

**AI-DRIVEN DESIGN OF CHEMICALLY RECYCLABLE POLYMERS TO
REPLACE COMMODITY PLASTICS**

A Dissertation
Presented to
The Academic Faculty

By

Aubrey Toland

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

Georgia Institute of Technology

December 2025

© Aubrey Toland 2025

**AI-DRIVEN DESIGN OF CHEMICALLY RECYCLABLE POLYMERS TO
REPLACE COMMODITY PLASTICS**

Thesis committee:

Dr. Rampi Ramprasad
School of Materials Science and Engineer-
ing
Georgia Institute of Technology

Dr. Seung Soon Jang
School of Materials Science and Engineer-
ing
Georgia Institute of Technology

Dr. Will Gutekunst
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Keith Hearon
CEO
Nuceptive Labs, Inc

Dr. Aaron Stebner
School of Materials Science and Engineer-
ing
Georgia Institute of Technology

Date approved: November 21, 2025

ACKNOWLEDGMENTS

I'd first like to acknowledge my advisor, Professor Rampi Ramprasad, for his mentorship and guidance throughout my PhD. Your continuous support and belief in my success, along with your profound passion for science have and will continue to serve as a constant inspiration for me. I have enjoyed being a part of your excellent research group tremendously, learning as much as I possibly can and I can only hope to become as great a scientist as you over many decades that will span my career.

To my committee members, I am honored to have your guidance throughout my journey as a graduate student. Professor Gutekunst your profound knowledge of polymer chemistry has helped steer me away from wrong turns many times throughout my research. Professor Stebner, your class cemented in me the importance of treating machine learning tools as precisely tools, always being vigilant to not trust blindly, letting science, not AI hype lead the way. Professor Jang, the class project of coding up a simple molecular dynamics simulator from the ground up reignited a needed spark for my passion for science, and I am truly grateful for that. Dr. Hearon, you have provided invaluable career advice before, during, and hopefully after my time as a PhD student. From working till 3 am in lab together when I was an undergraduate student, to introducing me to Rampi and his amazing group here at Georgia Tech, you have been fundamental in guiding me as a researcher.

To Huan, Lihua and Chiho I am very much indebted to you for your early support in me as I figured out how to be a graduate student and more importantly a researcher. Your mentorship laid a strong foundation for me to excel. To the many amazing people of the Ramprasad group it has been a pleasure to have been part of such an amazing research group together, and I have always enjoyed our conversations, whether they be scientific inclined or simply what we got up to the past weekend. In particular I would like to give a thank you to my office mates; Joe, Janhavi, Shivank, Brandon, Anagha you all made work so much more enjoyable each and every day. I am thankful for your friendship, support

and help.

I am also thankful to mentors I was lucky enough to have access to during undergrad. Dr. Cottrill you provided me with a fantastic first experience into the world of research and showed me what a truly dedicated researcher looks like. Professor Olivetti the 6 month stint of research I did in your lab was an amazing opportunity for me to learn first hand what research in the field of computational materials science looks like.

I also must mention the support of the Office of Naval Research through a Multidisciplinary University Research Initiative (MURI) Grant (N00014-20-1-2586) which has funded the major project I have had the privilege of working on. I also am extremely gratefully for support I received through the Graduate Research Fellowship Program (GRFP) from the National Science Foundation (NSF).

Lastly, I would be remiss if I did not acknowledge my friends and family. I have always been extremely lucky in this respect throughout my life. At every stage of my life I feel I have had strong friendships that have helped me find so much joy in life. To my friends from home, Mark and Matt in particular, you have provided me countless laughs and profound life advice all the same. To my college friends, truly too many to name, but to name a few; Thad, Graham, John, Skylar, Ryan, and so many more, you have helped me in so many important ways and made me into a better person. To my Atlanta friends, David, Andrew, Graham, Anshul, Zach you made life outside of the grind of graduate school so fun and I will cherish all of the memories we made in Atlanta together.

I am proud to say I have a fantastic family that I am lucky to be so close with from my two parents and brother who have always supported me through all of life, to all my aunts, uncles and cousins who have and I know will continue to help me in countless ways. To my soon to be sister in law Georgia, thank you for always bringing the high energy fun and in turn making my brother a better person. There are too many cousins to name all of them but I do want to give a shout out to Zoe for making time to come to my defense, you are the closest thing I have to a sister (but Georgia is doing her best to give you a run for your

money). To my brother Owen, you always pushed me to be better while still remaining one of my closest support systems. From trampoline basketball to learning to function as an adult, you have always driven me to achieve more and showed me how to enjoy life at the same time. I could not have learned from a better role model than you throughout my life. Lastly, to my Mom and Dad, I cannot thank you enough for the life you have provided me, the lessons you have taught me, and the unconditional love you have given me. I know for a fact if it was not for your support throughout my entire life I would have never thought of myself as a good student capable of anything near a PhD.

TABLE OF CONTENTS

Acknowledgments	iii
List of Tables	ix
List of Figures	x
List of Acronyms	xiv
Summary	xvi
Chapter 1: Introduction and Background	1
1.1 A Need to Replace Commodity Plastics	2
1.2 Polymer Thermodynamics and Ceiling Temperature	2
1.3 Ring Opening Polymerization Polymers and Chemical Recycling	4
1.4 Computational Tools for Calculating and Predicting Thermodynamics of ROP Polymers	5
1.5 Commodity Plastics and the Design Algorithms to Replace Them	8
Chapter 2: Methodology	11
2.1 Ab Initio Method to Calculate ΔH^{ROP}	11
2.2 Creating Machine Learnable Representations of Polymers and Monomers: Fingerprinting	13

2.2.1	Hierarchical Fingerprints for Polymers: Polymer Genome Fingerprints	13
2.2.2	Machine Learned Fingerprints for Polymers	14
2.2.3	Other Fingerprinting Schemes for Polymers and Molecules	14
2.3	ML Algorithms	15
2.3.1	Gaussian Process Regression	16
2.3.2	Support Vector Machine	16
2.3.3	Random Forest and Boosted Random Forest	17
2.3.4	Deep Neural Networks	18
2.3.5	Multi-task Learning	19
2.4	Machine Learning Force Fields	20
2.4.1	Covariance-normalized Equivariant Transformer	20
2.5	Virtual Polymer Design	22
2.5.1	Virtual Forward Synthesis	22
2.5.2	Genetic Algorithm	23
2.5.3	BRICS	24
2.5.4	SA Score	24
2.5.5	Fitness Function to Determine Promise of Polymer Candidates	25
2.6	Natural Language Processing Augmented Data Capture	26
Chapter 3: Development of ΔH^{ROP} Multitask Model: An Essential Tool to Discovering Depolymerizable Polymers		28
3.1	Introduction	28
3.2	Creation of ΔH^{ROP} Dataset	28

3.3	Multitask Model Training, ML Algorithm Selection, and Results	29
3.4	Production ΔH^{ROP} Multitask Model	31
Chapter 4: PolyGET: A Machine Learned Force-Field for ROP Polymers		33
4.1	Introduction	33
4.2	Poly24 Dataset	38
4.3	Force-centric Training of PolyGET	40
4.4	PolyGET Validation	42
Chapter 5: Virtual Polymer Design of Depolymerizable Polymers		51
5.1	Introduction	51
5.2	Predictive Models Leveraged	53
5.3	Virtual Forward Synthesis	55
5.4	Genetic Algorithm	62
Chapter 6: Conclusion, Outlook and Future Work		75
6.1	Conclusion	75
6.2	Future Work	78
References		82

LIST OF TABLES

3.1	Summary of the ΔH^{ROP} data generated, accumulated, and used herein. . .	29
3.2	Summary of four cases used in evaluating the ML algorithms, which are different in the training data.	30
3.3	RMSE, given in kJ/mol, and R^2 obtained from SVM, RF, BRF, and GPR for different cases described in the text.	31
4.1	Polymers in our used Poly24 dataset for training PolyGET. We classify the polymers into 4 broad categories: cycloalkanes, lactones, ethers, and others.	40
5.1	Table of target properties for the different polymers of interest we wish to replace using generative algorithms.	53
5.2	Sample of successful candidates for each target application and their properties from VFS workflow. For each property column the prediction is shown normally and then the uncertainty for the given prediction is shown in parentheses	62
5.3	Sample of successful candidates for each target application and their properties from the GA algorithm.	69

LIST OF FIGURES

1.1	Flowchart for the general process to design new, depolymerizable polymers to replace targeted commodity plastics.	10
3.1	Correlations between $\Delta H_{\text{expt}}^{\text{ROP}}$ and $\Delta H_{L=N}^{\text{ROP}}$, shown for (a) $L = 3$, (b) $L = 4$, (c) $L = 5$, (d) $L = 6$, and (e) $L = \infty$. In the plots r corresponds to the Pearson correlation between $\Delta H_{\text{expt}}^{\text{ROP}}$ and $\Delta H_{\text{comp}}^{\text{ROP}}$ and indicates how well correlated the variables are for a given L	29
3.2	Parity plot for the pre-production model where 10 % of the $\Delta H_{\text{expt}}^{\text{ROP}}$ data was withheld. Blue data points represent the test data, while grey data points represent the train data (which contain both $\Delta H_{\text{comp}}^{\text{ROP}}$ and $\Delta H_{\text{expt}}^{\text{ROP}}$)	32
4.1	Illustration of PolyGET. (1) We prepare Poly24, a database consisting of DFT simulation data for 24 types of polymer types. (2) We propose a Equivariant Transformer architecture called CNET to predict atomic forces and energy for the polymers. (3) The CNET model is optimized on Poly24 with force-centric optimization, and predicts accurate energy with linear adjustment. (4) The resulting model can extrapolate to robust and DFT-level accurate simulations of large loop polymers. (5) We test PolyGET in applications such as ring-opening enthalpy prediction, achieving performance equal to or surpassing DFT, while enabling much longer MD simulations for large polymers.	33
4.2	Accuracy of force predictions of PolyGET examined in all loop sizes up to loop 25. Each figure shows the scatter plot of predicted energies forces versus DFT calculations from the aggregation of all loop sizes for each polymer family. For each loop, 100 samples are uniformly sampled from the PolyGET trajectory. In the upper left, we show the histogram of the error, where the x-axis represents values of absolute error and y-axis represents frequency. In the bottom right we annotate the Mean Absolute Error (MAE) in eV/Å and Cosine Distance (CosD). The consistent high accuracy and correlation shows that PolyGET can extrapolate to larger polymer systems without performance deterioration.	43

4.3	Visualization of simulation statistics. In each subplot, we visualize two subplots, where the upper one shows the potential energy plot, and the bottom one shows the RMSD value of conformations during the simulation, which measures at each timestamp how far the conformation deviates from the initial conformation. The potential energy plots show a convergence to equilibrium for all simulations. The RMSD further shows the stability of simulations, as the conformations remain in certain proximity to the initial conformation while exploring different conformational spaces, as illustrated by (i) and (j).	44
4.4	Comparison of ΔH^{ROP} calculations done using both PolyGET model and ab-initio methods, i.e. DFT. In (a) we show the calculation done for poly(ethylene) and in (b) we show the calculation done for poly(ethylene oxide). The blue dots represent the ΔH^{ROP} calculated for a given system size from PolyGET model, the blue line shows the exponential fit said data, and the red Xs ΔH^{ROP} calculated from DFT based trajectories. Inset in the top left of the figure the ΔH^{ROP} values for experimental measurement, DFT based calculations and PolyGET calculations are shown respectively. Inset in the plots in the top right we see the polymerization reaction for the polymer being simulated.	47
4.5	Parity plots for both forces (a) and energies (b) for 16 chemistries from the Poly24 dataset.	49
4.6	Parity plot for enthalpy calculations made using PolyGET as the force field to run molecular dynamics. This method results in an RMSE of 7.1 kJ/ml and an R^2 of 0.964	50
5.1	In part (a) we see three general guidelines to follow for designing new polymers. First a screening criteria is defined, next ML models are developed for properties of interest to perform said screening, the design search space is decided and a candidate list is generated, and lastly, the generated polymers have their properties predicted and screened for the application at hand. In (b) we see the radar plot for the polymers of interest that we see to replace PET, PS, PP, and HDPE.	53
5.2	Parity plot illustrating model performance across different properties. The first set of values in the top left showcase counts and RMSE on the model test data of known polymers, which are represented by red circles. The second set delineates the dataset size and model performance specifically on known ROP polymers, denoted by black stars. For ΔH^{ROP} , dots depict training data, while black stars indicate test data, as the model was exclusively trained on ROP polymers.	55

5.3	Histograms of (a) mechanical and thermal property predictions and (b) their associated uncertainties	56
5.4	Schematics for the multi-stage screening process to achieve sustainable polymer candidates for various commodity plastic applications.	57
5.5	(a) Example of GA Initialization. The scaffold is a five-member cycloalkane with functionalization positions denoted by the labeled asterisks (*). New molecules are created by bonding the atom adjacent the * to the atoms adjacent the * in the R-groups displayed in the center. Some example monomers are displayed at the bottom that could be created from this scaffold an these R-groups. (b) Schematic illustration of the GA workflow, depicting the key processes of crossover, mutation, polymerization, property prediction, and fitness evaluation/selection. Selection is based on the predicted properties of the polymer as well as molecular complexity of the monomer, while crossover and mutation is performed on the monomer structure.	64
5.6	Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PET. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.	70
5.7	Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PS. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.	71
5.8	Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PP. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.	72
5.9	Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for HDPE. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.	73

5.10 Stacked histogram plots of polymer fitness for each application of interest:
(a) PET, (b) PS, (c) PP, (d) HDPE. Each polymer class used in the GA is represented by a different color as indicated by the legend to the right. As can be seen the histogram plots get progressively more right skewed from (a) to (d) as a result of the screening criteria being less and less difficult to meet from PET to HDPE. 74

LIST OF ACRONYMS

E Young's modulus

L integer number of monomer repeat units

R^2 coefficient of determination

T_c ceiling temperature

T_d degradation temperature

T_g glass transition temperature

T_m melting temperature

ΔG change in Gibbs' free energy

ΔH change in enthalpy

ΔH^{ROP} ROP polymer ΔH

ΔH_{calc}^{ROP} computational data for ΔH^{ROP}

ΔH_{exp}^{ROP} experimental data for ΔH^{ROP}

ΔS change in entropy

ΔS^{ROP} ROP polymer ΔS

ϵ_b elongation at break

σ_b ultimate tensile strength

BRF boosted random forest

BRICS breaking of retrosynthetically interesting chemical substructures

CRM Chemical Recycling to Monomer

DFT density functional theory

DNN deep neural network

EPA United States Environmental Protection Agency

GA Genetic Algorithm

GPR Gaussian process regression

HDPE high-density polyethylene

LDPE low-density polyethylene

LOOCV leave one out cross validation

MD molecular dynamics

ML machine learning

MLFF machine learned force field

PE polyethylene

PET polyethylene terephthalate

PP polypropylene

PS polystyrene

PVC polyvinyl chloride

RF random forest

RMSE root mean square error

ROP Ring Opening Polymerization

ROP polymers polymers synthesized via ring opening polymerization

SA Score Synthesis Accessibility Score

SMILES molecular-input line-entry system

SVM support vector machine

VFS Virtual Forward Synthesis

SUMMARY

Addressing the global plastic waste crisis requires a new paradigm of polymeric material that can be depolymerized back to monomer, enabling true chemical recycling. Polymers synthesized via Ring Opening Polymerization (ROP) have shown promise in the fact that they tend to have the necessary thermodynamics to be depolymerizable but lack the mechanical and thermal robustness needed for commercial adoption. This challenge provides an ideal opportunity for AI-driven design to develop such sustainable materials. Herein multiple machine learning (ML) models for relevant polymer properties work in tandem with generative algorithms to optimize across various necessary objectives for creating industry relevant and sustainable polymers. One crucial property in determining the depolymerizability tendencies of polymers is the change in enthalpy (ΔH) of polymerization. To handle this property, a ML algorithm to predict ΔH , that utilizes both experimental and ab initio data for enhanced accuracy, has successfully been developed, and continues to be improved so that polymers can efficiently be screened for the potential to be depolymerizable. In addition, current mechanical and thermal ML polymer property predictors have also been retrained and improved to better account for the ROP chemical space. Moving forward, this work identifies robust screening criteria to identify recyclable polymers with the potential to replace conventional food packaging plastics such as polyethylene terephthalate (PET), high-density polyethylene (HDPE), polystyrene (PS), and polypropylene (PP). These criteria are then be put to action using two generative algorithms, Virtual Forward Synthesis (VFS) and a Genetic Algorithm (GA) to screen through millions of hypothetical polymers synthesized via ring opening polymerization (ROP polymers). VFS screens commercially available monomers to discover promising ROP polymers that can be synthesized today, while the GA looks to the future to discover new potential polymers, pushing the boundaries of truly recyclable plastics. Close collaborations with experimentalists to create the most promising polymers from this work have been in place and it is

the true goal that recommended polymers from this work result in tangible progress in the creation of sustainable plastics for a circular economy.

CHAPTER 1

INTRODUCTION AND BACKGROUND

Plastic has become a ubiquitous material in our society. It plays an essential role in transporting goods around the world and extending the shelf life of food, among among many other uses of great importance to society. This prolific use along with poor solutions for plastic products at the end of their life cycle has led to a global crisis of plastic waste as it contaminates every ecosystem on earth and has negative impacts on human and environmental health [1]. Plastic has seen industrial level production since about the 1950s and of the 7 billion metric tonnes of plastic waste produced since that time only 10% of that plastic waste has been actually recycled [2]. Instead plastic waste is either leaked to the environment, landfilled or incinerated. According to the United States Environmental Protection Agency (EPA), in 2018 35.7 million tons of municipal solid waste was plastic, and only 8.7% of this plastic waste was recycled. The rest was either landfilled (75.6%) or incinerated for energy (15.7%) [3]. Neither of these options are ideal. Landfilling plastic waste tends to lead to environmental damage as degrading chemicals leach into the environment and plastic waste leaves the landfill to pollute the surrounding environments [4]. Whereas, incineration of plastic releases toxic products as fumes and greenhouse gases [5, 6]. While it can be common belief that this problem can be solved with the catchy slogan of "reduce, reuse and recycle", this antiquated tag line requires a new "R" to lead to a world where plastic can be produced and used in harmony with the environment; that new "R" is redesign, specifically molecular redesign. This thesis proposal is meant to be a small step in the direction of designing plastics from a molecular level to be capable of full chemical recycling so that a circular economy of plastic can be achieved one day.

1.1 A Need to Replace Commodity Plastics

There is a massive problem in the way current commodity plastics are recycled. In plastic recycling there can be thought of two major categories of recycling, mechanical recycling and chemical recycling. In mechanical recycling mechanical and thermal processing of plastic waste leads to new products of lower quality due to radical chain scission and coupling [7, 8]. This avenue of recycling is not sustainable as plastic waste can only be recycled in this manner so many times before a plastic of no value is achieved and has to be land-filled or incinerated. Mechanical recycling accounts the majority of plastic recycling that occurs. [2]

On the other hand chemical recycling is a sustainable process where plastic waste is processed such that it returns back to original chemical feedstocks, i.e. monomers, so that it can be repolymerized into products of the same value. This is also known as Chemical Recycling to Monomer (CRM) and this avenue of recycling is advantageous as polymers without degradation to original properties can be recreated. [9, 10, 11, 12] Thus, in theory, CRM leads to sustainable plastic production where plastic products of the same value can be created indefinitely. Currently chemical recycling is possible with some plastics but only accounts for approximately 14% of plastic collected for recycling. While chemical recycling does suffer from some similar setbacks as mechanical recycling, such as collection and sorting issues, the increase in value recovery of plastic products that chemical recycling provides, indicates that it has a much greater potential to become economically viable with producing virgin plastic from petroleum sources when compared to mechanical recycling [11, 13].

1.2 Polymer Thermodynamics and Ceiling Temperature

To design novel polymers capable of CRM, polymers that can be depolymerized via external stimulus are necessary. One promising avenue in this direction are polymers that can

be triggered via thermal stimulus and thus an important property called ceiling temperature (T_c) becomes relevant. T_c is defined by the International Union of Pure and Applied Chemistry (IUPAC) as "temperature above which, in a given chain polymerization, polymer of high molar mass is not formed" [14]. T_c is further mathematically defined as follows:

$$T_c = \frac{\Delta H_m^\circ}{\Delta S_m^\circ + R * \ln \frac{[M]_0}{c^\circ}} \quad (1.1)$$

Where ΔH_m° and ΔS_m° are the enthalpy change and entropy change per mole of monomer reacted, respectfully, R is the ideal gas constant, $[M]_0$ is the original monomer concentration, and c° is $1 \text{ mol} * \text{dm}^{-3}$. For simplicity of notation we will refer to ΔH_m° and ΔS_m° as ΔH and change in entropy (ΔS), moving forward.

It should be noted that in the vast majority of cases of polymerization, the reaction is enthalpically driven and must overcome an entropic barrier. In terms of the mathematical definition of T_c , this means that both ΔH and ΔS are negative. It becomes clear that this represents an enthalpic drive and entropic barrier to overcome when one considers change in Gibbs' free energy (ΔG) where $\Delta G = \Delta H - T\Delta S$. A negative ΔG indicates that a reaction is thermodynamically allowed whereas a positive ΔG means a reaction is thermodynamically forbidden. Thus a negative ΔH drives a reaction towards being thermodynamically allowed but a negative ΔS drives a reaction to be thermodynamically forbidden. It is important to note that when both ΔH and ΔS are negative T_c is positive. On the other hand if ΔH and ΔS have opposing signs T_c will be mathematically defined to be negative, but in practice it means a T_c does not exist for this polymer since a negative temperature is not possible. In the rare case that both ΔH and ΔS are positive, a floor temperature is achieved where polymerization only occurs above this temperature, and depolymerization occurs below this temperature (i.e. the opposite of a T_c). Only a few known polymer systems have a floor temperature which include S_8 , $(Me_2SiO)_4$, and Se_8 [15] as well as select other polymer systems [16]. Therefore, due to how rare a floor

temperature is for polymers, only polymers that exhibit a T_c are considered for this work.

In practice T_c is important when considering CRM because when polymer is subjected to a temperature well above its respective T_c it can depolymerize back to monomer, and thus be CRM. For many current commercial plastics this is an issue as the degradation temperature (T_d), or the temperature at which a polymer begins to chemically break down and lose its desirable physical properties is below the T_c . An example of this is polyethylene (PE), where thermolysis of PE leads to only 25% monomer yield with the rest of products being tar, gas and waxes as the polymer degrades significantly above its T_c . [17] Thus, in order to achieve truly sustainable plastic recycling, molecular redesign of polymer systems are necessary to design polymers that can depolymerize to monomer economically enabling CRM at scale.

1.3 Ring Opening Polymerization Polymers and Chemical Recycling

When designing polymers to be depolymerizable on demand it is essential to design for a T_c that balances the need for a high enough T_c for the polymer to have the necessary thermal robustness for a given application, while still possessing a low enough T_c that it is economically feasible to depolymerizable the polymer at scale. Further, the T_c must be well below the T_d so that the main product is monomer when it is thermally triggered to depolymerize. Thus to design a polymer to truly take advantage of CRM, a delicate balance of polymer energetics must be struck.

When considering what class of polymers holds the most promise to be compatible with these necessary thermodynamic requirements, ROP polymers have emerged in literature as a leading class of promising polymers [10, 18, 19, 20, 21]. ROP polymers tend to exist in a space thermodynamically where they have low enthalpic drive (small negative value for ΔH) and also a small entropic barrier (also a small negative value for ΔS) to polymerize, resulting in polymers that can polymerize with high yield but also depolymerize at reasonable temperature. Further, many different heterocycles are known to polymerize mak-

ing it so ROP polymers cover a wide chemical space including polyalkanes, polyalkenes, polyethers, polyamides, polycarbonates, polyesters, polythioesters, polythioethers, polythionoesters and polythiocanes [22, 23, 24, 21, 18, 25]. This is important as we consider that polymers of the future need not just be depolymerizable but also have the mechanical and thermal robustness to replace current polymers used in commodity plastics. Therefore, this research will focus on exploring the design space of ROP polymers and thus we will specifically refer to thermodynamic quantities moving forward as such, ROP polymer ΔH (ΔH^{ROP}) and ROP polymer ΔS (ΔS^{ROP}). By restricting the design space to ROP polymers, we are maximizing the potential for discovering a polymer capable of undergoing CRM.

