MACHINE LEARNING BASED MODELS FOR THE DESIGN OF
SOLID POLYMER ELECTROLYTES

A Dissertation
Presented to
The Academic Faculty

By

Shruti Venkatram

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Materials Science and Engineering

Georgia Institute of Technology

May 2021

© Shruti Venkatram 2021
MACHINE LEARNING BASED MODELS FOR THE DESIGN OF SOLID POLYMER ELECTROLYTES

Thesis committee:

Dr. Rampi Ramprasad, Advisor
School of Materials Science and Engineering
Georgia Institute of Technology

Dr. Roshan Joseph
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Sunderasan Jayaraman
School of Materials Science and Engineering
Georgia Institute of Technology

Dr. Seung Soon Jang
School of Materials Science and Engineering
Georgia Institute of Technology

Dr. Blair Brettmann
School of Chemical and Biomolecular Engineering
School of Materials Science and Engineering
Georgia Institute of Technology

Date approved: April 28, 2021
ACKNOWLEDGMENTS

My time as a graduate student in Georgia Tech and as an immigrant in the United States has been a bittersweet journey with very memorable highs and some soul-crushing lows. I have several people to thank for being an integral part of this experience that has shaped so much of the person I am today.

First, my advisor Prof Rampi Ramprasad for giving me a new lease of life as a graduate student and for always being present and supportive every step of the way. Under his mentorship, I learned immensely valuable lessons on selecting good research problems, thinking through them deeply and most importantly, being a more patient scientific writer (which I am still working on!). His drive, ambition and passion for research is inspiring. During the last three years, my discussions with him about research and beyond have been extremely enjoyable and informative. One such discussion, which I think about fairly often and has constantly kept me motivated was where he told me, ’You know, there is no free lunch in this world. If you work hard and are extremely good at what you do, you can control what you work on. But you cannot escape working hard.’ I’m indebted to him for being a great advisor and am thankful to him for making me the researcher I am today.

Next, I’d like to thank my committee for being generous with their time and insightful comments which have helped make my thesis better. To Prof Blair Brettmann for her valuable experimental insights during the course of the solvent work. To Prof Sunderasan Jayaraman for his support and perceptive questions during the course of the bulk of the polymer electrolyte work. To Prof Joseph for his valuable expertise on machine learning and statistics used in this work. To Prof Jang for introducing me to the wonderful world of computational materials as well as providing comments related to his expertise in computational materials that has made this work better.

I’d like to thank my both my sponsors and collaborators, the Kolon Center for Lifestyle Innovation (KCLI) as well as Prof Sundersan Jayaraman and Dr. Sungmee Park for
facilitating a smooth collaboration with Kolon. I’d also like to thank the Office of Naval Research (ONR) for funding the solvent work in this thesis through the Multi University Research Initiative (MURI). My collaborators in the MURI, Prof Blair Brettmann and Manali Banerjee for taking the time to explain unfamiliar experimental methods and for having several meaningful scientific discussions.

I have been fortunate to have wonderful current and previous labmates who have been instrumental in shaping my research especially Dr. Lihua Chen and Dr Rohit Batra for their honest feedback on several occasions as well as Dr. Chiho Kim for his help throughout my PhD.

This acknowledgment would be remiss without thanking my friends and family. To my friends – Ronak, Harsh, Parth, Avoni, Priya, Akanksha and Advait, thank you for being by my side and making my trips back home so much more special. To my first friend in grad school, Connor, thank you for sticking by my side and being such a wonderful friend during the hard times. Also, thanks to my other friends at Georgia Tech, Shreyas, Kavya and Aniruddha.

To my parents, Rahul and my grandmother for always being wonderful sources of support and encouragement during the entire journey. It would not be possible without your constant encouragement, help and delicious food! Also, my uncle, Venu Mama, thank you for all the wonderful home-made snacks that has made my days so much better. Finally, to Viraj, the love of my life and the wind beneath my wings – I’m so grateful for your unwavering support and companionship. Thank you for being so patient, kind and empathetic. The last eight years have been a learning experience but I always had you by my side to pause for a second, smell the roses and keep working on the next thing.
# TABLE OF CONTENTS

Acknowledgments ......................................................... iii

List of Tables ......................................................... ix

List of Figures ......................................................... xi

Summary ................................................................. xiv

Chapter 1: Introduction and Background ................................. 1

1.1 Towards safer Li-ion batteries .................................... 1

1.2 Lithium ion battery basics and associated safety issues .......... 2

1.3 The quest for discovering new polymer electrolytes ............... 4

1.4 Polymer informatics: An ML-driven discovery ecosystem .......... 5

1.5 Thesis objectives, outline and outcomes ............................ 7

Chapter 2: Methodology ................................................. 10

2.1 Dataset development ................................................. 10

2.1.1 Experimental datasets .................................... 11

2.1.2 Computational dataset .................................... 11

2.1.3 Proxy properties ........................................ 12

2.2 Fingerprinting .................................................... 13
2.3 Machine learning algorithms ............................................. 15
  2.3.1 Gaussian process regression ........................................ 15
  2.3.2 Multi-fidelity learning .............................................. 18
  2.3.3 Deep learning using neural networks ............................. 19

Chapter 3: Polymers as electrolytes – Diversity and requirements ........ 22
  3.1 Introduction ............................................................... 22
  3.2 Diversity and chemistry of polymers used in SPEs ................. 22
  3.3 Requirements of a polymer host in an SPE .......................... 25
  3.4 Summary ................................................................. 34

Chapter 4: Methodology for the rapid screening of polymer electrolytes .... 35
  4.1 Introduction ............................................................... 35
  4.2 Identifying relevant properties ....................................... 36
    4.2.1 Identifying proxies for the ionic conductivity ............... 37
    4.2.2 Finalized screening criteria ...................................... 44
    4.2.3 Dataset development .............................................. 45
  4.3 Machine learning methods ............................................. 47
    4.3.1 Polymer fingerprinting and featurization .................... 47
    4.3.2 Machine learning algorithms .................................... 47
  4.4 Polymer chemical space for screening .............................. 48
  4.5 Summary ................................................................. 48

Chapter 5: Surrogate models, screening and identifying new SPE candidates . 50
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3.1 Hildebrand model baseline – Gaussian process regression</td>
<td>73</td>
</tr>
<tr>
<td>7.3.2 Setup for SN v1 and SN v2</td>
<td>73</td>
</tr>
<tr>
<td>7.4 Results and discussion</td>
<td>79</td>
</tr>
<tr>
<td>7.5 Summary</td>
<td>82</td>
</tr>
<tr>
<td><strong>Chapter 8: Summary, impact and future work</strong></td>
<td>84</td>
</tr>
<tr>
<td>8.1 Summary</td>
<td>84</td>
</tr>
<tr>
<td>8.2 Future work</td>
<td>85</td>
</tr>
<tr>
<td>8.2.1 Generating ionic conductivity data using computational methods</td>
<td>85</td>
</tr>
<tr>
<td>8.2.2 Inverse design of SPEs with generative models</td>
<td>87</td>
</tr>
<tr>
<td>8.2.3 SPE retrosynthesis</td>
<td>87</td>
</tr>
<tr>
<td>8.2.4 Further developments for the solvent model</td>
<td>88</td>
</tr>
<tr>
<td>8.3 Impact – Polymer Genome online platform</td>
<td>88</td>
</tr>
<tr>
<td>8.4 Outlook</td>
<td>90</td>
</tr>
<tr>
<td><strong>Appendices</strong></td>
<td>92</td>
</tr>
<tr>
<td>Appendix A: Multi-fidelity model to predict the crystallization tendency of polymers</td>
<td>93</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>103</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1.1 Comparison of ionic conductivities of common electrolyte systems namely ceramics, liquids and polymers. ......................................................... 4

3.1 An extensive list of polymer properties which contribute to the development of a high-performing and safe solid polymer electrolyte. This table contains an overarching property type and associated polymer properties. A short summary of the purpose of the property type is also listed. ............................. 26

4.1 Dataset of 19 polymers, their corresponding ionic conductivity values and associated salt concentrations. .......................................................... 39

4.2 List of 13 proxies used instead of the ionic conductivity, the associated training dataset size, machine learning algorithms used and the cross-validation root mean square error (RMSE) on 100% of the training data. . . . 42

4.3 Final list of pertinent polymer properties and proxies for the ionic conductivity used to screen for new polymer electrolytes. Corresponding screening values and reason for inclusion in the ML workflow is listed. . . . 44

4.4 A summary of the down-selected polymer properties relevant for SPEs with their associated data set source and size. ............................................. 45

5.1 Complete list of 16 identified SPEs using our proposed ML workflow. The corresponding ML-predicted ESW, Tg, Td, LOI and TS values are also given. 53

6.1 Solvents and non-solvents with generic compounds and specific organic compounds for poly(ethylene oxide). ........................................... 60

6.3 Dataset for the assessment of the Hansen solubility model [113, 114] . . . . 62

7.1 Summary of the baselines and the models used in this work. . . . . . . . . . 72
7.2 Dataset details and comparison between SN v1 and SN v2. . . . . . . . . . 74
7.3 Fingerprinting methods and comparison between SN v1 and SN v2. . . . . 76
7.4 Neural network architecture and comparison between SN v1 and SN v2. . . 78

A.1 Representative features strongly correlated to the crystallization tendency of polymers. R represents an arbitrary chemical group of C, O, N or H elements.[143] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97
LIST OF FIGURES

1.1 Comparison of key physical and chemical features of common electrolyte systems namely ceramics, liquids and polymers. 3

1.2 Our proposed workflow for identifying new SPE materials. 7

2.1 Hierarchical fingerprints used to represent polymers in the machine learning-based modeling pipeline. 14

2.2 Representative figure of Gaussian process regression with associated prediction uncertainties. 16

2.3 Representative figure of Gaussian process regression with associated prediction uncertainties. 18

2.4 Representative figure of a fully connected neural network. 20

3.1 A visualization of the principal data set of 13,388 polymers where the positions of eight common solid polymer electrolytes are labeled. Color bars are used for encoding the fraction of sp3 bonded C atoms. The visualization was created by projecting the polymer data sets onto a 2D space spanned by PC0 and PC1, two first principal axes obtained by a principal component analysis. 24

3.2 Visualization of crystalline and amorphous sections of a polymer and associated ion conduction tendencies. 28

3.3 Energy diagram of the electrolyte interface with anode and cathode. ESW of the electrolyte is determined by its reduction and oxidation potentials, controlled by the conduction band maximum (CBM) and the valence band minimum (VBM), respectively. $\mu_A$ and $\mu_A$ are the electrochemical potentials of the anode and the cathode, respectively. 31
4.1 Proposed data-driven and machine learning based workflow for the rapid screening of a large chemical space to identify new solid polymer electrolytes. 35

4.2 Proxies for the ionic conductivity identified by correlating the ionic conductivity with several polymer properties predicted using prior ML models using a) Pearson correlation coefficients and b) relative importance analysis. [28] 43

4.3 Dataset distributions for the finalized properties used in the screening workflow for new SPE materials. 46

5.1 Visualized performance of ML models via parity plots for 5 polymer properties used to screen for new SPEs. RMSE CV , test is the average of the cross-validation test error of the models created when the 100% is used to train the model. These models are deployed on Polymer Genome. [27] 51

5.2 Plot of our machine learning-based workflow to screen for new polymer electrolytes is shown. An initial chemical space of 13,388 polymers is parsed to identify candidates which satisfy five property values illustrated in the figure. A partial list of proposed SPEs is illustrated. A complete list of proposed SPEs is presented in Table S2. 52

6.1 Solubility criterion of (a) the Hildebrand solubility parameter(δ), where the solvent Hildebrand values are within ± 2 MPa 1/2 of the polymer Hildebrand value (δ polymer) and the non-solvent Hildebrand values are over 2 MPa 1/2 and, (b) the Hansen solubility model, where the three axes represent the dispersion (δD), polar (δP) and hydrogen bonding (δH) components; solvents fall within a sphere of radius R = 8 MPa 1/2 and non-solvents outside the sphere. 58

6.2 Performance of the Hildebrand solubility machine-learning model developed using Gaussian process regression (GPR) for the prediction of unknown polymer Hildebrand solubility values. 63

6.3 Performance of the Hildebrand parameter for 75 polymers across polymer classes for solvents and non-solvents. 66

6.4 Performance of the Hildebrand and Hansen solubility parameters for 25 polymers for solvents and non-solvents. 67

7.1 Names of 58 solvents used in SN v2 along with their frequency of occurrence in the SN v2 dataset. 75
7.2 PCA plot of 58 solvents where the PC$_0$ and PC$_1$ components of the fingerprint vector are plotted. Similar polymers are clustered and marked on the figure. ................................................................. 77

7.3 Neural network architecture for the prediction of good solvents and non-solvents for polymers. ................................................................. 79

7.4 Per-class accuracy of the neural networks for the Hildebrand model, SN v1 and SN v2 for in-domain and out-of-domain test sets. The per-class accuracy of both, solvent and non-solvent classes are shown and the overall accuracy is presented. The error bars are from 5 statistical runs that vary the out-of-domain solvents. ................................................................. 80

7.5 Per-solvent average accuracy of the neural networks for the Hildebrand model, SN v1 and SN v2 for in-domain and out-of-domain test sets. The error bars are from 5 statistical runs that vary the out-of-domain solvents. 82

8.1 GUI-based polymer draw tool implemented in Polymer Genome. While a polymer is drawn, its smiles string is updated on-the-fly and when the drawing is done, it will be copied to the search tool by clicking button “Predict Properties.” The ladder polymer sketched in this figure is poly(naphthalene-2,3:6,7-tetrayl-6,7-dimethylene) ................................................................. 89

A.1 Graphical summary of the high and low fidelity datasets against the chemical space of polymers. Two leading components, PC1 and PC2, are produced by principal component analysis, and assigned to axes of the plot. 94

A.2 Prediction capability of the group contribution method for crystallization tendencies as a function of the high- and low-fidelity datasets. ................. 99

A.3 Prediction accuracy for machine learning models trained using single-fidelity GPR and multi-fidelity co-kriging methods. Part (a) comprises of learning curves for the single-fidelity GPR model and the multi-fidelity models with 407 and 429 low-fidelity points where the test sets consists of 22 randomly selected points and averaged over 50 runs, with error bars illustrating 1σ deviation. Figures (b) and (c) illustrate example parity plots with 107 high-fidelity train points and 22 test points for the GPR and co-kriging models respectively. The co-kriging model uses 429 low-fidelity points. ................................................................. 100
SUMMARY

With the prolific popularization and development of lithium-ion batteries, safety issues associated with the use of flammable organic electrolytes have increasingly garnered more attention. A promising alternative to organic liquid electrolytes are solid polymer electrolytes (SPEs) which demonstrate low flammability, good processability and no leakage issues.

However, presently known SPE candidates fall short of the required performance requirements, which are reliant on meeting a variety of material property requirements, such as polymer amorphicity, high ionic conductivities at room temperature, large electrochemical stability windows (4V vs Li+/Li), high Li ion transference, moderate tensile strength and thermal stability. Parsing the expansive polymer chemical space for viable SPE candidates which meet the aforementioned criteria is a non-trivial task.

My work involves the use of data-driven methods and machine learning methods to build predictive models of a variety of polymer properties relevant for the SPE application. Development of such predictive models require the collection and curation of the requisite data (from computational and experimental sources) for polymers spanning large enough chemical space, followed by the actual building of the machine learning models. These predictive models are then used to rapidly screen a large candidate space of 13,388 synthesizable polymers to identify new polymers and subsequent chemical rules for promising and reliable SPEs. Finally, we also develop models to assist in the synthesis of polymer electrolytes, specifically for identifying solvent and non-solvents for SPEs using machine learning methods to train a massive database. The work presented in this thesis highlights the ability of combining computational and experimental data with data-driven methods to accelerate the design and discovery of high-performing and reliable solid polymer electrolytes.
1.1 Towards safer Li-ion batteries

The 1970s oil embargo sparked the first wave of interest in alternative energy sources to develop advanced batteries for electric automobiles.[1] Today, energy requirements for varied industrial applications have propelled major research efforts towards developing materials for rechargeable batteries. With the rapid exhaustion of non-renewable fossil fuels and aggravation of environment problems, alternative energy sources are the need of the hour. Replacing gasoline with alternative sources in electric and hybrid electric vehicles in the automobile industry is an ongoing challenge area in the research community. Also, the recent rapid proliferation of handheld electronic and communication devices, such as smart phones, tablets, and laptop computers, has increased the demand for rechargeable batteries sharply. Along with this expanding market comes an increasing demand for improved battery performance: lighter weight, longer cycling life, faster recharging, and lower cost.

In 1976, Texas-based oil company Exxon published the first paper on a working lithium-ion battery (LIB) prototype. [2] Lithium, the lightest element of the periodic table has several advantages over its predecessor, the nickel-cadmium battery. Along with being much smaller and lighter, LIBs display high energy densities making it an ideal candidate for powerful and portable batteries. At the time, the battery was used to power digital watches; 50 years later, LIBs propelled a revolution in the electronics and transportation business. The most prominent successes of this technology is the popularization of super-slim handheld electronics and electric cars with significant mileage. Lithium ion batteries (LIBs) have several advantages, such as the high specific energy density, high working
voltage, low self-discharge rate, fast charge/discharge, long lifetime and no memory effect. However, LIBs, if utilized incorrectly, leads to potentially life-threatening safety issues which have been at the forefront of its drawbacks. [3, 4] Accidents related to fires and explosions of LIBs occur frequently worldwide, especially involving cell phones, laptops, electric vehicles, and airplanes. Some have been seriously threatening to human life and health and have led to numerous product recalls by manufacturers. For example, the Tesla electric car battery fire, the Boeing 787 Dreamliner battery problems and the Samsung Note 7 fires are some notable safety issues associated with LIBs. These incidents have grounded the need for safety to be of utmost priority while designing a lithium ion battery.

1.2 Lithium ion battery basics and associated safety issues

At the crux of LIBs’ safety issues lies the conventional liquid electrolyte which is inherently flammable. Usually, a lithium ion battery comprises of an anode, cathode, separator, electrolyte, and two current collectors viz. positive and negative. The anode and cathode store the lithium. The electrolyte carries positively charged lithium ions from the anode to the cathode and vice versa through the separator. The movement of the lithium ions creates free electrons in the anode which creates a charge at the positive current collector. The electrical current then flows from the current collector through a device being powered (cell phone, computer, etc.) to the negative current collector. The separator blocks the flow of electrons inside the battery. While the battery is discharging and providing an electric current, the anode releases lithium ions to the cathode, generating a flow of electrons from one side to the other. When plugging in the device, the opposite happens: Lithium ions are released by the cathode and received by the anode.

Typically, electrolytes consist of an organic solvent, typically a combination of ethylene carbonate, dimethyl carbonate, and diethyl carbonate, with a lithium salt, such as LiPF$_6$, LiBF$_4$, LiBOB [lithium bis(oxalato)borate], or LiClO$_4$. A thermal runaway reaction is a catastrophic exothermic reaction inside a battery initiated by a number of possibilities.
viz. overcharging, excessive external temperatures, internal wiring defects or short circuits. Since most conventional battery electrolytes are linear, organic liquids like the mixture of ethylene carbonate and dimethyl carbonate, the thermal runaway reactions causes them to combust and culminates in fires and explosions. Battery scientists and engineers have since developed several methodologies to make LIBs safer. These include methodologies like developing more stable electrode materials to prevent dendrite formation and evolution of combustible gases, functionalities to prevent overheating and heat accumulation like overcharge additives and thermo-responsive polymer switching (TRPS) as well as building more heat-resistant separators and electrolytes with flame-resistant additives. However, a definitive way to combat the flammability would be to eliminate the root cause – the liquid electrolyte and replace it with a solid electrolyte. [5, 6, 7, 8, 9]

Figure 1.1: Comparison of key physical and chemical features of common electrolyte systems namely ceramics, liquids and polymers.
Table 1.1: Comparison of ionic conductivities of common electrolyte systems namely ceramics, liquids and polymers.

