Contents lists available at ScienceDirect

Materials Science & Engineering R

journal homepage: www.elsevier.com/locate/mser

Polymer informatics: Current status and critical next steps

Lihua Chen [a], Ghanshyam Pilania [b], Rohit Batra [c], Tran Doan Huan [a], Chiho Kim [a], Christopher Kuenneth [a], Rampi Ramprasad [a,*]

[a] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[c] Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, USA

ARTICLE INFO

Keywords:
Polymer informatics
Machine learning
Deep learning
Polymer design and discovery
Polymer synthesis

ABSTRACT

Artificial intelligence (AI) based approaches are beginning to impact several domains of human life, science and technology. Polymer informatics is one such domain where AI and machine learning (ML) tools are being used in the efficient development, design and discovery of polymers. Surrogate models are trained on available polymer data for instant property prediction, allowing screening of promising polymer candidates with specific target property requirements. Questions regarding synthesizability, and potential (retro)synthesis steps to create a target polymer, are being explored using statistical means. Data-driven strategies to tackle unique challenges resulting from the extraordinary chemical and physical diversity of polymers at small and large scales are being explored. Other major hurdles for polymer informatics are the lack of widespread availability of curated and organized data, and approaches to create machine-readable representations that capture not just the structure of complex polymeric situations but also synthesis and processing conditions. Methods to solve inverse problems, wherein polymer recommendations are made using advanced AI algorithms that meet application targets, are being investigated. As various parts of the burgeoning polymer informatics ecosystem mature and become integrated, efficiency improvements, accelerated discoveries and increased productivity can result. Here, we review emergent components of this polymer informatics ecosystem and discuss imminent challenges and opportunities.

1. Introduction

Over the course of less than a century, polymers have become pervasive in everyday life and high-technology [1–10]. Mass production of niche polymers, such as polyethylene, polypropylene and polystyrene, has outstripped the production scale of iron and steel, which have been the staple materials for millennia [11]. Different parts of the practically infinite chemical space of polymers display a dizzying variety of distinctive properties, which can be tuned exquisitely through control of their chemical and morphological structure [1,2]. Extensive efforts have been devoted to searching the chemical space and tinkering with their structure and chemistry to optimize their properties for specific applications. Traditional intuition-driven and/or trial-and-error approaches have already revealed the promise that the polymer class of materials holds. Nevertheless, given the vastness of the chemical and structural space, new methods are required to effectively and efficiently search this space to identify optimal, application-specific solutions.

The field of polymer informatics attempts to address this daunting search problem by the utilization of modern data- and information-centric approaches, inspired by emerging artificial intelligence (AI) and machine learning (ML) methods [12–19]. Polymer informatics efforts are seeing heightened activity and successes in recent years [19–25], but many of the ideas and concepts have gradually taken shape over a period of decades [19–21,24,26].

Fig. 1 illustrates the essential elements of polymer informatics. The first vital ingredient is the polymer data, derived from experiments and (high-throughput) computations. Unlike hard materials, only limited well-organized/clean polymer data is available to be used for ML or AI-based techniques, e.g., in polymer handbooks [27] and online repositories [28]. Large volumes of experimental data remain trapped in the scientific literature, which is occasionally mined via laborious manual excerption. An emerging alternative approach is natural language processing (NLP) to continuously and dynamically extract polymer data, but significant future efforts are needed to effectively and

* Corresponding author.
E-mail address: rampi.ramprasad@mse.gatech.edu (R. Ramprasad).

https://doi.org/10.1016/j.mser.2020.100595

Available online 26 December 2020

accurately extract polymer data from literature. Another important resource of polymer data is high-throughput computations using density functional theory (DFT) [29–31] and classical molecular dynamics (MD) simulations [32–39]. The recent development of autonomous computational agents, composed of machine learning modules and high-throughput computations, holds great promise for polymer data generation [40].

The second important component of polymer informatics is a suitable framework to create machine-readable polymer representations. Linear notations are commonly adopted to describe the chemical information of polymers, for instance, using Simplified Molecular-Input Line-Entry System (SMILES) [41]. With SMILES as input, polymers are either directly fingerprinted using hierarchical polymer fingerprints [22,23] or molecular fingerprints [42,43] that are widely used in cheminformatics. Alternatively, optimal fingerprint representation (or latent knowledge) of polymers can be obtained using variational [44] or graph autoencoders [45,46]. Designing fingerprints that fully capture not just the chemical and morphological information of polymers, but also how they were synthesized and processed is one of the most challenging parts of polymer informatics.