1.4 Computational Tools for Calculating and Predicting Thermodynamics of ROP Polymers

In order to discover polymers with the necessary thermodynamic properties to be depolymerizable as well as the necessary thermal and mechanical properties to perform up to the standard of the convention plastics they are intended to replace, many hypothetical molecular designs must be explored. From just the ROP chemical space alone there are millions of hypothetic polymers that can be polymerized from monomer feedstock available today. [26] Further, when we look past just the ROP polymers derived from chemical feedstocks that are available today the design space expands to billions. In order to search this space efficiently, computational tools become essential as experimental methods require too much time to feasibly search the chemical space. This can be seen when just considering the main experimental method used to calculate both ΔH^{ROP} and ΔS^{ROP} ; nuclear magnetic resonance (NMR) performed at different temperatures. NMR can be performed this way either through individual batch experiments that are quenched at the desired temperature, or through in situ measurement with variable temperature NMR (VT-NMR). The data from these experiments is used to generate a van't Hoff plot from which the ΔH^{ROP} and ΔS^{ROP}

can be calculated. In practice this experimental method takes hours to a day for one chemistry, but that is with an optimized workflow where a suitable solvent for the given polymer is already known, and the polymer’s synthesis route is well established. If the synthesis route is unknown, it can take days, weeks or even months to understand how to polymerize a new ROP polymer of interest. While other experimental methods exist to measure ΔH^{ROP} and ΔS^{ROP} , namely calorimetry, this method is considered to be less accurate and is thus less popular in contemporary literature. Further, while using calorimetry is faster than NMR, this method on it’s own is still too slow to search the massive chemical space, and still has the same bottleneck of polymer synthesis and optimization. Therefore, given the chemical space of over a billion hypothetical polymers experimental trial-and-error based approaches are impractical. This opens the need for computational tools of varying degrees of cost and accuracy. In the context of this research, these tools include ML predictions and ab initio based molecular dynamics (MD).

When it comes to screening for depolymerizable polymers computationally, the more readily tuned thermodynamic property between ΔH^{ROP} and ΔS^{ROP} is ΔH^{ROP} . Further, the available literature data for ΔS^{ROP} has been shown to be calculated from temperature variable NMR inconsistently making any sort of data driven method extremely difficult [27]. On the other hand, ΔH^{ROP} is closely related to the monomer ring size (i.e. the ring strain), so it can be measured accurately [28, 29] and computed at reasonable levels of fidelity [30], and unlike ΔS^{ROP} , it is not a strong function of measurement variables such as solvent used. This second point, that ΔH^{ROP} can be computed with high fidelity is very important, as the amount of experimental data available in literature is scarce making traditional data driven methods difficult. As such, a state-of-the-art computational method [30] was employed to generate computational ΔH^{ROP} , so that the experimental dataset could be expanded to cover a larger ROP chemical space. To make a clear distinction between experimental and computational data moving forward, we use the following nomenclature: experimental data for ΔH^{ROP} (ΔH_{exp}^{ROP}) and computational data for ΔH^{ROP} (ΔH_{calc}^{ROP}).

This method involves uses ab initio MD simulations powered by density functional theory (DFT) for both monomer and polymer systems of varying size to compute ΔH_{calc}^{ROP} , and is described in detail in section 2.1. Thus, DFT calculations for ΔH_{calc}^{ROP} are an important computational tool to allow for generating data in new chemical spaces that have yet to be explored experimentally. Further, these ab initio based MD simulations also can be leveraged as a method to validate less computationally expensive ML predictions without taking as long as experimental validation.

The other computational tool of importance to this research is ML, which has been proven to show significant promise in the field of polymer informatics [31, 32, 33, 34]. In the context of this research there are two flavors of ML used, one is direct property prediction and the other is emulation of ab initio MD calculations. Property prediction is the use of ML algorithms to make direct predictions of polymer properties such as ΔH^{ROP} . These ML models require experimental data to train and can make predictions for a given polymer in fractions of a second. This is essential when designing novel polymers and millions to billions of polymers potentially need to be screened. The other flavor of ML covered in this research is emulating DFT based ab initio MD with ML. This type of model is referred to as a machine learned force field (MLFF). For MLFFs atomic forces, energies and/or stresses calculated by DFT are the ground truth that a deep learning model is trained on. This allows for MD at near ab initio accuracy that runs for orders of magnitude more steps and for much larger atomic systems than is feasible with DFT. By using this combination of computational tools, tools of lower levels of accuracy and cost can be employed first to screen down from a massive number of polymer designs to something more reasonable. Then the more promising polymer designs can use computational tools of higher accuracy and computational cost for further screening.

1.5 Commodity Plastics and the Design Algorithms to Replace Them

It is a common misconception that since the current plastic recycling chain does not currently work sustainably that we can solve this issue through just reduction of plastic use. This is not the case because for certain products such as food barriers and packing material, plastic is essential to the global economy. Plastic in food packaging is essential for extending the shelf life of foods allowing more people to be fed. Plastic packaging in general is essential to the global shipping of goods due to plastic's low density combined with its useful toughness. Plastic products can actually save fuel transporting goods and thus reduce greenhouse gases. This can be exemplified in a 2005 study [35] that found by using PET bottles instead of glass or metal energy consumption with regard to the transport of these goods was reduced by 52% in Europe. This is one simple, but demonstrable example of why plastic is extremely important to the global economy. Thus, the question becomes how can essential commodity plastic be replaced by chemically recyclable polymers.

The first step to designing new, chemically recyclable plastics to replace commodity plastics is to identify the commodity plastics that are most promising to replace. As previously noted, packaging plastic is a promising class of commodity plastic to work towards replacing because it is a plastic product that is very important and it makes up just over 40% of plastic waste [36]. The most common polymers used for food packaging are PET, HDPE, low-density polyethylene (LDPE), polyvinyl chloride (PVC), PP, and PS [36]. The combination of these plastics account for just over 60% of primary global plastic production in 2019 [37]. Of these, only PS and PET are able to be chemically recyclable [38], but neither have been able to be economically viable as the recycled material is more expensive than the respective virgin plastic. Therefore, given the volume at which these polymers are produced and the value they provide the world, they are the most promising polymers to pursue to replace. This research focuses on discovering depolymerizable alternatives for PET, PS, PP, and HDPE.

To this end this research will leverage two generative algorithms to work towards discovering new promising polymers. These algorithms include VFS and a GA. Figure 1.1 shows the general framework for the design process to create novel depolymerizable polymers to replace conventional plastics. First, a target commodity plastic is selected to work towards replacing, and a screening criteria is based on the polymer corresponding to this commodity plastic application. As mentioned, this research is first focusing on PET. The properties of interest to create the screening criteria, which can be found in Chapter 5, are ΔH^{ROP} , glass transition temperature (T_g), melting temperature (T_m), T_d , ultimate tensile strength (σ_b), Young's modulus (E), and elongation at break (ϵ_b). T_g , T_m , and T_d have been identified as the properties of interest to assess a potential polymer's thermal robustness while σ_b , E and ϵ_b are used to assess mechanical robustness.

Next hypothetical polymers are designed via a generative algorithm and these polymers' properties are predicted using ML. The screening criteria is enforced with a fitness function to parse through the generated polymers and select the most promising candidates. These most promising candidates are then selected to have their predicted ΔH^{ROP} validated by MD either via DFT or MLFF depending on whether the chemistry is expected to be within the training data of the MLFF. Polymers whose thermodynamic properties are validated by MD are then sent to chemists for synthesis and experimentation. If a polymer can be synthesized, then experimentation can be performed to discover its true properties. The computational and experimental data can then be stored and used to retrain and/or fine-tune the ML property prediction models so they are even more accurate. If the properties are experimentally validated to be competitive with the commodity plastic the screening criteria was derived from, a success case has been identified and the novel, depolymerizable polymer could replace it allowing for a chemically recyclable alternative to the original commodity plastic.

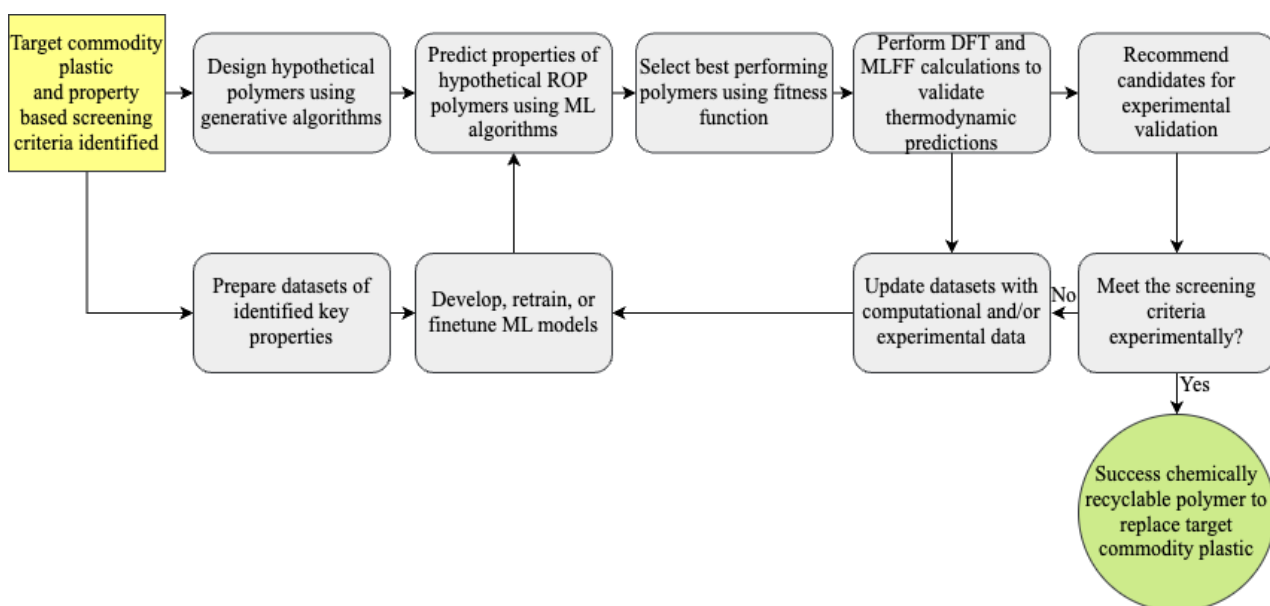


Figure 1.1: Flowchart for the general process to design new, depolymerizable polymers to replace targeted commodity plastics.

CHAPTER 2

METHODOLOGY

The goal of this research, in general, is to use computational tools to design novel polymers that possess the necessary thermodynamic properties to be depolymerized on demand ΔH^{ROP} , and have enough mechanical (E, σ_b) and thermal (T_g, T_m, T_d) robustness to be a good candidate to replace current commodity plastics. Specifically we first focus on replacing PET due to how important it is to society and how widely it is used, but plan to extend the work to other polymers used in the food packaging industry. As a result this research will encompass the development of ML models, needed fingerprinting for said models, ab initio MD calculations for ΔH^{ROP} , MLFF for ROP polymer dynamics, and finally leveraging of generative algorithms to discover depolymerizable polymers. This methodology section will introduce and explain how each of these various tools are implemented and work.

2.1 Ab Initio Method to Calculate ΔH^{ROP}

ΔH_{calc}^{ROP} was generated using the multistep procedure involving both classical MD and ab initio MD [30]. This method consists of three main steps; atomic model construction for both monomer and polymer, configurational space searching and sampling, enthalpy calculation via extrapolation. First in model construction, a series of closed loops comprised of an integer number of monomer repeat units (L) were constructed using a package called Polymer Structure Predictor [39, 40]. These loops are representations of the monomer system and polymer systems of various sizes. As $L \rightarrow \infty$, the loop approaches the true polymer limit, and $L = 1$ represents the monomer. The computations were generally performed for $L = \{1, 3, 4, 5, 6\}$. A classical MD simulation using an empirical Reax force field [41] was performed for each monomer/polymer model, thoroughly exploring the con-

figuration space while preserving the atomic connectivity. Using classical MD, trajectories of over 1 ns, or 1 million steps, were generated and thousands of snapshots were obtained. These snapshots were then sampled to maximize the atomic position diversity of the sample set using the Smooth Overlap of Atomic Positions (SOAP) [42, 43] fingerprinting method to vectorize each atomic configuration and sample vectors with the greatest distance from each other. The set of maximally diverse atomic configurations for each atomic model were then used to run ab initio MD simulation. The purpose of this step, using classical MD, is to provide a set of maximally diverse initial atomic structures to run ab initio MD on.

Next, a room-temperature ab initio MD simulation was performed for each sample, obtaining the lowest-energy equilibrated trajectory. The L -dependent estimation of ΔH_{calc}^{ROP} was then computed as $\Delta H_L^{ROP} = \frac{1}{L} * \langle E_L \rangle - \langle E_1 \rangle$, where $\langle E_L \rangle$ and $\langle E_1 \rangle$ are the potential energies at equilibration of the ab initio MD trajectories of the polymer model ($L > 1$) and monomer model ($L = 1$), respectively, while $\langle \cdot \rangle$ symbolizes the average over the ensemble of the microstates. Finally, in the extrapolation step ΔH_{calc}^{ROP} was computed as the limit as $L \rightarrow \infty$ (or, equivalently, $\frac{1}{L} \rightarrow 0$) of ΔH_L^{ROP} , i.e. $\Delta H_{comp}^{ROP} \equiv \lim_{L \rightarrow \infty} \Delta H_L^{ROP}$. In practice ΔH_{calc}^{ROP} was computed by assuming that ΔH_L^{ROP} depends linearly on $1/L$ and then making suitable extrapolations using a linear fit of ΔH_L^{ROP} vs. $\frac{1}{L}$ to the limit of $\frac{1}{L} \rightarrow 0$.

The central idea of this computational scheme is that polymers are soft materials; thus, they are naturally not locked at any specific atomic configuration but rather switch across multiple microstates continuously and rapidly. Therefore, this scheme was designed to thoroughly explore the configuration space at two levels of speed and accuracy: The first is in a less accurate but faster fashion, using a Reax force field with Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [44]. The second is using DFT with Vienna Ab initio Simulation Package (VASP) [45].

2.2 Creating Machine Learnable Representations of Polymers and Monomers: Fingerprinting

In order to be able to train ML models or subsequently make ML predictions polymer fingerprinting is necessary. The first step is to encode a given polymer into a simplified molecular-input line-entry system (SMILES) string. This SMILES string then has features, or numerical values to represent the physical structure encoded in the SMILES string, extracted. In this research two types of fingerprinting are used: hierarchical based fingerprints and machine learned fingerprints. The following sections will describe both.

2.2.1 Hierarchical Fingerprints for Polymers: Polymer Genome Fingerprints

The hierarchical features extracted from a smiles string in this work are motif-based topological fingerprints [46], with three levels of descriptors. The first level is a count of the number of atomic triplets present in the polymer repeat unit. The second level is block level descriptors which cover larger, known groupings of atoms such as carboxyl groups, benzene rings, phenyl rings etc. The third level covers quantitative structure-property relationship (QSPR) to describe molecular features such as molecular chi indices and Van der Waals volume. This is referred to commonly in this research as PolymerGenome fingerprints.

Additional features can be added on top of these PolymerGenome fingerprints. This specifically is used to train the multi-task ΔH^{ROP} model as will be discussed in section chapter 3. The nature of ΔH^{ROP} requires these additional features as ΔH^{ROP} data can come from either experimental or computed sources and for ΔH_{calc}^{ROP} data can be derived from models of various size L , as described in section 2.1. Thus to account for this a one hot encoding was added to specify whether the data is experimental or computational, and $\frac{1}{L}$ was added as a feature, where $\frac{1}{L} = 0$ is set for experimental data, to inform the model of the size of the calculation the data is derived from. Further, in order to differentiate

different monomers that polymerize to the same polymer, such as all simple cycloalkanes of various ring size polymerizing to poly(ethylene), an additional feature of ring size was also added which simply provides an integer number for number of atoms in the ring that opens during the ring opening polymerization reaction. These additional features are simply concatenated to the PolymerGenome fingerprint.

2.2.2 Machine Learned Fingerprints for Polymers

On the other hand we have Machine-learned fingerprints for polymers. These are data-driven polymer representations derived automatically through machine learning models rather than being predefined by human-designed descriptors such as with the PolymerGenome fingerprints. In contrast to PolymerGenome fingerprints that rely on explicit structural rules, i.e. the explicit atomic triplets, block level descriptors and QSPR descriptors, machine-learned fingerprints are obtained directly from the SMILES string via the training of the ML model at hand. In this work, two different neural networks are used which learn their own fingerprint representation. These models are PolyGNN and PolyBERT (both described below in subsequent subsections of this methodology section. During training, the given model learns to encode the relevant chemical and structural information into a continuous, high-dimensional latent space optimized for a given predictive task (in the case of this work the predictive task is always property prediction). This learned representation captures both local and global molecular features in a way that is tuned to the underlying data and objective, enabling more flexible, transferable, and often more informative embeddings than fixed-rule fingerprints.

2.2.3 Other Fingerprinting Schemes for Polymers and Molecules

While not presented in this work, there are other fingerprinting schemes present in literature that have been used to represent molecules as well as polymers, so for completeness, these methods will briefly be described. These other fingerprinting schemes include Mor-

gan, RDKit, and topological torsion fingerprints. Morgan fingerprints are circular molecular fingerprints that encode the local atomic environment around each atom up to a defined radius by iteratively hashing atom neighborhoods. [47] They capture molecular structure in a fixed-length binary or count vector and form the basis of Extended-Connectivity Fingerprints (ECFP). The RDKit fingerprint represents molecules by enumerating all possible linear subgraphs within a specified bond-length range and hashing their atomic and bonding features—such as atom type, bond order, and connectivity—into a fixed-length bit vector. Each hashed substructure is mapped to a position in the vector via a modulus operation, producing a compact encoding of the molecule’s topological features that reflects its overall structural diversity. [48] Lastly, in topological torsion fingerprinting sets of four bonded non-hydrogen atoms in a chain are selected to represent torsional angles in a molecule. The algorithm encodes the atomic type, number of non-hydrogen branches, and the number of π electron pairs in the set into bits of a vector. This process repeats for all sets until the entire molecule or polymer repeat unit is covered to generate the feature vector. [49]

2.3 ML Algorithms

The plethora of ML algorithms used in this research will be described from a very high level so the reader can understand the very basics about how these algorithms work. For a more in depth understanding, especially covering the math involved for how these ML algorithms work, the reader is encouraged to check out the reference listed for each ML algorithm. Further it should be noted that the Gaussian process regression (GPR), support vector machine (SVM), random forest (RF) and boosted random forest (BRF) are all used in the development of the ΔH^{ROP} ML model. section 3.3 shows how GPR was the best ML Algorithm for this task and thus was the ML algorithm that is used for this developed model. On the other hand, thermal and mechanical properties were predicted using a few different ML algorithms before one was selected as best performing for candidate screening. These include deep neural network (DNN) [50], GPR, PolyBERT, and PolyGNN.

2.3.1 Gaussian Process Regression

GPR is a powerful, non-parametric Bayesian method for regression tasks in machine learning. In GPR, we assume that the observed data points are drawn from a Gaussian process. Formally, a Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian (or normal) distribution. In practice, this means that for any set of inputs, we assume the outputs (or function values) follow a multivariate Gaussian distribution. This approach allows us to predict outcomes with an associated measure of uncertainty. The core idea is that we model the joint distribution of the target variables (outputs) based on the inputs by specifying a mean function and a covariance (kernel) function, which captures the relationships between data points. When making predictions, we use the training data to update our beliefs about the distribution of functions, effectively giving us a new Gaussian distribution that describes the likely values of the target variable at unseen points. This means a prediction is determined by using Bayes' rule to determine the conditional probability of the new observation given the old observations (training data). The choice of kernel is crucial because it defines the smoothness, periodicity, or other characteristics of the function we expect. GPR is computationally intense as it requires inverting large matrices, but it is highly flexible, providing not only point estimates but also uncertainty measures, making it especially valuable in areas requiring robust confidence intervals. The complexity, flexibility, and ability to inherently predict uncertainty make it an ideal choice for smaller datasets. For a more in depth description, including the essential mathematical formulations for how kernel and mean functions are defined as well as the basis of how Bayes' rule is used to make predictions on new data see [51].

2.3.2 Support Vector Machine

SVM are supervised machine learning algorithms used for classification and regression tasks, though they are most commonly associated with classification. The core idea behind SVM is to find the optimal hyperplane that separates data points of different classes with

the largest possible margin, minimizing the risk of inaccuracy. In a simple two-dimensional setting, this hyperplane is a line that divides the space into two regions, each corresponding to a value for prediction.

To achieve this, SVM focuses on the data points closest to the decision boundary, known as support vectors. These support vectors are critical because they define the margin and are the only points that directly influence the position of the hyperplane. SVM aims to maximize this margin, as a larger margin typically leads to better generalization on new data.

When data is not linearly separable, SVM uses a technique called the kernel trick to map data into a higher-dimensional space where a linear separation might be possible. Common kernels include the Radial Basis Function kernel, polynomial kernel, and linear kernel. By transforming data into this higher-dimensional space, SVM can draw complex, non-linear decision boundaries in the original feature space, making it versatile for many types of problems. [52]

2.3.3 Random Forest and Boosted Random Forest

In the context of regression, RF works by averaging the predictions of multiple decision trees to provide a continuous output value rather than a class label. Each decision tree in the forest is trained to predict a numerical value, and the final prediction is the average of these predictions from all trees, which smooths out individual errors and increases accuracy. Decision trees are typically created via a top-down approach where each descriptor is evaluated to assess how well it can predict the items' value.

RF for regression uses two types of randomness:

1. **Bootstrap Sampling:** Each tree is trained on a different random subset of the original training data, which is created by drawing samples with replacement. This approach, called bagging (bootstrap aggregating), helps each tree learn a slightly different relationship between the features and target variable, reducing the risk of overfitting to

any specific data points.

2. Random Feature Selection: When creating each split within a tree, Random Forest randomly selects a subset of features to consider, rather than using all available features. This randomization decreases the correlation among trees, making the ensemble more diverse and less likely to converge on the same errors.

In regression, RF provides not only accurate predictions but also a measure of feature importance, showing how frequently each feature contributes to decision splits. [53]

BRF, often referred to as Gradient Boosted Trees or Gradient Boosting, is a variation on traditional RF that builds trees sequentially rather than in parallel, where each new tree corrects the errors of the previous ones. Unlike RF, where trees are independent and aggregated at the end, boosting algorithms add trees iteratively, each one focusing more heavily on the data points that previous trees struggled to predict. This approach enables the model to learn from its mistakes and build increasingly accurate predictions.

BRF is often more powerful and accurate than traditional RF but can be more prone to overfitting if not carefully regulated. Despite being computationally intensive, BRF models are highly effective, especially in applications requiring fine-grained adjustments and high predictive accuracy, and they are commonly used in competitive machine learning settings. [54]

2.3.4 Deep Neural Networks

DNN are a subset of machine learning models inspired by the structure and function of the human brain. They consist of layers of interconnected nodes, or neurons, which process data in a way that allows the network to learn complex patterns and representations. A typical DNN has three main types of layers:

1. Input Layer: This layer takes in raw data, which in the context of this work is polymer fingerprints and passes it on to the next layers.

2. Hidden Layers: These are intermediate layers where most of the processing happens. Each neuron in a hidden layer transforms the input it receives using mathematical functions (usually nonlinear) to detect patterns and features. Deep networks have many hidden layers, which is why they are called "deep" learning.
3. Output Layer: This layer produces the final output of the network, which in the context of this work is a polymer property prediction.

DNN learn by adjusting the weights of the connections between neurons through a process called backpropagation, which uses gradient descent to minimize the error between predicted outputs and actual labels. In general deep learning neural networks require large amounts of data making them difficult to use in many materials science context, but when enough data is present, they have shown significant promise in the field of polymer informatics [50].

2.3.5 Multi-task Learning

Multi-task learning leverages multiple correlated tasks to enhance the accuracy and generalizability of predictions. This research incorporates multi-task learning for two different scenarios. The first is for building ML models that use both computational and experimental data. This is what was done when developing the ΔH^{ROP} s model and the increase in accuracy is discussed in detail in section 3.3. This shows how combining lower fidelity computational data with greater chemical diversity and high fidelity experimental data via multi-task learning can result in a model with increased accuracy and generalizability. The other scenario of using multi-task in this work is for correlated experimental properties. This method is what the thermal and mechanical properties predictor leverages. The models come from multi-task Kuenneth et al [50] and demonstrate that multi-task learning improves prediction accuracy and generalizability of ML models by utilizing inherent property correlations, thus outperforming single-task models based only on experimental data.

2.4 Machine Learning Force Fields

MLFFs serve to bridge the accuracy of ab initio MD and the computational efficiency of classical force fields. These MLFF models take atomic configurations from MD simulations as input and predict either the potential energy [55, 56, 57, 58, 59, 60], the atomic forces [61, 62, 63, 64, 65] or both. One class of MLFF models is energy-based, such as GAP [55, 56], SNAP [58, 59], and Deep Potential [60]. Another class of MLFF models is force-based, which directly predict atomic forces using model outputs. [66, 61, 62, 63, 64, 65]. Since atomic forces are inherently defined for individual atoms, whereas energy which is globally defined, using force-based MLFFs is particularly efficient for MD simulations where forces are essential for each step to update atomic positions. ML models in this line have achieved high-accuracy force field predictions, with errors as low as ≈ 0.01 eV/Å for elemental materials such as Cu and Al [66, 61, 62, 63, 64], and ≈ 0.05 eV/Å for small organic molecules in the MD17 dataset [65]. The MLFF used in this research is a force-based MLFF where the gradients of the neural network model are used in the loss function to represent the force of each input atom. The neural network model used for the MLFF is a covariance-normalized Equivariant Transformer, which will be described in the following section.