<table>
<thead>
<tr>
<th>Electrolyte system</th>
<th>Ionic conductivity (S/cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ceramics</td>
<td>$10^{-3}$ - $10^{-5}$</td>
</tr>
<tr>
<td>Liquids</td>
<td>$10^{-2}$ - $10^{-4}$</td>
</tr>
<tr>
<td>Polymers</td>
<td>$10^{-4}$ - $10^{-7}$</td>
</tr>
</tbody>
</table>

Broadly, the three most common classes of electrolytes are ceramics, liquids and polymers. Comparison of key physical and chemical features of liquid, polymer and ceramic electrolytes are presented in Figure 1.1. Liquid electrolytes exhibit superior ionic conductivity while pure polymer electrolytes have the lowest ionic conductivity among the 3 classes. The ionic conductivities of liquid and solid-state electrolytes are presented in Table 1.1. While high-performing, liquid electrolytes and separators have associated acute safety issues elucidated above. On the contrary, solid-state electrolytes have superior mechanical and thermal stability making them safer than liquids. Additionally, they also eliminate the need for a separator thereby limiting the possibility of dendrite formation which is another safety concern for liquid electrolytes. Further, solid-state electrolytes, especially polymers are more electro-chemically stable than their liquid counterparts. These safety issues have made it imperative to explore and identify new electrolyte materials, in particular, solid-state electrolyte materials. [10, 11, 12]

1.3 The quest for discovering new polymer electrolytes

In 1973, Wright et. al. [13] published a seminal work outlining promising electrical conductivities of crystalline complexes of alkali metal salts with poly(ethylene oxide). Since, a gradual development of polymer electrolytes ensued and firmly established SPEs as a worthy alternative to organic liquids; a detailed timeline of the development of solid polymer electrolytes is elucidated in Zhou et. al.’s recent review.[14] Conventionally, a solid polymer electrolyte comprises of 2 parts, a polar neutral polymer host and a lithium salt. This system is commonly known as a double-ion conducting polymer electrolyte.
which serves as a medium for ion transport. While several variations of this system exist, this work will mainly focuses on the double-ion conducting system.

Surprisingly, in the four decades following the development of the first polymer electrolyte, the gold standard for polymer matrices for SPEs is still poly(ethylene oxide) (PEO). This can be attributed to its ability to form stable solid complexes that exhibit reasonable ionic conductivities at room temperature. Additionally, PEO is versatile during processing and can be used with ease as not just dry polymer electrolyte but as a part of a composite SPE with ceramic fillers or liquid plasticizers as well as gel co-polymer SPEs with other polymers. Another similar polymer host for SPEs is poly(propylene) oxide (PPO) which also belongs to the poly(ether) family. Some other common polymer hosts include polar polymers like poly(vinyl chloride), poly(acrylonitrile), poly(dimethyl siloxane), poly(methyl methacrylate), poly(vinylidene flouride), poly(vinyl acetate) and poly(ethylene imine). [15, 16, 17, 18, 19, 20]

The quest to find new polymer hosts is tricky; hosts need to possess threshold values of multiple properties such as polymer amorphicity, high ionic conductivities at room temperature, large electrochemical stability windows (4V vs Li+/Li), high Li ion transference, moderate tensile strength and thermal stability. However, finding new hosts experimentally is an elaborate and expensive process leading to unexplored polymer chemistries. In the last decade, data-driven methods have shown tremendous promise for polymer discovery given that they enable rapid, rationally-driven screening of large chemical spaces to aid and assist experimental work. Such methods could be very advantageous to finding new polymer hosts for SPEs.

1.4 Polymer informatics: An ML-driven discovery ecosystem

Past works on using data-driven and machine learning methods to discover new electrolytes have been staggered. However, data-driven methods for polymers have been on an uptick for the last decade. While polymers are ubiquitous owing to their on-demand physical and
chemical property manipulation, most research and development was experimentally-led until a decade ago. In 2011, the Materials Genome Initiative in 2011 spearheaded one of the first synergistic rapid materials development by combining theory, computations and experiments which saw successes in polymer science specifically self-assembly in polymer films.[21, 22] Subsequently, the use of computations and machine learning to efficiently develop, design and discover polymers, a domain colloquially known as polymer informatics has made tremendous strides.[23, 24, 25, 26]

Recently, there has been a surge of research activities in the design and discover of polymers using data-driven methods and machine learning. One example is Polymer Genome, an online polymer informatics driven platform which contains an amalgamation of over 20 predictive models for pertinent polymer properties like the glass transition temperature, bandgap and dielectric constants at various frequencies.[27] Different properties have varied sources, quantities and quality of datasets and associated performances. Properties developed on computational data usually show higher predictive accuracies than those which are collected on experimental data because of the associated noise in the experimentation process. However, the RMSE values, elucidated in Huan et. al.’s recent paper is accurate enough to be able to virtually screen through numerous possible candidates rapidly or to support and augment the experimental design of experiments.[28] This method has been successfully used in the past for identifying materials for polymer capacitors which require high band gaps and dielectric constants.

In the data-driven search for electrolytes, most large-scale past works have been focused on inorganic solid electrolytes mainly via the Electrolyte Genome project which used an automated workflow to calculate the electron affinities and ionization potentials of over 4830 potential electrolyte candidates which were then rapidly screened for new electrolyte materials.[29] On a similar note, Ahmad et al. screened over 12,000 inorganic solids to find six new stable electrolyte materials.[30] Within the polymer electrolyte realm, Liu et. al. recently presented a work focused on collecting a dataset and training a machine
learning model on ethylene/oxide and Li ion ratios.[31] Another work by Hatakeyama-Sato et. al. demonstrated an ML prediction scheme for the ionic conductivity and perimitivity of single-ion conducting polymers using a dataset of 240 experimentally determined values.[32] Xie et. al.’s recent work used MD simulations to screen for SPE candidates from 53362 hypothetically generated polymers using computed ionic conductivity values.[33] However, there has been no past work that focuses on screening a very large and diverse experimental polymer chemical space using a more holistic screening criteria to find high-performing and safe SPEs.

1.5 Thesis objectives, outline and outcomes

The primary objective of this work is to identify new double-ion conducting homopolymers which show promise as electrolyte materials for Li-ion batteries using data-driven approaches as well as develop ML models to help in the synthesis of the identified polymers particularly for predicting solvents and non-solvents. An overarching workflow of this work is shown in Figure 1.2. The following issues have been addressed in this thesis:

- Identify pertinent polymer properties and their associated threshold values required for a good SPE.
- Curate datasets and build surrogate ML models for these identified polymer properties.
- Use these surrogate models to screen a large polymer space.
- Propose new SPE candidates for Li-ion batteries.

Figure 1.2: Our proposed workflow for identifying new SPE materials.
• Identifying pertinent polymer properties that affect and influence solid polymer electrolyte (SPE) performance

• Developing a data-driven workflow with the aforementioned polymer properties to rapidly screen for robust and reliable SPEs.

• Assist in the synthesis of the identified SPEs particularly for solvents and non-solvents using data-driven methods

• Identify chemical rules and patterns governing the ML models to ensure scientifically backed rational design and discovery

These issues have been addressed in the following chapters in this manner:

• Chapter 2
  
  I provide a detailed explanation of the general workflow of developing predictive models and their associated theoretical frameworks. Although the background is well-documented in literature, given the interdisciplinary nature of this work which may be viewed by researchers from different scientific communities, I have provided background for the methods utilized.

• Chapter 3
  
  Chapter 3 details the identification of relevant descriptors and proxies for SPEs. I discuss the overall requirements of SPEs detailed by the community and assess and identify relevant properties I use to screen for new SPEs. Further, I also discuss the need for proxy properties for certain cases where dataset availability is limited. Following the identification, I present the curated datasets that will be used to build the predictive models.

• Chapter 4
  
  In this chapter, we select pertinent polymer properties and proxies relevant to SPE
design and curate datasets for them. Further, we also identify the pertinent screening criteria for associated polymer properties and determine a large screening space of 13,388 polymers from which we will determine our new SPE candidates.

• Chapter 5
In this chapter, I present a workflow and associated ML models using the models developed in the previous chapter to screen and identify new SPEs. The pertinent polymer properties for these polymers are presented along with chemical guidelines and intuitions to help polymer chemists synthesize the polymers.

• Chapter 6
This chapter elucidates models to help identify pertinent solvents and non-solvents for polymers including the identified SPEs. I demonstrate the performance of the historically popular single-value Hildebrand parameter and explain its merits and drawbacks.

• Chapter 7
Following the findings of the previous chapter where the need for a model based on only data surpassing the parameter is explained, I detail our methodology which overcomes these drawbacks by using a neural network model trained on a massive solvent and non-solvent dataset for polymers and its performance.

• Chapter 8
I summarize the findings of the work and detail its impact as a part of a large informatics platform called Polymer Genome. Further, I provide a broad outlook of the critical next steps and their promises and challenges for identifying safe electrolytes using data-driven methods.
CHAPTER 2
METHODOLOGY

The essential ingredients of a data-driven workflow can be divided into three distinct parts. First, systematic and continuous accumulation of (experimental and computational) polymer data is required. The datasets are acquired from a variety of literature sources or are generated using computations in a high-throughput and consistent manner. Following this, the polymer data is converted into machine readable form via a fingerprinting step that encodes features of the polymer at a variety of hierarchical length scales in a numerical fingerprint vector. The last step in this pipeline is the learning step, during which the polymer fingerprint vectors are mapped onto the corresponding polymer property values, using one of many algorithms; this step is essentially a function finding exercise, i.e., the best hypothesis function that links the fingerprint vectors and the property values is identified following robust statistical practices. The hypothesis function, one for each property for which data are available, constitutes a “surrogate” model for the instantaneous prediction of the corresponding polymer property.

2.1 Dataset development

The datasets developed in this thesis spans over a wide range of homopolymers. The chemical space of the polymers spans over a wide range but predominantly included the following atomic species: H, B, C, N, O, F, Si, P, S, Cl, Br, and I. Given that this work is he first of its kind, for this iteration, co-polymers, polymer blends, polymers with additives, and cross-linked polymers are not considered in this study and are critical next steps following this work.
2.1.1 Experimental datasets

The majority of the property prediction models in this thesis utilizes experimentally measured data. The experimental dataset is manually curated from a plethora of published works including published journals, printed handbooks, and online repositories.[34, 35, 36, 37, 38, 39] In some cases where the data is sparse, theoretical (group contribution-based) methods are used to append the existing data to make predictive models. The size of the dataset (discussed in Chapter 3) varies by property with the largest being the solvent/non-solvent dataset comprising of 6595 polymers followed by the glass transition temperature dataset with 5076 polymers to substantially smaller ones like the limiting oxygen index which comprises of 101 polymers.

2.1.2 Computational dataset

For certain electronic properties viz. the electrochemical stability window (ESW), the dataset was generated using density functional theory (DFT). Density-functional theory (DFT) is a computational quantum mechanical modelling method used in physics, chemistry and materials science to investigate the electronic structure (or nuclear structure) (principally the ground state) of many-body systems, in particular atoms, molecules, and the condensed phases. Using this theory, the properties of a many-electron system can be determined by using functionals, i.e. functions of another function. In the case of DFT, these are functionals of the spatially dependent electron density. DFT is among the most popular and versatile methods available in condensed-matter physics, computational physics, and computational chemistry. This theory has been reviewed vastly and is very reliable for ESW computations.

The ESW dataset was built computationally using density functional theory (DFT) on single-chain polymer structures and comprises of 3881 polymers. The bandgap was computed using the hybrid Heyd Scuseria Ernzerhof (HSE06) electronic exchange-correlation functional.[27, 28]
2.1.3 Proxy properties

In materials science, it is common to encounter properties which have limited experimental data where appending the dataset with theoretical or computational methods are not possible. In this work, the ionic conductivity is one such property. In such cases, proxy properties can be used instead of the actual property. To identify proxies, a correlation analysis is conducted with the ionic conductivity and several other pertinent polymer properties. Two analysis used in this work are the Pearson correlation co-efficient and the relative importance analysis.

**Pearson correlation analysis**

In statistics, the Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the co-variance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the co-variance, such that the result always has a value between -1 and 1. 1 corresponds to a strong positive correlation and -1 corresponds to a strong negative correlation.[40] As with co-variance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. The Pearson correlation ($r$) is given by:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2(y - m_y)^2}}$$  \hspace{1cm} (2.1)

where $m_x$ is the mean of the vector $x$ and $m_y$ is the mean of the vector $y$.

**Relative importance analysis**

Relative importance analysis is a useful technique to calculate the relative importance of predictors (independent variables) when independent variables are correlated to each other. It is an alternative to multiple regression technique and addresses multicollinearity while it calculates the importance rank of variables. It creates a set of new independent variables
that are the maximally related to the original independent variables but are uncorrelated to each other. Because these new transformed independent variables are uncorrelated to each other, the dependent variable can be regressed onto this new set of independent variables producing a series of standardized regression coefficients. [41]

The actual calculations are complex and are usually performed with software. Johnson et. al. outlines the general steps as:

- Transform the predictor variables into a set of orthogonal (uncorrelated) variables. These variables are maximally related to the predictor variables from an unweighted least squares perspective.

- Regress the dependent variables on the new set of transformed variables.

2.2 Fingerprinting

Polymer structures are three dimensional in structure and therefore, not machine readable as structure files. In practice, it is essential for them to be represented numerically via fingerprinting or descriptors. These fingerprints are usually representative of polymers’ structure and capture chemo-structural information while satisfying mathematical requirements like being invariant with respect to transformations.

A popular format of representing polymer repeat units is the simplified molecular-input line-entry system, commonly known as SMILES. While these were initially developed for molecules, they have now been extended to include polymer structures and are used as input for the fingerprinting process. In this work, all the polymers are linear and homopolymers i.e. the polymer repeat unit has two connecting points to make a long chain. While writing a SMILES string for complicated strings is cumbersome, a GUI-based polymer draw tool is implemented on www.polymergenome.org.

The SMILES is the starting point for the polymer fingerprinting which is then converted to numerical fingerprint vectors. The early version of the fingerprint was fairly rudimentary
in terms of scale and number of features but has now significantly expanded. The current fingerprinting scheme has up to 3000 fingerprint components on different length scaled to adequately capture the atomistic and the larger scale properties of the polymers. The fingerprinting scheme is illustrated in Figure 2.2. The finest-level components are atomic triples $A_iB_jC_k$, comprised of an $i$-fold coordinated atom of species $A$, a $j$-fold coordinated atom of species $B$, and a $k$-fold coordinated atom of species $C$, joined together in this order. At the next (block) level, pre-defined fragments such as cyclopentane and cyclohexane are identified from the polymers and then their occurrence is normalized in the fingerprint components. At the (highest) chain level, characteristic features of the polymers such as the length of the longest side chain, the distance between two specific blocks, etc. are captured.

This unique fingerprinting scheme was designed to capture a plethora of length-scale based structural relationships which in turn govern different polymer properties. For example, the glass transition temperature $T_g$ characterizes the processes that involve the motion of long polymers chains, thus the most relevant fingerprint components for $T_g$ should be some long length-scale features such as the length of the longest side chain.

Figure 2.1: Hierarchical fingerprints used to represent polymers in the machine learning-based modeling pipeline.
On the other hand, the ESW is essentially determined by the block-scale of the polymers like specific polar blocks and conjugations which affect this property. This fingerprinting scheme is expandable i.e. it can be changed to include specific features which affects certain polymer properties like the temperature and the frequency.

2.3 Machine learning algorithms

2.3.1 Gaussian process regression

Once a fingerprinting scheme is selected and all materials in a database have been fingerprinted, the regression problem is to find a function that takes a material’s fingerprint as input and returns its associated target property as output. The materials fingerprints and input used to fit the model is called the training data which is then used to map the function known as the surrogate model or the ML model. These models instantly predict new properties for unknown materials by first fingerprinting them and then using the surrogate model to make new predictions.

Several strategies for arriving at the mapping function exploit different regression algorithms, including linear regression, kernel ridge regression (KRR), GPR, decision trees, random forest, support vector machine regression and NNs. These algorithms mostly vary in the type and complexity of the function space that is searched to fit the training data. In this thesis work, we use Gaussian process regression since it is able to quantify the uncertainty associated with every prediction. An illustration of GPR learning algorithm is given in Figure 2.3.

Gaussian process regression (GPR) is a non-parametric, Bayesian approach to regression that is making waves in the area of machine learning. GPR has several benefits, working well on small datasets and having the ability to provide uncertainty measurements on the predictions.

Unlike many popular supervised machine learning algorithms that learn exact values for every parameter in a function, the Bayesian approach learns a probability distribution
over all probable values. Assuming an initial linear function: \( y = wx + \epsilon \), the Bayesian approach works by specifying a prior distribution, \( p(w) \), on the parameter, \( w \), and reassigning probabilities based on observed data using Bayes’ Rule:

\[
p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)} \tag{2.2}
\]

The updated distribution \( p(w|y,X) \), called the posterior distribution incorporates information from both the dataset and the prior distribution. The prior and likelihood is usually assumed to be Gaussian for the integration to be tractable. Using that assumption and solving for the predictive distribution, we get a Gaussian distribution, from which we can obtain a point prediction using its mean and an uncertainty quantification using its variance.
Gaussian process regression is non-parametric (i.e. not limited by a functional form), so rather than calculating the probability distribution of parameters of a specific function, GPR calculates the probability distribution over all admissible functions that fit the data. However, similar to the above, we specify a prior (on the function space), calculate the posterior using the training data, and compute the predictive posterior distribution on our points of interest.

In GPR, we first assume a Gaussian process prior, which can be specified using a mean function, $m(x)$, and co-variance function, $k(x, x')$:

$$f(x) \sim GP(m(x), k(x, x'))$$  \hspace{1cm} (2.3)

More specifically, a Gaussian process is like an infinite-dimensional multivariate Gaussian distribution, where any collection of the labels of the dataset are joint Gaussian distributed. Within this GP prior, we can incorporate prior knowledge about the space of functions through the selection of the mean and co-variance functions.

GPR with a rational quadratic kernel uses the aforementioned Bayesian framework, wherein a Gaussian process is used to obtain the mapping from the polymer to its associated % crystallinity based on the available training set and the Bayesian prior, incorporated using the kernel function. In this case, the kernel function between two materials with features $x$ and $x'$ is given by:

$$k(x, x') = \sigma^2 \left( 1 + \frac{(x - x')^2}{2\alpha l^2} \right)^{-\alpha}$$  \hspace{1cm} (2.4)

Here, the three hyper parameters $\sigma$, $l$ and $\alpha$ are the variance, the length-scale parameter and the expected noise in the data, respectively. These hyper parameters were determined during the training of the models by maximizing the log-likelihood estimate. Further, 5-fold cross validation was adopted to avoid overfitting. [42] In this thesis, the GPR package from scikit-learn was used to make the models.
For all models, the root mean square error (RMSE and R-squared) values were used to evaluate and compare the performances of the models. To estimate the prediction errors on unseen data, learning curves were generated by varying the size of the training and the test sets. Additionally, for each case, statistically meaningful results were obtained by averaging RMSE results over 50 runs with varying training and test splits.

2.3.2 Multi-fidelity learning

![Co-kriging model with multi-fidelity information fusion](image)

Figure 2.3: Representative figure of Gaussian process regression with associated prediction uncertainties.