Using the numerical polymer fingerprints and target property data as input, we move to the third part of the polymer informatics: polymer property prediction and design. In the former, various machine learning algorithms, e.g., non-linear regression [22], multi-fidelity information fusion [47,48] and deep neural networks [43,49], can be applied to learn the relationship between polymer fingerprints and their target property, resulting in a surrogate property prediction model for polymers. The developed surrogate model can instantly predict various properties of new polymers defined by the user. Another key benefit of the polymer informatics ecosystem is to accelerate the discovery of polymers with target properties for various applications. Several polymer design algorithms have been proposed, e.g., screening candidates based on the ML predicted properties from a huge list of enumerated polymers, iteratively selecting the next interesting polymer using active learning, and producing desired hypothetical polymers using genetic or generative deep learning algorithms. These design approaches have significantly accelerated the polymer design process for capacitors [50–52], membrane separation [42], organic solar cells [53], among others.

Once the desired polymer candidates are proposed, the next step is to validate the polymers via computational methods and physical synthesis. The former is manageable using AI-automated data generation agents that control computational workflows (but are applicable to only those properties that are accessible through computations). The latter is a challenge, as the synthesis of the selected polymer candidates is not straightforward. Chemical reactions, precursors, reagents and processing conditions (temperature, pressure and solvents) must be identified for each polymer to successfully synthesize them. Attempts are being made to expedite this process by using AI-assisted synthesis planning and robotic/autonomous (retro)synthesis. Although computer-aided synthesis design for molecules was recently accomplished [54,55], there remains lots of scope and challenges for polymer synthesis planning, which is expected to blossom rapidly in the next several years. Moreover, the data obtained from these synthesized polymers and/or from their computations can be added into existing polymer repositories to re-optimize ML models and re-design or re-imagine the next experiments.

In this paper, we review these emergent components of the polymer informatics ecosystem and discuss imminent challenges and opportunities. In Section 2, we discuss protocols available for polymer data generation, acquisition and management. It is followed by a survey of various schemas for polymer representations in Section 3. Next, we move on to review machine learning algorithms utilized and adapted for polymer property prediction (Section 4) and design for various applications (Section 5). We then list several representative application examples that have benefited and may benefit from the polymer informatics philosophy in Section 6 and identify critical next steps that the community will need to address and surmount in the near future in Section 7.

2. Data generation, acquisition and management

The central tenet of polymer informatics is that if a sufficient volume of polymer data can be appropriately generated or curated, it can facilitate discovery/design of functional polymers with targeted performance. Below we discuss how polymer data can be accumulated from the literature or generated using high-throughput and autonomous computations.