2.4.1 Covariance-normalized Equivariant Transformer

As mentioned, the backbone for the model used to train the MLFF in this research is a Covariance-normalized Equivariant Transformer. To understand what this is from a high level each aspect (transformer, equivariance, and covariance normalized) will be described. A transformer is a deep learning model architecture designed to handle sequential data, but unlike traditional recurrent networks, it processes entire sequences in parallel. Transformers have become very popular in natural language processing and have seen huge success in models such as BERT and GPT. In the context of this work, the sequence the model is

learning is atomic configurations where each atom is labeled by its position and atom type. The core component of a transformer is its attention mechanism, specifically self-attention, which allows the model to focus on different parts of the input sequence as needed. Rather than processing data step-by-step, or in atomic data bond by bond, the transformer uses self-attention to dynamically weigh the relevance of each element in a sequence relative to every other element. This mechanism enables the model to capture long-range dependencies, meaning it can understand relationships between atoms that are potentially far apart in a molecule. This is a very simplistic view and for more information, including the mathematical formation of attention in transformer models we refer the reader to Vaswani et al. [67]

Now as mentioned transformers process data as sequences without being aware of spatial or rotational relationships, but molecules should be invariant to spatial manipulations such as rotations or translations. Thus, equivariant transformers are a special type of transformer that is designed to respect certain symmetries or invariances in data, meaning it processes inputs in a way that directly incorporates these properties. This is achieved by modifying the attention mechanism, such as through specialized kernels or constraints on the weights, so the model "knows" how different transformations affect the data and adjusts its outputs accordingly. By explicitly enforcing these symmetries, equivariant transformers can learn more efficiently and generalize better in tasks where specific transformations or invariances are relevant, such as molecular data. This approach reduces the need for extensive data augmentation and can lead to more interpretable and robust models for symmetry-rich data. Equivariant transformer based models have shown promise in molecular modeling with the development of TorchMD-NET[68, 69], but this model encounters exponential gradients and forces as the number of model parameters increases. This gradient explosion issue presents a challenge to scaling up the model to capture more complex chemistries such as those found in the polymer dynamics data we are trying to model.

This issue is solved by employing a Covariance Layer Normalization technique. Co-

variance Layer Normalization is a technique used to improve the stability and performance of transformer models by normalizing features in a way that accounts for their covariance structure, which is often important in tasks involving complex, high-dimensional data. In transformers, layer normalization is typically applied to stabilize the hidden representations in each layer, improving training efficiency and helping the model converge. Then by combining all of these methods we get to a covariance-normalized Equivariant Transformer that is used as the backbone for the MLFF in this work. The MLFF used in this work was produced in a collaboration with the Chao Zhang group at Georgia Tech and is referred to as Polyget. [70]

2.5 Virtual Polymer Design

Finally, with polymer property predictors defined we can move to generative algorithms to produce novel polymers to screen for promising properties. In this work two generative algorithms will be leveraged, VFS and a GA).

2.5.1 Virtual Forward Synthesis

In VFS, ring-based small molecules are found and run through digital "reactions" that mimic real-world ROP. These digital reactions are achieved via SMILES arbitrary target specification (SMARTS) to generate hypothetical polymers. These SMARTS provide reaction templates and they have been designed to mimic a plethora of known ROP reactions. With the SMARTS reaction templates in place a database of potential molecules is required. The databased used herein comes from five different sources: ZINC15, ChemBL, compounds sourced from literature, an eMolecules database dump from December 19th, 2020, and data from a VWR database, harvested through a webscraper. This created a list of over 30 million potential monomers, and the SMARTS reaction template was used to search which of these molecules are amenable to the various forms of ROP by doing a substructure search for the necessary ring structures for ROP reactions documented in literature. Then,

the monomers with the appropriate substructures are run through the matching reaction template to generate a hypothetical polymer. This polymer's properties are predicted, and the polymer is then screened to see if it would be a suitable candidate for synthesis.

2.5.2 Genetic Algorithm

A GA is a computational attempt to copy evolution as a design process. First, crossover and mutation operations are performed on a parent data set to create a new child data set, then each child is scored according to some fitness function. Then the best performing children from the current generation are selected based on their fitness scores and become the parents of the next generation. This cycle continues until a set of prescribed target criteria is achieved [71, 72].

In the context of ROP polymers, a set of known heterocycles that undergo ROP are defined as scaffolds onto which different functional groups can be attached. These functional groups and scaffolds act as the "genes" that are optimized over the generations. When a new child is created by crossover, VFS with a known chemical reaction template is employed to create the polymer that corresponds to the ring monomer monomer. The functional groups to be used as genes were determined from a procedure called breaking of retrosynthetically interesting chemical substructures (BRICS)[73] on a database of 14,000 known existing polymers to generate chemical fragments (more information on the BRICS algorithm can be found below in subsection 2.5.3. ML property predictors are used to predict the properties of the current generation, and the polymers with the highest fitness function (typically, a linear combination of scaled property values) are used to create linear child polymers, passing on the successful scaffolds and functional groups. This process repeats for an arbitrary number of generations, and the polymers that achieve some screening criteria are evaluated if they are likely to be synthesizable.

2.5.3 BRICS

BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) is a rule-based algorithm that decomposes molecules into chemically meaningful fragments using predefined SMARTS patterns. The method identifies specific bond types between functional groups that correspond to synthetically relevant disconnections. When a matching pair of substructures is found to be connected via atoms adjacent to the designated “L” dummy atoms, the molecule is cleaved at that bond, producing fragments that retain dummy atoms marking the points of disconnection. These fragments can subsequently be recombined by reconnecting corresponding dummy atoms, enabling the generation of new, synthetically plausible molecular structures. In this work, the BRICS algorithm is used via the RDKit implementation.

2.5.4 Synthesis Accessibility Score (SA Score)

Both VFS and GA incorporate a crucial metric known as SA Score. This metric assumes a critical role in the evaluation of synthetic complexity, wherein a score is derived through a computation involving fragment contributions and a complexity penalty. To determine fragment contributions, a 2-radius Morgan fingerprint was applied to over a million molecules sourced from PubChem. Each bit of the fingerprint is assigned contribution scores based on its prevalence within the dataset. The molecular complexity score is generated through the inclusion of penalties associated with atom count, chiral center count, ring spiro atom count, and ring bridge atom count [74]. Additionally, the RDKit implementation incorporates a symmetry score, correlating highly symmetrical molecules with enhanced synthesizability. These scores span a scale from 1 to 10, with a rating of 10 denoting the most challenging molecules to synthesize and a rating of 1 signifying the least challenging ones.

2.5.5 Fitness Function to Determine Promise of Polymer Candidates

In order to access how well a given generated polymer is for a the desired application, a quantitative method is required to measure how well a polymer conforms for a given screening criteria. For both the VFS and GA algorithms a fitness function which uses the following three criteria was employed.

1. **Enthalpy Transformation:** The target range for ΔH was set between -10 kJ/mol and -20 kJ/mol. To ensure uniform weighting between values less than -20 kJ/mol and greater than -10 kJ/mol, ΔH values were adjusted using the following equation:

$$\Delta H_i^t = \begin{cases} \Delta H_i + 30 & \text{if } \Delta H_i < -20 \\ 10 & \text{if } -20 \leq \Delta H_i \leq -10 \\ \Delta H_i * -1 & \text{if } \Delta H_i > -10 \end{cases}$$

where ΔH_i^t stands for the transformed ΔH value for polymer i and ΔH_i stands for the predicted ΔH value for polymer i .

2. **Clipping of Predicted Properties:** Predicted properties exceeding the target thresholds were clipped to the targets. This process is mathematically represented by the following equation:

$$k_i^t = \min(k_i, k_{target})$$

where k_i^t stands for the transformed property value for polymer i , k_i represents the predicted property k for polymer i and k_{target} represents the target value desired for the given property k . As a reminder, in the context of this work the predicted property k can be E , σ_b , ϵ_b , T_g , T_m , or T_d . This equation thus means that the value for a given property, k remains unchanged if the value is below the target, or becomes the target if it exceeds it. This equation allows the fitness function to prioritize polymers that satisfy all criteria over those that excel in only a few aspects.

3. Normalization and Fitness Calculation: The adjusted property values were normalized within the range of 0 to 1 using a simple MinMaxScaler. A composite fitness value for each polymer was then calculated by taking the average of these normalized properties, as described in the following equation:

$$\theta = \frac{1}{n} \sum_k \frac{k_i^t - k_{min}^t}{k_{target} - k_{min}^t}$$

where θ is the fitness function value for a given polymer, n is the number of properties screened, k again is the polymer properties of interest (E , σ_b , ϵ_b , T_g , T_m , or T_d), k_i^t is the min clipped transformed property, k_{min}^t is the minimum value of the clipped transformed predicted property and k_{target} is the target value for the given property for a given application. This use of normalization of each property as well as using the product of all the normalized properties means that the value of the fitness function is a harmonic mean of how close each target property is to being met. If a polymer has a fitness function value of 1 all target properties are met.

2.6 Natural Language Processing Augmented Data Capture

Training any ML property predictor requires experimental data with manual data extraction being the most straight forward way to compile such a dataset. That being said, manual data capture is extremely time consuming and as such a method which augments traditional manual data extraction from literature with Natural Language Processing techniques has been developed [75]. This process works by starting from a corpus of millions of HTML/XML formatted articles acquired from top scientific journals. The procedure then occurred in four steps, including (1) document parsing, converting original documents to a format that is suitable for Natural Language Processing, (2) coarse-grained filtering, where appropriate keywords were used to downselect to several thousands of articles from the initial set, (3) extracting useful information from the downselected papers, and (4) validating the extracted data by domain experts. In this procedure, step (3) includes three substeps, i.e., (3a) target paragraph identification, (3b) material name identification and linking to

property, and (3c) extraction from the literature. In (3a), heuristic rules were employed to identify candidate paragraphs. They included searching for paragraphs containing property names including abbreviations and relevant units. In (3b) an name entity recognition filter is applied to best identify the material name and link it to the property identified by the heuristic filters from (3a). In (3c) two models are leveraged to finally extract the linked material names and properties which are materialsBERT (an NER based extraction) and chatGPT-3.5 (a large language model based extraction). Finally in step 4, a domain expert is required for validation that the correct property has been extracted with the right linking to materials name and further to manually generate the necessary SMILES string for the given polymer extracted. Data extracted using this method has been made available publicly at <https://polymerscholar.org/>. This method expedites data extraction and specific to this work has led to an expansion of the ΔH_{exp}^{ROP} dataset as will be discussed in section 3.2.

CHAPTER 3

DEVELOPMENT OF ΔH^{ROP} MULTITASK MODEL: AN ESSENTIAL TOOL TO DISCOVERING DEPOLYMERIZABLE POLYMERS

3.1 Introduction

As has been described, ΔH^{ROP} is one of two fundamental thermodynamic properties for screening polymers for depolymerizability. To this end, a state-of-the-art ML predictor was trained leveraging multi-task modeling so both ΔH_{calc}^{ROP} and ΔH_{exp}^{ROP} could be used in training. The two data sources were incorporated to address the limitations inherent to both: while experimental data serve as the ground truth, they are often scarce and expand slowly. In contrast, computational data, though inherently approximate due to practical constraints, can be generated at scale, grown rapidly, and cover novel chemical spaces beyond experimental reach. This multitask approach works by simultaneously training on both ΔH_{calc}^{ROP} and ΔH_{exp}^{ROP} and leveraging correlations between them to enhance model robustness and generalizability to new chemical spaces.

3.2 Creation of ΔH^{ROP} Dataset

To train such a model comprehensive datasets of ΔH_{exp}^{ROP} and ΔH_{calc}^{ROP} were extracted from literature and generated, respectively. For ΔH_{exp}^{ROP} , manual extraction accounted for about 76% of the dataset while the method from section 2.6 was used to rapidly extract data to it's current size. This resulted in a dataset with 124 ΔH_{exp}^{ROP} , and the natural language processing method described is continuously used to aid in keeping the dataset up to date at all times. ΔH_{calc}^{ROP} was generated using the method outlined in section 2.1. This has resulted in a total of 428 ΔH_{calc}^{ROP} data points spread across different calculation sizes within the set $L = \{3456\}$ (again as described in section 2.1). Table 3.1 shows the exact breakdown of

Table 3.1: Summary of the ΔH^{ROP} data generated, accumulated, and used herein.

Category	Number	$\Delta H_{L=3}^{ROP}$	$\Delta H_{L=4}^{ROP}$	$\Delta H_{L=5}^{ROP}$	$\Delta H_{L=6}^{ROP}$
Polymers w/ $\Delta H_{\text{expt}}^{ROP}$ only	56	—	—	—	—
Polymers w/ $\Delta H_{\text{comp}}^{ROP}$ only	84	83	26	28	25
Polymers w/ both $\Delta H_{\text{expt}}^{ROP}$ & $\Delta H_{\text{comp}}^{ROP}$	68	66	42	45	35
Polymers w/ either $\Delta H_{\text{expt}}^{ROP}$ or $\Delta H_{\text{comp}}^{ROP}$	193	149	68	73	60

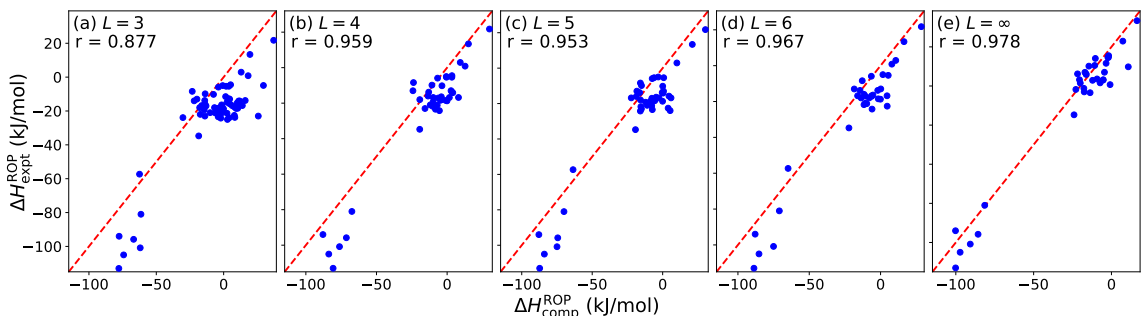


Figure 3.1: Correlations between $\Delta H_{\text{expt}}^{ROP}$ and $\Delta H_{L=N}^{ROP}$, shown for (a) $L = 3$, (b) $L = 4$, (c) $L = 5$, (d) $L = 6$, and (e) $L = \infty$. In the plots r corresponds to the Pearson correlation between $\Delta H_{\text{expt}}^{ROP}$ and $\Delta H_{\text{comp}}^{ROP}$ and indicates how well correlated the variables are for a given L .

the overlap of $\Delta H_{\text{exp}}^{ROP}$ and $\Delta H_{\text{calc}}^{ROP}$ datasets. Further as can be seen in Figure 3.1 there is significant correlation between $\Delta H_{\text{exp}}^{ROP}$ and $\Delta H_{\text{calc}}^{ROP}$, and this correlation increases as the system size, L , increases. This is important as this provides evidence for why combining these data streams has led to improved $\Delta H_{\text{exp}}^{ROP}$ predictions as will be shown in the next section.

3.3 Multitask Model Training, ML Algorithm Selection, and Results

To test the best algorithm suited to predict $\Delta H_{\text{exp}}^{ROP}$ four different ML algorithms, GPR, RF, SVM, and BRF (see section 2.3 for descriptions on all of these algorithms) were used in a custom leave one out cross validation (LOOCV) scheme. In this LOOCV scheme a held-out polymer, for which $\Delta H_{\text{exp}}^{ROP}$ is available, is targeted and predicted by four ML models trained on four different training cases, designed to systematically examine and reveal the role of $\Delta H_{\text{calc}}^{ROP}$, the subsequent benefit of multi-task learning, and the performance of the

developed models. These four cases are summarized in Table Table 3.2. In the first case, only the available experimental data were used for training, so the model is “effectively” a single-task (ST) model, and so, this case is named ST. The next three cases are MT1, MT2, and MT3, which are designed to gradually supply the multi-task learning algorithms with selected subsets of ΔH_{calc}^{ROP} and consequently gradually improve ML models. Among the three multi-task (MT) cases, MT1 does not include computed data of any size (L) for the held-out polymer. This simulates the case when there is no computational data available for the polymer of interest being predicted. The MT2 case assumes that there is minimal computational data available, i.e. just corresponding to $L = 3$, in the training data for the held-out polymer. Finally, the MT3 case represents the situation where plenty of computational data is available for the held-out polymer being predicted.

Table 3.3 shows two error metrics, i.e., the root mean square error (RMSE) and the coefficient of determination (R^2) obtained by using SVM, RF, BRF, and GPR for all 4 cases. There were 68 polymers where data existed for both experimental and computational data for all cases, so the RMSEs and R^2 metrics shown represent the results for 68 different training splits for the 4 different cases. These results demonstrated that by combining computed data and experimental data via multi-task learning, the ML models improved in accuracy. In terms of RMSE and R^2 , the best algorithm to learn our ΔH_{exp}^{ROP} data is GPR. Both of these results are not surprising as multitask learning has been shown to improve model accuracy in polymer informatics [50] and GPR has widely been shown in

Table 3.2: Summary of four cases used in evaluating the ML algorithms, which are different in the training data.

Case	Training data
ST	Experimental data only
MT1	Experimental data + computed data, excluding all ΔH_{comp}^{ROP} computed for the held-out polymer
MT2	Experimental data + computed data in which only $\Delta H_{L=3}^{ROP}$ computed for the held-out polymer is included
MT3	Experimental data + all computed data, including $\Delta H_{L=N}^{ROP}$ for all N computed for the held-out polymer

Table 3.3: RMSE, given in kJ/mol, and R^2 obtained from SVM, RF, BRF, and GPR for different cases described in the text.

Model type	ST		MT1		MT2		MT3	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
RF	8.3	0.89	10.7	0.87	10.0	0.85	8.8	0.88
SVM	17.1	0.55	11.2	0.81	10.5	0.83	9.2	0.88
BRF	9.3	0.87	9.4	0.87	9.7	0.86	9.0	0.88
GPR	12.2	0.77	9.2	0.87	8.8	0.88	8.0	0.90

the literature to be a promising ML algorithm for small datasets, especially when it comes to polymer informatics. [33, 34, 76, 77, 78, 79] Using GPR, RMSE is reduced from 12.2 kJ/mol for the ST case, again trained only on experimental data, to 9.2 kJ/mol for MT1, to 8.8 kJ/mol for MT2 and 8.0 kJ/mol for MT3. Therefore, GPR was selected for the eventual development of the predictive ML "production" model of ΔH_{exp}^{ROP} .

3.4 Production ΔH^{ROP} Multitask Model

In a final test of the model accuracy, all data was used to train a GPR model except 10 randomly chosen ΔH_{exp}^{ROP} were withheld from training. The results of this final model are visualized in Figure 3.2. The model had a training RMSE of 2.96 kJ/mol and a test RMSE of 8.28 kJ/mol. While this model does still show signs of overfitting, it still shows test accuracy that is high enough to screen polymers for promise of depolymerizability. Further, since GPR is being used, the model inherently has an uncertainty for each prediction that can help identify if a prediction is expected to be more or less accurate. If uncertainty is high for a prediction of a polymer of interest, ΔH_{calc}^{ROP} can be generated for said polymer to improve the model performance. Moving forward, this model will be used to screen for desirable ΔH_{exp}^{ROP} .

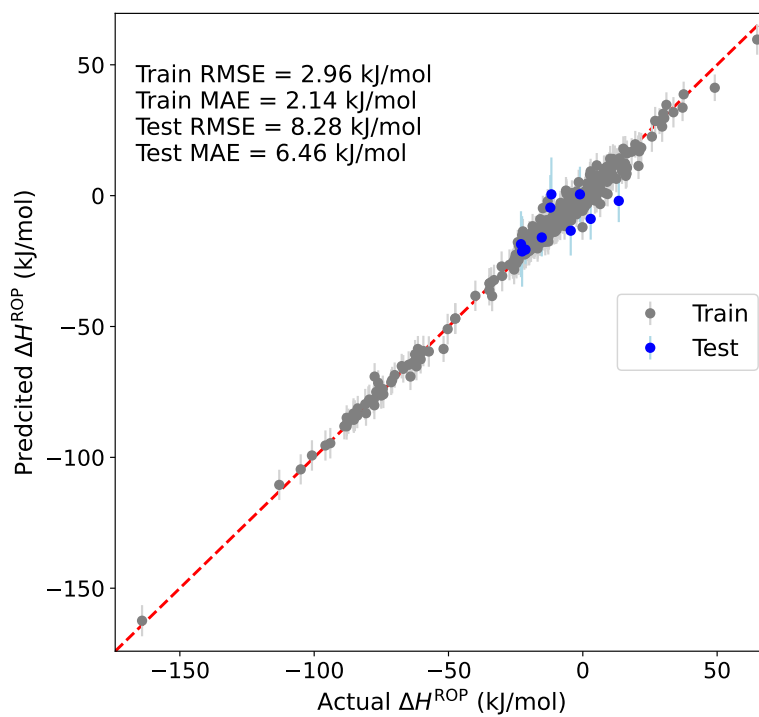


Figure 3.2: Parity plot for the pre-production model where 10 % of the $\Delta H_{\text{expt}}^{\text{ROP}}$ data was withheld. Blue data points represent the test data, while grey data points represent the train data (which contain both $\Delta H_{\text{comp}}^{\text{ROP}}$ and $\Delta H_{\text{expt}}^{\text{ROP}}$)

CHAPTER 4

POLYGET: A MACHINE LEARNED FORCE-FIELD FOR ROP POLYMERS

4.1 Introduction

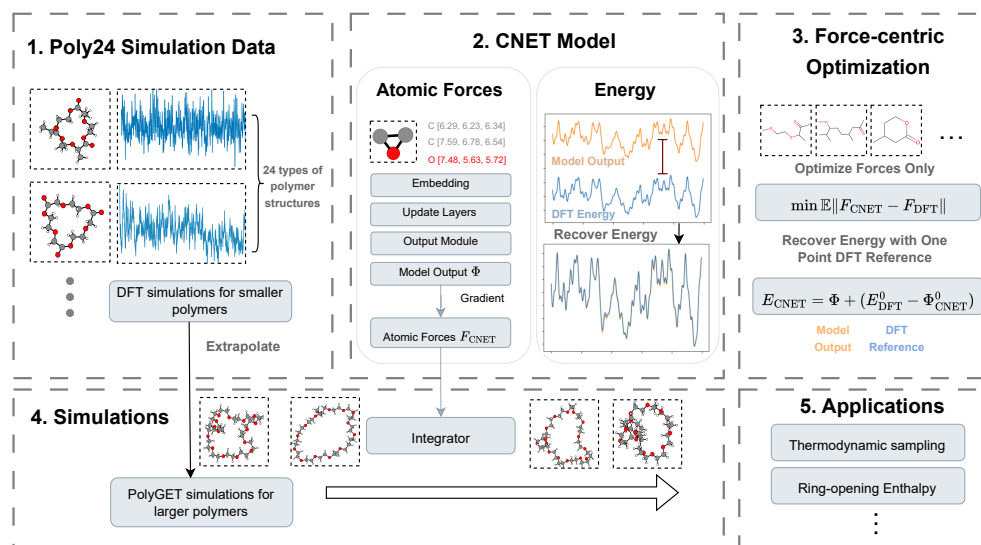


Figure 4.1: Illustration of PolyGET. (1) We prepare Poly24, a database consisting of DFT simulation data for 24 types of polymer types. (2) We propose a Equivariant Transformer architecture called CNET to predict atomic forces and energy for the polymers. (3) The CNET model is optimized on Poly24 with force-centric optimization, and predicts accurate energy with linear adjustment. (4) The resulting model can extrapolate to robust and DFT-level accurate simulations of large loop polymers. (5) We test PolyGET in applications such as ring-opening enthalpy prediction, achieving performance equal to or surpassing DFT, while enabling much longer MD simulations for large polymers.

While ab-initio MD has proven to be very useful in augmenting ΔH^{ROP} data to train a more accurate ML model, the calculations done provide much more data the model in Chapter 3 is not fully utilizing. Thus we look to leverage that data further in this chapter with the development of an MLFF for ROP polymers. Outside of the work shown herein to calculate ΔH^{ROP} , molecular dynamics (MD) [80, 81] is a key computational method in polymer science [82, 83, 84], with many applications ranging from studying polymer

conformations and dynamics to investigating polymer-solvent interactions and designing novel polymer materials. Starting from an initial configuration, MD simulates the evolution of a polymer structure over millions of spatial and temporal steps to extract desired properties. Due to its computational efficiency, MD allows for the simulation of polymers at length and time scales that are beyond the capabilities of ab initio methods, which rely on quantum mechanical calculations, such as DFT [85, 86]. The reliability of MD simulations hinges critically on the accurate evaluation of atomic forces, which drive the structural propagation. Although ab initio methods like DFT provide high precision in force calculations, their computational cost restricts their application to systems containing fewer than thousands of atoms. To overcome the limitations of DFT and enable simulations of larger systems, classical force fields (or potentials) [87, 88, 89, 90, 91] have been developed. These include ReaxFF [92], OPLS [93, 94], and GAFF [95, 96, 97], which offer a more computationally efficient solution by computing potential energy and atomic forces using empirical functional forms and parameters. However, despite being orders of magnitude faster and more scalable than DFT, these classical force fields often lack the accuracy and transferability required for complex systems, limiting their applicability in certain scenarios.