This model is utilized for the crystallization tendency prediction where the high-fidelity dataset comprises of experimental data for the crystallization tendency and the low-fidelity dataset is calculated using the group contribution methods for the heat of fusion. The multi-fidelity information fusion model uses a co-kriging method to effectively utilize the high and low fidelity datasets simultaneously.[43, 44] The flexibility of the co-kriging approach allows it to have a variable number of low and high-fidelity data points. For all cases whose high-fidelity value is known, the respective low-fidelity value is also available. In analogy to GPR, the CK model assumes the high-fidelity data to be a realization of the Gaussian process $Z_{hi}$, which is further defined as the sum of a low-fidelity process $Z_{lo}$ scaled by a factor $\rho$ plus another independent Gaussian process $Z_d$ which captures the difference.
between the available low- and high-fidelity data points. Therefore,

\[ Z_{hi}(x) = \rho Z_{lo}(x) + Z_d(x) \]  

For all models, the root mean square error (RMSE and R-squared) values were used to evaluate and compare the performances of the GPR and the co-kriging models. To estimate the prediction errors on unseen data, learning curves were generated by varying the size of the training and the test sets. Additionally, for each case, statistically meaningful results were obtained by averaging RMSE results over 50 runs with varying training and test splits.[27, 44]

2.3.3 Deep learning using neural networks

Deep learning uses multiple layers of simple but nonlinear modules to transform raw data (such as element type or atomic positions) into a more abstract representation that can be used to learn complex functions (such as the potential-energy surface). Fully connected NNs are the most basic type of deep learning models, wherein multiple layers of neurons are combined to transform the data, with each neuron being a weighted combination of its respective inputs, followed by a nonlinear activation unit (Fig. 4d). In modern applications of NNs, the number of layers of a model can be very large to achieve a better representation power, hence, the name deep learning. Unfortunately, the higher representation power is also accompanied by a need for larger training datasets to attain good accuracy. The NN weights are learned by optimizing a loss function measuring the difference between the NN outputs and the desired targets. NNs can also be viewed as a model with nonlinear basis functions in which the basis functions are themselves learned from the data. By contrast, for other nonlinear models, such as kernel methods, the nonlinear basis functions are typically fixed beforehand and only weighted combinations of these basis functions are learned from the data. A representative figure of a simple fully connected NN is shown in Figure 2.4.

Briefly, a neural network is defined as a computing system that consist of a number of
simple but highly interconnected elements or nodes, called ‘neurons’, which are organized in layers which process information using dynamic state responses to external inputs. This algorithm is extremely useful, as we will explain later, in finding patterns that are too complex for being manually extracted and taught to recognize to the machine. In the context of this structure, patterns are introduced to the neural network by the input layer that has one neuron for each component present in the input data and is communicated to one or more hidden layers present in the network; called ‘hidden’ only due to the fact that they do not constitute the input or output layer. It is in the hidden layers where all the processing actually happens through a system of connections characterized by weights and biases (commonly referred as $W$ and $b$): the input is received, the neuron calculate a weighted sum adding also the bias and according to the result and a pre-set activation function (most common one is sigmoid, $\sigma$, even though it almost not used anymore and there are better ones like ReLu), it decides whether it should be ‘fired’ or activated. Afterwards,
the neuron transmit the information downstream to other connected neurons in a process called ‘forward pass’. At the end of this process, the last hidden layer is linked to the output layer which has one neuron for each possible desired output. The main strength of machine learning algorithms is their ability to learn and improve every time in predicting an output. But what does it mean that they can learn? In the context of neural networks, it implies that the weights and biases that define the connection between neurons become more precise; this is, eventually, the weights and biases are selected such as the output from the network approximates the real value $y(x)$ for all the training inputs.

In this work, a slightly modified version of the multilayer perceptron neural network algorithm is utilized for this particular classification (good solvent or nonsolvent) problem. This model was built using PyTorch. The neural network, as depicted in Figure 7.3, consists of two input branches, one for the solvent fingerprints and the other for the polymer descriptors. The solvent branch of the neural network consists of 3 hidden layers, each with 100 neurons, and the polymer branch of the neural network consists of three hidden layers, each with 100 neurons. The solvent net and polymer net hidden layers are merged using a concatenation operation and passed on to a final set of hidden layers. There are 3 final hidden layers, each with 100 neurons. All of the hidden layers in the neural network are constructed with the rectified linear unit (ReLU) activation function. The ADAM gradient descent method is utilized for training the neural network weights along with a training batch-size of 256. we used the softmax activation function which returns an output layer with size equal to the number of classes to be predicted (2 in this case) and each neuron comprising of probabilities associated with the occurrence of solvent or non-solvent for that particular solvent-polymer pair. The class with the higher softmax values is the predicted class. The reason we chose softmax over sigmoid is that this architecture can accommodate in-between partial classes in the future.
CHAPTER 3
POLYMERS AS ELECTROLYTES – DIVERSITY AND REQUIREMENTS

3.1 Introduction

Four decades ago, a seminal work outlining the potential of poly(ethylene) oxide as a solid polymer electrolyte in conjunction with alkali salts was published.[13] Since, a gradual development of polymer electrolytes ensued and firmly established SPEs as a worthy alternative to organic liquids. At its core, a conventional double-ion conducting SPE comprises of a lithium salt in a neutral, polar polymer host which facilitates as a medium for Li-ion transportation between the electrodes. Several modifications of this conventional SPEs exist but the polymer host remains a critical component of a solid state battery. In the following section, we first delve into the chemistries of well-known SPEs. After this, we present an in-depth examination of requisite performance and safety-related properties for a polymer electrolyte.

3.2 Diversity and chemistry of polymers used in SPEs

Since the development of the first polymer electrolyte, the most-widely used and yardstick for polymer matrices for SPEs is still poly(ethylene oxide) (PEO). Another well-studied and widely used polymer is poly(propylene oxide) which has similar chemical and structural behaviour. The reason for the preference of PEO and PPO as the polymer host is due to the fact that they usually form stable dry complexes exhibiting a relatively higher ionic conductivity than other solvating polymers. The oxyethylene group: –CH2–CH2–O–, and the polar groups: –O–, –H–, –C–H–, in the polymer chains have the ability to dissolve/complex the ionic salts. A double-ion conducting SPE with PEO involves PEO chains coiled around Li+ ions, separating them from counter-anions. This favours the
dissolution of the Li-salt in the PEO matrix following a solvating mechanism which is approximately akin to that in liquid electrolytes. [15, 16, 17, 45, 18, 19, 20]

Another common polymer class used in SPEs are nitriles, specifically, polyacrylonitrile (PAN) and succinonitrile. This chemical class contains the polar and electron withdrawing nitrile group, that exhibit have dielectric constants of about 30. They also have a large electrochemical window, because of their low lowest unoccupied molecular orbital (LUMO) band energy. Recently, nitrile-based SPEs have received considerable attention as a competitor to PEO/PPO because of their strong coordination ability and good electrochemical properties. However, these electrolytes undergo severe passivation upon contact with lithium metal anodes or lithiated graphite, and poor cathodic stability, which limits their further application. Modification strategies are expected to improve the properties of these composites, like the tailored structural design of polymer matrix and the addition of film-forming fillers embedded into the matrix framework.[15, 16, 17, 45, 18, 19, 20]

Polysiloxanes are known as silicones have been studied as SPEs. With advantageous properties like backbone flexibility, chemical stability, thermal stability, and low toxicity, these are promising SPEs given that a wide variety of substituents bound to the silicon in the backbone of the alternating silicon and oxygen atoms. When polar unites are introduced to these polymers, they exhibit very low $T_g$s, which in turn greatly enhances the ionic conductivity with their high free volumes and added chemical inertness. The conductivity of these polymers is close to the practical use, but these materials suffer from the low mechanical strength as most low $T_g$ polymers. Cross-linking or networking of polysiloxane could be an effective way to improve the mechanical stability. The most used polymer in this class for SPEs is poly(dimethyl siloxane) (PDMS).[15, 16, 17, 45, 18, 19, 20]

Halogen containing polymers have also been evaluated for their use as SPEs. Their widespread use is attributed mainly to the presence of lone pair electrons at the halogen atom where Li-ion salts can be solvated. The dipole–dipole interaction between the
hydrogen and halogen atoms, which can stiffen the polymer backbone and perhaps make the halogen–salt complexation difficult, can be weakened by the addition of plasticizers. Poly(vinyl chloride) and poly (vinylidene fluoride) are used as SPEs.

Poly(ethylene oxide) and polymethyl methacrylate

Polyvinyl acetate

Polyacrylonitrile

Polyvinyl chloride

Polyvinylidene fluoride

Polydimethyl siloxane

Polyethylene imine

Polymethyl methacrylate

Polyethylene oxide

Poly(vinylidene fluoride)

Polydimethyl siloxane

Polyethylene imine

Figure 3.1: A visualization of the principal data set of 13,388 polymers where the positions of eight common solid polymer electrolytes are labeled. Color bars are used for encoding the fraction of sp3 bonded C atoms. The visualization was created by projecting the polymer data sets onto a 2D space spanned by PC0 and PC1, two first principal axes obtained by a principal component analysis.

Polyester-based SPEs have been used as SPEs because of the strongly polar groups \([-\text{O– (C=O–O–)}\]. These polymers include polyethylene carbonate (PEC), poly(trimethylene carbonate) (PTMC), poly(caprolactone) (PCL) and poly (propylene carbonate) (PPC). They can efficiently dissolve alkali metal salts and reduce the tendency for the formation of ion aggregates. The sequential polar carbonate groups may enhance the dielectric constant and increase the ion conductivity in polymer electrolytes.
The number of commonly used SPEs are minuscule proportional to the number of synthesizable and viable polymer hosts. Most polymers mentioned above are obvious and ubiquitous choices of polar polymers. This can be attributed to the labor intensive and expensive task of experimentally parsing a large number of potential SPEs experimentally. Further, from Figure 1.1, it is evident that a potential SPE needs to satisfy several property requirements. Before we present our workflow, we expound on the host requirements for an SPE in detail.

### 3.3 Requirements of a polymer host in an SPE

The requirements of an SPE can be broadly bifurcated as i) those which contribute to enhanced Li-ion battery performance via increased conductivity and, ii) that contribute to structural integrity and overall stability of the SPE structure to perform as a ion transport medium as well as a separator. A comprehensive list of these properties are listed in Table 1. These requirements have been explored in several previous works.[15, 16, 17, 45] The two set of categories, explained in detail below, are inversely correlated subsequently making SPE selection a complex optimization challenge. For improved performance, the relevant property types are the ionic conductivity, polarity and the amorphicity. Table 3.1 details the specific polymer properties which contribute to these and are further explained below:
Table 3.1: An extensive list of polymer properties which contribute to the development of a high-performing and safe solid polymer electrolyte. This table contains an overarching property type and associated polymer properties. A short summary of the purpose of the property type is also listed.

<table>
<thead>
<tr>
<th>Property type</th>
<th>Associated polymer property</th>
<th>Purpose of the property type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conductivity and polarity</td>
<td>Ionic conductivity</td>
<td>Determines the performance of an electrolyte based on ion movement between electrodes.</td>
</tr>
<tr>
<td></td>
<td>Dielectric constant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Refractive index</td>
<td></td>
</tr>
<tr>
<td>Polymer amorphicity</td>
<td>Glass transition temperature</td>
<td>Increased amorphicity contributes to more free volume leading to easier paths for ions to move.</td>
</tr>
<tr>
<td></td>
<td>Crystallinity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td></td>
</tr>
<tr>
<td>Electrochemical properties</td>
<td>Electrochemical stability window</td>
<td>Controls the open circuit voltage of an electrolyte and the cycling life of a battery.</td>
</tr>
<tr>
<td></td>
<td>Ionization potential</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Electron affinity</td>
<td></td>
</tr>
<tr>
<td>Thermal stability</td>
<td>Thermal degradation temperature</td>
<td>Determines the upper limit of a battery’s heat and temperature operating conditions.</td>
</tr>
<tr>
<td></td>
<td>Limiting oxygen index</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Melting temperature</td>
<td></td>
</tr>
<tr>
<td>Mechanical properties</td>
<td>Tensile strength at break</td>
<td>Ensures the physical integrity of the electrolyte.</td>
</tr>
<tr>
<td></td>
<td>Young’s modulus</td>
<td></td>
</tr>
</tbody>
</table>
• **Conductivity:** The ionic conductivity is the most important performance metric of a polymer electrolyte. In SPEs, Li-ion diffusion is promoted by diffusive motion along the polymer chain, together with intersegmental hopping and the collective motion of the entire polymer chain with the coordinated ion. It is reported that the dependence on temperature of the ionic conductivity of an SPE follows two pertinent conduction mechanisms – the Vogel–Tamman–Fulcher (VTF) type and the Arrhenius type. The empirical VTF type conduction mechanism can be expressed by:

\[
\sigma = \sigma_0 e^{\frac{\theta}{T - T_{VF}}}
\]  \hspace{1cm} (3.1)

where \(\sigma_0\) is the pre-exponential factor, which is related to the number of charge carriers, \(\theta\) is the pseudo-activation energy for the conductivity, and \(T_0\) is the equilibrium glass transition temperature (\(T_0 = T_g - 50\) K). The VTF equation was devised for describing the diffusion process in glassy and disordered materials. For polymer electrolytes, plots of \(\sigma\) versus \(1/T\) are typically nonlinear, which is indicative of a mechanism of conductivity that involves an ionic hopping motion coupled with relaxation/breathing and/or segmental motion of polymeric chains. VTF behaviour is generally observed in SPEs above the \(T_g\).

The Arrhenius equation, on the other hand, is most useful for Materials that exhibit linear Arrhenius variations indicate that ion transport occurs via a simple hopping mechanism decoupled from polymer chain breathing. This equation is expressed by:

\[
\sigma = \sigma_0 e^{\frac{E_a}{RT}}
\]  \hspace{1cm} (3.2)

where \(E_a\) is the activation energy, which can be calculated from nonlinear least-squares fitting of the data from plots of \(\log \sigma\) versus \(1/T\).

The ionic conductivity for polymers is usually in the order of \(10^{-4}\) to \(10^{-7}\) S/m and is
Figure 3.2: Visualization of crystalline and amorphous sections of a polymer and associated ion conduction tendencies.

reported as a measure of the temperature and the Li salt concentration. Polymeric conductivity, like all diffusive processes, is a structure-dependent phenomenon like the polymer polarity and amorphicity. [46, 15, 47]

• **Polymer amorphicity:** Ionic conductivity in polymers is typically attributed to the amorphous phase of the polymer and subsequently, the free volume above the glass transition temperature ($T_g$) as illustrated in Figure 3.2. [46, 15, 47] The detailed mechanism of ionic conduction is as follows: lithium ions are located at suitable coordination sites (e.g., –O– in polyethylene oxide, –CN in polyacrylonitrile, –NR in polyamide, etc.) in polar chains of the polymer. The polymer chains undergo constant local segmental motion, which results in the appearance of free volumes. Lithium ions migrate from one coordination site to new sites along the polymeric
chains or hop from one chain to another through these free volumes under the effect
of an electric field.

There have been some experimental works that demonstrate superior ionic
conduction in the crystalline region. In certain cases, the ionic conductivity in
static, ordered environments of the crystalline phase can be greater than that in the
equivalent amorphous material above $T_g$. In the crystalline phase of P(EO)$_6$–LiX
pairs of PEO chains fold to form cylindrical tunnels, within which the Li$^+$ cations
are located and coordinated by ether oxygens, while the anions are located outside
these tunnels in the inter-chain space and do not coordinate the cations. The Li$^+$ ions
can migrate from one site to another along these cylindrical tunnels without the aid
of segmental motion.[48]

However, the general consensus is that the key to superior ionic conductivity
in SPEs is by increasing the amorphous regions in the polymer. Amorphicity
cannot be discussed without the percent crystallinity of a polymer. Percent
crystallinity is the extent of long-range order in a polymer and is a process-
dependent parameter indicative of polymer structure-property relationships. The
% crystallinity also reflects the proportions of amorphous regions in a polymer
and is therefore, correlated to the capacity of an SPE to conduct ions. It is
also a catch-all parameter representative of several inherent polymer properties
by the virtue of being positively correlated to the $T_g$, mechanical and thermal
properties of a polymer. Akin to $T_g$, decreasing the polymer’s inherent tendency
to crystallize increases the ionic conductivity. Decreasing the $T_g$, crystallinity and
density enhances the segmental flexibility of polymer chain which is turns improves
the ionic conductivity by promoting easier ion transfer through diffusive motions or
intersegmental hopping.[49] However, the amorphicity needs to be carefully tuned
with the more stabilizing properties (detailed later in this section) to ensure overall
functionality of an SPE.
• **Polymer polarity**: The polymer polarity, commonly expressed via its dielectric constant or refractive index, is an active area of interest given that it provides an independent and alternative means to modulate the ionic conductivity by overcoming the intrinsic effect of polymer amorphicity. For solid polymer electrolytes, the polymer matrix should be able to dissolve/complex lithium ions. Polymers with sequential polar groups such as –O–, –S–, –N–, –P–, C=O, may dissolve lithium salts and form polymer-salt complexes. Furthermore, to facilitate the dissociation of inorganic salts in polymer hosts, the dielectric constant of the host polymer should be relatively high.

Increasing polarity is proposed to increase the ionic conductivity via ease of disassociation of Li salts in the polymer host. Past works have shown that the relationship between the polarity and ionic conductivity to be complicated.[50, 51, 52] While Barteau et al.’s work found a direct correlation between the ionic conductivity and the polarity, Wheatle et. al.’s work found that this correlation to be true only in certain regimes and an overarching trend could not be established.[53]

While the electrolyte performance is crucial, structural integrity and overall stability is necessary for a reliable battery and therefore, electrochemical, thermal and mechanical properties have to be accounted for while designing an SPE. These properties must be carefully chosen for some of them that impart tensile and thermal stability are detrimental to the ionic conductivity since they are heavily dependent on the non-amorphous/crystalline regions of the polymer[15, 16, 17, 45]:

• **Electrochemical properties**: The electrochemical stability window (ESW) is an important consideration for SPEs, controlling its open circuit voltage and, subsequently, the cycle life of Li-ion batteries. The ESW of the electrolyte, illustrated in Figure 3.3 is determined by its reduction and oxidation potentials, which is governed by the conduction band maximum (CBM) and the valence band minimum.
(VBM), respectively. For an SPE, a polymer should have a large ESW width, i.e., the energy gap between CBM and VBM should exceed the energy difference of the electrochemical potentials of the anode (\(\mu_A\)) and cathode (\(\mu_C\)). Additionally, (\(\mu_A\)) and (\(\mu_C\)) should be within the energy gap of the polymer electrolyte, i.e., \(\text{CBM} > \mu_A\) and \(\text{VBM} < \mu_C\).[54, 55] Failure to meet these requirements can lead to the formation of a passivation layer at the electrode/electrolyte interfaces which can block the transfer of Li+ ions to the electrolyte, thereby, reducing the cycling life of the battery. Additionally, other considerations are the electron affinity (EA) and ionization potential (IP) which are potential energy changes to form anions and cations respectively. For certain polymer electrolytes, the EA and IP were found to be related to the CBM and VBM respectively.

![Energy Diagram of the Electrolyte Interface](image)

**Figure 3.3:** Energy diagram of the electrolyte interface with anode and cathode. ESW of the electrolyte is determined by its reduction and oxidation potentials, controlled by the conduction band maximum (CBM) and the valence band minimum (VBM), respectively. \(\mu_A\) and \(\mu_A\) are the electrochemical potentials of the anode and the cathode, respectively.