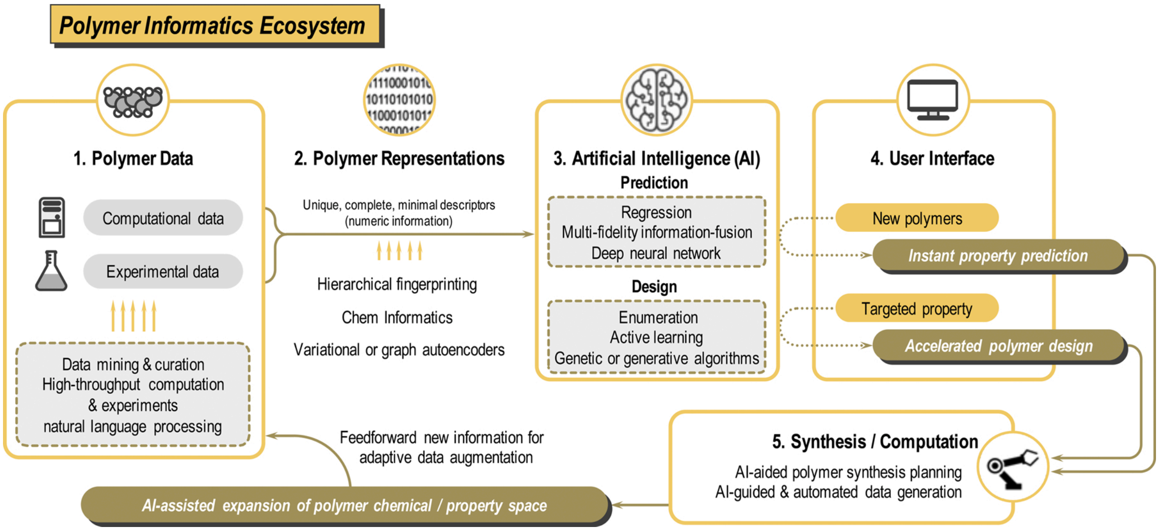

Polymer Informatics Ecosystem
1. Polymer Data
Computational data
Experimental data
Data mining & curation
High-throughput computation
& experiments
natural language processing
2. Polymer Representations
Unique, complete, minimal descriptors
(numeric information)
Hierarchical fingerprinting
Chem Informatics
Variational or graph autoencoders
3. Artificial Intelligence (AI)
Prediction
Regression
Multi-fidelity information-fusion
Deep neural network
Design
Enumeration
Active learning
Genetic or generative algorithms
4. User Interface
New polymers
Instant property prediction
Targeted property
Accelerated polymer design
5. Synthesis / Computation
AI-aided polymer synthesis planning
AI-guided & automated data generation
Feedforward new information for
adaptive data augmentation
AI-assisted expansion of polymer chemical / property space

Fig. 1. Essential elements of Polymer Informatics Ecosystem: (1) polymer data, derived from (high-throughput) computations and/or experiments (through manual or natural language processing-aided excerption); (2) polymer representations, transforming polymers into numerical fingerprints and making it amenable to ML/AI models; (3) developing surrogate models for polymer property prediction and design polymers with desired properties for specific applications; (4) Online user interfaces provide easy and quick public access to the developed surrogate models and/or the underlying polymer data; (5) AI-aided computational and synthesis validation, feeding new information to existing polymer repositories.

2.1. Scientific literature

A reliable and enormous data resource for polymer data is the scientific literature, including printed handbooks [56–59], online repositories [28] and journal articles. As listed in Table 1, polymer handbooks, such as the Polymer Handbook [56] and Properties of Polymers [58], are introductory materials containing chemical, property and synthesis information on polymers. More recently, several polymer databases have been digitalized, allowing for easy access to polymer data. A few representative databases include PoLyInfo supported by the National Institute for Materials Science of Japan (NIMS) [28], CROW Polymer Property Database [60], Polymers: A property database [61], CAMPUS [62], LANDOLT-BORNSTEIN [63] and Polymer Property Predictor and Database (NIST) [64]. In contrast to the field of inorganic materials, only a few computation-based property databases for polymers are available. This can be attributed to the high computational complexity of polymers due to their complicated physical and chemical structures. A good example of a database of computational data polymers is Khazana [65], which includes DFT computed band gap, dielectric constant, refractive index and charge injection barriers. The third important resource of polymer data is the ever-increasing corpus of published journal articles.

Timely dynamical extraction of polymer data from the literature in a machine-readable format can be challenging and is achieved using either the laborious manual excerption and/or machine-learning methods usually classified as NLP. Manual text excerption refers to the old-fashioned procedure of collecting data from the literature and entails laborious efforts for data extraction, validation, and normalization owing to the absence of standard journal policies for publishing polymer data. Nonetheless, researchers have painstakingly collected important information on polymer types, their chemical structures (repeat units), names and class, their properties (e.g., physical, thermal, mechanical, dielectric, physicochemical and solution properties), and their synthesis recipes (e.g., polymerization paths, reactants and reagents) [28,56–59]. Crucially, easy access to the resulting databases has been provided through online repositories, as discussed earlier.

Machine learning-based NLP has emerged as an alternative approach for information excerption from the literature in the last few years. NLP can be used to automatically scan the literature corpus and extract relevant polymer properties, which can be organized in a tabular fashion based on the NLP model predicted text relations. The use of NLP in materials science is still in its infancy, due to the difficulty in interpreting technical languages and incorporating domain knowledge. It is further complicated by the absence of standard journal policies for publishing scientific data. Several initial attempts have tried to use NLP to collect materials synthesis recipes, capture latent knowledge and to even predict potentially superior thermoelectrics [18,67,68]. These successes motivate the further application of NLP in polymer informatics, such as the extraction of properties, synthesis recipes or processing conditions from past literature. Despite initial success, many unique challenges are posed in the case of polymers, for example, non-uniform polymer names. More details are described in Section 7.

2.2. High-throughput and autonomous computational agents

High-throughput computations using first-principles theory and classical MD are important approaches to amass polymer data. However, this task is non-trivial because of the enormously complicated chemical and physical structures of polymers at the atomic scale; polymers usually display either amorphous or semi-crystalline phases. Given the expensive computational cost of first-principles computations, small length-scale models ($<$100 atoms) have been developed to represent polymers and approximate their physical, electronic and dielectric properties [29–31,40,51]. The computed results, however, generally suffer from certain accuracy issues depending on the methodology. To model polymers in the large-scale, classical MD with empirical force fields have been applied to study the structural dynamics [69–76] and diverse properties (e.g., dielectric, thermal, mechanical and ion transport properties) [32,34,35,37,38,71,77–82] of polymers. However, such parameterization schemes are also restricted by the availability of force fields and the high computational cost to simulate very large systems ($>$thousands of atoms).