To bridge the gap between ab initio accuracy and the computational efficiency of classical force fields, MLFFs have been developed [98, 99, 100, 101, 66]. These ML force field models take atomic configurations from ab initio based MD simulations as input and predict either the potential energy [55, 56, 57, 58, 59, 60] or the atomic forces [61, 62, 63, 64, 65] necessary for simulation. One class of MLFFs is energy-based, such as GAP [55, 56], SNAP [58, 59], and Deep Potential [60]. In these models, the energy contribution E_i of each atom i is predicted independently based on its local environment, and these contributions are summed to obtain the total potential energy $E = \sum_i E_i$, from which the atomic forces are subsequently derived. Recent Transformer-based force fields [69, 102, 103, 104, 105] have improved model expressivity by predicting E_i utilizing attention mechanisms

to model deep layers of interactions among atoms. When trained on well-known small organic molecule datasets such as QM9 [106, 107], MD17 [108], and ANI1 [109, 110], these MLFFs have demonstrated accuracy comparable to DFT. Another class of MLFFs is force-based, which directly predict atomic forces using model outputs, bypassing the need to model intermediate variables such as E_i [66, 61, 62, 63, 64, 65]. Since atomic forces are inherently defined for individual atoms, unlike the global energy E , these models can be particularly efficient for MD simulations where forces are more essential than energy. ML models in this line have achieved high-accuracy force field predictions, with errors as low as ≈ 0.01 eV/Å for elemental materials such as Cu and Al [66, 61, 62, 63, 64], and ≈ 0.05 eV/Å for small organic molecules in the MD17 dataset [65]. However, a significant challenge for force-based models is ensuring compliance with conservation laws; for example, the sum of independently predicted atomic forces may not be zero, violating momentum conservation—a property that is naturally preserved in classical and energy-based MLFFs.

Despite considerable advancements in MLFFs [100, 101, 66], the development of such force fields for polymers [111, 112, 113] remains less mature compared to those for other solid-state materials. This gap arises from the unique challenges posed by polymers, including their complex morphology, long-range dynamics, and the significant role of non-bonding interactions, which are challenging to capture even with DFT [114, 115]. The complex morphology of polymers, characterized by entanglements, chain folding, and hierarchical structures, requires force fields that can accurately model both local and global conformational changes. Long-range dynamics, such as chain repulsion and viscoelastic behavior, necessitate force fields that can maintain stability and accuracy over extended simulation distances and times. Furthermore, non-bonding interactions, including van der Waals forces and electrostatic interactions, play a crucial role in determining polymer properties, but their accurate description within DFT remains a challenge [114, 115]. Moreover, the scarcity of training data for polymers is another limiting factor for developing MLFFs for polymers [116, 117]. The high computational cost of DFT calculations for large poly-

mer systems, along with the need for extensive conformational sampling to capture the diverse morphologies of polymers, limits the availability of high-quality training data.

In this work, we introduce Polymer Force Fields with Generalizable Equivariant Transformers (PolyGET), a novel framework of machine learning forcefield designed for highly accurate and generalizable ML force fields for polymer simulations (Figure Figure 4.1). PolyGET offers two primary advantages over previous ML forcefield models: (1) Multi-molecule training and generalization. PolyGET utilizes a force-centric training paradigm, enabling the training of a general ML forcefield model on a chemically and spatially diverse polymer dataset and effective generalization to previously unseen polymers. (2) Accurate and robust simulations. Unlike previous MLFFs, which are unstable when employed for MD simulations, PolyGET achieves robust and precise MD simulations with force accuracy comparable to ab initio methods, while running orders of magnitude faster.

PolyGET achieves these benefits through several innovative designs. First, we introduce the Covariance-Normalized Equivariant Transformer (CNET) to enable unified and large-scale multi-chemistry training. Building on the Equivariant Transformer (ET) from the TorchMD-Net package [69], as ET is proposed from TorchMD-Net; we have since developed our own variant of the model. CNET models three-dimensional geometric interatomic relationships using vector-based latent feature representations of atoms. Like ET, CNET maintains equivariance to atomic positions, ensuring compliance with fundamental conservation laws. CNET enhances ET by incorporating covariance normalization, which allows the model to scale to deeper networks without the exponential increase in gradient and force norms. This capability enables the modeling of more complex interatomic interactions. As a result, larger CNET models can be trained on various polymer types with greater stability and improved performance compared to the original ET [69].

Second, PolyGET introduces a novel force-centric training paradigm that enables multi-chemistry training with a unified ML model. Unlike existing deep-learning approaches [69, 118] that jointly optimize potential energy and forces for individual chemistries, our

approach focuses exclusively on force optimization across multiple polymer chemistries. This strategy has demonstrated generalizability across various polymer chemistries, avoiding the optimization instabilities typically encountered in multi-chemistry settings [119]. By employing force-centric multi-chemistry training, PolyGET acquires robust and transferable knowledge of quantum mechanical interactions, enabling accurate and reliable MD simulations and the generalization to larger, unseen polymers. This approach ensures that the predicted forces are consistent with the true forces acting on the atoms, leading to more physically meaningful and stable simulations.

While the optimization of PolyGET involves only forces and not the potential energy, the model learns the potential energy surface (PES) by matching the negative gradient of the scalar model output with DFT forces in the training data. Thus, forces are directly predicted, along with potential energy with an arbitrary constant shift. Furthermore, the model maintains momentum conservation by predicting forces as the negative gradient of a scalar model output.

The novel designs of PolyGET contribute to its superior performance along multiple dimensions. We evaluate PolyGET on Poly24 [30, 25], a large-scale polymer simulation dataset encompassing 24 distinct polymer types and 6,552,624 conformations. Our model demonstrates state-of-the-art performance in force prediction, as well as robust and accurate MD simulations. Moreover, PolyGET enables faster and more robust MD simulations over long time horizons for polymers of large loop sizes, with accuracy comparable to DFT while requiring only small atomic systems from DFT as training data. Benefiting from this ability to generalize to large atomic polymer systems, our model facilitates accurately calculating ΔH^{ROP} , an important thermodynamic property for ring-opening polymers, with similar or better accuracy to established DFT methods [30].





























4.2 Poly24 Dataset

A significant portion of this work I specifically contributed to is the creation of the benchmark ab-initio dataset to train PolyGET. This is known as Poly24 and is made publically available at . Poly24 includes ab initio data for 24 unique ROP polymers. Although this data contains only 24 unique polymers, it contains a multitude of different trajectories as trajectories are produced for both monomer and polymer, and polymer is represented at multiple different sizes. Polymer models were obtained by multiplying the monomer repeat unit by a small integer L (e.g., $L = 3, 4, 5,$ and 6), and then forming a looped oligomer of size L by bonding the ends of the polymer chain to each other resolving any dangling bonds. These initial atomic models were readily generated using Polymer Structure Predictor (PSP) package [117]. For each monomer or polymer model 10 or more maximally diversified configurations were selected as the initial configurations for the DFT-based MD simulations, this was done to find lowest energies structures, and now is leveraged in this work to account for the diverse configurational space of each polymer.

All atomic models are non-periodic (i.e. gas in a vacuum), and thus we used the Γ -point version of the Vienna *Ab initio* Simulation Package (VASP) [120, 121]. We employed a plane wave basis set with kinetic energy up to 400 eV to represent the Kohn-Sham orbitals. The ion-electron interactions were computed using the projector augmented wave (PAW) method [122], while the exchange-correlation (XC) energies were computed using the generalized gradient approximation Perdew-Burke-Ernzerhof (PBE) functional [123].

Table Table 4.1 shows the 24 polymer chemistries that comprise this dataset, broadly classified into cycloalkanes, ethers, lactones, and others, following the classification in [124]. On average, the dataset includes 10 DFT trajectories for both monomers and polymers. The polymers are formed by polymerizing the monomers, resulting in a comprehensive collection of data for each polymer type. In total, we have 1,311 DFT trajectories and 6,552,624 molecular conformations. This large number of DFT-based molecular dynamics

trajectories and resulting molecular configurations allows for the deep learning architecture used as the backbone in the PolyGET model.

Index	Type	# Atoms	Monomer	Polymer
CK1	cyclopropane	9		
CK2	cyclobutane	12		
CK3	cyclopentane	15		
CK4	cyclohexane	18		
CK5	cycloheptane	21		
CK6	cyclooctane	24		
ETH	ethylene oxide	7		
LAC1	γ -butyrolactone	9		
LAC2	γ -butyrolactone	12		
LAC4	δ -valerolactone	15		
LAC6	β -methyl- δ -valerolactone	18		
LAC7	δ -caprolactone	18		
LAC8	δ -decalactone	30		
LAC9	(-)-Menthide	30		





















LAC10	n-alkane sub δ -valerolactone	18		
LAC11	α -Methylene- γ -butyrolactone	13		
LAC12	n-alkane sub δ -valerolactone	21		
LAC13	n-alkane sub δ -valerolactone	24		
LAC14	n-butyl δ -valerolactone	27		
LAC15	n-alkane sub δ -valerolactone	30		
LAC16	n-alkane sub δ -valerolactone	33		
LAC17	n-alkane sub δ -valerolactone	42		
OTH1	1,4-dioxan-2-one	13		
OTH2	3-methyl-1,4-dioxan-2-one	16		

Table 4.1: Polymers in our used Poly24 dataset for training PolyGET. We classify the polymers into 4 broad categories: cycloalkanes, lactones, ethers, and others.

4.3 Force-centric Training of PolyGET

This portion of the PolyGET work, was done by my collaborator Rui Feng in Chao Zhang’s group, but I include it here in this work as it is important to understanding how PolyGET works and how I contribute to validation, energy recovery, and ultimately determining the procedure to calculate ΔH^{ROP} using PolyGET. Here we address the challenges of generalizing ML force fields across multiple molecules by adopting a force-centric approach.

Traditional methods optimize both potential energy and atomic forces, often over-fitting to single molecules and failing to generalize. Instead, we drop the energy optimization and focus on minimizing the differences between predicted and ground-truth forces, defined as:

$$\mathbb{E}_{x, \mathbf{r}_0} \left[\mathbb{E}_{\mathbf{r} \sim \mathbb{T}^{\text{DFT}}(x, \mathbf{r}_0, \mathcal{T})} \left[\left\| -\nabla \Phi_{\theta}(\mathbf{r}, \mathbf{z}) - F(\mathbf{r}, \mathbf{z}) \right\|_2^2 \right] \right],$$

where $-\nabla \Phi_{\theta}(\mathbf{r}, \mathbf{z})$ are the predicted forces, and $F(\mathbf{r}, \mathbf{z})$ are the ground-truth forces. We find that dropping the energy in optimization significantly improves forces prediction accuracy when training on diverse types of polymer chemistry. By focusing exclusively on forces, which have consistent distributions across molecules, our model avoids optimization bottlenecks caused by mismatches in potential energy distributions and achieves robust performance across diverse molecules.

Although our model does not directly optimize potential energy, energy can still be recovered since force is simply the gradient of energy with respect to position. Thus, since we model each force in the loss function as the gradient of the model output with respect to the position (ultimately since positions are the model input), then by definition the model output Φ_{θ} , it is equivalent to ground-truth energy off by a constant. Therefore, we can perform a linear calibration from Φ_{θ} to ground-truth energy:

$$E^*(\mathbf{r}, \mathbf{z}) = \mathbf{W}_E \Phi_{\theta}(\mathbf{r}, \mathbf{z}) + \mathbf{b}_E,$$

where $(\mathbf{W}_E, \mathbf{b}_E)$ are parameters that can be determined by comparing Φ_{θ} with reference ground-truth energy. This ensures that energy predictions remain consistent with ground-truth data while benefiting from the force-centric training. In practice, we can determine \mathbf{b}_E by as little as a single point DFT calculation as reference at the start of the MD simulations, and leave \mathbf{W}_E as 1.

Now that it is understood that force centric training was used, and what that entails we can get into specifics of how PolyGET was trained. We trained our model on the aforemen-

tioned Poly24 benchmark dataset, which includes 24 polymer types with 6.5 million DFT-generated conformations. Training focused on smaller polymers (up to 210 atoms) with loops of up to 5 monomers, while testing evaluated larger polymers (up to 25 monomers). Our model demonstrated high fidelity to DFT data in MD simulations and accurately predicted thermodynamic properties, such as ring-opening enthalpy (ROE), showcasing its capacity for generalization across unseen molecular systems.

4.4 PolyGET Validation

Both energy and force prediction accuracy of PolyGET are barely affected by the number of monomers in the polymers tested, as shown in Figure Figure 4.2. From loop 5 to loop 25 poly(ethylene), the energy MAE changed from 0.024 to 0.031 eV/atom, and force MAE changed from 0.006 to 0.005 eV/Å. For poly(ethylene oxide), energy MAE changed from 0.008 to 0.040 eV/atom, and force MAE changed from 0.009 to 0.011 eV/Å. The non-increasing trend of force MAE from 5 loop to 25 loop suggests stable model performance when loop size increases. Such results show that PolyGET can extrapolate well to larger polymers, requiring only smaller polymers as training data that are easier to obtain from DFT calculations. This enables large-scale MD simulations on the level of DFT accuracy on such large polymers previously not possible with either DFT or classical forcefields.

In Figure 4.3, we present the potential energy and RMSD (Root Mean Square Deviation) profiles. Both the potential energy and RMSD exhibit fluctuations around a relatively constant value after a certain time point, indicating that the system’s conformation stabilizes. This stabilization suggests that the simulation has reached equilibrium, providing evidence of structural and energetic convergence. Here, Figure 4.3 (a)-(d) shows the potential energy and RMSD for poly(ethylene), and (e)-(h) shows for poly(ethylene oxide). The potential energy shows equilibrium, i.e. it converges to and fluctuates around a mean value, for all 8 simulations. RMSD is a metric to evaluate the stability of the simulations; at each point, RMSD is evaluated as the root-mean-square-distance of the pairwise atomic distance

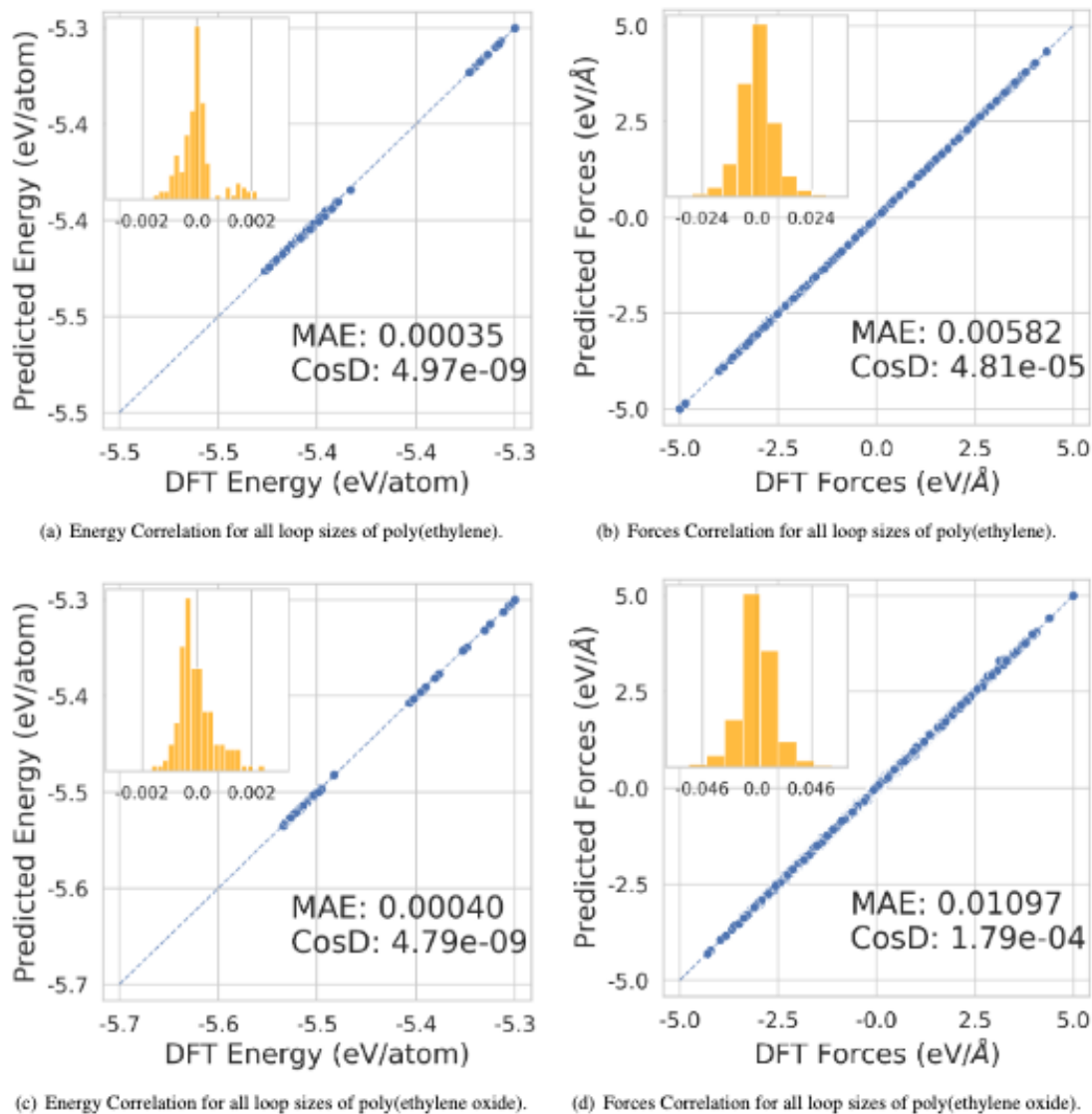


Figure 4.2: Accuracy of force predictions of PolyGET examined in all loop sizes up to loop 25. Each figure shows the scatter plot of predicted energies forces versus DFT calculations from the aggregation of all loop sizes for each polymer family. For each loop, 100 samples are uniformly sampled from the PolyGET trajectory. In the upper left, we show the histogram of the error, where the x-axis represents values of absolute error and y-axis represents frequency. In the bottom right we annotate the Mean Absolute Error (MAE) in eV/Å and Cosine Distance (CosD). The consistent high accuracy and correlation shows that PolyGET can extrapolate to larger polymer systems without performance deterioration.

matrices between the current conformation and the starting conformation. Namely, it evaluates at each point how much the current conformation deviates from the initial one. For all 8 simulations, we can see RMSD increasing at first, before converging to and fluctuating

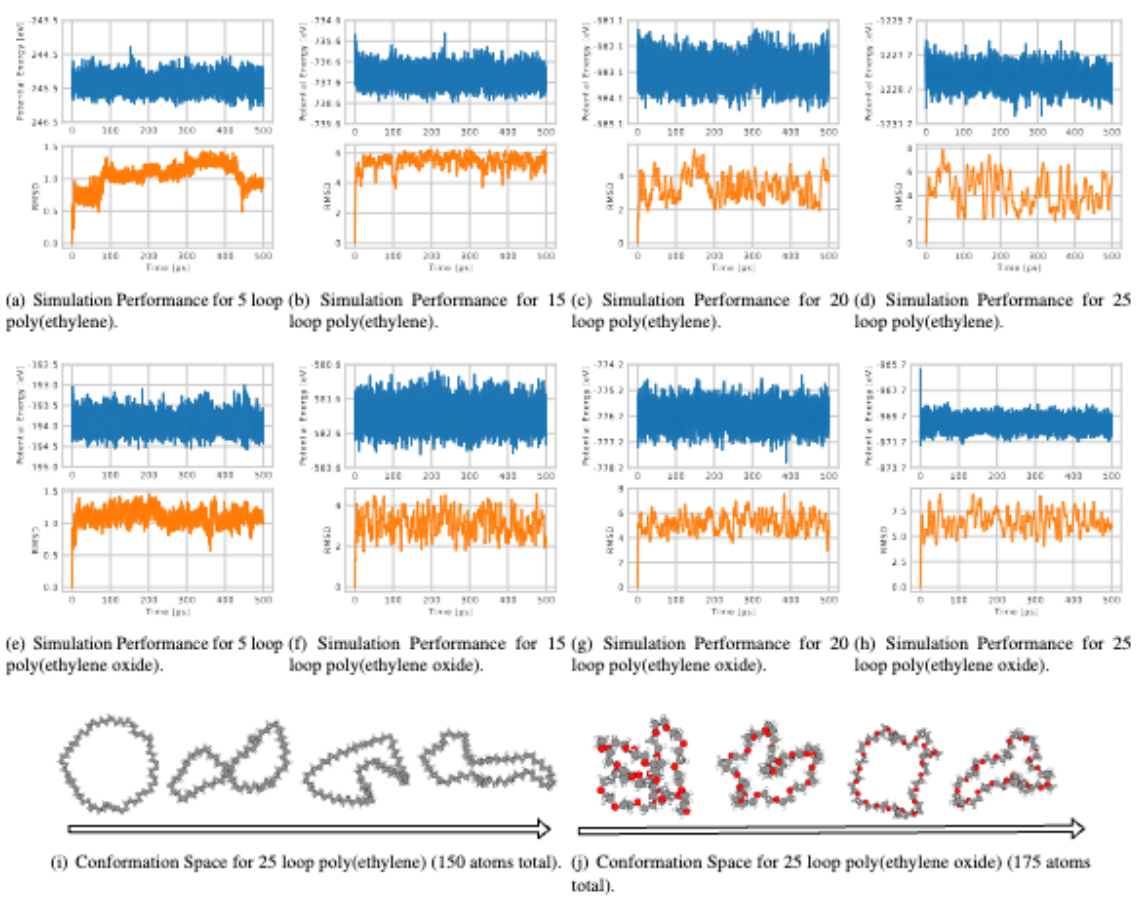


Figure 4.3: Visualization of simulation statistics. In each subplot, we visualize two subplots, where the upper one shows the potential energy plot, and the bottom one shows the RMSD value of conformations during the simulation, which measures at each timestamp how far the conformation deviates from the initial conformation. The potential energy plots show a convergence to equilibrium for all simulations. The RMSD further shows the stability of simulations, as the conformations remain in certain proximity to the initial conformation while exploring different conformational spaces, as illustrated by (i) and (j).

around a certain upper bound. This suggests that PolyGET simulations explore the conformation space for more diverse structures, while not exceeding a reasonable bound for stability, as illustrated in Figure 4.3 (i) and (j), where 4 sample conformations are visualized respectively for the two type of polymers, each consisting 25 monomers.

Energy conservation is a fundamental principle in molecular dynamics (MD) simulations, ensuring physically meaningful and realistic results. Figure Figure 4.3 shows potential energy plots from simulations, offering insights into the thermodynamic behaviors

of the simulated systems. The fluctuations around a mean potential energy in equilibrium states align with the canonical ensemble (NVT), where the system exchanges energy with a heat bath to maintain a constant temperature. Despite these fluctuations, the near-constant potential energy underscores 's simulation ability to accurately mimic energy exchange processes in physical systems at equilibrium, demonstrating physical realism and thermodynamic consistency.

The root-mean-square deviation (RMSD) plots in Figure 4.3 provide insights into the conformational stability and exploration capabilities of our simulations. The RMSD quantifies the structural deviation of polymer conformations from the initial configuration, showing an initial exploration of diverse conformational states followed by convergence and fluctuations around an upper bound. This balance between conformational exploration and structural stability is crucial for accurately capturing the complex energy landscapes and dynamics of polymer systems. Representative conformations visualized in Figure 4.3 (i) and (j) further illustrate the diversity of structures explored by Polyget simulations while maintaining structural integrity, highlighting its ability to navigate vast conformational spaces of large polymers and effectively sample relevant configurations for accurate thermodynamic property predictions.

We tested the PolyGET in a series of challenging MD simulations designed to calculate the ring-opening enthalpy ΔH^{ROP} , an important thermodynamic property of ROP polymers. [124]. These MD simulations include running NPT trajectories at 300 K for 1 ns for varying system sizes. These different system sizes include a system to represent the original ring monomer, and then cyclic oligomers to represent polymers of various sizes. The polymer systems contain 3, 4, 5, 6, 8, 10, 12, 15, 20 or 25 different monomer repeat units. We use the variable L to represent how many monomer repeat units are used to represent the polymer system. Then for each system size we can determine a size-dependent ROP enthalpy as $\Delta H_L^{ROP} = \frac{1}{L} \langle E_L \rangle - \langle E_{L=1} \rangle$, where $\langle E_L \rangle$ and $\langle E_{L=1} \rangle$ are the averaged energies of 10 minimized structures from the last 100,000 steps of the MD trajectory of the polymer

model of size L and the monomer model, respectively.

In Figure 4.4, we compare the results of running the procedure for calculating ΔH^{ROP} using PolyGET to drive MD simulations with those using DFT for ab-initio MD. The results demonstrate that by using PolyGET, we were able to calculate lower ΔH^{ROP} values for each polymer model size where both calculations from MD trajectories using PolyGET and those using DFT were available (polymer model sizes 3, 4, 5, and 6). This improvement is hypothesized to be due to finding sets of canonical states that are better equilibrated and then running minimization on these more equilibrated structures. Furthermore, by simulating larger polymer system sizes, we observed a more natural relationship between polymer system size and calculated ΔH^{ROP} that follows an exponential decay function of the form $\Delta H_L^{\text{ROP}} = a \exp(-bL) + c$, where L is the polymer system size. The exponential decay function exhibited high R^2 values of 0.99 for poly(ethylene) and 0.962 for poly(poly(ethylene oxide)), indicating a good fit to the data.

