene Energy Band Gap and Substitution Effects," *J. Chem.,* no. 296386, 2015.

[166] A. Kahn, "Fermi level, work function and vacuum level," *Mater. Horiz.,* vol. 3, no. 7, 2016.

[167] G. Zhang and C. Musgrave, "Comparison of DFT Methods for Molecular Orbital Eigenvalue Calculations," *J. Phys. Chem. A,* vol. 111, pp. 1554-1561, 2007.

[168] C.-G. Zhan, J. Nichols and D. Dixon, "Ionization Potential, Electron Affinity, Electronegativity, Hardness, and Electron Excitation Energy: Molecular Properties from Density Functional Theory Orbital Energies," *J. Phys. Chem. A,* vol. 107, pp. 4184-4195, 2003.

[169] M. Miao, S. Yarbro, P. Barton and R. Seshadri, "Electron affinities and ionization energies of Cu and Ag delafossite compounds: A hybrid functional study," *Phys. Rev. B.,* vol. 89, no. 045306, 2014.

[170] R. Partridge, "Vacuum-Ultraviolet Absorption Spectrum of Polyethylene," *J. Chem. Phys.,* vol. 45, no. 1685, 1966.

[171] P. Duchowicz, S. Fioressi, D. Bacelo, L. Saavedra, A. Toropova and A. Toropov, "QSPR studies on refractive indices of structurally heterogeneous polymers," *Chemometrics and Intelligent Laboratory Systems,* vol. 150, pp. 86-91, 2015.

[172] "Heterocyclic Compounds," 2007. [Online]. Available: https://www.slideshare.net/MUBOSScz/08-heterocyclic-compounds.

[173] H. Fujiwara, "Ellipsometry," in *Handbook of Optical Metrology: Principles and Applications, Second Edition*, Boca Raton, CRC Press, 2009, p. 641–660.

[174] J. Simpson and A. S. Clair, "Fundamental insight on developing low dielectric constant polyimides," *Thin Solid Films,* Vols. 308-309, pp. 480-5, 1997.

[175] H. J. Hwang, C. Li and C. S. Wang, "Dielectric and thermal properties of dicyclopentadiene containing bismaleimide and cyanate ester. Part IV," *Polymer,* vol. 47, pp. 1291-9, 2006.

LIST OF PUBLICATIONS

1. A. Mannodi-Kanakkithodi, G. M. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing, & R. Ramprasad, “Rational Co-Design of Polymer Dielectrics for Energy Storage.” *Adv. Mater. Prog. Rep.* **2016**, doi: 10.1002/ adma.201600377. (Chapter 3 and Chapter 4 of Thesis)

2. A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, & R. Ramprasad, “Machine Learning Strategy for Accelerated Design of Polymer Dielectrics.” *Sci. Rep.* **2016**, 6, 20952. (Chapter 5 of Thesis)

3. T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, & R. Ramprasad, “A polymer dataset for accelerated property prediction and design.” *Sci. Data.* **2016**, 3, 160012. *(A.M.K. and T.D.H. are equal contributing authors.)* (Chapter 6 and Chapter 7 of Thesis)

4. A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, T. Lookman, & J.E. Gubernatis, “Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers”, Comput. Mater. Sci. **2016**, 125, 92.

5. A. Mannodi-Kanakkithodi, G. Pilania, & R. Ramprasad, “Critical assessment of regression-based machine learning methods for polymer dielectrics”, Comput. Mater. Sci. **2016**,125, 123. (Chapter 5 of Thesis)

6. G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis & T. Lookman “Machine learning bandgaps of double perovskites.” *Sci. Rep.* **2016**, 6, 19375.

7. M. Misra, A. Mannodi-Kanakkithodi, T. C. Chung, R. Ramprasad, & S. K. Kumar, “Critical role of morphology on the dielectric constant of semicrystalline polyolefins”, J. Chem. Phys. **2016**, 144, 234905.

8. G. M. Treich, S. Nasreen, A. Mannodi-Kanakkithodi, R. Ma, M. Tefferi, J. Flynn, Y. Cao, R. Ramprasad, & G. A. Sotzing, “Optimization of Organotin Polymers for Dielectric Applications”, ACS Appl. Mater. Interfaces, **2016**, 8 (33), 21270–21277. (Chapter 4 of Thesis)

9. T.D. Huan, A. Mannodi-Kanakkithodi, & R. Ramprasad, “Accelerated materials property predictions and design using motif-based fingerprints.” *Phys. Rev. B.* **2015**, 92, 014106. (Chapter 5 and Chapter 6 of Thesis)

10. A. Mannodi-Kanakkithodi, C. C. Wang, & R. Ramprasad, “Compounds based on Group 14 elements: building blocks for advanced insulator dielectrics design.” *J. Mater. Sci.* **2015**, 50, 801. (Chapter 4 of Thesis)

11. A. F. Baldwin, R. Ma, A. Mannodi-Kanakkithodi, T. D. Huan, C. C. Wang, J. E. Marszalek, M. Cakmak, R. Ramprasad, & G. A. Sotzing, “Poly(dimethyltin glutarate) as a high dielectric polymer for energy storage applications.”, *Adv. Mater.* **2015**, 27, 346. (Chapter 4 of Thesis)

12. A. F. Baldwin, T. D. Huan, R. Ma, A. Mannodi-Kanakkithodi, N. Katz, R. Ramprasad, & G. A. Sotzing, “Organotin Polymers as High Dielectric Constant Materials.” *Macromolecules.* **2015**, 48, 2422. (Chapter 4 of Thesis)

13. A. Mannodi-Kanakkithodi and R. Ramprasad, “Rational Design of Polymer Dielectrics: An Application of Density Functional Theory and Machine Learning,” in Computational Materials Discovery, *(Under Review)* (Chapter 1 of Thesis)

14. A. Mannodi-Kanakkithodi, T.D. Huan & R. Ramprasad, “Mining Materials Design Rules from Data: The Example of Polymer Dielectrics.” *(In Preparation)* (Chapter 6 of Thesis)

15. G.M. Treich, M. Tefferi, S. Nasreen, A. Mannodi-Kanakkithodi, Z. Li, R. Ramprasad, G A. Sotzing, Y. Cao, “A rational co-design approach to the creation of new dielectric polymers with high energy density”, *IEEE Trans. Dielectr. Electr. Insul.* **2017**, 24, 732. (Chapter 3 and Chapter 4 of Thesis)