Balancing the trade-off between cost and accuracy, past efforts have led to the computation of the electronic, dielectric and optical properties (such as band gap, charge injection barriers and dielectric constant) of thousands of polymers using DFT [29–31,40,83]. A hierarchy of models, i.e., single-chain, pure crystal and amorphous, has been developed to represent realistic polymers, as shown in Fig. 2a). The simplest single-chain model is composed of only a chain of monomers in vacuum, while crystal and amorphous models represent the crystalline and amorphous regions of polymers, respectively. In spite of this simplification, the creation of correct low-energy crystalline structure of a polymer, especially for novel polymers, remains a major challenge [40]. To address this issue, Huan et al. developed a general computational workflow, referred to as polymer structure predictor (PSP), to predict crystal structures of linear polymers. In this workflow, a polymer is defined in terms of its chemical composition and atomic connectivity, using the SMILES notation (more details in Section 3). Reasonable single-chain and crystal models of the polymer can be predicted/created [40] using this scheme. Such efforts have led to formation of the largest dataset for polymers using computations, which can be accessed from https://khazana.gatech.edu. Some of the important computed properties include the crystal band gap, single-chain band gap, charge injection barriers, atomization energy, ionization energy, electron affinity, dielectric constant and refractive index [29,40]. Other researchers have spent extensive efforts on estimating thermal conductivity [78,79,84,

Table 1
Available polymer data resources.

Source	Name	Data type
Handbook	Polymer Handbook [56], Handbook of Polymers [57],	Empirical
	Properties of Polymers [58], Polymer Data Handbook [59]	Empirical
	Polymer synthesis: theory and practice [66]	Empirical
Online Repositories	PoLyInfo [28]	Empirical
	CROW Polymer Property Database [60]	Empirical
	Polymers: A property database [61]	Empirical
	CAMPUS [62]	Empirical
	LANDOLT-BORNSTEIN [63]	Empirical
	Polymer Property Predictor and Database (NIST) [64]	Empirical
	Khazana [65]	Computational
Published journal articles	Various	Empirical/Computational