- **Thermal stability:** SPEs have a distinct advantage of being more thermally stable than liquid electrolytes. The thermal stability can be quantified via thermal degradation temperature (\(T_d\)), melting temperature (\(T_m\)) and the limiting oxygen index (LOI).
The $T_m$ quantifies the temperature at which the polymer disorders. While the $T_g$ is the property of the amorphous region of the polymer, $T_m$ is characteristic of the crystalline regions. In thermodynamics, the phase changes during $T_g$ and $T_m$ transitions are described as second and first order transitions respectively. The semi-crystalline polymer shows both the transitions corresponding to their crystalline and amorphous regions. $T_m$ increases in the presence of double bonds, aromatic groups, bulky or large side groups are present in the polymer chain, because they restrict the flexibility of the chain. The branching of chains causes the reduction of melting point, as defects are produced because of the branching. [56, 57]

Thermal degradation temperature ($T_d$) can present an upper limit to the service temperature of SPEs as much as the possibility of mechanical property loss. In polymers, $T_d$ refers to damaging chemical changes which take place at elevated temperatures, without the simultaneous involvement of other compounds such as oxygen. Even in the absence of air, polymers will begin to degrade if heated high enough. Unlike thermal-oxidation, which can usually take place at less elevated temperatures, $T_d$ is much higher. The onset of thermal degradation dictates the maximum usable temperature for polymers. Three main reactions during nonoxidative thermal degradation of polymers are depolymerization, random chains scission, and side-group elimination. In some cases, a recombination process that leads to cross-linking or cyclization may occur during degradation. Usually, more than one degradation mechanism take place simultaneously. At high temperatures, the components of the long chain backbone of the polymer can break (chain scission) and react with one another (cross-link) to change the properties of the polymer. These reactions result in changes to the molecular weight (and molecular weight distribution) of the polymer and can affect its properties by causing reduced ductility and increased embrittlement, chalking, scorch, colour changes, cracking and general reduction in most other desirable physical properties. [56, 57]
A measure of the polymer flammability is the limiting oxygen index (LOI). Limiting oxygen index (LOI) is the minimum concentration of oxygen in a mixture of oxygen and nitrogen that is needed to support the flaming combustion of a material. It is expressed in volume percent (vol%). Materials with LOI values less than 21% are classified as combustible, but those with LOI greater than 21% are classed as self-extinguishing since their combustion cannot be sustained at ambient temperature without an external energy contribution. Materials with a high LOI value generally exhibit a better flame resistance. This method remains one of the most important screening for polymers to characterize both the ignitability and flammability resistance. [56, 57]

- **Mechanical properties:** The mechanical limits of an SPE is an important consideration given the multi-component nature of a lithium-ion battery. For this property, we consider the Young’s modulus and tensile strength at break.

Young’s modulus (E) is defined as the ratio of the stress applied to the material along the longitudinal axis of the specimen tested and the deformation or strain, measured on that same axis. The Young’s modulus \((E)\) is a measure of the polymer stiffness and is given by:

\[
E = \frac{\sigma}{\epsilon}
\]  

(3.3)

where \(\sigma\) is the tensile stress and \(\epsilon\) is the tensile strain. When a stretching force (tensile stress) is applied to an object, it extends, and its behavior can be obtained using stress-strain curve in the elastic deformation region from Hooke’s Law. The extension that a force produces is not only dependent on the material but also on other factors like dimensions of the object (e.g. length, thickness etc.). High dimensional stability guarantees processing feasibility and the isolation of the positive and negative electrodes.
3.4 Summary

In summary, the state-of-the-art polymers used as electrolytes were presented and an exhaustive list of all the properties that affect the performance of a good solid polymer electrolyte from the conductivity as well as thermal, mechanical and electrical stability have been discussed. Some interesting takeaways from this analysis are:

- The number of polymers commonly used as SPEs are minuscule compared to the vast number of synthesizable polymers. This can be attributed to the fact that finding new polymers is not a trivial task and experimentally parsing the polymer chemical space for new polymers is tedious.

- We identified 13 different properties that are useful for SPEs. Some contribute to the ionic conductivity (\(T_g\), dielectric constant, crystallization tendency) while some are necessary for the SPE integrity (electrochemical stability window, tensile strength, \(T_d\)). These two sets of properties work against each other since the conductivity is a function of the polymer amorphicity which in turn reduces the mechanical and thermal properties. Finding new SPEs is a complex optimization challenge.
CHAPTER 4
METHODOLOGY FOR THE RAPID SCREENING OF POLYMER ELECTROLYTES

4.1 Introduction

I. Identification of relevant properties and screening criteria
   Dataset curation
   Experimental data sources
   Computational data generation
   Integrated dataset for SPE screening and screening criteria

II. Machine learning modeling
   Hierarchical fingerprinting
   Model training using Gaussian process regression
   Machine learning surrogate models for screening

III. Screening for new polymer electrolyte candidates
   Decision on candidate space for screening
   Screen using screening criteria and surrogate models
   Identify new SPE candidates

Figure 4.1: Proposed data-driven and machine learning based workflow for the rapid screening of a large chemical space to identify new solid polymer electrolytes.

In the previous chapter, we found that the requirements of an SPE can be broadly bifurcated as i) those which contribute to enhanced Li-ion battery performance via increased conductivity and, ii) that contribute to structural integrity and overall stability of the SPE structure to perform as a ion transport medium as well as a separator. The two set of categories are inversely correlated subsequently making SPE selection a complex optimization challenge. We propose a machine learning strategy to tackle this problem to
find new SPEs.

The workflow of our proposed machine learning strategy is demonstrated in Figure 4.1. The workflow can be broadly demarcated into 3 steps of which the first two are presented in this section. In the first step, we determined a subset of pertinent properties presented in Table 3.1 that will be used in this work and curate datasets for them. We also discuss the concept of proxies for properties with sparse datasets and identify those. The second step is to train and develop machine learning surrogate models which will be used to screen a large polymer chemical space. The third step involved finalizing the screening criteria which will then be used to identify new SPE candidates. The machine learning models, screening and the proposed SPE candidates with relevant analysis are contained in Chapter 5.

4.2 Identifying relevant properties

An extensive list and discussion of relevant polymer descriptors was discussed in Chapter 3. However, in practice, the entire list is not used in tandem given that some features are more important or supersede others in practical use. Therefore, a subset of the properties in Table 1 need to be down-selected for the purpose of the work.

There are 6 properties we chose to retain from the larger dataset. From a performance perspective, the ionic conductivity and glass transition temperature were retained. While the ionic conductivity is the main performance metric for a battery, the glass transition temperature has been corroborated by numerous past works as a factor that affects the primary ion diffusion mechanism for a lithium ion battery electrolyte. While the polymer polarity is an active area of research for its contribution to battery performance, the trends for the dielectric constant and refractive index have not been firmly established and therefore, is not used in this work. Also, since the density is directly correlated to the $T_g$, we chose to forego this property as well.

From a safety and reliability point of view, the four properties retained were the
electrochemical stability window (ESW), the thermal degradation temperature \(T_d\), the limiting oxygen index (LOI) and the tensile strength at break. These properties span a good deal of breadth to ensure battery safety.

4.2.1 Identifying proxies for the ionic conductivity

For the ionic conductivity of double-ion conducting polymer hosts, we chose to curate data at room temperature which is the Li-ion battery utilization temperature and the Li salt as lithium perchlorate \((\text{LiClO}_4)\) since via our literature survey, we found this dataset to be the largest. With these parameters in mind, the dataset for the ionic conductivity comprises of only 19 polymers with a total of 33 salt concentration dependent ionic conductivity values. The dataset with the polymers and corresponding salt concentration is presented in Table 4.1. The lack of data can be attributed to the fact that most prior work use only a select number of hosts most predominantly PEO and pure polymer systems comprising of just the polymer and Li-salt are rare since they do not meet performance standards. The electrolyte performance is enhanced using active and passive filler materials, varying compositions of ingredients and adding plasticizers. While the total number of polymers are limited, the dataset is chemically diverse. Along with the well-known polyoxides (PEO, PPO), this dataset also contains polacrylics (PMMA, PAN), polyimines (PEI) and polyvinyls (PVA, PVC). Given the paucity of data, to incorporate the ionic conductivity contribution for our screening criteria, we propose using proxy polymer properties that affect the ionic conductivity. We propose a similar approach for this work by identifying proxies for the ionic conductivity using the large set of polymer properties identified in Table 3.1.

In Table 4.1., we present the curated dataset for the ionic conductivity with 19 polymers of varying chemistries and LiClO\(_4\) as the Li salt. Given the paucity of data, a reliable machine learning model cannot be built for the ionic conductivity. Therefore, to include ionic conductivity contributions, we identified proxies i.e. polymer properties which could be used in lieu of the the property. Some of these datasets were developed particularly for
this work ($T_d$, limiting oxygen index, crystallization tendency) but the rest are a part of an integrated Polymer Genome effort and details of the datasets, models and their subsequent applications are summarized in Tran et. al’s comprehensive paper.[28] To identify proxies, we conducted a correlation analysis of the ionic conductivity dataset with 13 previously identified and relevant polymer properties in Table 3.1.

For all the properties except the electrochemical stability window and the crystallization tendency, the datasets are entirely experimental from different sources mentioned in Section 2.1. The experimental dataset is manually curated from a plethora of published works including published journals, printed handbooks, and online repositories.[34, 35, 36, 37, 38, 39] The size of the dataset (discussed in Chapter 3) varies by property with the largest being the solvent/non-solvent dataset comprising of 6595 polymers followed by the glass transition temperature dataset with 5076 polymers to substantially smaller ones like the limiting oxygen index which comprises of 101 polymers. For the crystallization tendency, where the data is sparse, theoretical (group contribution-based) methods are used to append the existing data to make predictive models.
Table 4.1: Dataset of 19 polymers, their corresponding ionic conductivity values and associated salt concentrations.

<table>
<thead>
<tr>
<th>Index</th>
<th>Polymer</th>
<th>Conductivity (S/cm)</th>
<th>Salt conc. (%)</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2-(2-methoxyethoxy)ethanol functionalized poly(dichlorophosphazene)</td>
<td>1.70E-05</td>
<td>4%</td>
<td>[58]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.10E-05</td>
<td>12.50%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.70E-05</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.60E-05</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.30E-05</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poly(ethylene oxide)</td>
<td>1.00E-07</td>
<td>12.50%</td>
<td>[58]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.00E-07</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Poly(3-hydroxypropionic acid)</td>
<td>8.00E-07</td>
<td>1%</td>
<td>[59]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.00E-06</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.00E-06</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.00E-06</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Poly(ethylene pimelate)</td>
<td>5.00E-06</td>
<td>8%</td>
<td>[60]</td>
</tr>
<tr>
<td>5</td>
<td>Poly(methylhydroquinone-alt-4,4’-[oxybis(ethyleneoxy ethyleneoxy)] dibenzoic acid)</td>
<td>3.00E-06</td>
<td>4%</td>
<td>[61]</td>
</tr>
<tr>
<td>6</td>
<td>Poly(oxydiethylene pimelate)</td>
<td>6.00E-06</td>
<td>8%</td>
<td>[60]</td>
</tr>
<tr>
<td>7</td>
<td>Poly(propane-1,2-diol)</td>
<td>5.00E-07</td>
<td>12.50%</td>
<td>[62]</td>
</tr>
<tr>
<td>8</td>
<td>Poly(vinyl alcohol)</td>
<td>1.30E-09</td>
<td>1%</td>
<td>[63]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.00E-09</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.10E-08</td>
<td>15%</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.1 continued

<table>
<thead>
<tr>
<th></th>
<th>Polymer Name</th>
<th>Zeta Potential (mV)</th>
<th>Log d (nm)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Poly(vinyl butyral)</td>
<td>5.80E-10</td>
<td>12.50%</td>
<td>[64]</td>
</tr>
<tr>
<td>10</td>
<td>Poly(2,2’-[ethylenebis(oxyethylene oxy)]ethanol)-alt-(3,3’-sulfanediylpropanoic acid))</td>
<td>2.00E-05</td>
<td>5.50%</td>
<td>[60]</td>
</tr>
<tr>
<td>11</td>
<td>Poly[(3,3’-sulfanediylpropanoic acid)-alt-(2,2’-oxydiethanol)]</td>
<td>3.00E-06</td>
<td>5.50%</td>
<td>[60]</td>
</tr>
<tr>
<td>12</td>
<td>Poly[(3,3’-sulfanediylpropanoic acid)-alt-(3,6,9-trioxaundecane-1,11-diol)]</td>
<td>7.00E-06</td>
<td>5.50%</td>
<td>[60]</td>
</tr>
<tr>
<td>13</td>
<td>Poly[(3,3’-sulfanediylpropanoic acid)-alt-(ethylene glycol)]</td>
<td>7.00E-06</td>
<td>5.50%</td>
<td>[60]</td>
</tr>
<tr>
<td>14</td>
<td>Poly[(3,3’-sulfanediylpropionic acid)-alt-(triethylene glycol)]</td>
<td>5.00E-06</td>
<td>5.50%</td>
<td>[60]</td>
</tr>
<tr>
<td>15</td>
<td>Poly[2-(2-methoxyethoxy)ethyl methacrylate]</td>
<td>2.00E-08</td>
<td>3.30%</td>
<td>[65]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.00E-08</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.50E-07</td>
<td>8.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.60E-06</td>
<td>10%</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.1 continued

| 16 | Poly[\text{oxy}(1-chloromethylethylene)oxyethylene] | 7.00E-07 | 1\% | [66] |
| 17 | Polyacrylonitrile | 1.58E-09 | 15\% | [60] |
| 18 | Polyvinylacetate | 7.60E-07 | 10\% | [67] |
|   |                 | 2.90E-06 | 15\% | |
|   |                 | 6.20E-05 | 20\% | |
| 19 | Poly[\text{oxyethylene} oxymethylene-[5-(octadecyloxy)-1,3-phenylene]methylene] | 3.00E-08 | 50\% | [68] |

For the crystallization tendency model, we utilized two datasets, viz., a high-fidelity, experimental dataset for 107 polymers and a more diverse, less accurate low-fidelity dataset for 429 polymers which used group contribution methods. The high-fidelity dataset comprises of 107 polymers with associated \% crystallinity values. All values are curated from existing sources of experimental measurements like handbooks, published papers and online sources. [69, 34, 35, 36, 37, 38, 39] The ESW dataset, as mentioned previously, was previously calculated in Tran et al.’s work calculated using density functional theory (DFT).[28]

For each polymer contained in the ionic conductivity dataset, corresponding potential proxy values were predicted using machine learning models. All models except the crystallization tendency were developed using Gaussian process regression (GPR). The crystallization tendency used a multi-fidelity information fusion method called co-kriging. The details of the ML models including the test cross-validation error and the machine learning algorithms, presented in a previous work, are summarized in Table 4.2 and the
Table 4.2: List of 13 proxies used instead of the ionic conductivity, the associated training dataset size, machine learning algorithms used and the cross-validation root mean square error (RMSE) on 100% of the training data.

<table>
<thead>
<tr>
<th>Property type</th>
<th>Associated polymer property</th>
<th>ML algo.</th>
<th>Dataset size</th>
<th>RMSE&lt;sub&gt;CV&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conductivity and polarity</td>
<td>Dielectric constant</td>
<td>GPR</td>
<td>1193</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Refractive index</td>
<td>GPR</td>
<td>516</td>
<td>0.04</td>
</tr>
<tr>
<td>Polymer amorphicity</td>
<td>Glass transition temperature</td>
<td>GPR</td>
<td>5076</td>
<td>18.8 K</td>
</tr>
<tr>
<td></td>
<td>Crystallinity</td>
<td>CK</td>
<td>429/107</td>
<td>8.38%</td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td>GPR</td>
<td>890</td>
<td>0.03 g/cc</td>
</tr>
<tr>
<td>Electrochemical property</td>
<td>Electrochemical stability window</td>
<td>GPR</td>
<td>3881</td>
<td>0.26 eV</td>
</tr>
<tr>
<td></td>
<td>Ionization potential</td>
<td>GPR</td>
<td>371</td>
<td>0.21 eV</td>
</tr>
<tr>
<td></td>
<td>Electron affinity</td>
<td>GPR</td>
<td>371</td>
<td>0.18 eV</td>
</tr>
<tr>
<td>Thermal stability</td>
<td>Thermal degradation temperature</td>
<td>GPR</td>
<td>3545</td>
<td>28.03 K</td>
</tr>
<tr>
<td></td>
<td>Limiting oxygen index</td>
<td>GPR</td>
<td>101</td>
<td>3.73%</td>
</tr>
<tr>
<td></td>
<td>Melting temperature</td>
<td>GPR</td>
<td>2084</td>
<td>27.1 K</td>
</tr>
<tr>
<td>Mechanical properties</td>
<td>Tensile strength at break</td>
<td>GPR</td>
<td>672</td>
<td>4.75 MPa</td>
</tr>
<tr>
<td></td>
<td>Young’s modulus</td>
<td>GPR</td>
<td>629</td>
<td>120 MPa</td>
</tr>
</tbody>
</table>

Details about the algorithms are detailed in Section 2.3.

To find correlations, we use 2 methods. First, we use the Pearson correlation coefficient which analyses the relationship between the ionic conductivity and the proxy and assigns a value between -1 and 1. 1 corresponds to a strong positive correlation and -1 corresponds to a strong negative correlation.[40] Next, we use the relative importance analysis where we calculate the contribution a variable, in this case the potential polymer proxies makes to the prediction of a criterion variable by itself and in combination with other predictor variables. [41] The results of these correlations are presented in Figure 4.2.

From the figures, we find a significant similarity between the correlation analysis from both methods. From Figures 4.2.a. and b., we see that between the two methods, the
polymer properties which contributed the most to the ionic conductivity were the glass transition temperature ($T_g$), tensile strength at break and electrochemical stability window (ESW). These polymer properties align with the SPE host requirements in Section 3.1. However, we observe that the absolute values of the Pearson correlations are low (less than 0.6) and this could be because this analysis does not account for Li salt contribution. The polymer host is simply a medium for salt transport and our analysis focuses on the relationship between the host and ionic conductivity which past works have shown to be tenuous. However, the Pearson correlation analysis accurately shows that the ionic conductivity is inversely correlated to the $T_g$ since an increase in the $T_g$ would correspond to a decrease in the amorphous, ion-transporting regions of the polymer. Also, we see that the tensile strength and ESW are also accurately inversely correlated to the conductivity and amorphicity of a polymer. Therefore, in lieu of the ionic conductivity, for screening
purposes, we utilized the $T_g$, ESW and the tensile strength of the SPE host. A critical next step would be to create a reliable and diverse dataset for the SPE conductivity given that the existing literature is fairly fragmented in terms of SPE compositions to curate such a dataset.

Table 4.3: Final list of pertinent polymer properties and proxies for the ionic conductivity used to screen for new polymer electrolytes. Corresponding screening values and reason for inclusion in the ML workflow is listed.

<table>
<thead>
<tr>
<th>Polymer properties and proxies</th>
<th>Screening criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glass transition temperature</td>
<td>$&lt;213$ K</td>
<td>Increases amorphicity and negatively correlated to the ionic conductivity</td>
</tr>
<tr>
<td>Electrochemical stability window</td>
<td>$&gt;4$ eV vs Li/Li+</td>
<td>Ensures electrochemical stability and is negatively correlated to the ionic conductivity</td>
</tr>
<tr>
<td>Thermal degradation temperature</td>
<td>$&gt;423$ K</td>
<td>Ensures thermal stability of the electrolyte</td>
</tr>
<tr>
<td>Limiting oxygen index</td>
<td>$&gt;21%$</td>
<td>Quantifies inflammability of the electrolyte</td>
</tr>
<tr>
<td>Tensile strength at break</td>
<td>$&gt;25$ MPa</td>
<td>Maintains structural integrity and negatively correlated to the ionic conductivity</td>
</tr>
</tbody>
</table>

4.2.2 Finalized screening criteria

The next step is to ascertain the precise values to use to screen our polymer chemical space. The corresponding values for each property we used are summarized in Table 4.3. For most criteria, we choose polyethylene oxide (PEO) as the benchmark polymer to surpass. Therefore, the required $T_g$ for our screened polymers is lesser than 213 K. The ESW value required is greater than 4 eV which corresponds to several studies on PEO-based SPEs using linear sweep voltammetry. The thermal degradation temperature criteria of greater
than 213 K is based off of Yue et al’s recommendation for SPEs.[12] Tensile strength at break for PEO usually varies from 3.9 to 16.6 MPa, therefore, greater than 25 MPa was chosen to be the required value for the screened polymers. Finally, a LOI of 21% was chosen since at this temperature, under atmospheric conditions, a polymer is fairly inflammable.

4.2.3 Dataset development

From Table 4.3., we now have a finalized list of polymer properties which we want to develop ML models for. The first step here is to develop datasets for these models. For this work, the data for T\textsubscript{g}, T\textsubscript{d}, LOI and tensile strength are experimental and are manually curated from a myriad of published works and repositories. The ESW was computationally generated in a previous work by Tran et al. using density function theory (DFT).[27, 28] The dataset sizes and their sources are summarized in Table 4.4. A distribution of these datasets are presented in Figure 4.3.