Using the exponential decay function, we predicted the ΔH^{ROP} values for poly(ethylene) and poly(poly(ethylene oxide)) to be -106.9 kJ/mol and -100.8 kJ/mol, respectively. For poly(ethylene), this demonstrates an increase in accuracy for predicting ΔH^{ROP} when using MD trajectories from PolyGET compared to MD trajectories run by DFT. In the case of poly(ethylene oxide), we observed essentially the same results between trajectories powered by PolyGET and DFT. These findings highlight that by using PolyGET to run the MD trajectories, we obtained a more intuitive relationship between ΔH^{ROP} and polymer system size, and achieved either similar or even better results in terms of the actual computed ΔH^{ROP} .

To further show the effectiveness of the PolyGET model to run MD for ROP polymers, we ran long time-horizon MD for 16 of the 24 polymers in the Poly24 dataset. These 16 different MD simulations showed extremely high force and energy accuracy for larger loop sizes that are distinctly outside the training data, as shown in Figure 4.5. The average force RMSE for loop sizes 15 and greater is 11.562 eV/Å and the average energy RMSE is

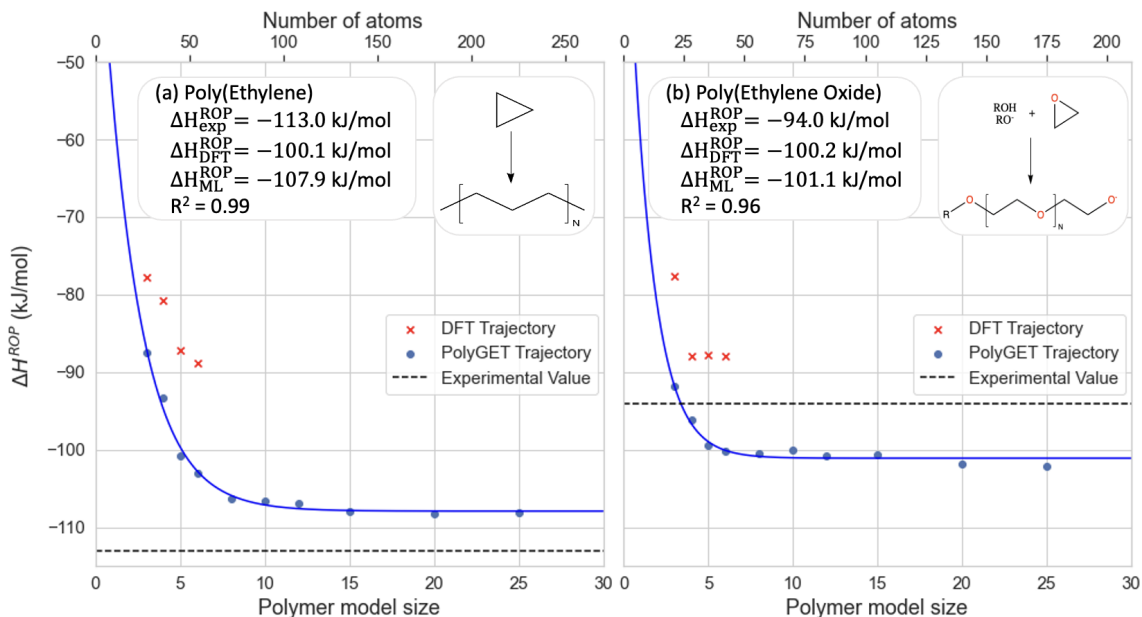


Figure 4.4: Comparison of ΔH^{ROP} calculations done using both PolyGET model and ab-initio methods, i.e. DFT. In (a) we show the calculation done for poly(ethylene) and in (b) we show the calculation done for poly(ethylene oxide). The blue dots represent the ΔH^{ROP} calculated for a given system size from PolyGET model, the blue line shows the exponential fit said data, and the red Xs ΔH^{ROP} calculated from DFT based trajectories. Inset in the top left of the figure the ΔH^{ROP} values for experimental measurement, DFT based calculations and PolyGET calculations are shown respectively. Inset in the plots in the top right we see the polymerization reaction for the polymer being simulated.

just under 0.2 meV/atom. These results showcase best-in-class force and energy accuracy. Further, since the training of PolyGET model only included atomic systems of up to loop size 5, these results further show that the PolyGET model can extrapolate to larger atomic systems with extremely high accuracy.

Going beyond force and energy validations, the MD trajectories obtained using PolyGET model for the 16 ROP polymer chemistries comprised within the Poly24 dataset were used to calculate ΔH^{ROP} . When calculating ΔH^{ROP} for a more diverse set of polymers than just poly(ethylene) and poly(ethylene oxide), we found that the same relationship between polymer system size and energy did not always exist. Namely that the exponential decay relationship does not hold specifically for polymers that have a ΔH^{ROP} closer to 0. Instead, it was found that the energy difference between monomer and larger loop sizes still

tended to saturate near the experimental value, and thus the average of the ring opening energy of loop sizes 15, 20 and 25 were used to calculate ΔH^{ROP} with high accuracy. It should be noted that this procedure provides approximately the same value as the method used previously for poly(ethylene) and poly(ethylene oxide). Figure Figure 4.6 shows the parity plot of calculating ΔH^{ROP} for all 16 chemistries that had MD trajectories ran using PolyGET model. Using PolyGET model we achieve an overall RSME of 7.1 kJ/mol and an R^2 of 0.964, showing how well PolyGET model can be used to calculate ΔH^{ROP} for polymers of interest that have DFT data for small loop models within the train set. This provides a path forward to expanding PolyGET model to be more generalizable. Smaller atomic systems that are computationally less expensive to run DFT can be computed, added to the training, and then the PolyGET model can be retrained to include these data. Further, we leave it as future work to test the ability of the model to be finetuned with additional data, and how much DFT data has to be generated for the model to be able to extrapolate to larger loop models such as loop 25. This provides a path forward to iteratively achieve a much more generalizable MLFF model for ROP polymers.

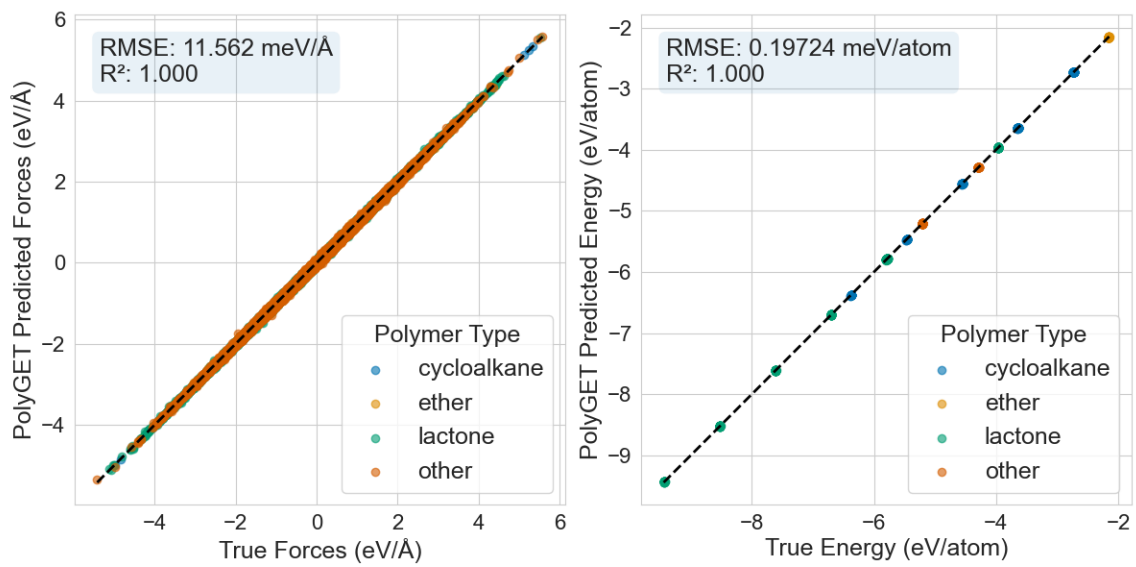


Figure 4.5: Parity plots for both forces (a) and energies (b) for 16 chemistries from the Poly24 dataset.

CHAPTER 5

VIRTUAL POLYMER DESIGN OF DEPOLYMERIZABLE POLYMERS

5.1 Introduction

Multi-objective optimization presents difficult challenges as optimizing one variable can be at the consequence of decreasing a separate correlated variable you also wish to optimize. Polymer design falls directly inline with this pitfall. Here the various objectives to consider are different polymer properties that must be optimized in order to create a polymer that is not only depolymerizable but also useful, i.e. has mechanical and thermal robustness. These encompass thermal properties, such as the T_g , T_m , T_d , which not only dictate stability at operational temperatures but also affect processing conditions, mechanical properties like E , σ_b , and ϵ_b , which play crucial roles in determining the polymer's ability to be durable enough for a given application, and thermal properties like ΔH^{ROP} and T_c which determine whether or not a plastic will be chemically recyclable. Optimizing all of these properties to get to a useful chemically recyclable polymer is a non-trivial endeavor.

This work, along with the work in section 5.4 focuses on discovering polymers to replace many of the most widely used plastics throughout industries for applications such as beverage containers, food containers, packaging materials, and wrapping plastics. The polymers focused on in this work are PET, PS, PP, and HDPE. These plastics make up BLANK percent of all plastic waste produced, making them the top candidates for replacing to reduce global plastic waste. Such substantial volumes are required for recyclable alternatives to make sense, as economies of scale often dictate economic incentives when it comes to recycling [2]. While PS and PET are technically recyclable, the overwhelming volume of both are not recycled due to prohibitive costs [125, 38]. As for PP, and HDPE, while both can be mechanically recycled neither are chemically recyclable making

them ideal candidates to search for chemically recyclable alternatives. Further, there is a substantial cost of not recycling these materials and landfilling them or even worse having them leak out to the environment. One major concern with leakage to the environment is microplastics. Microplastics resulting from plastic waste have been identified as potential immune system stimulants and toxic to freshwater organisms wreaking havoc on both humans and ecosystems across the globe [126, 127]. In Figure 5.1, a radar chart is shown to illustrate the properties of the polymers we desire to replace alongside the specific property targets set for our design (represented by orange lines). These targets, explicitly outlined in the first box of the informatics workflow of Figure 5.1, were carefully selected through a comprehensive analysis of each polymer's properties, coupled with considerations of the typical operating conditions of each application. For PET the application is beverage containers, for PS the application is food containers, while for PP and HDPE the application is packing materials and plastic wraps. All target properties for the given polymers that desired to be replaced can be seen in Figure 5.1 in the radar plot as well as clearly laid out in the Table 5.1.

Finally using the VFS scheme, detailed in subsection 2.5.1, 7 million potential ROP polymers were screened. Polymers were first screened for a ΔH^{ROP} amenable to depolymerization using both ML model prediction and the uncertainty of the model. This narrowed the search space to just over 2.5 million potential polymers. Then for each polymer and application identified to be replaced, the polymers were further screened for mechanical and thermal properties resulting in hundreds to thousands of potential replacements being discovered. Further, from this work, a polymer was identified to be synthesized which is currently underway for experimental validation. Then looking forward, past just commercially available starting monomers, the GA algorithm detailed in subsection 2.5.2 was employed. This resulted in a further hundreds to thousands of novel polymers with promise to replace the listed commodity plastics being discovered.

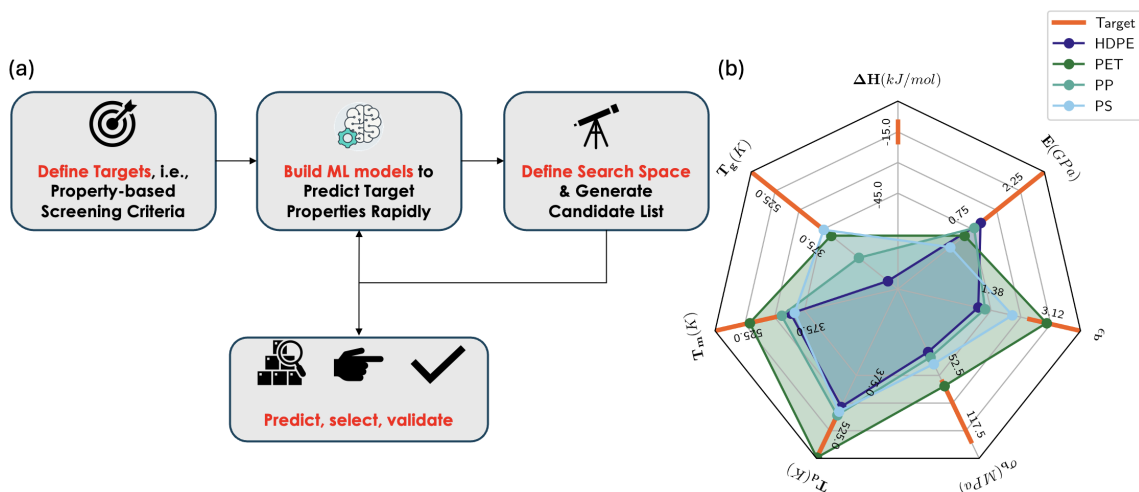


Figure 5.1: In part (a) we see the general guidelines to follow for designing new polymers. First a screening criteria is defined, next ML models are developed for properties of interest to perform said screening, the design search space is decided and a candidate list is generated, and lastly, the generated polymers have their properties predicted and screened for the application at hand. In (b) we see the radar plot for the polymers of interest that we see to replace PET, PS, PP, and HDPE.

Table 5.1: Table of target properties for the different polymers of interest we wish to replace using generative algorithms.

Polymer	Application	E (GPa)	ϵ_b	σ_b	T_g (K)	T_m (K)	T_d (K)
PET	Beverage Container	2.5	0.5	65	350	400	500
PS	Take-home Food Container	2	0.05	39	373	400	473
PP	Packaging Material	1.2	0.8	31	263	433	483
HDPE	Plastic Wrap	1	1	25	172	410	460

5.2 Predictive Models Leveraged

Previous work using this VFS workflow (CITE JOE'S PAPER) had used two subsets of previously developed and published models[25, 50]: a multi-task GPR model to predict ΔH^{ROP} based on the polymer and monomer[25], and a multi-task neural network trained on homo and copolymer data to predict all other properties[50]. This work found that mechanical and thermal properties were lacking in accuracy when it came to the ROP chemical space as seen in parity plots in Figure 5.2, used with permission from Kern et al. In these plots, the red points are data of known polymers while the ROP dataset for

known ROP polymer data is represented by black stars. The count and root-mean square error (RMSE) shown represents firstly the test dataset size and model performance on the test data, then on the known ROP polymers. For ΔH^{ROP} , the black stars and labeled RMSE and count indicate the performance on test ROP data and the size of the test dataset, while the red points represent the training data (an important distinction as all data for the ΔH^{ROP} model is data for ROP polymers). The ΔH^{ROP} model reported here is the same model described in detail in chapter 3. It is clear from these plots, that all models, except for the ΔH^{ROP} ML model, perform significantly worse when predicting for polymers in the ROP chemical space. Thus to remedy this, two multi-task DPNN models were trained using the Polymerize [128] platform; one for the thermal properties and one for the mechanical properties. The training data for these models were ensured to include all ROP data ensuring that performance for this class of materials is better.

These thermal and mechanical models were used to screen the about 2.3 million polymers from the original 7 million which were found to be in the -10 to -20 kJ/mol desired ΔH^{ROP} range. For those 2.3 million cases, histogram plots for both direct ML predictions and the associated predicted uncertainties are shown in Figure 5.3. In each plot the mean and standard deviation of the histograms are displayed. These values become relevant when defining what hypothetical polymers have low uncertainty when selecting best performing polymers to recommend for the chemists to make. In this work, low uncertainty is considered to be if the uncertainty is half a standard deviation below the average uncertainty for given property of interest. For example, For Young's modulus where the mean uncertainty is 0.39 GPa and the standard deviation of the uncertainty is 0.14, a prediction is for E is considered to have low uncertainty if the uncertainty is less than 0.32 GPa. This definition is important when down-selecting compromising candidates further to try and find cases where the predictive models are more confident the predictions being made are more likely to be correct. Further, we see from these plots that the predictions follow close to normal distributions and that while the uncertainties tend to be left skewed this can be expected as

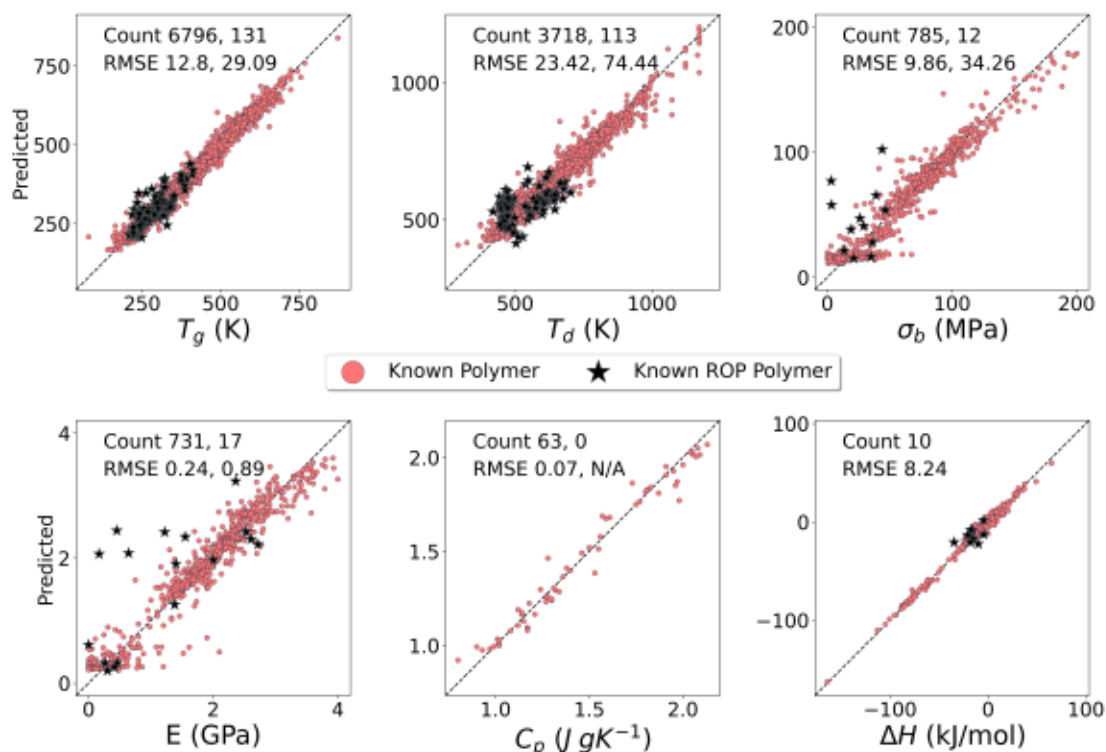


Figure 5.2: Parity plot illustrating model performance across different properties. The first set of values in the top left showcase counts and RMSE on the model test data of known polymers, which are represented by red circles. The second set delineates the dataset size and model performance specifically on known ROP polymers, denoted by black stars. For ΔH^{ROP} , dots depict training data, while black stars indicate test data, as the model was exclusively trained on ROP polymers.

there will be a small percentage of polymers that are far different from the training data. Moving forward these predictive models will be used, first in the VFS frame work, and next in conjunction with the GA. Results for both are found in the subsequent sections, section 5.3 and section 5.4.

5.3 Virtual Forward Synthesis

To initiate the search for viable polymer candidates, a VFS approach was employed to systematically construct a chemically diverse yet synthetically feasible library of hypothetical polymers. This methodology involves algorithmically generating polymers from a curated database of commercially available monomers using predefined reaction rules that cap-

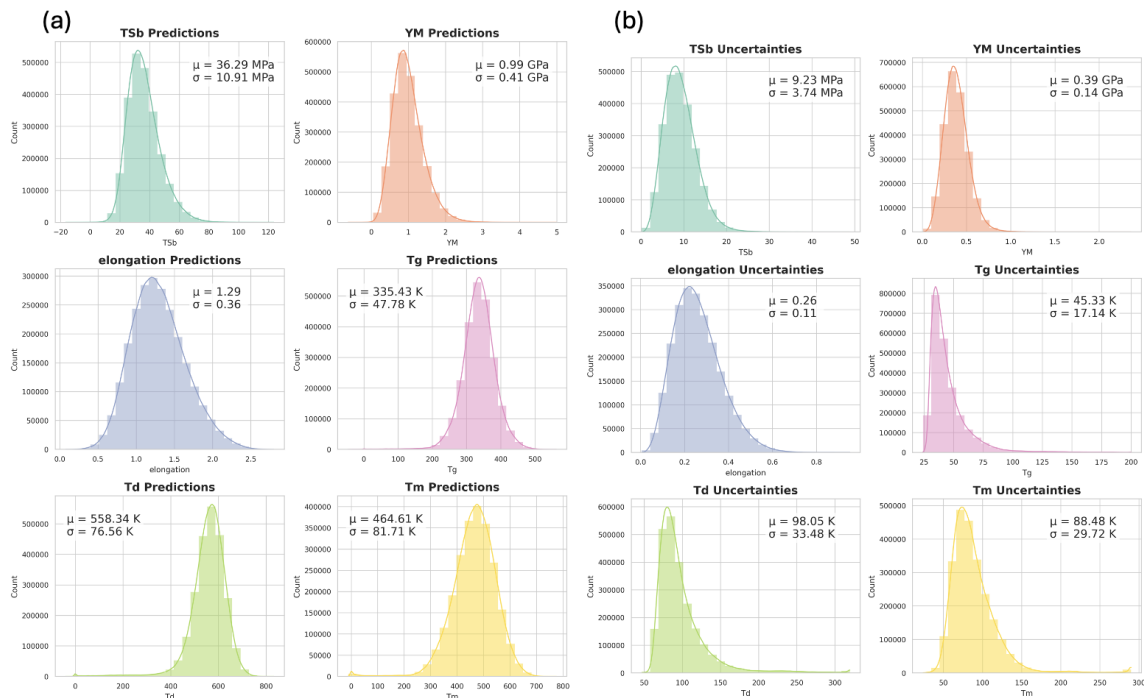


Figure 5.3: Histograms of (a) mechanical and thermal property predictions and (b) their associated uncertainties

ture established polymerization pathways. Building on prior frameworks such as the Open Macromolecular Genome and SMiPoly, this implementation extends the paradigm through deliberate use of ROP reactions and the incorporation of multi-step reaction sequences, enabling a more comprehensive exploration of polymer chemical space specifically for polymers amenable for depolymerization.

The VFS workflow was executed using RxnChainer, a component of the PolymRize™ platform[128], which automates the enumeration of polymerization reactions by applying expert-defined reaction templates. Encoding these transformations requires substantial expertise in both synthetic polymer chemistry and chem-informatics to ensure chemical validity and realistic synthetic accessibility. Through the systematic application of ROP rules to approximately 30 million commercially available molecules, RxnChainer produced a high-fidelity virtual library of roughly 7.4 million hypothetical polymer candidates. This dataset constitutes the chemically meaningful search space used in the subsequent computational screening via property prediction using previously trained ML models.

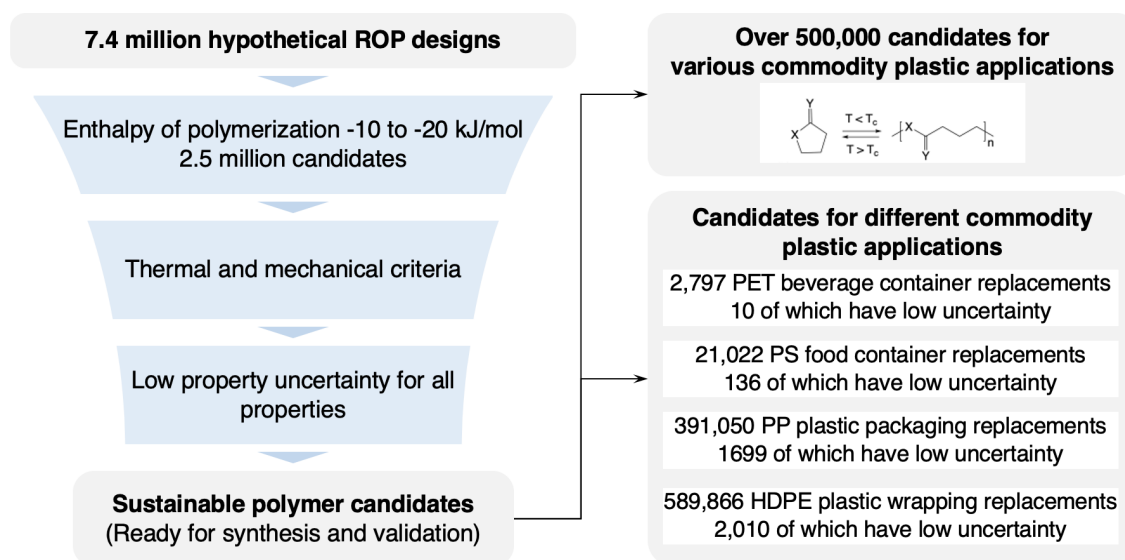


Figure 5.4: Schematics for the multi-stage screening process to achieve sustainable polymer candidates for various commodity plastic applications.