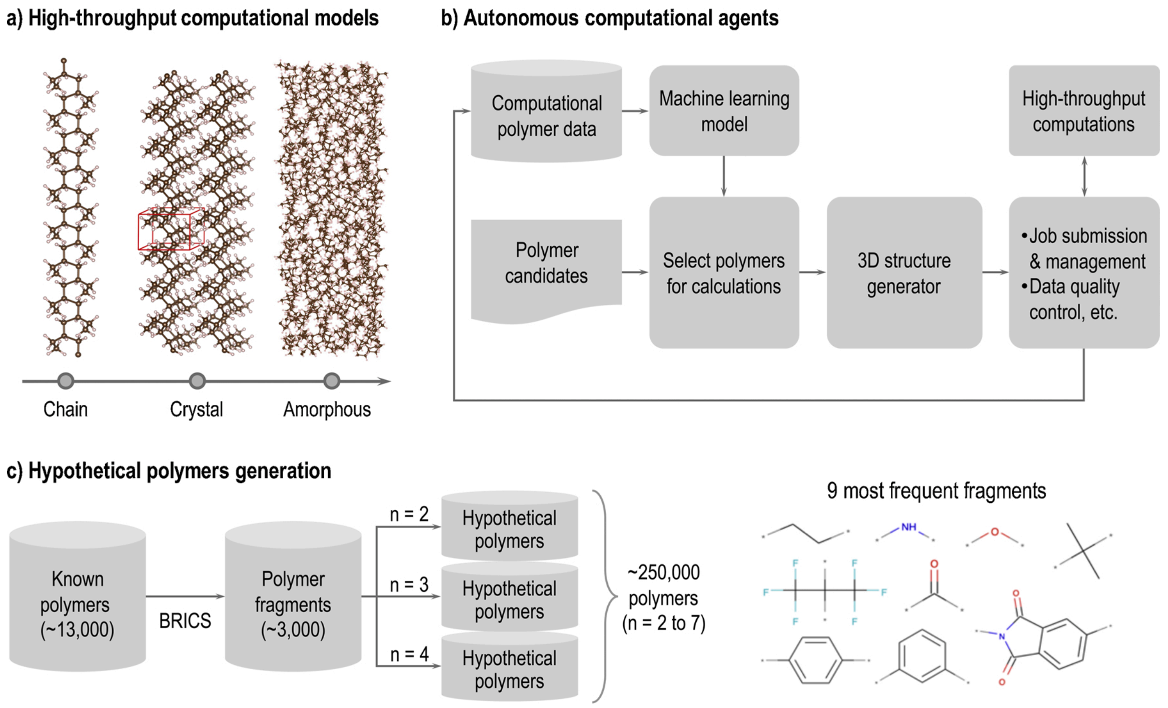

a) High-throughput computational models
Chain
Crystal
Amorphous
b) Autonomous computational agents
Computational polymer data
Machine learning model
High-throughput computations
Polymer candidates
Select polymers for calculations
3D structure generator
•Job submission & management
•Data quality control, etc.
c) Hypothetical polymers generation
Known polymers (~13,000)
BRICS
Polymer fragments (~3,000)
n = 2
n = 3
n = 4
Hypothetical polymers
Hypothetical polymers
Hypothetical polymers
~250,000 polymers (n = 2 to 7)
9 most frequent fragments
NH
O
F
N

Fig. 2. (a) Hierarchy of models to represent polymers, i.e., single-chain, pure crystal and amorphous phases. This figure is taken from Ref. [40] with permission from ACS Publications. (b) Autonomous computational agents to generate polymer data. (c) Design of hypothetical polymers using the BRICS scheme, along with some common polymer building blocks. '*' represents the possible linking position for each building block.

85], Young's modulus [82], tensile strength [34,80,81], and the lithium conductivity [33,36,37] of representative polymers using classical MD simulations. However, it is still challenging to compute these properties for a diverse range of polymers, especially those that have not been studied well.

Given the vast chemical space of polymers, a new strategy aided by an autonomous computational agent has been developed to dynamically select the next-candidate polymer with target properties [40]. As visualized in Fig. 2b), by utilizing the available (seed) computational dataset, machine-learning models are developed with the capability to instantly predict target properties for a large number of new polymers. Candidate polymers that meet the desired properties are selected, followed by a "3D structure" conversion step (involving generation of hierarchical models as shown in Fig. 2a)). The newly created structures are then modeled using high-throughput calculations, and the obtained results are added to the seed dataset iteratively. This autonomous platform is applicable for single or multiple polymer property predictions, and offers an efficient way to discover/design polymers with desired performances.

2.3. *Hypothetical polymers*

Data derived from experiments or the computations mentioned involve only known polymers. But how can we expand and explore beyond the *known* polymer chemical space? Variations of this question have already been tackled by different communities within chemical and materials sciences, such as drug discovery, inorganic solid state, metal-organic frameworks, 2D materials, etc., by exploiting various theoretical tools to construct databases of hypothetical molecules (e.g., ZINC) or materials (e.g., Materials Project). Further, computational tools, such as first-principles or classical methods, have been employed to estimate properties of these hypothetical cases, and screen promising candidates for future synthesis efforts. Some databases even estimate the synthesizability of a candidate using multiple models (e.g., based on free energy, synthetic accessibility scores, etc.) to ensure only realistic and plausible candidates are included. The successes of such molecule/materials databases are inspiring the creation of similar libraries for polymers.

In this regard, Batra et al. [44] devised an approach to explore the vast *unknown* chemical polymer space by generating new, but realistic, hypothetical polymers. As illustrated in Fig. 2c), they first obtained SMILES representation of ~12,000 polymers successfully synthesized in the past. Next, using the concepts of breaking of retrosynthetically interesting chemical substructures (BRICS) [86,87], polymer "building blocks" — each with two or more chain ends denoted by symbol `[*]`, e.g., `[*]c1ccc([*])cc1`, `[*]C(=O)[*]`, `[*]CC[*]` — were obtained along with their frequency of occurrence. Following this, hypothetical polymer SMILES strings were created by combining (at the [*] location) various numbers of building blocks, ranging from 2–7, resulting in a total of ~250,000 hypothetical polymers. This approach can be extended to create nearly infinite polymer candidates that may later be used to screen target properties. Care is taken to preserve the frequency of occurrence of different building blocks, making the constructed hypothetical SMILES dataset realistic and representative of the initial collected empirically known polymers. Moving forward, different chemical constraints or block neighbor restrictions can be introduced to limit the possible combination of building blocks, and generate more realistic/synthesizable polymers.

3. Polymer representation

Once the polymer structural, chemical, property and synthesis data are collected from the aforementioned resources, it should be processed/transformed to make it amenable to AI/ML based methods. Depending on the target polymer property or the input data type, different polymer representations may be chosen. Below, following a short discussion on the group contribution method, we discuss some of the more recent and successful polymer representation methods.