Table 4.4: A summary of the down-selected polymer properties relevant for SPEs with their associated data set source and size.

<table>
<thead>
<tr>
<th>Property type</th>
<th>Associated polymer property</th>
<th>Dataset source</th>
<th>Dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polymer amorphicity</td>
<td>Glass transition temperature</td>
<td>Experimental</td>
<td>3545</td>
</tr>
<tr>
<td>Electrochemical properties</td>
<td>Electrochemical stability window</td>
<td>Computational via DFT</td>
<td>3881</td>
</tr>
<tr>
<td>Thermal stability</td>
<td>Thermal degradation temperature</td>
<td>Experimental</td>
<td>5076</td>
</tr>
<tr>
<td></td>
<td>Limiting oxygen index</td>
<td>Experimental</td>
<td>101</td>
</tr>
<tr>
<td>Mechanical properties</td>
<td>Tensile strength at break</td>
<td>Experimental</td>
<td>672</td>
</tr>
</tbody>
</table>

For the LOI, T\textsubscript{d}, T\textsubscript{g} and tensile strength at break, experimental datasets of varying sizes are curated from several sources.[34, 35, 36, 37, 38, 39] In cases where multiple values
for a polymer are reported, the median value is calculated and used for training the ML model. Of these datasets, the \( T_d \) and \( T_g \) are the most diverse comprising of 3545 and 5076 polymers respectively. The diversity of this dataset can be attributed to the fact that both properties are commonly measured and widely reported polymer properties. The LOI is the least diverse dataset comprising of 101 polymers given the specific nature of this property. An in-depth description of these datasets are presented in a previous work.[28]

The ESW dataset was built computationally using density functional theory (DFT) on single-chain polymer structures and comprises of 3881 polymers. The bandgap was computed using the hybrid Heyd Scuseria Ernzerhof (HSE06) electronic exchange-correlation functional and was reported in previous works.

Figure 4.3: Dataset distributions for the finalized properties used in the screening workflow for new SPE materials.
4.3 Machine learning methods

4.3.1 Polymer fingerprinting and featurization

A hierarchical fingerprinting method was used to capture descriptors that control the polymer properties and was described in a previous work. The fingerprinting scheme comprises of four hierarchical levels of descriptors. The first level is at the atomic scale comprising of the count of atomic triples (e.g., O2-C3-C4, representing of a two-fold coordinated oxygen, a three-fold coordinated carbon, and a four-fold coordinated carbon). The second set of fingerprint components captures a population of pre-defined chemical building blocks such as cyclopentane and cyclohexane. The final fingerprinting level includes morphological features such as the topological distance between rings, fraction of atoms that are part of side chains and length of the largest side chain. This multilevel fingerprinting scheme is one of the first of its kind and is able to numerically represent varying levels of polymer properties.[27, 28]

4.3.2 Machine learning algorithms

For the glass transition temperature ($T_g$), limiting oxygen index (LOI), tensile strength (TS), thermal degradation temperature ($T_d$) and the electrochemical stability window, Gaussian process regression was used to map the polymers to the property.

Gaussian process regression (GPR) with a rational quadratic kernel uses a Bayesian framework, wherein a Gaussian process is used to obtain the mapping from the polymer to its associated polymer property based on the available training set and the Bayesian prior, incorporated using the kernel function. In this case, the kernel function between two materials with features $x$ and $x'$ is given by:

$$k(x, x') = \sigma^2 \left( 1 + \frac{(x - x')^2}{2\alpha l^2} \right)^{-\alpha}$$  (4.1)

Here, the three hyper parameters $\sigma$, $l$ and $\alpha$ are the variance, the length-scale parameter
and the expected noise in the data, respectively. These hyper parameters were determined during the training of the models by maximizing the log-likelihood estimate. [42]

4.4 Polymer chemical space for screening

To screen for new SPE candidates, a wide polymer chemical space is required. For this work, we use a chemical space comprising of 13,388 synthesizable, double-ion conducting homopolymers which have been curated from numerous literature sources. The chemical space of the polymers spans over a wide range but predominantly included the following atomic species: H, B, C, N, O, F, Si, P, S, Cl, Br, and I. While some previous works have used larger hypothetical polymers to screen for new candidates, no previous work has used such a diverse screening space to search for new polymer electrolyte materials. [31, 33]

As mentioned previously, the most common polymer hosts include 9 polar polymers like polyethylene oxide (PEO), polypropylene oxide (PPO), poly(vinyl chloride), poly(acrylonitrile), poly(dimethyl siloxane), poly(methyl methacrylate), poly(vinylidene flouride), poly(vinyl acetate) and poly(ethylene imine). From this plot, two things are evident – first, that the total number of commonly used SPEs is a minuscule fraction of the polymer space and is severely unexplored and second, that the size of the polymer chemical space is difficult to navigate experimentally. Therefore, with our rationally-driven machine learning based approach, we show a methodology to rapidly screen this chemical space and present promising SPE candidates.

4.5 Summary

In this chapter, we select pertinent polymer properties and proxies and curate datasets for them. Further, we also identify the pertinent screening criteria for associated polymer properties and determine a large screening space of 13,388 polymers from which we will determine our new SPE candidates. The main takeaways are:
• We find that from the exhaustive list of 13 polymer properties presented in Chapter 3, only 6 are relevant. These properties are the ionic conductivity, $T_g$, ESW, $T_d$, limiting oxygen index and the tensile strength at break.

• The ionic conductivity had extremely sparse data with only 19 polymers, therefore we found proxy properties from the exhaustive list of 13 properties using correlation techniques to use instead of the ionic conductivity. These proxy properties were found to be the ESW, $T_g$ and the tensile strength at break.

• A final list of 5 properties were chosen to screen new SPEs namely - $T_g$, ESW, $T_d$, limiting oxygen index and the tensile strength at break. The identified screening criteria for these properties are determined in relation with poly(ethylene) oxide or the industry safety standards for them to operate in a battery.

• To screen for new SPE candidates, a wide polymer chemical space comprising of 13,388 synthesizable, double-ion conducting homopolymers have been curated from numerous literature sources.
In the previous section, five polymer properties and proxies were finalized to be used in our workflow for SPE design. These properties, shown in Table 3, are the glass transition temperature ($T_g$), electrochemical stability window, thermal degradation temperature ($T_d$), the limiting oxygen index and tensile strength at break. Further, datasets and ML models were developed for these properties.

In the data-driven search for electrolytes, most large-scale past works have been focused on inorganic solid electrolytes mainly via the Electrolyte Genome project which used an automated workflow to calculate the electron affinities and ionization potentials of over 4830 potential electrolyte candidates which were then rapidly screened for new electrolyte materials.[29] On a similar note, Ahmad et al. screened over 12,000 inorganic solids to find six new stable electrolyte materials.[30] Within the polymer electrolyte realm, Liu et. al. recently presented a work focused on collecting a dataset and training a machine learning model on ethylene/oxide and Li ion ratios.[31] Another work by Hatakeyama-Sato et. al. demonstrated an ML prediction scheme for the ionic conductivity and permittivity of single-ion conducting polymers using a dataset of 240 experimentally determined values.[32] Xie et. al.’s recent work used MD simulations to screen for SPE candidates from 53362 hypothetically generated polymers using computed ionic conductivity values.[33] However, there has been no past work that focuses on screening a very large and diverse experimental polymer chemical space using a more holistic screening criteria to find high-performing and safe SPEs.
5.1 Models development

For each polymer property of interest (detailed in Section 2.1.), an ML model was developed and implemented using the dataset detailed in Section 4.1. Figure 4.4 visualizes the performance of each model using parity plots which represents the accuracy of the predicted values against the dataset values. Additionally, the figure details the RMSE error values for these models. Further details and dataset references of these models are presented in our previous work and all the models are implemented on the Polymer Genome online platform. [28]

For all models, the root mean square error (RMSE and R-squared) values were used to evaluate and compare the performances of the GPR and the co-kriging models. To estimate
the prediction errors on unseen data, learning curves were generated by varying the size of
the training and the test sets. Additionally, for each case, statistically meaningful results
were obtained by averaging RMSE results over 50 runs with varying training and test splits.
Further, 5-fold cross validation was adopted to avoid overfitting.[27, 42]

5.2 Screening for new SPE candidates

Figure 5.2: Plot of our machine learning-based workflow to screen for new polymer
electrolytes is shown. An initial chemical space of 13,388 polymers is parsed to identify
candidates which satisfy five property values illustrated in the figure. A partial list of
proposed SPEs is illustrated. A complete list of proposed SPEs is presented in Table S2.

To screen for new SPEs, our screening workflow is presented in Figure 5.1. The
models developed previously were used to make predictions for the five screening polymer
properties on a massive dataset comprising of 13,388 synthesizable polymers curated from
various literature sources. Subsequently, we used the screening criteria in Table 3 to
identify new SPEs. We found 16 new potential SPE candidates which satisfy all five
Table 5.1: Complete list of 16 identified SPEs using our proposed ML workflow. The corresponding ML-predicted ESW, \( T_g \), \( T_d \), LOI and TS values are also given.

<table>
<thead>
<tr>
<th>SMILES</th>
<th>ESW (eV)</th>
<th>( T_d ) (K)</th>
<th>( T_g ) (K)</th>
<th>TS (MPa)</th>
<th>LOI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyphosphazenes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCC)OCC)</td>
<td>5.40</td>
<td>543.85</td>
<td>182.33</td>
<td>27.47</td>
<td>31</td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCCC)OCCC)</td>
<td>5.31</td>
<td>569.66</td>
<td>185.54</td>
<td>29.48</td>
<td>28</td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCCCC)OCCCC)</td>
<td>5.30</td>
<td>571.90</td>
<td>185.43</td>
<td>27.53</td>
<td>26</td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCCC)OCCCCC)</td>
<td>5.26</td>
<td>563.57</td>
<td>182.32</td>
<td>26.13</td>
<td>24</td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCCCCC)OCCCCC)</td>
<td>5.29</td>
<td>551.36</td>
<td>185.13</td>
<td>21.05</td>
<td>23</td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCCC)OCCCC)</td>
<td>5.28</td>
<td>569.21</td>
<td>183.97</td>
<td>28.15</td>
<td>27</td>
</tr>
<tr>
<td>Fluorine-containing polymers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>([\ast]N=P(\ast)(OCC(F)(F)C(F)F)OCC(F)(F)C(F)F)</td>
<td>5.96</td>
<td>511.79</td>
<td>211.91</td>
<td>37.05</td>
<td>34</td>
</tr>
<tr>
<td>([\ast]CC(F)(F)C(F)C(F)C(F)C(F)COCOCO[\ast])</td>
<td>7.19</td>
<td>712.42</td>
<td>198.85</td>
<td>46.93</td>
<td>29</td>
</tr>
<tr>
<td>O([\ast])COCC(F)(F)C(F)C(F)C(F)C(F)C(F)C(F)C(F)C([\ast])</td>
<td>7.50</td>
<td>747.11</td>
<td>207.08</td>
<td>42.51</td>
<td>38</td>
</tr>
<tr>
<td>O([\ast])COCC(F)(F)C(F)C(F)C(F)C(F)C([\ast])</td>
<td>7.32</td>
<td>731.69</td>
<td>199.77</td>
<td>44.17</td>
<td>34</td>
</tr>
<tr>
<td>Other polar polymers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>([\ast]CCCSSCCCCOCO[\ast])</td>
<td>5.90</td>
<td>618.56</td>
<td>199.29</td>
<td>22.01</td>
<td>25</td>
</tr>
<tr>
<td>([\ast]CCIC1CC1CO[\ast])</td>
<td>6.11</td>
<td>482.84</td>
<td>184.14</td>
<td>36.93</td>
<td>23</td>
</tr>
<tr>
<td>([\ast]COCOCOCOCO(=O)CCCSSCCCCCC(=O)O[\ast])</td>
<td>5.51</td>
<td>544.99</td>
<td>211.37</td>
<td>29.11</td>
<td>22</td>
</tr>
<tr>
<td>([\ast]CCCCCOCC1(CO[\ast])CC1)</td>
<td>6.05</td>
<td>619.68</td>
<td>199.96</td>
<td>33.62</td>
<td>22</td>
</tr>
<tr>
<td>([\ast]C<a href="C">Si</a>(C)<a href="%5Cast">Si</a>(C)C)</td>
<td>4.54</td>
<td>673.30</td>
<td>201.91</td>
<td>21.40</td>
<td>37</td>
</tr>
</tbody>
</table>

property requirements. The entire list of these candidates are listed in Table 5.2. with their corresponding screening criteria values.

5.3 Chemical insights about identified SPEs

All 16 identified candidates in this work, unsurprisingly, have oxygen moieties, more specifically polyether groups. Several past works have shown that the oxygen atoms in the polymer is critical for ion transport via intersegmental and collective motion of the chain via the polar polyether groups. This finding aligns also with Xie et al.’s recent molecular dynamics driven predictive work where all the identified SPEs had polyether-based backbones.[33] However, the screening criteria in this work is more holistic in that it factors the thermal and mechanical stability of the polymer host. Therefore,
all the polymers have more varied chemistries and include polar groups namely sulfur, fluorine, nitrogen and silicon that contribute to higher thermal resistance and increased crystalline. The identified candidates can be broadly classified into three sections: first, polyphosphazenes or \(-\text{N}=\text{P}-\) containing polymers, second, fluorine containing polymers and finally, a third set of polar polymers which contain polar moieties like oxygen and sulfur.

Of the SPE candidates, 8 are polyphosphazenes which are a class of polymers which have a \(-\text{N}=\text{P}-\) backbone and the phosphorous atom comprises of two substituents. A well studied SPE of this class is poly[bis((methoxyethoxy)ethoxy)phosphazene] (MEEP) which demonstrated higher ionic conductivities than PEO at room temperature but had poor mechanical properties. [70, 71, 72] However, our predicted polyphosphazenes have tensile strength values ranging from 21-29 MPa. Additionally, the nitrogen and phosphorous components contribute to inflammability which is demonstrated by the predicted LOI values which are between 24-34%.

Another four of our identified SPEs are fluorine containing polymers with structures similar to perfluoropolyethers which have been synthesized in the past as SPEs.[73, 74, 75] The remaining 4 polymer have polar moieties like silicon, sulfur and oxygen. The distinct advantage of a fluorine, sulfur or silicon-containing polymer is its superior thermal stability with an LOI of greater than 23% and \(T_d\) greater than 480 K.

### 5.4 Summary

This chapter is the culmination of the machine learning based workflow to identify new SPE candidates. Some salient points are:

- In the previous chapter, 5 properties were chosen to screen new SPEs namely \(-T_g\), ESW, \(T_d\), limiting oxygen index and the tensile strength at break. Further, datasets were curated for these properties. In this chapter, we successfully developed reliable machine learning models for these properties with reasonable errors.
• Using these models, we made predictions for these properties on a large dataset of 13,388 polymers which was our curated experimental screening space.

• After applying our chosen screening criteria to the polymer space, we find 16 new SPE candidates which satisfies all our performance and safety requirements. The identified candidates can be broadly classified into three sections: first, polyphosphazenes or \(-\text{N=P-}\) containing polymers, second, fluorine containing polymers and finally, a third set of polar polymers which contain polar moieties like oxygen and sulfur.

• Some key chemical rules that we discovered in this process was the presence of polar moieties specifically, polyether groups. Several past works have shown that the oxygen atoms in the polymer is critical for ion transport via intersegmental and collective motion of the chain via the polar polyether groups.

• However, our screening criteria also optimizes for thermal and mechanical stability. Therefore, we find nitrogen, phosphorous and sulfur moieties as well which contribute to high thermal stability and contribute to the crystallinity which in turn makes them mechanically stable.

The biggest advantage of a workflow demonstrated here is the versatility and speed in comparison to a completely experimental method. This mechanism can be used in tandem with experiments to rationally guide SPE exploration in a holistic manner.
CHAPTER 6
SOLVENT SELECTION FOR SOLID POLYMER ELECTROLYTES: ASSESSING SOLUBILITY PARAMETERS

6.1 An introduction to SPE manufacturing

In the previous chapter, we developed a workflow to identify new SPE candidates and proposed 16 new SPE candidates that are predicted to be high-functioning and safe. In Chapter 6 and 7, we work on assisting in the synthesis of SPEs specifically for identifying solvents and non-solvents for their manufacturing.

For SPE manufacturing, three different routes could be used: powder-based processing, wet chemical processing, and high-viscosity processing. Of these, wet chemical processing shows most promise for large-scale SPE manufacturing. In wet chemical processing, the polymer particles are solubilizes in a solvent to obtain a suspension with a target viscosity. SPEs are then formed by solution casting, electrophoretic deposition, or coating process. After the SPE formation, the excess solvent is removed either via heat or a non-solvent. Finding a suitable solvent or non-solvent for a polymer system is time-consuming and riddles with trial-and-error. Therefore, to alleviate this process, data-driven methods could accelerate the time it takes to identify new solvents.

In this Chapter, In this study, we critically assess the performance of both the Hildebrand and Hansen models by directly comparing their predictions against actual experimental data (collected from the Polymer Handbook) on solvents and non-solvents for a benchmark set of polymers.
6.2 Hildebrand and Hansen models

The need to identify suitable (non-)solvents for polymers has led to the development of quantitative models of polymer-solvent compatibility governed by the notion of “like dissolves like” [76, 77]. In essence, a solvent with a cohesive energy density similar to that of a polymer is expected to be a good solvent for the polymer. The cohesive energy density of a condensed substance is a measure of the intramolecular bond strength, and defined as the energy necessary to isolate unit volume of the molecules making up the substance from each other.

Two widely used measures of solvent-polymer compatibility are the Hildebrand [78, 77, 79, 80] and the Hansen [76, 78, 77, 79] solubility parameters. The Hildebrand model utilizes a single parameter, $\delta$, defined as the square-root of the cohesive energy density, to determine whether a substance is a good solvent or non-solvent for a polymer. Solvents with $\delta$ values different from that of a polymer by over 2 MPa$^{1/2}$ are deemed non-solvents, and those within $\pm$ 2 MPa$^{1/2}$ of the polymer $\delta$ value are considered good solvents; the factor 2 MPa$^{1/2}$ has been determined based on empirical considerations. Figure 1(a) pictorially depicts the Hildebrand solubility criterion.

The Hansen model utilizes 3 parameters, $\delta_D$, $\delta_P$ and $\delta_H$, to quantify solvent-polymer compatibility. The 3 parameters represent the dispersion, polar and hydrogen bonding components, respectively, of the Hildebrand parameter $\delta$, such that, $\delta^2 = \delta_D^2 + \delta_P^2 + \delta_H^2$. If visualized as a 3-dimensional plot, the axes being $2\delta_D$, $\delta_P$ and $\delta_H$, the polymers and solvents are points in this space. Those solvents within a sphere of radius $R = 8$ MPa$^{1/2}$ centered at a point corresponding to a polymer are deemed as good solvents for that polymer, and those solvents falling outside the sphere are considered non-solvents. Again, the factor 8 MPa$^{1/2}$ has been determined based on empirical considerations. Figure 1(b) pictorially depicts the Hansen solubility criterion.