First using the ΔH^{ROP} predictive ML model, we screened the database specifically for polymers with a moderate to low uncertainty for predicted ΔH^{ROP} (less than 20 kJ/mol) and where the predicted ΔH^{ROP} was either in the range of -10 to -20 kJ/mol or the uncertainty put the polymer within this desired range. This range of ΔH^{ROP} is desirable as it shows high promise for the polymer to be both able to depolymerize due to the likelihood of a low T_c , but still able to form polymer with high molecular weight. It should be noted that T_c depends on the ratio of both ΔH and ΔS , but as mentioned in section 1.4, ΔS is extremely hard to establish computational tools for, including both ab initio MD based calculations and direct ML predictions due in large part to differences in literature for how ΔS is calculated from temperature variable NMR [27]. Thus while we shoot for -10 to -20 kJ/mol as a ΔH^{ROP} , we leave this as a less rigid requirement, allowing for moderate uncertainty to allow polymers outside of this range into the screening as it is still possible that they possess the necessary thermodynamics to be depolymerizable and thus chemically recycled. From this first step of screening for ΔH^{ROP} alone, the candidate pool of ROP polymers dropped from about 7.4 million to just over 2.3 million candidates. After this, six mechanical and thermal properties (3 each, E , σ_b , ϵ_b , T_g , T_m , and

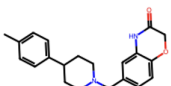
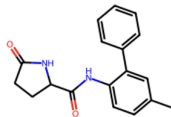
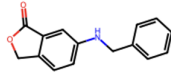
T_d) were predicted for all 2.3 million candidates remaining candidates. Then using the target properties laid out in Table 5.1 as well as visually shown in the radar plot of part (b) of Figure 5.1, we screen down for each commodity plastic application of interest. We further use the definition of low uncertainty outlined in section 5.2 to down select to polymers that have sufficiently low uncertainty for both thermal and mechanical properties to be the most promising cases. This process is visualized in Figure 5.4. This figure further shows that 2,797 potential candidates for PET based plastic beverage containers were discovered 10 of which fit the definition set forth for low uncertainty. For PS based take home food containers 21,022 potential polymer candidates were discovered with 136 of said candidates meeting the criteria for low uncertainty. For PP based packaging material close to 400,000 different promising candidates were discovered with 1,699 of those having low prediction uncertainty, and lastly for HDPE based plastic wrapping for shipping logistics well over 500,000 candidate replacements were identified with 2,010 of those having all thermal and mechanical property predictions with low uncertainty. This shows the power and versatility of the VFS method as a multitude of polymer candidates can be identified and screened for various different applications that desperately need sustainable alternatives to help limit the amount of plastic waste society continues to produce. With this being said, identifying polymers with higher thermal and mechanical robustness is a significantly difficult task as can be seen in the small percent of polymers found to be able to replace PET and PS. Less than 0.04 % of polymers in the VFS database were found to have the potential to replace PET and this drops down to 0.0001 % when we try to find polymers whose predictions have low uncertainty. For PS there are only 0.28 % of polymers that meet the desired criteria and this drops down to less than 0.002 % when we search for low uncertainty candidates. Then for applications that require less thermal and mechanical robustness, i.e. PP based plastic packaging and HDPE based plastic wrapping, the number of successful candidates for replacements explodes. For PP over 5 % of polymers in the ROP chemical space were identified as good replacements and when search for low uncertainty about 0.02 % poly-

mers satisfied all conditions. Lastly, for the application that required the least stringent thermal and mechanical properties, HDPE plastic wrapping, almost 8 % of all polymers in the ROP chemical space were found and when the more stringent uncertainty requirements were applied less than 0.03 %.

This highlights a few important aspects of the design issue and the use of VFS as a solution. Firstly, that even in the least stringent of thermal and mechanical criteria, the success rate of a brute force method such as VFS was only 5 % elucidating how impossible brute force would be in a laboratory setting, as here predictions take fractions of a second, but optimizing polymerization from a desired chemical pathway takes weeks to months per polymer and then characterization takes further days in a best case scenario. It should be noted that computational methods will never replace the experimental work, but the idea is that by searching for the 5 % of polymers in the least stringent case and the 0.0001 % of polymers in the most stringent case, methods such as this can make experimentalists exponentially more productive in the lab as it helps narrow in on the most promising polymer candidates quickly. Secondly, while the percent of promising polymers is quite low, the sheer number of promising candidates can be staggeringly high. For instance, over 500 thousand promising polymers found for replacing HDPE plastic wrapping is simply too many potential structures to send to the chemists to even look through. This is where the use of model uncertainty to further narrow the list can be extremely beneficial, as it can allow us to only suggest polymers that the trained ML models are more confident about their predictions. In the case of HDPE, 2,010 polymer structures is a much more reasonable number of polymers to expect a chemist to be able to look through and select promising polymers that they see a clear synthetic pathway towards making.

In Table 5.2 we can see a small sample of successful candidates for all target applications. The table shows an image of the ROP monomer discovered as well as the predicted ΔH^{ROP} as well as thermal and mechanical properties. In addition to these properties, you will see a column for SA Score which estimates how complex a given molecule is to syn-

thesize. The score goes from 1 to 10, with 1 being the most simple cases and 10 being the most complex (further details for SA Score can be found in subsection 2.5.4). As can be seen all candidates shown in Table 5.2 have an SA Score of 3 or less meaning they are more likely to be synthesizable. All candidates shown are in Table 5.2 fell under the category of low uncertainty except for the third candidate shown as a potential replacement for PET. This candidate did not hit the stringent low uncertainty metric for all categories, but PET had only 10 that did fall under this category and many of those ten were much more complex, making synthesis less feasible. Further, it should be noted that a wide range of different ROP chemical classes show promise as candidates. Promising candidates shown include lactam, lactone, ROMP and cyclic ether monomers, showing a wide range of potential chemistries that show promise. It is also seen that the presence of fused rings, bulky pendant groups such as benzene rings, and the presence of nitrogen tend to be common motifs among successful candidates. These candidates, along with the thousands of others that hold high promise have been shared with chemist collaborators in the hopes that a few can be validated in the lab.

Monomer	Application	ΔH^{ROP} (kJ/mol)	T_g (K)	T_m (K)	T_d (K)	E (GPa)	σ_b (MPa)	ϵ_b	SA Score
	PET Alternative	-16.68	428.8	570.1	628	2.55	66.53	1.075	2.15
	PET Alternative	-11.10	425.8	642.3	624.5	2.83	65.24	0.89	2.41
	PET Alternative	-8.41	419.5	529.2	632.8	3.072	67.93	1.001	1.91

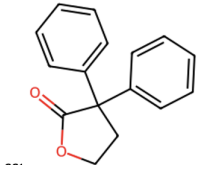
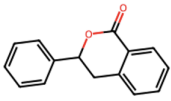
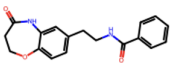
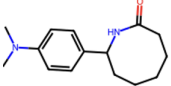
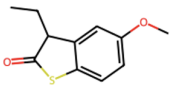
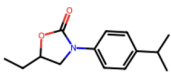
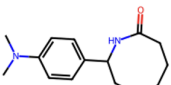
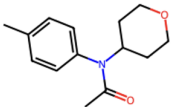
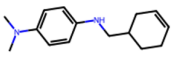
	PS Alternative	-13.68	385	413.8	590.7	2.77	51.19	0.71	2.27
	PS Alternative	-8.348	382.7	457.3	613	2.87	49.26	0.66	2.28
	PS Alternative	-9.80	421.2	585.6	605.9	2.165	53.48	0.89	2.00
	PP Alternative	-14.09	358	552.4	682.6	2.57	44.45	1.28	2.60
	PP Alternative	-9.90	421.6	477.2	599.6	1.24	37.83	1.75	3.00
	PP Alternative	-8.97	343.2	501.1	584.2	1.45	39.66	1.36	2.47
	HDPE Alternative	-14.09	358	552.4	682.6	2.57	44.45	1.28	2.60
	HDPE Alternative	-12.63	360.1	490.9	615.8	1.07	37.02	1.35	2.36
	HDPE Alternative	-19.41	318.6	497.8	581.1	1.10	28.82	6.56	2.67

Table 5.2: Sample of successful candidates for each target application and their properties from VFS workflow. For each property column the prediction is shown normally and then the uncertainty for the given prediction is shown in parentheses

5.4 Genetic Algorithm

Moving forward from the efforts of the VFS algorithm, a GA algorithm designed specifically for ROP polymers was employed to discover novel ROP monomers outside of what is directly from currently available online databases. When designing new to the world ROP monomers one simple approach would be to come up with a list of different ring structures known to be able to undergo ROP and then a list of different functional groups to attach to said ring structures and then use a brute force method with VFS to produce a huge array of hypothetical polymers to make ML predictions on. The issue with this is that the number of hypothetical polymers that can be produced is so high it is not feasible to even make ML predictions for every such polymer. To make this point clear, take the example of just looking at ring structures, which we will refer to as scaffolds, for ring sizes 4, 5 and 6 where functional groups can just be added to 2 set locations for each scaffold. Considering the relevant chemistries in this work, which includes alkenes, ethers, thioethers, esters, thioesters, thionoesters, thiocanes, amides, and carbonates, this equates to 27 different scaffolds each with two bonding sites (assuming the simplest case where bonding sites are always the same on a given scaffold). Now when considering functional groups, the BRICS algorithm described in subsection 2.5.3 was employed on the approximately 7.4 million monomers leveraged in the VFS work previously, to identify chemical motifs of interest that can serve as functional groups with high likelihood to be synthesizable since they are being generated from commercially available polymers. This resulted in an initial 29,030 identified functional groups. Further screening was done to remove any functional groups that are known to compete with various ROP pathways, such as alcohols, amines

charged species and more, as well as removing duplicate functional groups only differentiable by stereochemistry as the ML models do not account for this nuance. This resulted in 11,682 different functional groups to consider in our design problem. Now this would mean that for all 27 potential scaffolds that $11,682^2$ hypothetical polymers exist, meaning that a total of over 3.68 billion polymers would have to be explored. This is not a feasible amount of polymers to make ML predictions from either a time or compute resource standpoint.

To address this issue, the GA was created by Atasi and Kern [129] specifically in the context of designing ROP polymers. The workflow for the GA is shown in Figure 5.5. In short, the GA designs novel monomers that have the necessary chemistry to potentially undergo ROP and then VFS creates the polymers from these monomers where then ML predictors are used to assess the hypothetical polymers' ability to depolymerize as well as mechanical and thermal robustness. The design of these monomers and subsequent polymers by the GA involves three steps:

1. **Scaffold Generation:** A molecular scaffold is constructed, comprising a base structure capable of ROP and featuring variable functionalization sites, as seen in Figure 5.5 (a). The new monomers will be created by bonding functional groups to the atom adjacent the labeled asterisks.
2. **Functional group Creation:** A global list of functional groups is created, serving as chromosomes for the GA optimization process. These R-groups are connected to the scaffold by bonding the atom adjacent the asterisks seen in Figure 5.5 (a) "R-groups" to the adjacent atoms seen in "Scaffold". Some examples are shown in the "Monomers".
3. **Reaction SMARTS Definition:** reaction SMARTS are defined to virtually polymerize the molecule using VFS, generating polymers suitable for ML property prediction, as can be seen in Figure 5.5 (b) "Polymerization" section. Reprinted with

permission [129].

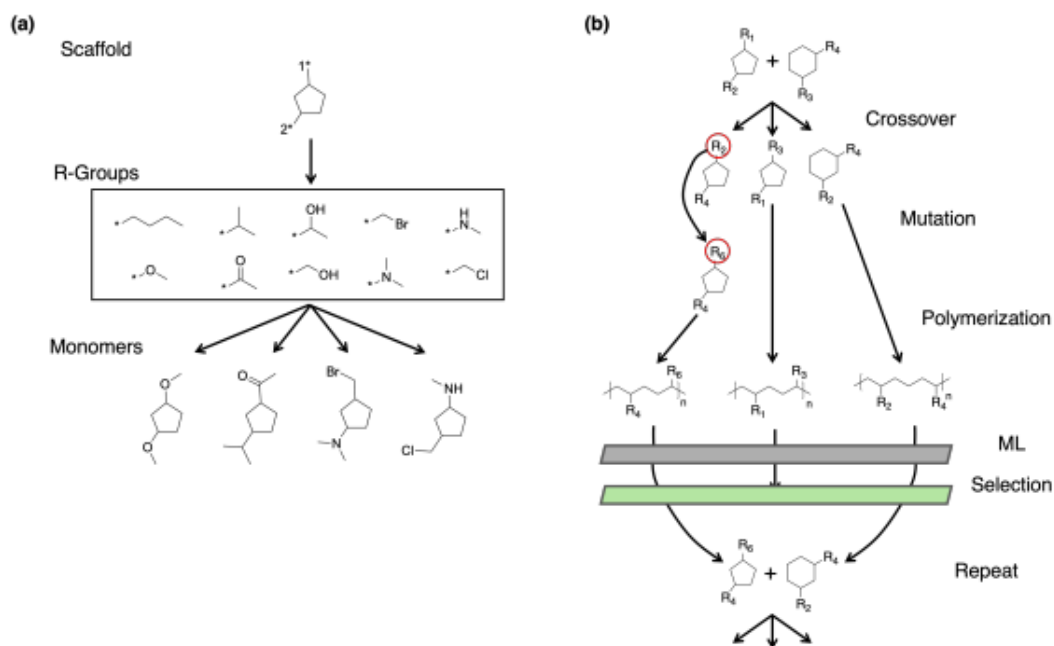


Figure 5.5: (a) Example of GA Initialization. The scaffold is a five-member cycloalkane with functionalization positions denoted by the labeled asterisks (*). New molecules are created by bonding the atom adjacent the * to the atoms adjacent the * in the R-groups displayed in the center. Some example monomers are displayed at the bottom that could be created from this scaffold an these R-groups. (b) Schematic illustration of the GA workflow, depicting the key processes of crossover, mutation, polymerization, property prediction, and fitness evaluation/selection. Selection is based on the predicted properties of the polymer as well as molecular complexity of the monomer, while crossover and mutation is performed on the monomer structure.

The GA creates an initial population starting from the given scaffolds and randomly sampling the functional groups present and then using the above steps to create polymers to screen. Just like in the previously mentioned VFS work, the polymers are screened using the models described in section 5.2, and the fitness of each polymer is determined using the screening criteria for the application of interest as well as the prediction uncertainty of the model. The uncertainty is factored in by using the lower bound of a one-sided 75 % confidence interval calculated from the model prediction and predicted uncertainty. This done using the following equation: $\theta_{75} = \theta - 0.67 * \theta_{uncertainty}$ where θ_{75} is the 75 %

confidence interval of a one-tail z-test, θ is the direct ML prediction, and $\theta_{uncertainty}$ is the ML prediction uncertainty. It should be noted that the value of 0.67 is the z-score that comes directly from table of z-scores for normal distributions. This θ_{75} is then fed into the fitness function for each thermal and mechanical property making it so that the GA directly takes into account each properties predicted uncertainty. The overall fitness function is the same used and described in section 5.3. The top 100 polymers with the highest values from the fitness function of a given generation are then selected to be parents for the next generation.

The parent polymers are randomly paired and generate a predetermined number of children. Each child polymer randomly inherits one of its parents' scaffolds, and then crossover occurs, combining the parents' R-groups into a pool. The child's functionalization sites are then randomly assigned R-groups from this pool. Subsequently, a small subset of child polymers undergo mutation, where some functionalization sites are replaced with random functional groups from the global list. If a child polymer has been previously encountered in a prior generation, mutation continues until a unique molecule is generated. This generation of polymers and monomers is then hierarchically fingerprinted (as described in subsection 2.2.1) and undergo property prediction using ML models previously described. The top performing polymers of that generation then become the "parents" for the next generation. This iterative process, illustrated in Figure 5.5 (b), repeats for a set number of generations, driving the evolution of new polymers that better fit the desired criteria.

For each commodity plastic of interest to replace a separate GA run was conducted. This was to account for the different screening criteria that application required, which again is outlined in Table 5.1. Each run utilized the same hyper-parameters which are as follows: Initially, a population of 300 molecules was randomly generated. In each subsequent generation, the first 100 top-performing polymers, identified through their fitness function, were chosen. These top polymers were then randomly paired to form 200 families, capable of producing up to 3 offspring each, although fewer were generated if no

unique combinations were feasible. Approximately 7.5 % of these child polymers would have one R-group mutated. Moreover, to enrich the diversity of generated polymers, each polymer's SMILES representation was stored in a Python set. If a previously encountered SMILES was identified, the corresponding monomer underwent mutation until a unique polymer emerged, ensuring that mutation did not result in the same polymer being screened twice, while still making sure mutation led to novel exploration of the chemical space.

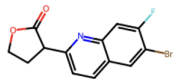
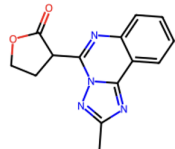
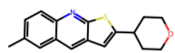
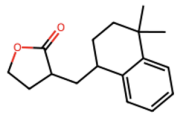
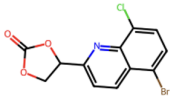
First, this GA approach was applied to finding polymer replacements for PET based beverage containers. Figure 5.6 shows the progression of the 8 properties over the generation number. These plots show averages so while it may seem that the line plot does not reach the shaded target region, there were 46 polymers found that hit all desired targets while still adhering to the stringent case set forth for low uncertainty. These 46 promising PET replacements not only meet the screening criteria, but also the more stringent requirement that the lower bound of the 75 % confidence interval of each thermal and mechanical property is above the desired threshold.

The GA was also applied for discovering replacements for PS based take home food containers, PP plastic packaging, and HDPE plastic wrapping. Evolution of property plots for each case can be seen in the following three figures (Figure 5.7, Figure 5.8, Figure 5.9). For PS 475 successful replacement polymers were discovered, for PP 2,122 successful replacement polymers were discovered, and for HDPE 2,504 successful polymer replacements were found. This not only expands the number of successful polymers found where predictions also have low uncertainty, but also does so with many less predictions made making the GA extremely efficient. The total number of predictions made for each GA run for each application of interest ranged from 30,388 to 37,668, which is significantly less than the millions of predictions that was done in the VFS workflow. This is more aptly captured in the success rate of the GA compared to VFS. For GA the success rate for PET was 0.15 % whereas for the strict uncertainty requirement in VFS the success rate was 0.0001 %. In the case of PS, GA had a success rate of 1.32 % compared to 0.002 %. For PP and

HDPE the trend continued, where for the GA PP improved to a success rate of 6.17 % from 0.02 % and HDPE improved to 6.65 % from 0.02 %. Thus, for the GA we see an improvement in success rate of three orders of magnitude for PET and PS and an improvement of over 2 orders of magnitude for PP and HDPE. This highlights one of the large benefits that GA has over VFS, which is its efficiency. Lastly, stacked histograms for all applications are shown in Figure 5.10. In this figure we can see how for each successive application, as the screening criteria becomes less and less difficult to achieve, the plot becomes more and more right skewed, meaning more and more polymers of high fitness for the given criteria are discovered.

Through using both the VFS and GA approaches many potential candidate polymers to replace multiple commodity plastics of interest were discovered. With this said though, the two approaches have their own pros and cons. VFS leverages curated databases of synthetically realized molecules as its generative substrate, thereby biasing the search toward chemically simple, synthetically accessible architectures; this empirical grounding often enhances downstream feasibility and interpretability of the resulting polymers. However the reliance on enumerative sampling from known chemical fragments confers substantial inefficiency: successful candidates can emerge with success rates as low as 0.0001 %, as was seen in the search for successful PET replacements with low uncertainty. This reflects the inherently sparse mapping between known molecular space and the stringent performance criteria of high-value polymer applications. In contrast, the GA approach exploits evolutionary operators such as mutation and crossover to traverse chemical space more adaptively, enabling two-order-of-magnitude higher efficiency in identifying promising candidates. This results in far fewer property predictions and thus less computational cost. Yet this efficiency comes at the expense of synthetic parsimony, as GA-generated polymers often exhibit increased structural complexity and reduced resemblance to established chemical motifs, raising concerns about practical synthesizability despite their algorithmic optimality. This increase in complexity can be seen in two places. First in the

evolution of SA Score plots found within Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9, and in the difference monomers found in both Table 5.2 and Table 5.3. The monomers shown in Table 5.3 were selected to be well representative of what is seen in the successful polymers and observations of complexity can be seen from immediate observation. The successful polymers from the GA approach have a higher propensity for halogen groups, bulky fused rings side groups, and specifically nitrogen containing ring groups. This significant increase in complexity is a large drawback of the GA approach, and should be addressed as is mentioned in the future work of the next chapter. Regardless, this work culminates in the discovery of thousands of potential replacements for various commodity plastics of essential importance to the global economy.

Monomer	Application	ΔH^{ROP} (kJ/mol)	T_g (K)	T_m (K)	T_d (K)	E (GPa)	σ_b (MPa)	ϵ_b	SA Score
	PET Alternative	-7.67	379.2	594.1	607.5	3.174	75.32	0.984	3.16
	PET Alternative	-6.652	411.5	513.5	574.4	2.585	69.06	1.037	3.12
	PET Alternative	-6.44	401.7	473.7	660.8	2.565	84.99	0.812	2.613
	PS Alternative	-9.686	386.3	542.3	569.2	2.347	57.14	1.09	3.48
	PS Alternative	-12.29	417.1	563.2	612.8	2.549	53.76	1.028	3.301

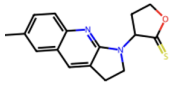
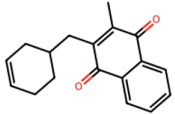
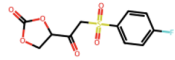
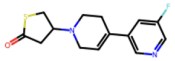
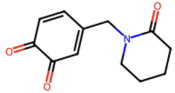
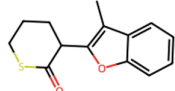
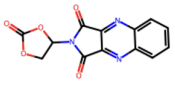
	PS Alternative	-12.39	392.2	588.3	668	2.543	72.83	1.143	3.31
	PP Alternative	-11.96	398.2	434.1	592.3	1.593	40.37	1.021	3.06
	PP Alternative	-17.53	377.6	442.1	576.7	2.119	44.14	0.93	3.07
	PP Alternative	-10.56	326.1	479	660.1	2.25	51.15	1.475	3.56
	HDPE Alternative	-16.58	342.7	419.3	695	1.033	35.86	1.055	2.65
	HDPE Alternative	-19.35	332.3	432	540	2.54	56.89	1.32	3.26
	HDPE Alternative	-10.6	290	434.6	524.7	3.32	103	1.001	3.35

Table 5.3: Sample of successful candidates for each target application and their properties from the GA algorithm.

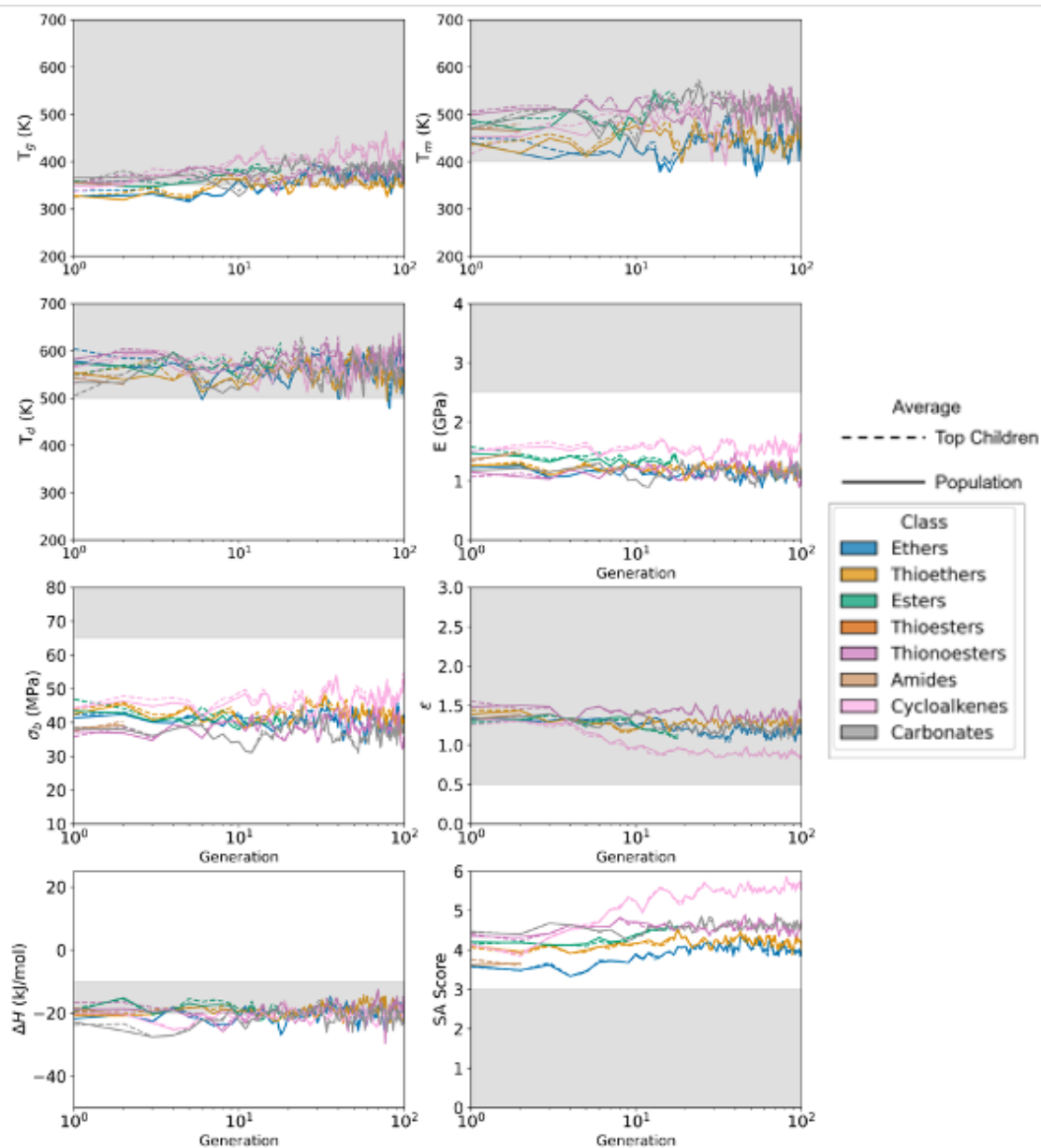


Figure 5.6: Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PET. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.