The group contribution technique developed by Van Krevelan and coworkers assumes that a polymer property can be estimated as a weighted sum of contributions arising out of its constituting fragments (referred to as quantitative structure-property relationships (QSPR) fingerprints) [88]. Using this and subsequent variations of this method, models describing a range of polymer properties have been developed with/without machine learning, including glass transition temperature, dielectric constant, refractive index, electrical conductivity, thermal conductivity, gas and aqueous diffusion, and intrinsic viscosity [88–91]. However, the developed models rely on the available fragment library and have little predictive capabilities for new polymers containing chemical fragments outside this pre-defined library.

To efficiently encode chemical information of molecules into machine-readable format, line notations have been designed to describe molecules using a line of text strings. Examples of such approaches include SMILES, the Wiswesser line notation (WLN), SYBYL Line Notation (SLN) and IUPAC International Chemical Identifier (InChi). SMILES is one of the most popular methods to represent molecules, because it is both human-readable and machine-friendly [41]. Further, various molecular fingerprinting algorithms have been developed to transform SMILES of small molecules into numerical vectors. Avalon, Daylight and Extended-Connectivity [93] fingerprints are examples of common fingerprinting algorithms that can be accessed through the open-source RDkit library [87]. Within these fingerprints, the presence or absence of substructures within a molecule is encoded into binary vectors, which can be used as inputs to data-driven models. SMILES representations of molecules can also be utilized (or transformed as graphs) in generative neural networks for fingerprint learning and molecular generations [46], but can also be directly used as input language in text-based machine learning algorithms [94,95]. The use of SMILES and similar line notations for molecules has transformed data-driven research in chem- and bio-informatics.

In contrast to small molecules, polymers are macromolecules composed of many repeat units, and require unique ways to capture their structural information. As illustrated in Fig. 3a), in modern data-driven models, SMILES of oligomers with several repeat units (< 5) have been applied to represent polymers, which can be fingerprinted using regular molecular fingerprinting algorithms [42,43,85]. The ML models developed using such oligomer fingerprints can predict various properties of polymers fairly well, although the effect of polymer morphology on the target property is excluded. In a similar development, modified SMILES representations for polymers have been developed which represent endpoints or connection points of repeat units using special symbols. For example, polyethylene is represented as [*]CC[*], where CC is the repeat unit of polyethylene and * represents the connecting points between repeat units [22,23]. Furthermore, a hierarchy of hand-crafted fingerprints for polymers have been developed to capture the connectivity and morphology information of polymers in order to improve the property prediction accuracy [22,96–99]. Fig. 3a) shows details of the hierarchical fingerprint, including the (1) atomic-level, (2) block-level, and (3) chain-level components. At the atomic-level, the number fraction of atomic-level fragments within the polymers, defined by the generic label "$A_iB_jC_k$", are considered. The block-level fingerprint components are the number fraction of pre-defined building blocks that constitute the polymers, such as C_6H_4, C=O, CH_2 and CO. Chain-level features capture information at the morphological scale, including the length of the longest or shortest side chains with or without rings. Further, QSPR fingerprints, such as the volume to surfaces ratio and van der Waals surface area, are also considered. Using this approach, models to predict many properties of polymers have been developed, including band gap, glass transition temperature and dielectric constant [22,23]. Detailed examples are provided in Section 4.

Additionally, BigSMILES has been recently developed for describing macromolecules, e.g., homo-and co-polymers [92]. It is an extension to SMILES, expressed as {RepUnit1, RepUnit2, RepUnit3, …}. Here, RepUnit1, RepUnit2, RepUnit3 are a list of (same or different) repeat units within polymers, with random positions. For example, Poly (ethylene-co-propylene) is denoted by {CC, CC(C)}, as shown in Fig. 3b). In this representation, branched, network and terminal group information of polymers may be also incorporated. However, there are no available fingerprinting algorithms to transform BigSMILES into numerical vectors yet.

Molecular structures can also be represented as a graph via an input of SMILES, where atoms and bonds are represented by nodes and edges, respectively. Such a method has been widely utilized for molecular structure generation and property prediction in cheminformatics, bioinformatics and materials science with great success [100–104]. Since it is challenging to use graphs to represent polymers, due to its large-scale morphology, researchers have attempted to use oligomers (including less than 5 repeat units) to label polymer graphs with atom-based or substructure/motif based-methods [45,46,105]. Motifs refer to larger size substructures. However, because monomers of many polymers are large and complicated, this leads to monomer generation failures using small substructures. Further, polymers are composed of large numbers of monomers, and it is not clear how one can incorporate large-scale morphological information as graphs. These critical issues are discussed in Section 7.