In this study, we critically assess the performance of both the Hildebrand and Hansen
Figure 6.1: Solubility criterion of (a) the Hildebrand solubility parameter ($\delta$), where the solvent Hildebrand values are within $\pm 2 \text{ MPa}^{1/2}$ of the polymer Hildebrand value ($\delta_{\text{polymer}}$) and the non-solvent Hildebrand values are over $2 \text{ MPa}^{1/2}$ and, (b) the Hansen solubility model, where the three axes represent the dispersion ($\delta_D$), polar ($\delta_P$) and hydrogen bonding ($\delta_H$) components; solvents fall within a sphere of radius $R = 8 \text{ MPa}^{1/2}$ and non-solvents outside the sphere.
models by directly comparing their predictions against actual experimental data (collected
from the Polymer Handbook) on solvents and non-solvents for a benchmark set of polymers. The model predictions were based on the criteria described above and portrayed in Figure 6.1. A challenge we had to overcome in this study is that the Hildebrand and Hansen solubility parameters are available only for a limited number of polymers (although the solubility parameters are available for numerous solvents). For those polymers for which experimental data on solvents and non-solvents was available, but the Hildebrand parameter was not available, a previously developed machine-learning (ML) model [27] was used first to predict the Hildebrand parameter, followed by the application of the criterion discussed above to predict solvents and non-solvents. The Hansen model values were only available for 25 polymers - too few to train a ML model. Thus, Hansen model assessments were performed only for these 25 polymers.

6.3 Dataset

6.3.1 Experimental data on solvents and non-solvents

Experimental data on solvents and non-solvents for polymers are well-documented in the polymer handbook by Brandup et. al. [81] In some cases, solvents and non-solvents are specified as generic organic substances like alcohols, esters, ethers, ketones, acids and hydrocarbons. However, solubility models are defined for specific organic compounds. Therefore, to apply solubility models, generic compounds (e.g., alcohols) found in data sources were replaced by specific common organic solvents (e.g., methanol, ethanol, etc.) to allow for the application of solubility models. For the example of polyethylene oxide, the generic groups tabulated in the Polymer Handbook as (non-)solvents were replaced by specific substances as follows (highlighted in bold in Table 6.1):

- **Alcohols**: methanol, ethanol, t-butanol, methyl isobutyl carbinol, diethylene glycol, 2-ethyl hexanol, furfuryl alcohol, benzyl alcohol
Table 6.1: Solvents and non-solvents with generic compounds and specific organic compounds for poly(ethylene oxide).

(a) Solvents and non-solvents from literature: generic compounds in bold.

<table>
<thead>
<tr>
<th>Solvents</th>
<th>Non-solvents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aliphatic hydrocarbons</strong>, water (hot), ethers, dioxane</td>
<td>Acetonitrile, <strong>alcohols</strong>, benzene, chloroform, cyclohexanone, dimethyl formamide, <strong>esters</strong>, water (cold)</td>
</tr>
</tbody>
</table>

(b) Expanded solvents and non-solvents with specific solvents: generic compounds replaced by specific solvents in bold. [81, 82, 83]

<table>
<thead>
<tr>
<th>Solvents</th>
<th>Non-solvents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetonitrile, methanol, ethanol, t-butanol, methyl isobutyl carbinol, diethylene glycol, 2-ethyl hexanol, furfuryl alcohol, benzyl alcohol, benzene, chloroform, cyclohexanone, dimethyl formamide, ethyl acetate, propyl acetate, t-butyl acetate, hexyl acetate, isoamyl acetate, water (cold)</td>
<td>Hexane, heptane, pentane, water (hot), tetrahydrofuran, diethyl ether, ethylene glycol dimethyl ether, dipropylene glycol methyl ether, diethylene glycol monomethyl ether, dioxane</td>
</tr>
</tbody>
</table>

- **Esters**: ethyl acetate, propyl acetate, t-butyl acetate, hexyl acetate, isoamyl acetate
- **Aliphatic hydrocarbons**: hexane, heptane, pentane
- **Ethers**: tetrahydrofuran, diethyl ether, ethylene glycol dimethyl ether, dipropylene glycol methyl ether, diethylene glycol monomethyl ether

All the generic substances present in this study and their associated specific organic solvents used for the purpose of the assessment of the solubility parameters are listed in the supplementary information.

6.3.2 Hildebrand and Hansen solubility parameters

To assess the solvent prediction capability of the Hildebrand model, a benchmark polymer dataset (shown in Table 6.2) of 75 polymers classified into 4 categories, viz., 12 amorphous non-polar polymers, 17 amorphous polar polymers, 21 semi-crystalline non-polar polymers and 25 semi-crystalline polar polymers was used. Since the Hildebrand solubility
Table 6.2: Dataset for the assessment of the Hildebrand solubility model across different polymer classes [84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 81]

<table>
<thead>
<tr>
<th>Polymer class</th>
<th>Polymers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amorphous non-polar</td>
<td>Atactic polypropylene, atactic polystyrene, polyphenylene oxide, cis-polyisoprene, poly(1,3-butadiene), atactic poly(1-butene), atactic poly(cyclohexylethylene), poly(a-methylstylene), low mol. wt. poly(phenyl acetylene), atactic poly(cyclohexenylethylene), poly(4-acetoxy styrene), poly(1-hexene)</td>
</tr>
<tr>
<td>Amorphous polar</td>
<td>Polyvinyl chloride, polyether imide, polycarbonate, polymethyl methacrylate, polyurethane, atactic polycrylic acid, polyethyl methacrylate, polyvinyl methyl ether, poly(isobutyl vinyl ether), poly(tert-butyl vinyl ether), polyacetaldehyde, poly(2,6-dimethyl -1,4-phenylene oxide), polyvinyl acetate, poly(2-chlorobutadiene), poly(4-hydroxystyrene), poly(n-propyl methacrylate)</td>
</tr>
<tr>
<td>Semi-crystalline non-polar</td>
<td>Isotactic polypropylene, isotactic polystyrene, polymethylpentene, high density polyethylene, polyphenylene sulfide, polyallene, poly(1,2-butadiene), high mol.wt. poly(phenylacetylene), polyacetylene, cyclo-poly(methylene) polyisobutene, isotactic poly(cyclohexenylethylene), poly(2-tert-butyl-1,3-butadiene), poly(dicyclopentadiene), poly(5,7-dimethyl-1,6-octadiene), poly(1,4-diphenyl-1,3-butadiene), isotactic poly(1-butene), isotactic poly(4-methyl-1-pentene), poly(cyclopentene), poly(acenaphthylene), poly(1,3-cyclohexadiene)</td>
</tr>
<tr>
<td>Semi-crystalline polar</td>
<td>Polycrylonitrile, nylon 6,6, polyethylene terephthalate, polyoxymethylene, polyvinyl alcohol, isotactic polycrylic acid, poly(1-methoxybutadiene), poly(2-chlorobutadiene), poly(2-chloromethylbutadiene), polyethylene oxide, poly(e-caprolactone), poly(e-caprolactam), poly(vinyl methyl ketone), poly(vinyl fluoride), poly(vinylidene chloride), poly(vinyl bromide), poly(bis(methylthio)acetylene), poly(p-chlorostyrene), poly(styrenesulfonic acid), poly(methacrylonitrile), poly(vinyl formal), poly(vinyl trifluoroacetate), poly(2-vinylpyridine), poly(propylene oxide), poly(urethane)</td>
</tr>
</tbody>
</table>
parameters were not available for all polymers (although the solubility parameters are available for most solvents), the Hildebrand values for the polymers in the dataset was predicted using a previously developed Gaussian process regression (GPR) ML model. [27]

113 polymers were used to train the machine learning model, and all of them are plotted in Figure 6.2. The model performance was assessed using a learning curve in Kim et al’s work. However, the final model used for the prediction of the solubility parameter uses all 113 data points to ensure maximum diversity, therefore there is no validation set. The root mean square error for the model prediction is 0.49 MPa$^{1/2}$ and the model performance plot is shown below in Figure 6.2. Each polymer in this dataset has corresponding experimentally determined solvents and non-solvents which are documented in the polymer handbook by Brandup et al. [84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 81]

The evaluation of the solvent prediction capabilities of the Hansen model was conducted using 25 polar and non-polar polymers (shown in Table 6.3). The size of the polymer dataset was constrained by the scarcity of data for the Hansen parameter. The Hansen parameter for polymers was collected from the official Hansen parameter website.[113, 114] For comparison, the dataset was also assessed using the Hildebrand model, where the parameter values where obtained using the ML model described above. Each polymer in this dataset has corresponding experimentally determined solvents and non-solvents which are documented in the polymer handbook by Brandup et al. [81]
Figure 6.2: Performance of the Hildebrand solubility machine-learning model developed using Gaussian process regression (GPR) for the prediction of unknown polymer Hildebrand solubility values.

The Hildebrand and Hansen parameter values for all the solvents and non-solvents were obtained from literature. [81, 114]

### 6.4 Assessment procedure

To assess the Hildebrand model, the benchmark polymer dataset and its associated experimental solvents and non-solvents (explained in the Dataset section) were evaluated using the miscibility criteria depicted in Figure 1(a). Similar Hildebrand parameters indicate miscibility i.e. the difference between the Hildebrand parameters of a miscible solvent-polymer pair should not exceed $\pm 2 \text{ MPa}^{1/2}$. For non-solvent-polymer pairs, the difference between the Hildebrand parameters must be greater than $\pm 2 \text{ MPa}^{1/2}$. The (non-)solvent prediction accuracy was calculated using this miscibility criteria for each (non-)solvent-polymer pair. [78, 79]

Likewise, the Hansen model is evaluated using the aforementioned dataset (in the
Dataset section) and its associated solvents and non-solvents. The miscibility criteria depicted in Figure 6.1(b) is used to assess the model where miscible polymer-solvent pairs are within $\pm 8 \text{ MPa}^{1/2}$ and polymer-non-solvents are greater than $\pm 8 \text{ MPa}^{1/2}$. Each (non-)solvent-polymer pair was evaluated using this criterion. [76, 114, 115, 113]

For every polymer in the Hildebrand and Hansen datasets, the (non-)solvent prediction accuracy was calculated as a percentage of (non-)solvents correctly predicted by the solubility models. Then, to estimate the performance of these models across the polymer space, the population statistics for each dataset was determined using a well-established statistical re-sampling method called bootstrapping. The bootstrap method is a statistical technique for estimating the mean and standard deviation of a population by averaging estimates from a smaller dataset. In this method, larger datasets, representative of the population, is constructed by randomly duplicating entries in the smaller dataset. This allows a given sample to be included in the larger dataset more than once and is commonly referred to as sampling with replacement. These larger datasets can then be used to calculate the population statistics - namely, the mean and the standard deviation. Utilizing this method, larger datasets, representative of the polymer population was constructed for both the Hildebrand and the Hansen datasets. The larger datasets comprised of 1000 samples and was constructed by randomly sampling each smaller original dataset with replacement. These datasets were used to assess the (non-)solvent prediction accuracy for each polymer and the associated error bars arises from the bootstrapping process. Although we did not explicitly account for the uncertainty from the ML predictions, the machine learning model RMSE is low at 0.49 MPa$^{1/2}$ and we believe the bootstrapping process captures the statistical uncertainty that arises from the (non-)solvent prediction variability in the dataset. Further, we also believe that the prediction trends will remain unchanged given the inherent capability of the Hildebrand parameter to capture only the Van der Waal’s forces for a given system and the lack of data for the Hansen parameter for polymers.
6.5 Results and discussion

6.5.1 Assessment of the Hildebrand solubility parameter across different polymer classes

The performance of the Hildebrand parameter for solvents and non-solvents for the 4 distinct classes of polymers are shown in Figure 6.3. The Hildebrand parameter was originally developed for amorphous, non-polar polymers and its prediction capability is known to falter for semi-crystalline and polar polymers. [79] Here, we evaluated its prediction capabilities quantitatively across different polymer classes divided into 4 classes: amorphous polar, amorphous non-polar, semi-crystalline polar and semi-crystalline non-polar polymers. The Hildebrand approach predicts solvents for non-polar polymers more accurately than for polar polymers. Solvent prediction for amorphous non-polar polymers and semi-crystalline non-polar reaches accuracies of 77% and 72%, respectively. However, the performance deteriorates for polar polymers, at 59% for amorphous polar polymers and 57% for semi-crystalline polar polymers. At greater than 78%, the parameter performance is uniformly satisfactory for non-solvent prediction for all polymer classes.

The performance trend of the parameter for solvents can be explained by the nature of interactions captured by the Hildebrand parameter and the interactions at play for various classes of polymers. The Hildebrand parameter essentially captures the total cohesive interactions which includes the dispersion forces, dipole interactions and hydrogen bonding interactions for a molecule. For non-polar polymers which do not have noteworthy polar or hydrogen bonding contributions, the total cohesive energy suffices as the primary criterion for solubility. For polar polymers, specific interactions like hydrogen bonding plays a crucial role which the Hildebrand parameter, being a single-parameter, is unable to explicitly capture. For instance, polycarbonate (PC) and polymethyl methacrylate (PMMA) are well-known polar amorphous polymers. Acetone is a solvent for PMMA but is a non-solvent for PC, and yet the three systems have similar Hildebrand parameters. Acetone is a polar aprotic solvent, therefore does not have hydrogen atoms which can form hydrogen
bonds during solvation. Therefore, the PMMA-acetone solubility can be explained by a polymer-specific polar or hydrogen-bonding interaction that is absent in the PC and acetone interaction, which the single-parameter Hildebrand model is incapable of predicting. [79, 80, 116] Additionally, solvent and non-solvent temperatures also play an important role in solubility by altering the free energy of mixing which the parameter does not consider. For non-solvents, the Hildebrand parameter has a reasonable level of accuracy since the criterion is based on exclusion.

Despite its disadvantages, the Hildebrand parameter can be used to predict solvents for non-polar polymers with reasonable accuracy and predict non-solvents for all classes of polymers. Additionally, the Hildebrand parameter has an expansive dataset for polymers and we have utilized this to train a ML model which predicts the parameter values for a new polymer or polymers absent from the dataset. This development has led to a user-friendly ML based screening tool to predict solvents and non-solvents for a queried polymer using the Hildebrand parameter. This tool is implemented at www.polymergenome.org.
6.5.2 Assessment of the Hansen solubility parameter and its comparison to the Hildebrand parameter

The performance of the Hansen and Hildebrand solubility parameters for 25 polymers and their associated solvents and non-solvents are shown in Figure 6.4. We find, as expected that the average solvent prediction accuracy of the Hansen model at 69% is higher compared to the Hildebrand model which has an accuracy of 60%. The non-solvent accuracy of both methods is equivalent at 76%.

![Figure 6.4: Performance of the Hildebrand and Hansen solubility parameters for 25 polymers for solvents and non-solvents.](image)

The Hansen parameter was developed to overcome a major inadequacy of the Hildebrand parameter, which is the inclusion of the polar and hydrogen bonding contributions.[114] However, the difference between the accuracies of the parameters for polymer systems, especially polar polymers is not striking. We find that the Hildebrand parameter is equivalent to or outperforms the Hansen parameter in solvent predictions for 12 polar polymers, which constitutes 48% of the Hansen dataset. The reasonable non-
solvent accuracy implies that the parameters can be used interchangeably for this purpose.

The Hansen parameter’s lack of significant improvement can be attributed to two reasons. First, the parameter is heavily biased towards its dispersion component which capture the non-polar interactions of a molecule. The dispersion component has a multiplier of 2 attached to it which diminishes the contributions of the polar and hydrogen-bonding parameters. Second, polymer solubility is a complex process involving swelling, diffusion and solvation processes which are often dependent on the polymer structure and other physical attributes like the solvation temperature and solvent concentration. [80] These parameters are not represented in the Hansen or the Hildebrand models and their inclusion could be used to develop a better solubility model.

### 6.6 Drawbacks of parametric models

The heuristic nature of the Hildebrand and Hansen models’ renders it easy to use but results in significant drawbacks and limitations. Polymer solubility is an inherently complex process and depends on a multitude of chemical, morphological, thermodynamic and kinetic factors. Quantifying such a phenomenon using a limited number of manually fitted parameters is an exceptionally challenging endeavor[117]. For instance, both the Hildebrand and Hansen solubility parameters are completely unable to capture the specific nature of the interaction of water with polymers.[113, 80] Another major drawback is the limited availability of the model/parameter values for polymers, especially the Hansen model. [76] Indeed, in some cases, conflicting values of the Hansen parameters are reported for the same polymer.[118] Moreover, for a newly designed polymer, it is difficult to obtain a measurement of the Hildebrand parameter or an estimate of the Hansen solubility parameter. Other solubility models, like the chi parameter, have specific values/parameters for polymer-solvent pairs at given temperatures and concentrations, making it difficult to generalize to unexplored chemical space. [119]

To overcome these drawbacks, the solvent/non-solvent data could be directly used
to train an ML classifier and therefore bypass the requirement of obtaining intermediary solubility parameters. Such an approach is likely to possess significant advantages over the Hildebrand and Hansen criteria due to its ability to learn from a larger chemical space without the requirement of generating empirically derived parameters. Modern ML algorithms, like deep-neural networks, could capture the highly non-linear, multivariate mapping between the characteristics of a polymer and its solubility in a given solvent.

6.7 Summary

This study critically examines and establishes a predictive baseline for two well-established solubility models (based on the Hansen and Hildebrand solubility parameters) widely utilized to characterize polymer solubility in solvents. This comparative study could be used to make informed decisions while utilizing the models for solvent predictions as well as be used as a baseline for new polymer solubility models. Key findings from this study are summarized below:

- We find that the Hildebrand model has a better solvent prediction accuracy for non-polar polymers than for polar polymers. The non-solvent prediction accuracy is uniformly high for all polymer classes. The higher solvent prediction accuracy for non-polar polymers can be attributed to the accurate representation of Van der Waal’s forces by the Hildebrand parameter. The lower solvent prediction accuracy for polar polymers is because polar and hydrogen bonding contributions are lacking in the single-parameter Hildebrand model.

- We find that the Hildebrand and Hansen models have similar prediction accuracies for solvents and non-solvents for polymers. Considering that the Hansen model was developed as an improvement of the Hildebrand model, the insignificant difference in both models’ performance can be attributed to the scarcity of available data for the Hansen model and the complexity of polymer solubility which both models do not
capture factors like temperature, concentration or polymer molecular weight are not considered in both models.

Traditional solubility parameters provide a simple and powerful gateway to understand polymer solubility by utilizing parameters which attempt to capture forces which are pertinent to solvation. This is particularly true of the Hildebrand approach which may be widely used for a large portion of the polymer chemical space (e.g., using ML methods, as has been done at https://www.polymergenome.org). However, the limited number of parameters such models contain are insufficient to capture the full complexity of polymer-solvent interactions. In the future, we aim to bypass the use of traditional solubility parameter based models and utilize experimental data as-is with numerous potentially relevant features and ML algorithms to determine a unique structure property mapping of solvent and polymer structures to their miscibility.
CHAPTER 7

SOLVENT SELECTION FOR SOLID POLYMER ELECTROLYTES USING DEEP LEARNING METHODS

7.1 Need for machine learning based models and prior work

In the previous chapter, we evaluated the performance of the Hildebrand and Hansen solubility parameters for polymers and found it to severely lacking especially for polar polymers. This can be attributed to the simplistic nature of both models wherein the the Hildebrand and Hansen use one and three parameters respectively to determine polymer-solvent similarity. For a phenomenon as complex as solubility, using empirical values to determine behavior is very challenging.

To overcome these limitations, we introduce a novel framework in which room-temperature polymer–solvent compatibility is determined using a deep learning model trained on experimental data in contrast to more traditional aforementioned models. The first version of this work called SolNet v1 (SN v1), by Chandrasekaran et al.[120] accumulated a massive data set of over 4500 polymers and their corresponding solvents/nonsolvents. The data set is used to train a deep neural network binary classifier, which takes as input the polymer and solvent descriptors and as output whether the solvent is a good solvent or a nonsolvent (for that particular polymer). SN v1 was shown to overwhelmingly outperform the Hildebrand approach for all classes of polymers and solvents. In addition to the vastly improved prediction accuracy, this approach enables the visualization of solvents and polymers in terms of chemically relevant low-dimensional latent space vectors. It is important to note that these latent space representations have been obtained intrinsically via the ML algorithm, without the need to input any empirically or manually derived parameters.
Table 7.1: Summary of the baselines and the models used in this work.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Models</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1</td>
<td>Hildebrand solubility model</td>
<td>Single-parameter empirical model traditionally used for solubility prediction. Values calculated using cohesive energy density measurements.</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>SolNet v1 (SN v1)</td>
<td>First deep neural network trained directly on solvent-polymer solubility data.</td>
</tr>
<tr>
<td>This work</td>
<td>SolNet v2 (SN v2)</td>
<td>Improved on SN v1 with respect to dataset and featurization technique.</td>
</tr>
</tbody>
</table>

The following chapter is the next iteration of this model called SolNet v2 (SN v2) has the same overall classification objective and uses a similar framework. However, SN v2 uses a much larger polymer-solvent chemical space by training the model with 50% more polymer-solvent pairs. Also, SN v2 is more generalizable in its prediction capability in that the fingerprinting method used for the solvents is extendable to solvents outside the training space.[120, 121]

7.2 Approach

The main aim of this work is to make a state of the art deep learning framework to classify a polymer-solvent pair as a good solvent or a poor solvent. The latest model in this work is SolNet v2 (SN v2). Therefore, this model is compared against several baselines to evaluate its performance summarized in Table 7.1. The first baseline is the Hildebrand model, which is an empirical one-parameter model that has been traditionally used to predict solubility. The second is the first version of our deep learning framework which is SolNet v1 (SN v1).