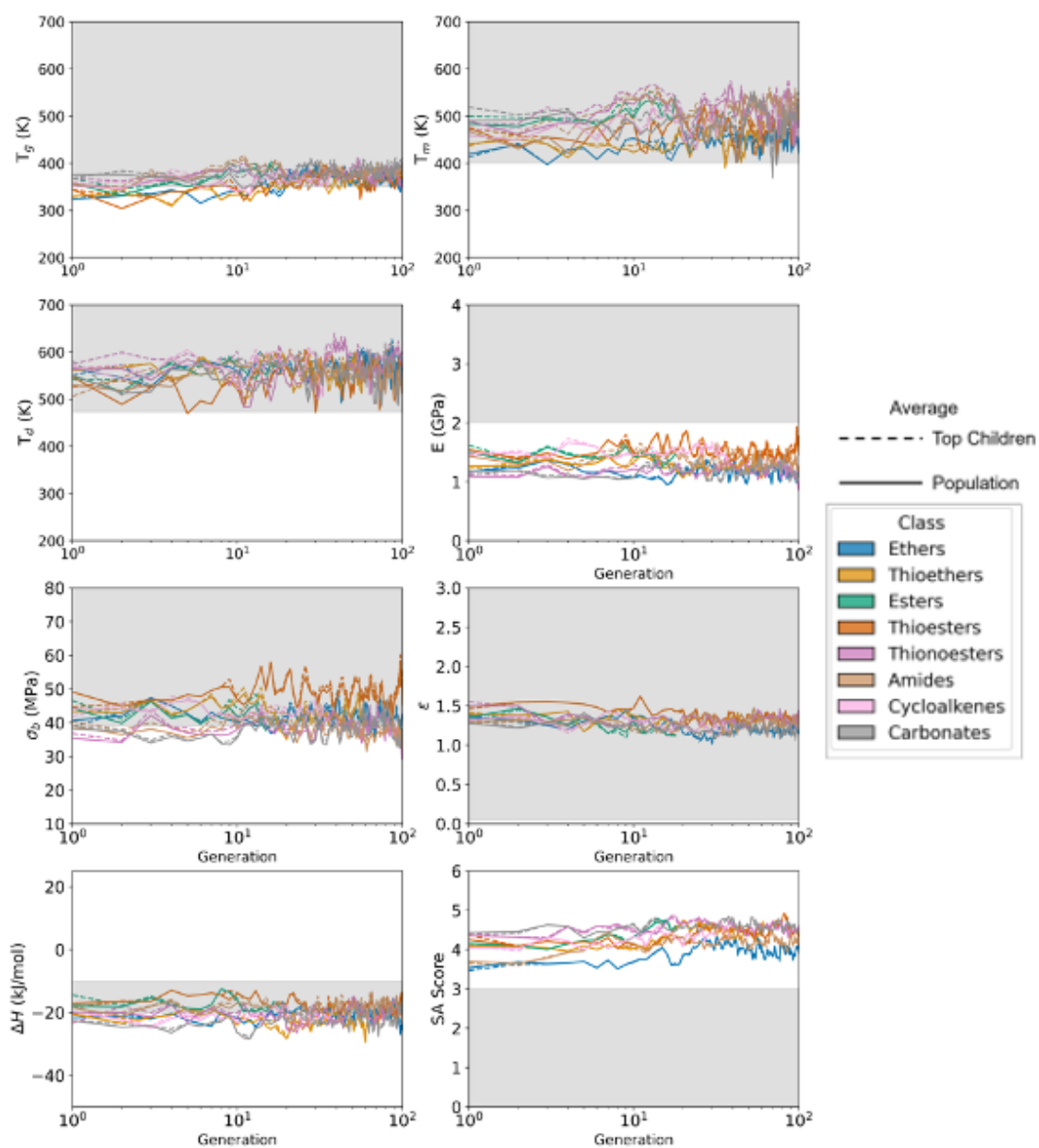


Figure 5.7: Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PS. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ , ΔH^{ROP} , SA Score.

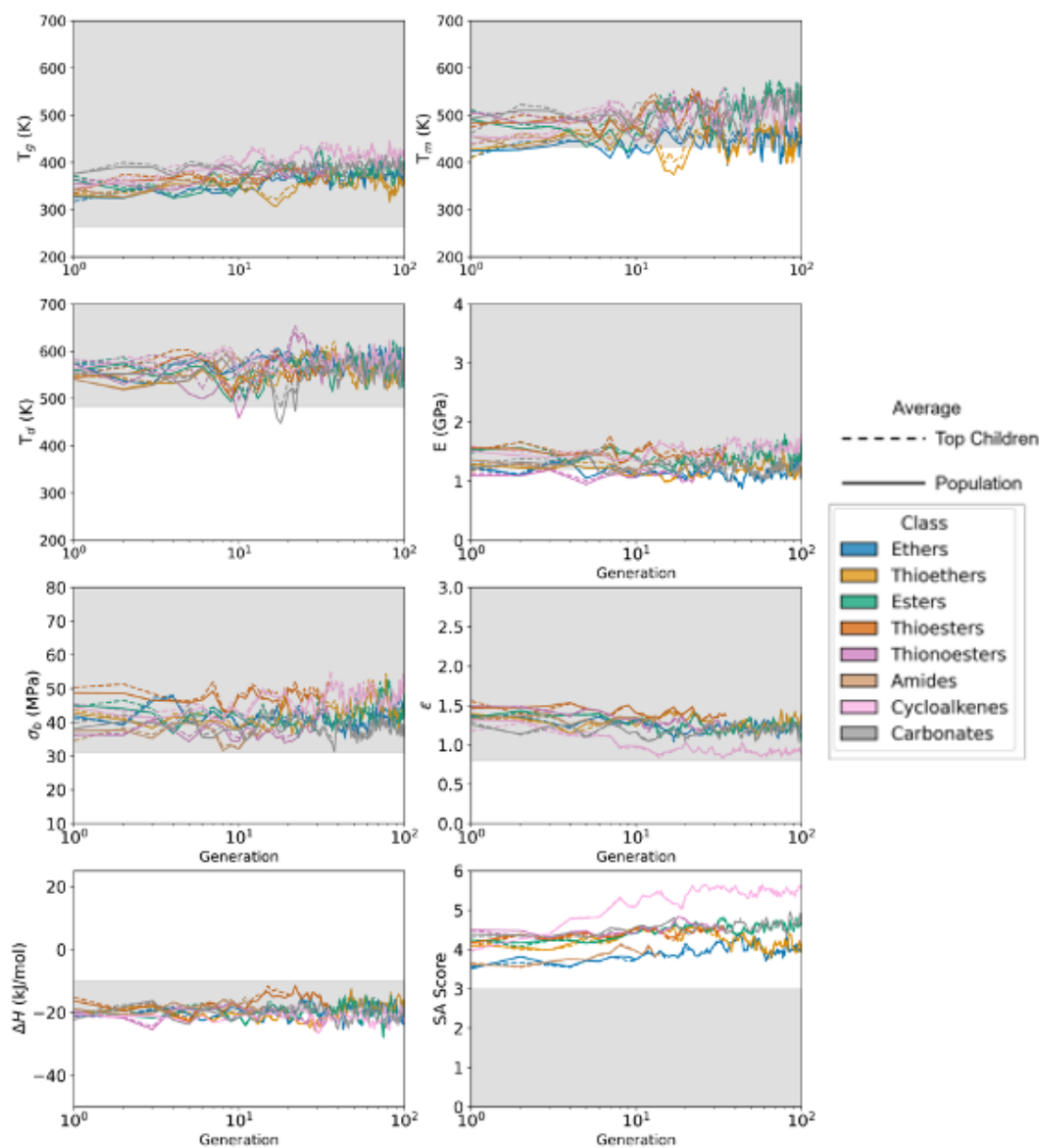


Figure 5.8: Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for PP. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.

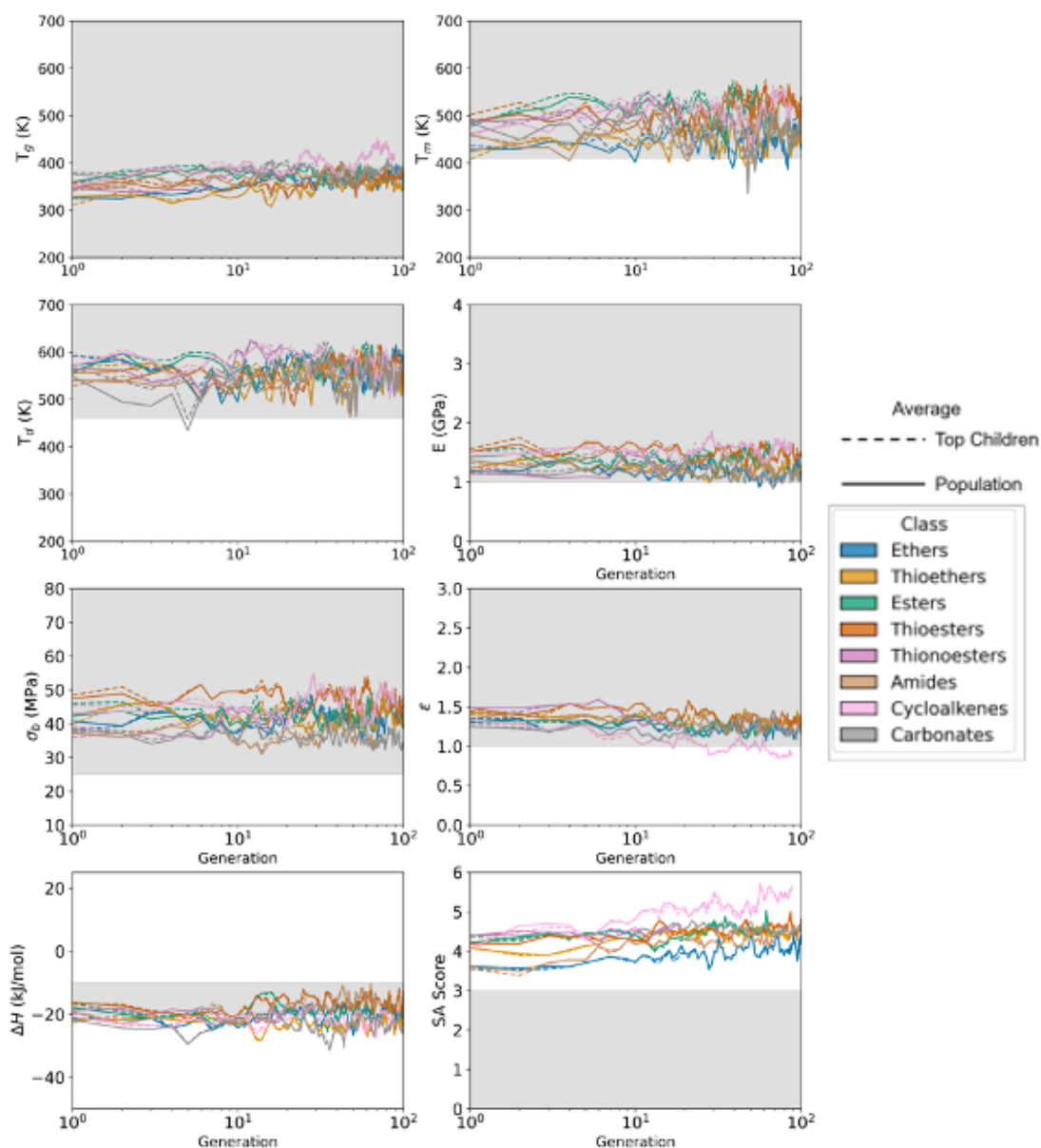


Figure 5.9: Line plots showing the change in average properties for top children polymers (dashed lines) and the total population (solid lines) over generations for one run of the GA on each class of monomers for HDPE. The top 100 polymers, selected based on their fitness function values, serve as parents for the next generation. The gray region indicates the target property range. In order the plots go T_g , T_m , T_d , E , σ_b , ϵ_b , ΔH^{ROP} , SA Score.

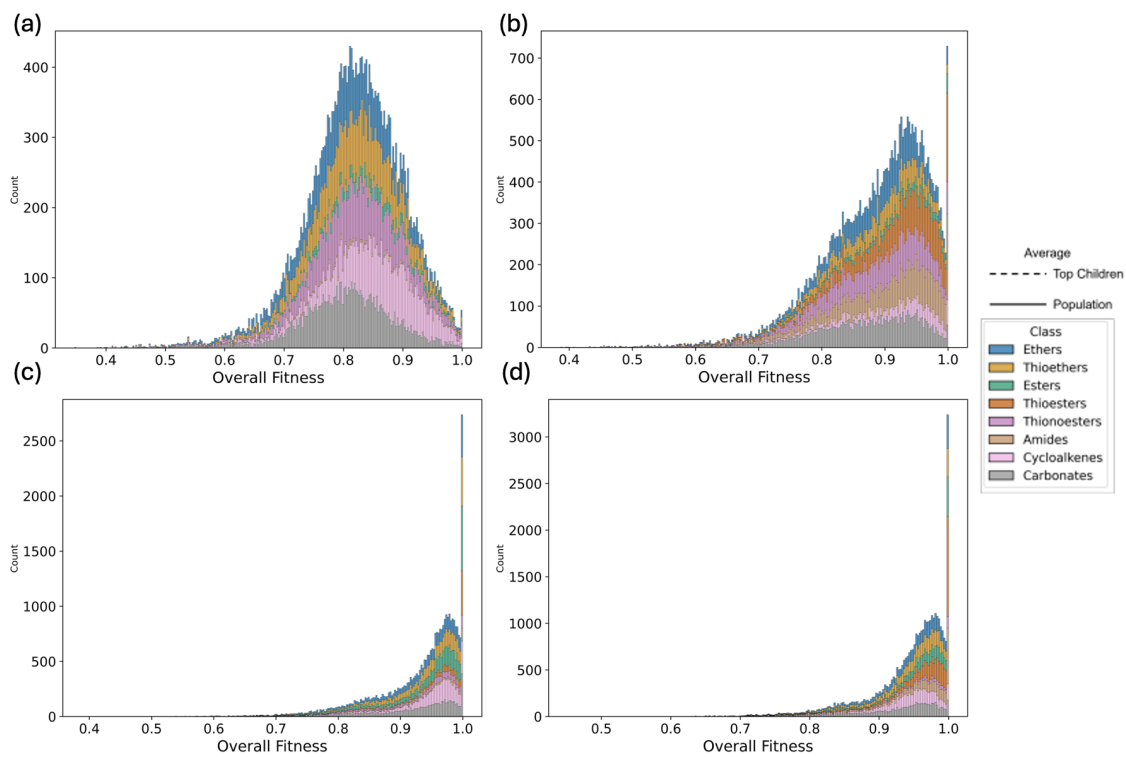


Figure 5.10: Stacked histogram plots of polymer fitness for each application of interest: (a) PET, (b) PS, (c) PP, (d) HDPE. Each polymer class used in the GA is represented by a different color as indicated by the legend to the right. As can be seen the histogram plots get progressively more right skewed from (a) to (d) as a result of the screening criteria being less and less difficult to meet from PET to HDPE.

CHAPTER 6

CONCLUSION, OUTLOOK AND FUTURE WORK

6.1 Conclusion

As has been outlined and discussed throughout this thesis, plastic production, use and subsequent "handling" at end of use is fully unsustainable in its current form. Further, this issue cannot be solved by simply using less plastic, as plastic is essential to the workings of the global economy. Further, current widespread used plastics are either not chemically recyclable or doing so is simply too expensive to be economically feasible, and herein lies the motivation to designing new-to-the-world polymers that have the necessary thermodynamics to be chemically recyclable while still having the thermal and mechanical robustness necessary to be industrially useful. This thesis works to serve as an important step in the direction of computational tools towards designing these much needed sustainable options for commodity plastics. In the first part of this thesis we show how tools are developed to access the chemically recyclability of a polymer using computational tools. In the second part of this thesis we look towards adding these leveraging these developed tools as well as previously developed data driven methods and generative algorithms to recommend the most promising polymers to chemist collaborators for synthesis. Throughout this work we harness the potential of ROP polymers as this broad class of polymers have been shown in literature to have the necessary thermodynamics to be chemically recyclable as well as significant chemical diversity to allow for a huge potential number of different designs and chemistries to pursue. This work culminates in the recommendation of many promising polymers for various industrially relevant applications.

In Chapter 3, we describe the work done to train a multi-task ML model capable of prediction ΔH^{ROP} with high accuracy. This model became pivotal in all downstream work

as it is the quintessential data driven method developed for screening polymers for chemical recyclability. Further, this model was trained such that it can take in both experimental data as well as computational data. This allowed for a huge expansion of available training data, making the model as accurate as possible. In addition, the use of computational data allows the computational chemist to expand the training data to the most pertinent chemistries of interest so the model can better extrapolate to the experimental ΔH^{ROP} for novel polymers. This work showed how computational ΔH^{ROP} data alone could significantly improve the models ability to predict the desired experimental value, and this provided an important feedback loop where first computational data could be created for chemistries of interest in a high throughput manner. This computational data that was developed throughout this thesis then also became the training data for the work done in the next chapter.

In Chapter 4, this thesis helped create a first of its kind MLFF for ROP polymers which was called PolyGET. The dataset used for training PolyGET is referred to as Poly24, a novel dataset of ab initio MD trajectories for 24 ROP polymers which contain hydrogen, carbon and oxygen. PolyGET provides unprecedented accuracy for both forces and energies allowing for larger atomic systems to run for much longer time horizons in near ab initio accuracy MD trajectories. This work then showed how not only can PolyGET be used to run simulations that would not be feasible to run using purely DFT due to time and cost considerations, but by using these larger atomic systems in simulations that ran for longer time horizons, better ΔH^{ROP} calculations could be achieved. Further, these calculations could be reached in about 12 hours, whereas the original DFT methodology could take more than a week. This work expands upon the work done in Chapter 3 in that, the computational data embedded into the ML model for ΔH^{ROP} only used the final energy differences used in calculating ΔH^{ROP} , but from just 24 of these calculations there are over 6.5 million trajectory steps, each one with an energy value and three force values for each atom in the simulation. In other words, PolyGET leverages much more of the data generated from calculating ΔH^{ROP} using DFT. This provides a first step in the direction

of developing a generalizable MLFF for ROP polymers but much more work is left to be done and will be discussed in further detail in the subsequent section.

Next in Chapter 5, this thesis leveraged two already developed generative algorithms for ROP polymers, VFS and GA. This work built upon these previous works in two ways. First, the mechanical and thermal property predictors were retrained to include the relevant ROP data as previous work had shown that these models perform poorly in this chemical space of interest. Secondly, the retrained models were ensured to have easily interpretable uncertainty quantification, and these uncertainties were used in both methods to recommend polymers the ML property predictors had lower uncertainty for. This is essential, as inherently the ML models are being asked to predict for better performing polymers and thus searching in uncharted territory. By using uncertainty metrics in the selection process, one can be more confident in the polymer recommended to chemist collaborators for synthesis. In the VFS method we leveraged a database of molecules from several online sources, in the idea that these reported molecules are more likely to either be commercially available or more readily synthesizable. From this work thousands of polymers were shown to have high potential for replacing polymers used in essential commodity plastics such as beverage containers, take home food containers, packaging and plastic wrap. Next the GA was used as a means to relax the constraint of only considering molecules previously reported. Again this method was used the newly trained thermal and mechanical property predictors as well the uncertainty quantification that accompany them. By using the GA, again thousands of potential chemically recyclable polymers that hold significant promise to replace commodity plastics are discovered.

To summarize broadly this thesis has achieved several outcomes in the direction of developing computational tools to accelerate the design of chemically recyclable polymers. First, a state of the art ML predictor of ΔH^{ROP} was developed and published which provides the backbone of screening polymers for chemical recyclability. In this work I was the main driver for which I was able to publish as first author. Secondly, a first of its

kind MLFF for ROP polymers was developed in collaboration with Rei Feng in the Chao Zhang Group. In this work I developed the training data along side Huan Tran, provided all needed validation, and developed a novel, more accurate method to calculate ΔH^{ROP} using PolyGET. For this work I will co-first author on a soon to be submitted paper with Rui Feng. Lastly, thousands of novel polymers were discovered that have the potential to replace many essential commodity plastics. This work leveraged previous work of both my own, the multi-task ΔH^{ROP} ML model, and that of Joseph Kern and Chureh Atasi [26, 129]. I expanded upon the initial work that was done by using mechanical and thermal models that had been retrained to ensure that mechanical and thermal properties of ROP polymers were included

6.2 Future Work

While the entirety of this endeavor has remained rooted in theoretical exploration, it has yielded tangible outcomes of notable significance. Beyond the design insights discussed in Chapter 5, several of the proposed designs are currently undergoing active investigation by polymer chemists with the intent of practical synthesis. Furthermore, the ΔH^{ROP} model and the associated data were made accessible to the group, allowing it to be integrated in both the original and current VFS and GA work, as well as published for others to use freely. The dataset for PolyGET, Poly24 has also been made publicly available, and the paper as well as code for PolyGET is currently in the process of being submitted to be published. All of these models and methods can be used in the future to further design polymers tailored to other applications and be expanded upon to make even more robust.

However, as is always the case in scientific research, this work is but one step to be continued by further research in the future. In the case of the work surrounding PolyGET, there is a lot of exciting future work to be conducted. First and foremost, PolyGET needs to be retrained so that the training data includes significant amounts of chemistries that contain sulfur and nitrogen. Preliminary work has been done in this direction but the resulting

model had significant difficulty handling the forces resulting from the inclusion of nitrogen atoms. We theorize this could be due to two main reasons: first that nitrogen bonds in more complex arrangements than carbon, hydrogen, oxygen and sulfur, and two that nitrogen is the least represented atom in the dataset attempted to be used to expand PolyGET to sulfur and nitrogen containing chemistries. These issues can be resolved by carefully selecting the nitrogen containing chemistries used in training to have a fair and equal distribution of the different bonding that nitrogen exhibits. Also more efforts can be done to expand the dataset to contain significantly more nitrogen containing chemistries. Once PolyGET is able to handle nitrogen and sulfur, it can be used to simulate the vast majority of ROP polymers of interest. It would be a natural next step then to see how well PolyGET can extrapolate to new, never before seen chemistries and iteratively test how little of data the model can be finetuned on to be able to handle said new chemistry. For instance, if just monomer calculations with finetune training are enough for PolyGET to improve its performance for a new out of training domain chemistry, this would provide a path to a high throughput method to make PolyGET much more generalizable. With a more generalizable version of PolyGET it would be interesting to see how finetuning on direct property prediction would perform? Could a finetuned version of PolyGET to directly predict ΔH^{ROP} rather than calculate it in a time consuming manner, be significantly more accurate than the traditional data driven approach to produce a ΔH^{ROP} property predictor shown in Chapter 3? Further, it is very worth while work to figure out how to scale PolyGET to much larger atomic systems to simulate other properties of interest such as T_g or mechanical properties. These types of simulations require on the order of magnitude of tens of thousands of atoms and preliminary work in this direction showed that the current implication of PolyGET would require over 100 GBs of GPU memory, which at the time was cost prohibitive to further test. These steps could lead to a foundational model, like what is seen in the natural language space with Large Language Models, but for the polymer chemistry space.