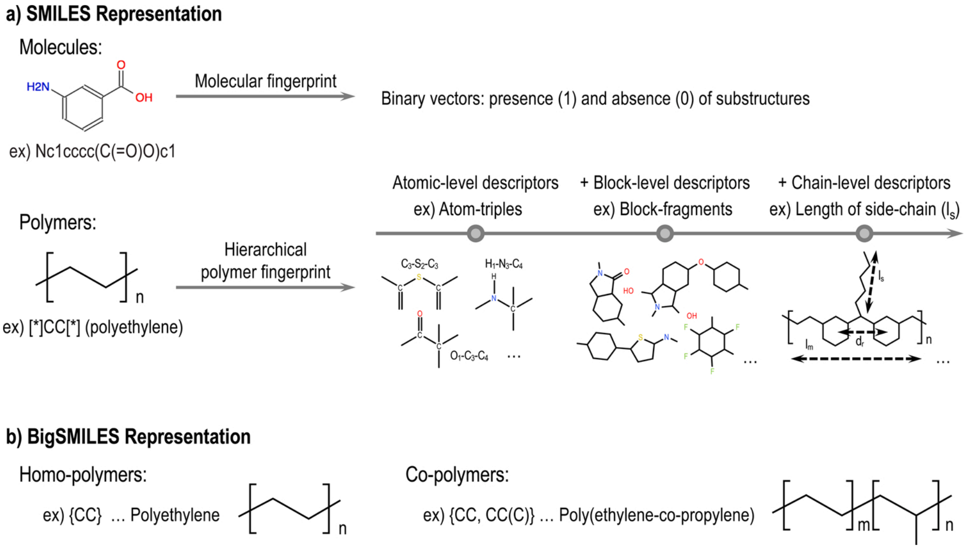


Fig. 3. Polymer representations: SMILES and BigSMILES [92]. Using the input of SMILES, molecular [42] and hierarchical polymers fingerprints [22,23] were developed to numerically represent polymers.

4. Property prediction schemes

The selection of suitable learning algorithms to map polymer fingerprints and properties is a critical step. Depending on the complexity of the target property, the volume and the nature of the available datasets, various learning algorithms have been applied, such as linear or non-linear regression algorithms, multi-fidelity information fusion and deep neural networks.

4.1. Linear/Non-linear regression

The linear regression algorithm assumes that the property being modeled is a linear function of the fingerprints, which is the simplest method to build machine learning models. For polymers, various property prediction models have been developed using group contribution methods [88], multiple linear regression [89,90], and least-squares regression [21,107,108], with QSPR or quantum-chemical fingerprints. These models are limited by the neglect of the non-linear relationships between polymer fingerprints and their properties. To overcome this issue, non-linear regression algorithms have been employed, such as support vector machine (SVM), kernel ridge regression (KRR) and Gaussian process regression (GPR). For instance, Yu et al. used SVM to train glass transition prediction models using QSPR fingerprints and experimental property values [89,109], while KRR has been applied to develop models for a series of high-throughput computed polymer properties (e.g., band gap and dielectric constant) [110].

In recent years, the GPR algorithm has been broadly utilized to build machine learning models for polymer property prediction [22,23,42,52,111]. As illustrated in Fig. 4a), the key advantage of GPR is that predicted uncertainties are provided by learning a generative and probabilistic distribution with the mean representing the prediction and the confidence interval estimating the uncertainty. Fig. 5a) illustrates four representative GPR models, including chain band gap, glass transition temperature, frequency-dependent dielectric constant and gas permeability. These models were trained using 3881 DFT computed, and 5076, 1193 and 1779 experimental values, respectively. 5-fold cross-validation (CV) was employed to avoid model overfitting. R^2 and RMSE denote the coefficient of determination and the root-mean-square error, respectively. $RMSE_{100\%}$ and $RMSE_{CV,test}$ respectively denote the RMSE errors on the whole dataset used for model training or on the test subset during 5 fold-CV. In the case of the gas permeability model, 6 gases, i.e., O_2, N_2, CH_4, He, CO_2 and H_2, were considered and numerically represented using one-hot encoding. Likewise, in the dielectric constant model the frequency value (at 9 different frequencies) was used as a feature to obtain a frequency-dependent dielectric constant model. We note that the developed GPR models exhibit very high R^2 and acceptable $RMSE_{CV,test}$ with respect to the wide property range of the training datasets. Additionally, the performance of these models has been further validated by using systematic analysis, involving the effect of feature reduction, various levels of train-test splits (i.e., learning curves) and validation on completely unseen datasets.