The Methods section below details the datasets and methodology used to set up these baselines and the current framework. Following this, the results and analysis are detailed in the following sections. Finally, we address some drawbacks of SN v2 and the critical next steps.
7.3 Methods

In this section, we first detail the methodology used to establish the Hildebrand model baseline. Next, we talk about the dataset, featurization and neural network model used in SN v1 and SN v2 and compare and contrast them.

7.3.1 Hildebrand model baseline – Gaussian process regression

As mentioned previously, given the popularity of the Hildebrand model and its data availability for a large number of polymers and solvents, we decided to use this as a baseline for this work. Since the Hildebrand solubility parameters were not available for all polymers (although the solubility parameters are available for most solvents), the Hildebrand values for the polymers in the dataset was predicted using a previously developed Gaussian process regression (GPR) ML model. [27]. 113 polymers were used to train the machine learning model, and all of them are plotted in Figure 6.2. Through this process, we are able to obtain a machine learning model to predict the solubility parameter of all of the polymers in our data set of 6282 polymers. For the case of the 24 solvents, the Hildebrand parameter was readily available from multiple sources in the literature.[114]

Following the procedure we used in the previous chapter, if the absolute value of the difference in Hildebrand parameters of the solvent and polymer was less than 2 MPa$^{1/2}$, the solvent was classified as a good solvent for that particular polymer. If the difference was greater than 2 MPa$^{1/2}$, the solvent was classified as a nonsolvent for that polymer.

7.3.2 Setup for SN v1 and SN v2

Dataset

The dataset used in both SN v1 and SN v2 are entirely experimental and are manually curated from a plethora of published works including published journals, printed handbooks, and online repositories.[34, 35, 36, 37, 38, 39] The chemical space of the
Table 7.2: Dataset details and comparison between SN v1 and SN v2.

<table>
<thead>
<tr>
<th>Dataset components</th>
<th>SolNet v1 (SN v1)</th>
<th>SolNet v2 (SN v2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of polymers</td>
<td>4595 polymers</td>
<td>6282 polymers</td>
</tr>
<tr>
<td>Number of solvents</td>
<td>24 solvents</td>
<td>58 solvents</td>
</tr>
<tr>
<td>Polymer-solvent pairs</td>
<td>11,958</td>
<td>19,655</td>
</tr>
<tr>
<td>Polymer-non-solvent pairs</td>
<td>8,469</td>
<td>10,378</td>
</tr>
</tbody>
</table>

polymers spans over a wide range but predominantly included the following atomic species: H, B, C, N, O, F, Si, P, S, Cl, Br, and I. Copolymers, polymer blends, polymers with additives, and cross-linked polymers are not considered in this study. We also limit this study to the investigation of room-temperature solubility and do not consider partial solubility or high-temperature solubility.

The differences between the datasets for SN v1 and SN v2 are listed in Table 7.2. SN v2 has a dataset comprising of almost 1600 more polymers than SN v1 and also more than twice as many solvents. SN v1 used 4595 polymers and 24 solvents with a total of 20,427 unique polymer-(non-)solvent pairs to train the ML model. SN v2 massively increased this dataset by incorporating 6282 polymers with 58 solvents and 30,388 polymer-(non-)solvent pairs. While the datasets are different, the SN v1 baseline is implemented via the solvent fingerprint method detailed in the next section. The datasets when we implement SN v1 as a baseline is the same since the crux of that method lies in its unique fingerprinting methodology.

58 solvents occur in the dataset with varying frequencies. The solvents and their frequencies are demonstrated as a histogram in Fig 7.1. As expected some solvents like chloroform, THF and DMF occur very frequently with 2865, 2817 and 2196 pairs respectively. On the other end of the spectrum, is 1,2,4–trichlorobenzene and 1,1,2-trichloroethane with 12 pairs each. The frequencies are a function of their popularity experimentally with chloroform, THF and DMF being versatile, easy solvents for a large variety of polymers.
Figure 7.1: Names of 58 solvents used in SN v2 along with their frequency of occurrence in the SN v2 dataset.

**Fingerprinting**

For the polymers in the SN v1 and SN v2 datasets, a hierarchical fingerprinting method was used to capture descriptors that control the tendency to crystallize in polymers and was described in a previous work.[27] The fingerprinting scheme comprises of 663 features and four hierarchical levels of descriptors. The first level is at the atomic scale comprising of the count of atomic triples (e.g., O2-C3-C4, representing of a two-fold coordinated oxygen, a three-fold coordinated carbon, and a four-fold coordinated carbon). The second set of fingerprint components captures a population of pre-defined chemical building blocks (e.g., -C6H4-, -CH2-, -C(=O)-). The third hierarchical level comprises of quantitative structure-property relationship (QSPR) descriptors, such as van der Waals surface area, topological surface area and the fraction of rotatable bonds, implemented in the RDKit cheminformatics library. The fourth and last fingerprinting level includes morphological features such as the topological distance between rings, fraction of atoms that are part of
Table 7.3: Fingerprinting methods and comparison between SN v1 and SN v2.

<table>
<thead>
<tr>
<th>Fingerprinting</th>
<th>SolNet v1 (SN v1)</th>
<th>SolNet v2 (SN v2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solvent fingerprinting</td>
<td>One-hot encoding</td>
<td>Hierarchical fingerprinting</td>
</tr>
<tr>
<td>Generalizability</td>
<td>No - restricted to only those in the dataset</td>
<td>Yes - can be used to make predictions on solvents outside the dataset</td>
</tr>
</tbody>
</table>

The crux of this work lies in the difference between SN v1 and SN v2 for solvent fingerprinting. The difference in the methods used are detailed in Table 7.2. In SN v1, given the limited number of solvents, one-hot encoding was used to represent the solvents. One-hot encoding is a technique to represent categorical variables using a binary representation.

As mentioned previously, the implementation of SN v1 as a baseline assumes the dataset comprises of the expanded dataset of 6282 polymers and 58 solvents. This is because the crux of SN v1 lies in its unique one-hot fingerprinting method and to ascertain the value of the method, it is necessary to compare it with the same dataset. So while SN v2 features an expanded dataset in its iterative upgrade, the baseline for SN v1 will be trained on 6282 polymers and 58 solvents. When the SN v1 baseline is implemented, the fingerprint vector for a particular solvent would have a dimensionality of 58, of which 57 components would be zero and only one component would possess the integer value of one. With one-hot encoding, a significant drawback is its lack of generalizability. Since the fingerprint size is defined by the solvent dataset size, it cannot incorporate solvents outside the dataset severely limiting its usability.

To overcome this drawback, in SN v2, we use a fingerprinting method akin to the polymer fingerprinting wherein atomic and block-level features for solvents and captured and a feature space of 63 features captures the solvent space. Unlike one-hot encoding, this fingerprint is extendable to solvents outside the dataset. To capture the fingerprint diversity and ensure its veracity, we used a principle component analysis (PCA) plot (Figure 7.2)
Figure 7.2: PCA plot of 58 solvents where the PC\textsubscript{0} and PC\textsubscript{1} components of the fingerprint vector are plotted. Similar polymers are clustered and marked on the figure.

plotted using the solvent fingerprint for the PC\textsubscript{0} and PC\textsubscript{1} axes. From this plot, we find that similar solvent cluster together in the fingerprint space. Nitrogen containing solvents like DMF, DMAc and formamide are clustered together and chlorine containing solvents like chloroform and chloromethane are clustered together. The scatter in some areas of the plot can be attributed to the fact that the PCA plot only accounts for around 60% of the variance.

Neural network details

SN v1 and SN v2 use almost identical model architectures except for some differences in the way that the results are predicted (see Table 7.4). For both models, a slightly modified version of the multilayer perceptron neural network algorithm is utilized for this particular classification (good solvent or nonsolvent) problem. This model was built using PyTorch. The neural network, as depicted in Figure 7.3, consists of two input branches, one for the solvent fingerprints and the other for the polymer descriptors. The solvent branch of the
Table 7.4: Neural network architecture and comparison between SN v1 and SN v2.

<table>
<thead>
<tr>
<th>Model architecture</th>
<th>SolNet v1 (SN v1)</th>
<th>SolNet v2 (SN v2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN output layer activation function</td>
<td>Sigmoid activation function</td>
<td>Softmax activation function</td>
</tr>
<tr>
<td>Out-of-domain test set</td>
<td>No - restricted to solvents in the dataset</td>
<td>Yes - can be used to make predictions on solvents outside the dataset</td>
</tr>
</tbody>
</table>

The neural network consists of 3 hidden layers, each with 100 neurons, and the polymer branch of the neural network consists of three hidden layers, each with 100 neurons. The solvent net and polymer net hidden layers are merged using a concatenation operation and passed on to a final set of hidden layers. There are 3 final hidden layers, each with 100 neurons. All of the hidden layers in the neural network are constructed with the rectified linear unit (ReLU) activation function. The ADAM gradient descent method is utilized for training the neural network weights along with a training batch-size of 256.

The difference between SN v1 and SN v2 lies in the final output layer which decides the classification results. SN v1 consisted of a single neuron with the sigmoid activation function wherein an ideal good solvent is denoted by an activation value of “1” and an ideal nonsolvent is denoted by an activation value of “0”. The threshold to differentiate between a nonsolvent and a good solvent is set at 0.5. In SN v2, we used the softmax activation function which returns an output layer with size equal to the number of classes to be predicted (2 in this case) and each neuron comprising of probabilities associated with the occurrence of solvent or non-solvent for that particular solvent-polymer pair. The class with the higher softmax values is the predicted class. The reason we chose softmax over sigmoid is that this architecture can accommodate in-between partial classes in the future. The current solvent and non-solvent predictions present only 2 extremes of the solubility scale and when the dataset becomes amenable to expand to more classes, the softmax function is far more flexible. Additionally, there is no disadvantage of using the softmax over a sigmoid activation function for this classification problem.
Figure 7.3: Neural network architecture for the prediction of good solvents and non-solvents for polymers.

Additionally, to evaluate the generalizability of the solvent fingerprint, the dataset is first divided into in-domain and out-of domain where the latter comprises of 15% of the solvent dataset. Here, 9 random solvents are chosen from 58 solvent list and all the associated solvent-polymer pairs are held out during training to test the performance of the SN v2 hierarchical fingerprinting method. For the remaining in-domain solvents and polymers, a randomized 80%/10%/0% train/validation/in-domain test split is used. The validation set is used to determine the optimal number of epochs. The in-domain test set tests the performance of the model for solvents in the dataset. Further, 5 separate seeds are used for each baseline and setting for statistical analysis.

7.4 Results and discussion

As mentioned in the previous section, in this work, we evaluate the performance of SN v2 and compare and contrast it with the Hildebrand model and SN v1. To do so, all models are evaluated on the same dataset comprising of 6282 polymers and 58 solvents. The Hildebrand model uses the Hildebrand parameters for the polymer-solvent pairs. SN v1 uses one-hot encoding for the solvent fingerprinting while SN v2 uses a hierarchical methodology similar to polymer fingerprinting for solvents. Further, the results are presented on 2 test sets – the first is the in-domain test set where 85% of the solvents are randomly chosen and the training data sees instances of all in-domain solvents. The
remaining solvents called out-of-domain solvents are completely held out of the training data and are primarily used to evaluate the generalizability of the solvent fingerprinting method. There are 2 accuracy metrics used in this work – the first is the per-class accuracy also known as the recall of the model and the second is the overall accuracy. The per-class accuracy is defined as:

\[
\text{Per class accuracy} = \frac{\text{Number of correctly predicted pairs}}{\text{Total number of experimentally accurate pairs}} \times 100 \tag{7.1}
\]

The overall accuracy is the same metric added for both classes. The results for all the baselines and above metrics are presented in Figure 7.4.

From Figure 7.4, the in-domain accuracies of SN v2 and SN v1 are very similar in that the solvent accuracies are 90% while the non-solvent accuracies are 86%. The Hildebrand model performs the worst among the three models, with a solvent accuracy
52% and non-solvent accuracy of 80%. The performance of the Hildebrand model is in-line with the results presented in the previous chapter especially Figures 6.3 and 6.4. The model historically underperforms for polar polymers which forms a bulk of synthesizable polymers. The similar performances of SN v1 and SN v2 can be attributed to the fact that solvents usually pick a class. For instance, THF, chloroform, NMP is always a good solvent while methanol is always a non-solvent. Therefore, the performances are similar.

The core difference in the SN v2 methodology lies in the out-of-domain test set. Here, the Hildebrand model performs similarly to in-domain but SN v1 is incapable of making predictions for out-of-domain solvents because of the nature of the one-hot encoding fingerprinting method. SN v2 performs better than the Hildebrand model for out-of-domain solvents but the performance dips in comparison, as expected, to the in-domain solvents given that the solvent-polymers pairs are completely held out.

The error bars for the in-domain and out-of-domain are very variable. The in-domain test set shows small errors since the training data encompasses all solvents used. For the out-of-domain test set, the error bars depend on which solvents get picked. If a solvent that occurs very frequently in the dataset gets held out (chloroform, DMF), the accuracy of the model is extremely low since the training data is affected.

While these accuracy values discussed above are for individual classes, Figure 7.5 shows the per-solvent average accuracy values. The per-solvent accuracy value is the weighted average (by number of samples) of the individual solvent accuracies. A detailed breakdown of the per-solvent accuracy for each of the individual 58 solvents is provided in Appendix B.

Figure 7.5 shows similar trends to per-class accuracies with the Hildebrand model as the lowest performing model with similar performances of the SN v1 and SN v2 for in-domain test sets. For out-of-domain, SN v2 is the best model while SN v1 cannot compute these values.
7.5 Summary

In summary, we have developed an improved iteration of a novel deep-learning framework capable of accurately predicting good solvents and nonsolvents for massive dataset of polymers. In the previous chapter and past work,[121] we critically examined the accuracy of the Hildebrand and Hansen criteria for a number of polymers and solvents (for which these criteria were available). The Hildebrand criteria were shown to have an accuracy of 60% for solvents and 76% for nonsolvents, whereas the Hansen criteria were shown to have an accuracy of 67% for solvents and 76% for nonsolvents.[121]

In this version, SN v2 outperforms the Hildebrand model and also improves on some of the drawbacks of SolNet v1 by implementing a hierarchical solvent fingerprinting method which can be generalized and used to make predictions on solvents outside the dataset. The purely data-driven framework described in this work can be systematically improved as it is exposed to an even larger quantity and diversity of data. Further, future works can also focus on including concentration and temperature influences for the whole solubility scale.
including partial solvent classes.
8.1 Summary

Through this work, I established a machine learning based workflow for the discovery of new polymer electrolytes. This work is the first of its kind in terms of finding new double-ion conducting polymer electrolytes which not only meet performance and safety criteria but also are experimentally synthesizable. Further, I also established a deep learning methodology to find solvents and non-solvents for their synthesis. The main findings of this work are summarized below:

• **Identifying and curating properties pertinent to SPEs**
  
  We identified 13 properties relevant to the SPE performance and reliability and finalized 6 core properties namely the ionic conductivity, electrochemical stability window, $T_g$, $T_d$, limiting oxygen index and the tensile strength at break. Further, we curated datasets for each property and through this process found that we needed to identify proxy properties for the ionic conductivity which were found to be the $T_g$, ESW and tensile strength at break. With this, we have 5 properties which we identified to be useful for SPE design.

• **Training machine learning models to predict these properties for large datasets**

  We used Gaussian process regression (GPR) to train models for the 5 aforementioned properties and found that the trained models had reasonable error metrics. Further, these models could be used to make predictions on a large dataset of 13,388 experimentally synthesizable polymers.

• **Finalizing a screening criteria and identifying 16 new SPE candidates**
Using past known SPEs and expert-determined requisites, we finalized an extensive screening criteria to screen high-performing and safe SPEs. We found 16 candidates from our initial list of polymers which includes functional groups like polyphosphazenes, fluorine-containing polymers and other polar polymers.

- **Developing deep learning models to predict solvents and non-solvents for SPE synthesis**

We developed an improved iteration of a novel deep-learning framework capable of accurately predicting good solvents and non-solvents for massive dataset of polymers called SolNet v2. SN v2 outperforms the Hildebrand model and also improves on some of the drawbacks on its predecessor, SolNet v1 by implementing a hierarchical solvent fingerprinting method which can be generalized and used to make predictions on solvents outside the dataset. The purely data-driven framework described in this work can be systematically improved as it is exposed to an even larger quantity and diversity of data. Further, future works can also focus on including concentration and temperature influences for the whole solubility scale including partial solvent classes.

### 8.2 Future work

This work demonstrates a workflow which can be used to discover and design new SPEs however, there are several future works which address some of the challenges faced in this work and present other opportunities to develop the next generation of safe battery materials.

#### 8.2.1 Generating ionic conductivity data using computational methods

An obvious drawback of this work is the sparsity of standardized ionic conductivity data for polymers and a critical next step is to develop computational methods to develop this dataset. Over the past decades, theoretical and simulation methods in the field of SPEs
have been widely applied to investigate detailed ion conduction mechanisms as well as to provide guidelines on design parameters.

Coarse-grained (CG) simulations, where a collection of atoms or monomers are combined as a single interaction site, are also often applied for more affordable ionic conductivity calculations or for understanding the general trends as a function of molecular features.[122, 123, 124] Several recently-developed strategies have been shown to significantly reduce the uncertainty of conductivity calculation and have a great potential to be used in tandem with high throughput all-atom simulations to screen ionic conductivity.[125, 126, 127, 33] Additionally, recent advances in new techniques regarding autonomous and systematic coarse-graining is very important for the future of integrating CG simulations with data-driven methods.[128, 129, 130, 131, 132] These methods provide prospects for systematic exploration of large polymer chemical space as well as generation of large ionic conductivity dataset from simulations.

Other than direct computation of ionic conductivity, several related properties can be obtained through modeling. First, self-diffusion constant of ions is one of the key factors relevant for optimizing ionic conductivity in an SPE, making it a reasonable screening proxy property. Moreover, segmental motion of the polymer backbone (related to the glass transition temperature \( T_g \)) is also highly related to ion diffusion behavior. Datasets of polymer segmental mobility can be generated via ab initio methods by calculating the energy barriers for intermolecular rotation or via classical MD simulations by measuring dynamic structure factor of polymers. Notably, a recent classical all-atom MD simulation has demonstrated the feasibility of direct \( T_g \) screening of a large number of polymers. [133]

Ion dissociation/correlation also plays an important role since it controls the number of charge carriers. Creating datasets of binding/dissociation energy from DFT calculations based on energy difference between individual ions and the neutral ion pair can allow for data-driven methods to filter out salts with good solubility in SPEs. In addition, ion dissociation can also be explored through characterization of ion solvation environment
in simulations, and recent development on automated classification of ion solvation sites using ML approaches will aid dataset generation for ion solvation properties.[134, 135]

8.2.2 Inverse design of SPEs with generative models

An alternative to the enumeration approach used in this work that relies on screening from a predetermined polymer candidate set is to directly generate polymers that satisfy property thresholds. This inverted approach directly yields promising candidates making it more general-purpose. Two approaches to inverse design are genetic algorithms (GA) and virtual autoencoders (VAE).