The work of Chapter 5 can be improved in multiple ways through future work. First,

retro-synthetic heuristics can be worked into both the VFS and GA frameworks to better determine if proposed polymers are truly synthesizable. This has been attempted to be accounted for using SA Score, but it has been found that just having a low SA Score alone is not necessarily that strong of an indicator that a given monomer targeted for ROP polymers is truly synthesizable. Thus combining the SA Score with fundamental heuristics of ROP is a straight forward way to improving the quality of candidates recommended to the chemists. For the VFS work in particular more work can be done to expand the initial set of 7.4 million molecules to include data scrapped from chemical vendors. This was done in part by scrapping VWR, but more work can be done to more updated chemical catalogues from more chemical vendors. Further, the VFS method can be better expanded to include multi-step synthesis for known chemical pathways to get to hetero-rings of interest. As for the GA method, work can be done to further downselect the functional groups used. Candidates from the GA tend to be overly complex as can be seen in the various SA Score evolution plots where the population average stays well above 4 on average. In addition to this, functional groups should be extracted from careful consideration of successful cases discovered in the VFS work. This way the best functional groups found during the brute force like method of VFS can be ensured to be used in the GA work to better discover less complex, thermally and mechanically robust polymers.

Of course, both VFS and GA hinges upon the robustness of the ML models, warranting caution, especially for accurate extrapolation to previously unseen chemical spaces. The refinement of these models, particularly through the integration of multi-fidelity techniques that combine lower fidelity MD and DFT data with experimental observations, assumes paramount importance in extending our reach into uncharted chemical territories. Thus further work can be conducted to expand the computational data not only for ΔH^{ROP} but also thermal and mechanical properties as these are the property predictors of least confidence in the screening process.

Nevertheless, the models and methods developed in this work provide a robust ground-

work for both innovative polymer design and polymer dynamics modeling. The advancement of polymer ML property predictors as well as their integration with design methodologies show immense potential for expediting the discovery of novel chemically recyclable commodity plastic replacements. The polymers candidates generated via VFS and GA are readily available for further exploration and hopefully experimental validation in the near future. Moreover, PolyGET lays an initial framework for applying MLFFs to polymer molecular dynamics, and the potential for foundational ML models in the space of polymer science. It is the true hope of the author that the work encapsulated within this thesis proves to be a helpful step in the direction of discovering polymer candidates capable of replacing commodity plastics, enabling a truly circular plastics economy and helping fix the global crisis that is plastic waste.

REFERENCES

- [1] R. Geyer, J. R. Jambeck, and K. L. Law, “Production, use, and fate of all plastics ever made,” *science Advances*, vol. 3, no. 7, 2017.
- [2] Ellen Macarthur Foundation, *The new plastics economy: Catalysing action*, <https://www.ellenmacarthurfoundation.org/the-new-plastics-economy-catalysing-action>, [Online; accessed July 2024], 2016.
- [3] US EPA, *Plastics: Material-specific data*, <https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/plastics-material-specific-data>, [Online; accessed July 2024], 2024.
- [4] G. Ozbay, M. Jones, M. Gadde, S. Isah, and T. Attarwala, “Design and operation of effective landfills with minimal effects on the environment and human health,” *Journal of Environmental and Public Health*, no. 1, 2021.
- [5] C. A. Bernardo, C. L. Simões, and L. M. C. Pinto, “Environmental and economic life cycle analysis of plastic waste management options. a review,” *AIP Conf. Proc.* 1779, no. 1, 2016.
- [6] J. Zheng and S. Suh, “Strategies to reduce the global carbon footprint of plastics,” *Nat. Clim. Chang.*, vol. 9, pp. 374–378, 2019.
- [7] P. Oblak, J. Gonzalez-Gutierrez, B. Zupančič, A. Aulova, and I. Emri, “Processability and mechanical properties of extensively recycled high density polyethylene,” *Polymer Degradation and Stability*, vol. 114, no. 1, pp. 133–145, 2015.
- [8] H. Hinsken, S. Moss, J.-R. Pauquet, and H. Zweifel, “Degradation of polyolefins during melt processing,” *Polymer Degradation and Stability*, vol. 34, no. 1, pp. 279–293, 1991, *Polymer Stabilisation Mechanisms and Applications*.
- [9] J. Hopewell, R. Dvorak, and E. Kosior, “Plastics recycling: Challenges and opportunities,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1526, 2009.
- [10] G. W. Coates and Y. D. Y. L. Getzler, “Chemical recycling to monomer for an ideal, circular polymer economy,” *Nat Rev Mater*, vol. 5, pp. 501–516, 2020.
- [11] A. Schwarz, T. Ligthart, D. G. Bizarro, P. D. Wild, B. Vreugdenhil, and T. van Harmelen, “Plastic recycling in a circular economy; determining environmental performance through an lca matrix model approach,” *Waste Management*, vol. 121, pp. 331–342, 2021.

- [12] G. P. Karmakar, *Encyclopedia of Materials: Plastics and Polymers*, M. Hashmi, Ed. Oxford: Elsevier, 2022, pp. 634–651, ISBN: 978-0-12-823291-0.
- [13] J. Nikiema and Z. Asiedu, “A review of the cost and effectiveness of solutions to address plastic pollution,” *Environmental Science and Pollution Research*, vol. 29, no. 17, pp. 24 547–24 573, 2022.
- [14] S. Penczek and G. Moad, “Glossary of terms related to kinetics, thermodynamics, and mechanisms of polymerization (iupac recommendations 2008),” *Pure and Applied Chemistry*, vol. 80, no. 10, pp. 2163–2193, 2008, [Online; accessed September 2020].
- [15] G. Odian, *Principles of Polymerization 4th edn.* Wiley, 2004.
- [16] X.-B. Meng et al., “Thermally stable and chemically recyclable poly(ketal-ester)s regulated by floor temperature,” *Journal of the American Chemical Society*, vol. 146, no. 22, pp. 15 428–15 437, 2024.
- [17] L. S. Diaz-Silvarrey, K. Zhang, and A. N. Phan, “Monomer recovery through advanced pyrolysis of waste high density polyethylene (hdpe),” *Green Chemistry*, vol. 20, no. 8, pp. 1813–1823, 2018.
- [18] P. Olsén, K. Odelius, and A.-C. Albertsson, “Thermodynamic presynthetic considerations for ring-opening polymerization,” *Biomacromolecules*, vol. 17, no. 3, pp. 699–709, 2016.
- [19] A. Tardy, J. Nicolas, D. Gimes, C. Lefay, and Y. Guillaneuf, “Radical ring-opening polymerization: Scope, limitations, and application to (bio)degradable materials,” *Chemical Reviews*, vol. 117, no. 3, pp. 1319–1406, 2017.
- [20] M. Hong and E. Y. Chen, “Chemically recyclable polymers: A circular economy approach to sustainability,” *Green Chemistry*, vol. 19, no. 16, pp. 3692–3706, 2017.
- [21] M. Hong and E. Y. Chen, “Completely recyclable biopolymers with linear and cyclic topologies via ring-opening polymerization of -butyrolactone,” *Nature Chem*, vol. 8, pp. 42–49, 2016.
- [22] W. Zhang, J. Dai, Y.-C. Wu, J.-X. C. S.-Y. Shan, Z. Cai, and J.-B. Zhu, “Highly reactive cyclic carbonates with a fused ring toward functionalizable and recyclable polycarbonates,” *ACS Macro Letters*, vol. 11, no. 2, pp. 173–178, 2022.
- [23] D. J. Saxon, E. A. Gormong, V. M. Shah, and T. M. Reineke, “Rapid synthesis of chemically recyclable polycarbonates from renewable feedstocks,” *ACS Macro Letters*, vol. 10, pp. 98–103, 2021.

- [24] E. V. Antonakou and D. S. Achilias, “Recent advances in polycarbonate recycling: A review of degradation methods and their mechanisms,” *Waste Biomass Valorization*, vol. 4, pp. 9–21, 2013.
- [25] A. Toland et al., “Accelerated scheme to predict ring-opening polymerization enthalpy: Simulation-experimental data fusion and multitask machine learning,” *The Journal of Physical Chemistry A*, vol. 127, no. 50, pp. 10 709–10 716, 2023.
- [26] J. Kern, Y. Su, W. Gutekunst, and R. Ramprasad, “An informatics framework for the design of sustainable, chemically recyclable, synthetically-accessible and durable polymers,” *arXiv*, 2024.
- [27] K. A. Stellmach, M. K. Paul, Y.-L. Su, R. Ramprasad, A. C. Engler, and W. R. Gutekunst, “Improving the accuracy of ceiling temperature measurements: Best practices and common pitfalls,” *Macromolecules*, no. 8, pp. 3729–3741, 2025.
- [28] A. Duda and A. Kowalski, *Handbook of Ring-Opening Polymerization*. John Wiley and Sons, Ltd., 2009, Chapter 1, pp 151.
- [29] K. A. Stellmach et al., “Modulating polymerization thermodynamics of thiolactones through substituent and heteroatom incorporation,” *ACS Macro Letters*, vol. 11, no. 7, pp. 895–901, 2022.
- [30] H. Tran, A. Toland, K. Stellmach, M. K. Paul, W. Gutekunst, and R. Ramprasad, “Toward recyclable polymers: Ring-opening polymerization enthalpy from first-principles,” *The Journal of Physical Chemistry Letters*, vol. 13, no. 21, pp. 4778–4785, 2022.
- [31] H. Tran et al., “Design of functional and sustainable polymers assisted by artificial intelligence,” *Nature Reviews Materials*, 2024.
- [32] L. Chen et al., “Polymer informatics: Current status and critical next steps,” *Materials Science and Engineering: R: Reports*, vol. 144, 2021.
- [33] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, “Polymer genome: A data-powered polymer informatics platform for property predictions,” *The Journal of Physical Chemistry C*, vol. 122, no. 31, pp. 17 575–17 585, 2018.
- [34] H. D. Tran et al., “Machine-learning predictions of polymer properties with polymer genome,” *Journal of Applied Physics*, vol. 128, no. 17, p. 171 104, 2020.
- [35] G. G. für umfassende Analysen GmbH, *The contribution of plastic products to resource efficiency*, <https://www.teppfa.eu/wp-content/uploads/HSEGUAFinalreportplasticsresource.pdf>, [83, A-1150 Vienna; pp 57–58], 2005.

- [36] US EPA, *Containers and packaging: Product-specific data*, <https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/containers-and-packaging-product-specific>, [Online; accessed July 2024], 2024.
- [37] Our World in Data, *Global primary plastic production by polymer*, <https://ourworldindata.org/grapher/plastic-production-polymer>, [Online; accessed July 2024].
- [38] Y. Miao, A. von Jouanne, and A. Yokochi, “Current technologies in depolymerization process and the road ahead,” *Polymers*, vol. 13, no. 3, p. 449, 2021.
- [39] T. D. Huan and R. Ramprasad, “Polymer structure prediction from first principles,” *The Journal of Physical Chemistry Letters*, vol. 11, no. 15, pp. 5823–5829, 2020.
- [40] H. S. K.-H. Shen, J. H. Montoya, H. Tran, and R. Ramprasad, “Polymer structure predictor (psp): A python toolkit for predicting atomic-level structural models for a range of polymer geometries,” *Journal of Chemical Theory and Computation*, vol. 18, no. 4, pp. 2737–2748, 2022.
- [41] M. A. Wood, A. C. T. van Duin, and A. Strachan, “Coupled thermal and electromagnetic induced decomposition in the molecular explosive hmx; a reactive molecular dynamics study,” *The Journal of Physical Chemistry A*, vol. 118, no. 5, pp. 885–895, 2014.
- [42] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, “Comparing molecules and solids across structural and alchemical space,” *Physical Chemistry Chemical Physics*, vol. 18, no. 20, pp. 13 754–13 769, 2016.
- [43] L. Himanen et al., “Dscribe: Library of descriptors for machine learning in materials science,” *Computer Physics Communications*, vol. 247, 2020.
- [44] S. Plimpton, “Fast parallel algorithms for short-range molecular dynamics,” *J. Comput. Phys.*, vol. 117, pp. 1–19, 1995.
- [45] G. Kresse and J. Furthmüller, “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set,” *Computational Materials Science*, vol. 6, pp. 15–50, 1996.
- [46] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, “Accelerated materials property predictions and design using motif-based fingerprints,” *Physical Review B*, vol. 92, no. 1, 2015.
- [47] H. L. Morgan, “The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service,” *Journal of Chemical Documentation*, no. 2, pp. 107–113, 1965.

- [48] RDKit., <https://www.rdkit.org/>, [Online; accessed October 2025].
- [49] R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, “Topological torsion: A new molecular descriptor for sar applications. comparison with other descriptors,” *Journal of Chemical Information and Computer Sciences*, no. 2, pp. 82–85, 1987.
- [50] C. Kuenneth, A. C. Rajan, H. Tran, L. Chen, C. Kim, and R. Ramprasad, “Designing exceptional gas-separation polymer membranes using machine learning,” *Sci. Adv.*, vol. 6, eaaz4301, 2020.
- [51] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2006.
- [52] S. Suthaharan, *Support Vector Machine*. Springer, 2016, vol. 36.
- [53] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: From early developments to recent advancements,” *Systems Science and Control Engineering*, vol. 2, no. 1, pp. 602–609, 2014.
- [54] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi, “Boosted random forest,” *IEICE TRANSACTIONS on Information and Systems*, vol. E98-D, no. 9, pp. 1630–1636, 2015.
- [55] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, “Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons,” *Phys. Rev. Lett.*, vol. 104, no. 13, p. 136403, 2010.
- [56] A. P. Bartók et al., “Machine learning unifies the modeling of materials and molecules,” *Sci. Adv.*, vol. 3, no. 12, e1701816, 2017.
- [57] V. L. Deringer, M. A. Caro, and G. Csányi, “A general-purpose machine-learning force field for bulk and nanostructured phosphorus,” *Nat. Commun.*, vol. 11, no. 1, pp. 1–11, 2020.
- [58] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, “Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials,” *J. Comput. Phys.*, vol. 285, pp. 316–330, 2015.
- [59] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, “Accurate force field for molybdenum by machine learning large materials data,” *Physical Review Materials*, vol. 1, no. 4, p. 043603, 2017.

- [60] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, “Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics,” *Phys. Rev. Lett.*, vol. 120, no. 14, p. 143 001, 2018.
- [61] V. Botu and R. Ramprasad, “Learning scheme to predict atomic forces and accelerate materials simulations,” *Phys. Rev. B*, vol. 92, no. 9, p. 094 306, 2015.
- [62] V. Botu and R. Ramprasad, “Adaptive machine learning framework to accelerate ab initio molecular dynamics,” *Int. J. Quant. Chem.*, vol. 115, no. 16, pp. 1074–1083, 2015.
- [63] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, “A universal strategy for the creation of machine learning-based atomistic force fields,” *npj Comput. Mater.*, vol. 3, no. 1, p. 37, 2017.
- [64] T. D. Huan, R. Batra, J. Chapman, C. Kim, A. Chandrasekaran, and R. Ramprasad, “Iterative-learning strategy for the development of application-specific atomistic force fields,” *J. Phys. Chem. C*, vol. 123, no. 34, pp. 20 715–20 722, 2019.
- [65] I. Ramzan, J. Kalayan, L. Kong, R. A. Bryce, and N. A. Burton, “Machine learning of atomic forces from quantum mechanics: An approach based on pairwise interatomic forces,” *Int. J. Quant. Chem.*, vol. 122, no. 21, e26984, 2022.
- [66] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, “Machine learning force fields: Construction, validation, and outlook,” *J. Phys. Chem. C*, vol. 121, no. 1, pp. 511–522, 2017.
- [67] A. Vaswani et al., “Attention is all you need,” *arXiv*, 2023.
- [68] S. Doerr et al., “Torchmd: A deep learning framework for molecular simulations,” *J. Chem. Theory Comput.*, vol. 17, no. 4, pp. 2355–2363, 2021.
- [69] P. Thölke and G. De Fabritiis, “Torchmd-net: Equivariant transformers for neural network based molecular potentials,” *arXiv preprint arXiv:2202.02541*, 2022.
- [70] R. Feng et al., “May the force be with you: Unified force-centric pre-training for 3d molecular conformations,” *NeurIPS*, 2023.
- [71] J. McCall, “Genetic algorithms for modelling and optimisation,” *Journal of Computational and Applied Mathematics, Special Issue on Mathematics Applied to Immunology*, vol. 184, no. 1, pp. 205–222, 2005.
- [72] X.-S. Yang, *Chapter 5 - Genetic Algorithms*. Oxford: Elsevier, 2014, pp. 77–87.

- [73] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey, "In the art of compiling and using 'drug-like' chemical fragment spaces," *ChemMedChem*, vol. 3, no. 10, pp. 1503–1507, 2008.
- [74] P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," *Journal of Cheminformatics*, vol. 1, no. 8, 2009.
- [75] P. Shetty et al., "A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing," *Npj Computational Materials*, vol. 9, no. 1, pp. 1–12, 2023.
- [76] D. Kamal et al., "Novel high voltage polymer insulators using computational and data-driven techniques," *J. Chem. Phys.*, vol. 154, no. 17, p. 174906, 2021.
- [77] J. W. B. et al., "Designing exceptional gas-separation polymer membranes using machine learning," *Sci. Adv.*, vol. 6, eaaz4301, 2020.
- [78] L. Chen et al., "Frequency-dependent dielectric constant prediction of polymers using machine learning," *npj Comput. Mater.*, vol. 6, no. 1, p. 61, 2020.
- [79] J. Nistane, L. Chen, Y. Lee, R. Lively, and R. Ramprasad, "Estimation of the flory-huggins interaction parameter of polymer-solvent mixtures using machine learning," *MRS Commun.*, vol. 12, pp. 1096–1102, 2022.
- [80] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023.
- [81] D. C. Rapaport, *The art of molecular dynamics simulation*. Cambridge university press, 2004.
- [82] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.
- [83] T. D. Huan et al., "Advanced polymeric dielectrics for high energy density applications," *Prog. Mater. Sci.*, vol. 83, p. 236, 2016.
- [84] H. Tran, R. Gurnani, C. Kim, G. Pilania, H.-K. Kwon, and R. Lively, "Polymer informatics comes of age: Design of functional and sustainable polymers assisted by artificial intelligence," submitted 2024.
- [85] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, B864–B871, 1964.

- [86] W. Kohn and L. Sham, “Self-consistent equations including exchange and correlation effects,” *Phys. Rev.*, vol. 140, A1133–A1138, 1965.
- [87] J. E. Jones, “On the determination of molecular fields.—ii. from the equation of state of a gas,” *Proc. R. Soc. A*, vol. 106, no. 738, pp. 463–477, 1924.
- [88] M. S. Daw and M. I. Baskes, “Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals,” *Phys. Rev. B*, vol. 29, no. 12, p. 6443, 1984.
- [89] M. S. Daw, S. M. Foiles, and M. I. Baskes, “The embedded-atom method: A review of theory and applications,” *Mater. Sci. Rep.*, vol. 9, no. 7-8, pp. 251–310, 1993.
- [90] J. Tersoff, “New empirical approach for the structure and energy of covalent systems,” *Phys. Rev. B*, vol. 37, no. 12, p. 6991, 1988.
- [91] A. C. Van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, “Reaxff: A reactive force field for hydrocarbons,” *J. Phys. Chem. A*, vol. 105, no. 41, pp. 9396–9409, 2001.
- [92] T. P. Senftle et al., “The reaxff reactive force-field: Development, applications and future directions,” *npj Comput. Mater.*, vol. 2, no. 1, pp. 1–14, 2016.
- [93] W. L. Jorgensen and J. Tirado-Rives, “The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin,” *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, 1988.
- [94] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, “Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids,” *J. Am. Chem. Soc.*, vol. 118, no. 45, pp. 11 225–11 236, 1996.
- [95] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, 2004.
- [96] C. J. Dickson, L. Rosso, R. M. Betz, R. C. Walker, and I. R. Gould, “Gafflipid: A general amber force field for the accurate molecular dynamics simulation of phospholipid,” *Soft Matter*, vol. 8, no. 37, pp. 9617–9627, 2012.
- [97] K. Sprenger, V. W. Jaeger, and J. Pfaendtner, “The general amber force field (gaff) can accurately predict thermodynamic and transport properties of many ionic liquids,” *J. Phys. Chem. B*, vol. 119, no. 18, pp. 5882–5895, 2015.

- [98] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, no. 14, p. 146401, 2007.
- [99] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: Recent applications and prospects," *npj Computational Materials*, vol. 3, no. 1, p. 54, 2017.
- [100] O. T. Unke et al., "Machine learning force fields," *Chemical Reviews*, vol. 121, no. 16, pp. 10142–10186, 2021.
- [101] I. Poltavsky and A. Tkatchenko, "Machine learning force fields: Recent advances and remaining challenges," *J. Phys. Chem. Lett.*, vol. 12, no. 28, pp. 6551–6564, 2021.
- [102] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, "Se(3)-transformers: 3d rotation equivariant attention networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1970–1981, 2020.
- [103] S. Batzner et al., "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nat. Commun.*, vol. 13, no. 1, pp. 1–11, 2022.
- [104] Y.-L. Liao and T. Smidt, "Equiformer: Equivariant graph attention transformer for 3d atomistic graphs," *arXiv preprint arXiv:2206.11990*, 2022.
- [105] F. Xie, T. Lu, S. Meng, and M. Liu, "Gptff: A high-accuracy out-of-the-box universal ai force field for arbitrary inorganic materials," *arXiv preprint arXiv:2402.19327*, 2024.
- [106] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Sci. Data*, vol. 1, no. 1, pp. 1–7, 2014.
- [107] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," *J. Chem. Inf. Model.*, vol. 52, no. 11, pp. 2864–2875, 2012.
- [108] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," *Nature communications*, vol. 9, no. 1, p. 3887, 2018.
- [109] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: An extensible neural network potential with dft accuracy at force field computational cost," *Chemical science*, vol. 8, no. 4, pp. 3192–3203, 2017.

- [110] C. Devereux et al., “Extending the applicability of the ani deep learning molecular potential to sulfur and halogens,” *Journal of Chemical Theory and Computation*, vol. 16, no. 7, pp. 4192–4202, 2020.
- [111] S. Mohanty et al., “Development of scalable and generalizable machine learned force field for polymers,” *Sci. Rep.*, vol. 13, no. 1, p. 17 251, 2023.
- [112] S. J. Hong et al., “First-principles-based machine-learning molecular dynamics for crystalline polymers with van der waals interactions,” *J. Phys. Chem. Lett.*, vol. 12, no. 25, pp. 6000–6006, 2021.
- [113] P. Gkeka et al., “Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems,” *J. Chem. Theory Comput.*, vol. 16, no. 8, pp. 4757–4775, 2020.
- [114] S. Grimme, “Semiempirical gga-type density functional constructed with a long-range dispersion correction,” *J. Comp. Chem.*, vol. 27, p. 1787, 2006.
- [115] K. Lee, É. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, “Higher-accuracy van der waals density functional,” *Phys. Rev. B*, vol. 82, 081101(R), 2010.
- [116] T. D. Huan and R. Ramprasad, “Polymer structure prediction from first principles,” *J. Phys. Chem. Lett.*, vol. 11, no. 15, pp. 5823–5829, 2020.
- [117] H. Sahu, K.-H. Shen, J. H. Montoya, H. Tran, and R. Ramprasad, “Polymer structure predictor (psp): A python toolkit for predicting atomic-level structural models for a range of polymer geometries,” *J. Chem. Theory Comput.*, vol. 18, no. 4, pp. 2737–2748, 2022.
- [118] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “Schnet-a deep learning architecture for molecules and materials,” *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241 722, 2018.
- [119] R. Feng et al., “May the force be with you: Unified force-centric pre-training for 3d molecular conformations,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [120] G. Kresse and J. Furthmüller, “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set,” *Comput. Mater. Sci.*, vol. 6, pp. 15–50, 1996.
- [121] G. Kresse and J. Furthmüller, “Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set,” vol. 54, pp. 11 169–11 186, 1996.

- [122] P. E. Blöchl, “Projector augmented-wave method,” vol. 50, pp. 17 953–17 979, 1994.
- [123] J. P. Perdew, K. Burke, and M. Ernzerhof, “Generalized gradient approximation made simple,” vol. 77, pp. 3865–3868, 1996.
- [124] H. Tran, A. Toland, K. Stellmach, M. K. Paul, W. Gutekunst, and R. Ramprasad, “Toward recyclable polymers: Ring-opening polymerization enthalpy from first-principles,” *J. Phys. Chem. Lett.*, vol. 13, no. 21, pp. 4778–4785, 2022.
- [125] Ellen Macarthur Foundation, *The global commitment 2021 progress report*, <https://www.unep.org/resources/report/global-commitment-2021-progress-report>, [Online], 2021.
- [126] R. Nugnes, M. Lavorgna, E. Orlo, C. Russo, and M. Isidori, “Toxic impact of polystyrene microplastic particles in freshwater organisms,” *Chemosphere*, vol. 299, p. 134 373, 2022.
- [127] J. Hwang, D. Choi, S. Han, S. Y. Jung, J. Choi, and J. Hong, “Potential toxicity of polystyrene microplastic particles,” *Scientific Reports*, vol. 10, no. 1, p. 7391, 2020.
- [128] Polymerize Matmerize, Inc., <https://polymrize.matmerize.com/>, [Online; accessed July 2025].
- [129] C. Atasi, J. Kern, and R. Ramprasad, “Design of recyclable plastics with machine learning and genetic algorithm,” *J. Chem. Inf. Model.*, vol. 64, no. 24, pp. 9249–9259, 2024.