Genetic algorithms (GA) based on the principle of natural selection works by treating the polymer as a combination of chemical blocks that can be selected and mutated to create a polymer with required properties. Virtual autoencoders (VAE) are unsupervised ML methods which uses an encoder unit to represent the polymer in a high-dimensional latent space and the decoder unit maps it back to valid polymer with desirable polymer properties. Kim et. al. and Batra et. al. used a GA and VAE respectively to design polymers with high glass transition temperature and large bandgaps for high-energy capacitors owing to their high thermal and electrical stability. This approach can be extended to electrolytes as well with properties elucidated in Table 3. [136, 137]

8.2.3 SPE retrosynthesis

While the methods above identify new SPE materials, their synthetic realization is a slow-moving process owing to the several considerations required to develop a polymer. While some are more procurement based like raw materials cost and unavailability, others are far more severe like monomer and solvent toxicity and harsh processing conditions. In the past decade, chemical retrosynthesis which identified raw materials and processing steps for a given material has moved from domain expertise to computer-aided. Most prominent works have been in the molecule domain with promising results.[138, 139, 140]
However, this area is still nascent for polymer synthesis owing to the inherent complexity of the polymerization techniques like the method of polymerization, catalysts, monomers and polymerization conditions like temperature and pressure. Currently, no repository comprising of polymerization information exists however, the data can be curated from literature sources. Further, natural language processing methods could potentially accelerate the data curation process. The development of a retrosynthesis methodology for SPEs will provide an end-to-end data-driven workflow from the identification of potential electrolyte candidates to providing a path for their synthesis.

8.2.4 Further developments for the solvent model

The solvent model presented in this work only captures room temperature solubility at a concentration of 5-30 mg/L with 2 classes namely, good solvents and non-solvents. In reality, solubility is a variable scale with several partial levels of solubility which is dependent on the temperature and concentration conditions of the polymer. Therefore, the next step in this direction could be using the chi parameter to develop models. The chi parameter is a miscibility metric for a particular polymer-solvent pair at a given temperature and pressure which could help make a more fine-grained model with more than 2 classes. Within the scope of SolNet version 2, experimental efforts to append to the existing dataset are underway.

8.3 Impact – Polymer Genome online platform

All the models used in this work are deployed and freely accessible on an online polymer informatics platform called Polymer Genome. This platform can be accessed at www.polymergenome.org. Users can easily query polymers of interest using an easy-to-use GUI interface and the polymer SMILES string. The platform uses the SMILES string, converts it into hierarchical fingerprints and predicts relevant properties using trained and deployed models. Polymer Genome returns results very rapidly from a few seconds to a
Figure 8.1: GUI-based polymer draw tool implemented in Polymer Genome. While a polymer is drawn, its smiles string is updated on-the-fly and when the drawing is done, it will be copied to the search tool by clicking button “Predict Properties.” The ladder polymer sketched in this figure is poly(naphthalene-2,3:6,7-tetrayl-6,7-dimethylene).

There are several methods to query a polymer namely by using its name, common abbreviation, the building block representation of its repeat unit, its class, its smiles string, and especially by drawing it using the implemented GUI-based polymer draw tool. Because writing a smiles string for a complex polymer is generally not straightforward and often very time-consuming, the polymer draw tool offers a very convenient and powerful method for graphically drawing the queried polymers. This tool handles linear and ladder polymers. Accepting the query for a polymer from users, Polymer Genome
returns its class, abbreviation, synonyms, and similar polymers, the 3D visualization of the repeat unit with atomic coordinates, and its predicted properties. Dozens of properties predicted are categorized into multiple groups, including electronic properties, dielectric and optical properties, mechanical properties, thermal properties, physical and thermodynamic properties, and solubility properties. A complete tutorial of the platform and details of the models are presented in a previous work by Tran et al. [28]

8.4 Outlook

Today, the time to develop a new material for a given application can be a staggering 10-20 years. Despite the vast contributions by the scientific community, utilizing the collective intuition of an entire field is a humanly impossible task. However, the need of the hour is to reduce the time-to-market as well as the development cost of new products by at least 50%.

A data-driven approach like the one presented here provides a cost and time-efficient method of rapidly screening through numerous possibilities and provides a rationally designed set of potential materials which can be used by experimental scientists to synthesize new materials for alternative energy applications. The most advantageous aspects of a data-driven workflow is the possibility that it can be developed in tandem with experimental input.

The polymer science community is at the cusp of a paradigm shift in the way that materials are discovered. Workflows such as this have shown the immense potential in a data-driven system to rapidly find materials with required target properties. However, there are some associated growing pains with this shift, the major one being the quest to organize, curate, and exploit experimentally and computationally generated data to make high-quality and standardized datasets for the community to utilize. Polymers currently occupy a shipment market of 500 billion dollars and is expected to grow rapidly in the coming years. Data-driven efforts such as this work could fundamentally alter and improve
the way materials are discovered and designed in a rapid, rationally-driven manner for a plethora of applications.
Appendices
APPENDIX A
MULTI-FIDELITY MODEL TO PREDICT THE CRYSTALLIZATION TENDENCY OF POLYMERS

A.0.1 High- and low-fidelity datasets

The datasets used in this study belong to two categories — high- and low-fidelity. The high-fidelity dataset comprises of accurate and explicit experimental values where the variance in the % crystallinity values for a polymer is only due to its processing and experimental measurement heterogeneity. The high-fidelity dataset comprises of 107 polymers with associated % crystallinity values. All values are curated from existing sources of experimental measurements like handbooks, published papers and online sources.[34, 35, 36, 37, 38, 39] The % crystallinity values in this dataset comprise of either explicit % crystallinity values via x-ray diffraction (XRD) and nuclear magnetic resonance (NMR) or are obtained via intensive properties like the heat of fusion and density measurements for polymers where pure-component (100% crystalline) values have been experimentally established. For cases where multiple values were reported, we used the median % crystallinity to train machine learning models. This value corresponds to the polymer’s tendency to crystallize at a high-fidelity level.

\[
\% \text{crystallinity} = \frac{\Delta H_{\text{fus}, \text{observed}}}{\Delta H_{\text{fus}, 100\% \text{crystalline}}} \times 100 \tag{A.1}
\]

The low-fidelity dataset comprises of % crystallinity values estimated using a combination of experimental and group-contribution methods. Therefore, the variance in this dataset arises from not just processing and experimental heterogeneity but also the inherent uncertainty of the group-contribution method. This makes the accuracy of this dataset lower than the high-fidelity dataset. The low-fidelity dataset includes 429 polymers.
For this dataset, the % crystallinity value for each polymer is calculated using Equation A.1 described above. Here, the heat of fusion $\Delta H_{fus,\text{observed}}$ is curated from differential scanning calorimetry (DSC) experiments from a variety of existing sources like polymer handbooks and prior published works. For polymers where multiple experimental heat of fusion values were reported, we used the median value to calculate the % crystallinity and train machine learning models. This value is the polymer’s tendency to crystallize at the low-fidelity level. However, the $\Delta H_{fus,\text{100\% crystalline}}$ is not available for all polymers. Therefore, we used group contribution methods using values established by Van Krevelen [34] to calculate the $\Delta H_{fus,\text{100\% crystalline}}$ values for this dataset. Additionally, we believe the spread of the calculated low-fidelity values to be similar to the distribution shown in Figure A.1 since the heat of fusion directly correlates to the polymer’s tendency to crystallize. For all polymers whose high-fidelity value is known, the low-fidelity value is also available.

Figure A.1: Graphical summary of the high and low fidelity datasets against the chemical space of polymers. Two leading components, PC1 and PC2, are produced by principal component analysis, and assigned to axes of the plot.

To demonstrate the diversity of the datasets, we compared both datasets using principal
component analysis (PCA) (Figure 3). Since the high-fidelity dataset is a subset of the low-fidelity dataset, we performed a principal component analysis (PCA) of the low-fidelity fingerprint vector (explained in the Methods section below). The horizontal and vertical axes are the first two principal components, PC₁ and PC₂. The high-fidelity dataset is less diverse than the low-fidelity dataset – the high diversity dataset includes polymers which belong to the class of polyolefins, polyvinyls and polyacrylics. However, the low-fidelity dataset also includes polymers in the polyimines and polyimides chemical space where the high-fidelity dataset is absent.

A.1 Methods

A.1.1 Feature set and dimensionality reduction

A hierarchical fingerprinting method was used to capture descriptors that control the tendency to crystallize in polymers and was described in a previous work.[27] The fingerprinting scheme comprises of four hierarchical levels of descriptors. The first level is at the atomic scale comprising of the count of atomic triples (e.g., O2-C3-C4, representing a two-fold coordinated oxygen, a three-fold coordinated carbon, and a four-fold coordinated carbon). The second set of fingerprint components captures a population of pre-defined chemical building blocks (e.g., -C6H4-, -CH2-, -C(=O)-). The third hierarchical level comprises of quantitative structure-property relationship (QSPR) descriptors, such as van der Waals surface area, topological surface area and the fraction of rotatable bonds, implemented in the RDKit cheminformatics library. The fourth and last fingerprinting level includes morphological features such as the topological distance between rings, fraction of atoms that are part of side chains and length of the largest side chain. Additionally, the melting temperature of each polymer was added as an additional fingerprint given its positive correlation with the heat of fusion (ΔHfus).[141]. For new polymers whose melting temperature is unknown, they can be predicted using the machine-learning based model implemented at www.polymergenome.org. Consequently,
this feature space includes 256 features.

To retain only relevant features, LASSO (least absolute shrinkage and selection operator) was performed on the initial 256-dimensional feature vector and both, the high and low-fidelity datasets with five-fold cross-validation.[142] LASSO is a shrinkage and selection method for linear regression which minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Since a single-fidelity GPR model is trained on the high-fidelity dataset as a baseline and a multi-fidelity co-kriging model is trained using datasets where the low-fidelity dataset is larger (with 429 polymers), feature engineering was performed on both datasets. For the high-fidelity dataset, LASSO retains 14 pertinent features (including the melting temperature) which are then used to train the single-fidelity GPR model. For the low-fidelity dataset, LASSO retained 97 important features (including the melting temperature) which was used to train the multi-fidelity co-kriging model. LASSO retains more features on the low-fidelity dataset since it is larger and more chemically diverse than the high-fidelity dataset.

Factors affecting the crystallization tendency of polymers

In addition to performing feature elimination to retain only relevant features, it is also valuable to analyze the retained features and its correlation to the tendency to crystallize. In this analysis, we study the 97 features retained after performing feature elimination on the low-fidelity dataset since the dataset is more diverse and representative of the polymer chemical space. The representative features are tabulated in Table A.1. As expected, there are positive correlations to atomic and block-level features like phenyl, carbonyl and anhydride groups which are known to increase the chain stiffness and therefore, its crystallization tendencies.[143] However, an interesting observation was that chain stiffening groups promote crystallization only when they are present in the main chain. In our dataset, chain stiffening groups like amide and sulfones were mostly present in the side chain (in polymers like acrylamides and certain sulfur containing polyoxides). Since
<table>
<thead>
<tr>
<th>Correlation with tendency to crystallize</th>
<th>Representative features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive correlation</strong></td>
<td></td>
</tr>
<tr>
<td>Atomic and block level</td>
<td><img src="https://example.com/carbon.png" alt="Carbon" />, <img src="https://example.com/phenyl.png" alt="Phenyl" />, <img src="https://example.com/carbonyl.png" alt="Carbonyl" />, <img src="https://example.com/anhydride.png" alt="Anhydride" /></td>
</tr>
<tr>
<td>Chain level</td>
<td>Relative number of atoms in the main chain</td>
</tr>
<tr>
<td>Special features</td>
<td>Melting temperature ($T_m$)</td>
</tr>
<tr>
<td><strong>Negative correlation</strong></td>
<td></td>
</tr>
<tr>
<td>Atomic and block level (in side chains)</td>
<td><img src="https://example.com/amide.png" alt="Amide" />, <img src="https://example.com/sulfone.png" alt="Sulfone" />, <img src="https://example.com/bulky.png" alt="Bulky pendant groups" />, <img src="https://example.com/terminal.png" alt="Terminal CH$_3$ group" /></td>
</tr>
<tr>
<td>Chain level</td>
<td>Length of longest side chain</td>
</tr>
</tbody>
</table>

Table A.1: Representative features strongly correlated to the crystallization tendency of polymers. R represents an arbitrary chemical group of C, O, N or H elements.[143] 

the crystallization is a function of the packing of the polymer, these functional groups were negatively correlated since the presence of chain stiffening functional groups reduce the packing including the aforementioned amides, sulfones as well as bulky ring-containing side groups. Since the majority of terminal CH$_3$ groups are present in side chains, this group is also negatively correlated to the tendency to crystallize. Further, there are correlations to block and chain-level features which propagated greater packing. For instance, it is positively correlated to the linear CH$_2$ chains, the length of the main chain (relative to the side chain) and negatively correlated to the length of the side chains (since they enhance branching, reducing the crystallinity). Additionally, the melting temperature is known to increase with the increase in the crystallinity [141], and we see a strong positive correlation between them. These chemical guidelines can help in rationally designing polymers with desirable crystallization tendencies.

97
A.1.2 Machine learning models

The single-fidelity machine learning model utilizes Gaussian process regression (GPR) with a rational quadratic kernel to map the high-fidelity dataset to its tendency to crystallize. GPR uses a Bayesian framework, wherein a Gaussian process is used to obtain the mapping from the polymer to its associated tendency to crystallize based on the available training set and the Bayesian prior, incorporated using the kernel function. In this case, the kernel function between two materials with features $x$ and $x'$ is given by:

$$k(x, x') = \sigma^2 \left( 1 + \frac{(x - x')^2}{2\alpha l^2} \right)^{-\alpha}$$

Here, the three hyperparameters $\sigma$, $l$, and $\alpha$ are the variance, the length-scale parameter and the expected noise in the data, respectively. These hyperparameters were determined during the training of the models by maximizing the log-likelihood. Further, five-fold cross validation was adopted to avoid overfitting.

The root mean square error (RMSE and R-squared) values were used to evaluate and compare the performances of the GPR and the co-kriging models. To estimate the prediction errors on unseen data, learning curves were generated by varying the size of the training and the test sets. For all ML models, i.e., GPR and co-kriging, the prediction accuracy was computed on a completely unseen and randomly chosen test set consisting of 22 data points (20% of the high-fidelity dataset). Additionally, for each case, statistically meaningful results were obtained by averaging RMSE results over 50 runs with varying training and test splits.[27, 44]
A.2 Results and discussion

A.2.1 Correlations between the low- and high-fidelity datasets

Group contribution methods have been commonly used to calculate a variety of intensive properties,[34] however, there has been no reported work on its prediction capability for % crystallinity. For this work, all polymers in the high-fidelity dataset have associated low-fidelity values calculated using 100% crystalline heat of fusion values using group contribution methods. Figure A.2 depicts a parity plot of 107 polymers and its associated high-fidelity and the low-fidelity values. The associated RMSE and $R^2$ are calculated as shown in Figure A.2.

![Parity plot of polymers](image)

**Figure A.2:** Prediction capability of the group contribution method for crystallization tendencies as a function of the high- and low-fidelity datasets.

This analysis yields two results. First, this analysis shows that for this limited dataset, the high- and low-fidelity datasets are quite closely correlated. The group contribution methods work especially well for polyolefins and polyvinyls. However, for polyesters and polyamides, the heat of fusion values waver from the experimental values as the complexity of the polymer (and thus, the number of groups) increases. Secondly, the good agreement between the high- and low-fidelity datasets make them excellent candidates for
multi-fidelity based machine learning using co-kriging. The co-kriging method utilizes the difference between the high- and low-fidelity values and a good correlation between the 2 values is beneficial to this model. Additionally, this quantifies the uncertainty error from the different methods of measurements.

A.2.2 Machine learning models — performance and comparison

Figure A.3: Prediction accuracy for machine learning models trained using single-fidelity GPR and multi-fidelity co-kriging methods. Part (a) comprises of learning curves for the single-fidelity GPR model and the multi-fidelity models with 407 and 429 low-fidelity points where the test sets consists of 22 randomly selected points and averaged over 50 runs, with error bars illustrating 1σ deviation. Figures (b) and (c) illustrate example parity plots with 107 high-fidelity train points and 22 test points for the GPR and co-kriging models respectively. The co-kriging model uses 429 low-fidelity points.

The high-fidelity dataset and a combination of the high- and low-fidelity datasets were used to train the machine learning models. The performance of machine learning models is evaluated using learning curves with varying test-train splits. The learning curve represents the learning trend of the model upon adding more data. For all the machine learning models in this work, the test set comprises of 22 points, which is around 20% of the high-fidelity dataset. Using only the high-fidelity dataset, a Gaussian process regression (GPR) model was trained using 14 pertinent features selected using LASSO. Using a combination of the high- and the low-fidelity datasets, co-kriging multi-fidelity machine learning models were trained using 97 important features selected using LASSO. The first multi-fidelity model
assumes the unavailability of the low-fidelity data for the test set, and therefore comprises of 407 low-fidelity points. The second multi-fidelity model accounts for the case where the low-fidelity values of the test set is available, and therefore uses all the 429 low-fidelity data points. The performance of the machine learning models can be evaluated from the learning curves presented in Figure A.3.(a), wherein average RMSE on the training and the test sets as a function of training set size are included. The error bars denote the $1\sigma$ deviation in the reported RMSE values over 50 runs. Figures A.3.(b) and A.3.(c) are parity plots which depict the prediction performance on individual cases included in the training and test sets for the single-fidelity GPR model and multi-fidelity co-kriging model with 407 multi-fidelity points which uses 85 training points and 22 test points. Along with learning curves, parity plots are a valuable comparison tool, in this case between the single and multi-fidelity model.

From Figure A.3.(a) we can make several pertinent conclusions. First, for all train-test combinations (Figures A.3.(a), A.3.(b) and A.3.(c)), the co-kriging multi-fidelity models outperforms the GPR single-fidelity model. The average RMSE of the GPR single-fidelity model which uses 85 training points is 17.04%. In comparison, the multi-fidelity models with 407 and 429 low-fidelity points has an average RMSE of 12.58% and 13.06% respectively. The performance of both the multi-fidelity models are quite similar, but as expected, the standard deviation for the model with 429 low-fidelity points is lower than that of the 407 low-fidelity points. The fairly consistent performance of the models can be attributed to the limited chemical diversity of the high-fidelity dataset where the learning is extremely rapid. This is also suggested from the similar performance trend displayed by both the single-fidelity and the multi-fidelity models. Additionally, the multi-fidelity models also utilize a large number of low-fidelity points in its training. With this, we have established that group contribution methods which were used in the low-fidelity dataset captures meaningful information that can be successfully used with co-kriging machine learning in this manner. Secondly, both multi-fidelity models perform
similarly. This implies that in order to make predictions for tendency of new polymers to crystallize under conventional synthesis and processing conditions, the low-fidelity crystallization tendency is not required. Therefore, this model does not require any further experimentation or information to make new predictions at the high-fidelity level of accuracy. Next, the RMSE of the multi-fidelity model can be discussed in comparison to the group contribution method’s prediction capability. The variance in the high- and low-fidelity datasets arises from 2 sources — the variance from different polymer processing conditions which was previously discussed and due to the group contribution method used to curate the low-fidelity dataset. The group contribution method is devoid of processing variability which is higher in value (13-15%) than its prediction error (10.11%). Additionally, the low-fidelity dataset is significantly more diverse and is not included in the group contribution analysis. This explains the higher RMSE of the multi-fidelity model compared to the group contribution method’s prediction error. Finally, we also believe that the multi-fidelity model captures the polymer processing variance; the processing variability established previously is 13 – 15% which is similar to the RMSE of the multi-fidelity models.
REFERENCES


[36] Polymer Database.