The GPR algorithm can be used to build accurate and reliable ML models for a single property while also providing prediction uncertainties. However, the GPR method has two issues: (1) it requires a manageable dataset size. Large datasets (>5000) become prohibitively expensive to train. (2) It does not have the capability to train multiple properties in one single model. Therefore, more advanced algorithms have been utilized to improve these issues, such as multi-fidelity information fusion and deep learning methods as discussed below.

4.2. Multi-fidelity information fusion approaches

It is quite common to encounter problems where several datasets have varying levels of accuracy, data generation cost and noise levels are present. Typically, the most precise experimental measurements (or computations) also tend to be the most time and resource intensive, the so-called high-fidelity (HF) data. However, polymer properties of interest can also be estimated via cheaper methods at lower accuracy. For instance, empirical trends, simple group-contribution methods and computationally demanding quantum mechanical simulations can generate this low-fidelity (LF) data. Given such a situation, a multi-fidelity (MF) information fusion model aims to consolidate all the available information from the varying fidelity sources to make the most accurate and confident property predictions at the highest level of fidelity [47,48,112–116]. Comparative studies have shown that the multi-fidelity approach performs better than any single-fidelity based method in terms of prediction accuracy, especially for small (high--fidelity) data sets. Typical strategies for MF learning are discussed in Ref. [47]. Among them, the Gaussian processes-based co-kriging regression method [117] is viewed as a powerful method and has been utilized to predict polymer properties, such as band gap and degree of crystallinity [48,118]. As shown in Fig. 4b), this MF approach is composed of two models: the Gaussian processes $Z_{LF}(\boldsymbol{x})$ of the low-fidelity function and the Gaussian processes $Z_d(\boldsymbol{x})$ related to the difference between the low-fidelity and the high-fidelity functions. The property prediction at the high-fidelity level ($Z_{HF}(\boldsymbol{x})$) is $Z_{HF}(\boldsymbol{x}) = \rho Z_{LF}(\boldsymbol{x}) + Z_d(\boldsymbol{x})$. Here, ρ is a scaling factor that quantifies the correlation between the two fidelities of data.

Fig. 5b) shows two successful examples of MF approaches being applied to polymer property predictions [48,106], i.e., the tendency to crystallize and the band gap. For the former, a MF model was trained

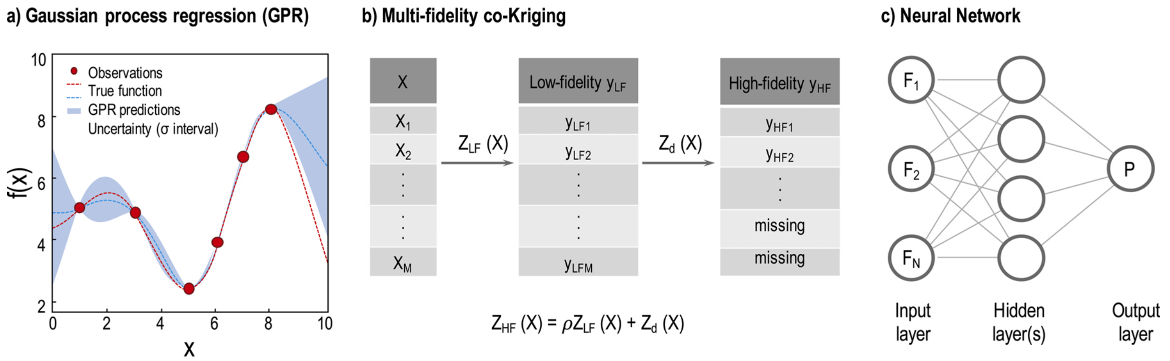


Fig. 4. (a) Gaussian process regression (GPR) model to learn the correlation between fingerprints and target property, providing predicted values and uncertainty (shaded regions). (b) Multi-fidelity (MF) co-kriging approach depends on two models: the Gaussian processes $Z_{LF}(\boldsymbol{x})$ of the low-fidelity (LF) function mapping the fingerprint space and low-fidelity property (y_{LF}) and the Gaussian processes $Z_d(\boldsymbol{x})$ to map the fingerprint space and difference between the low-fidelity and the high-fidelity (HF) functions. The property prediction at the high-fidelity level ($Z_{HF}(\boldsymbol{x})$) is $Z_{HF}(\boldsymbol{x}) = \rho Z_{LF}(\boldsymbol{x}) + Z_d(\boldsymbol{x})$, where ρ is a scaling factor that quantifies the correlation between the two fidelities of data. (c) General Neural network workflow, including input, hidden and output layers.

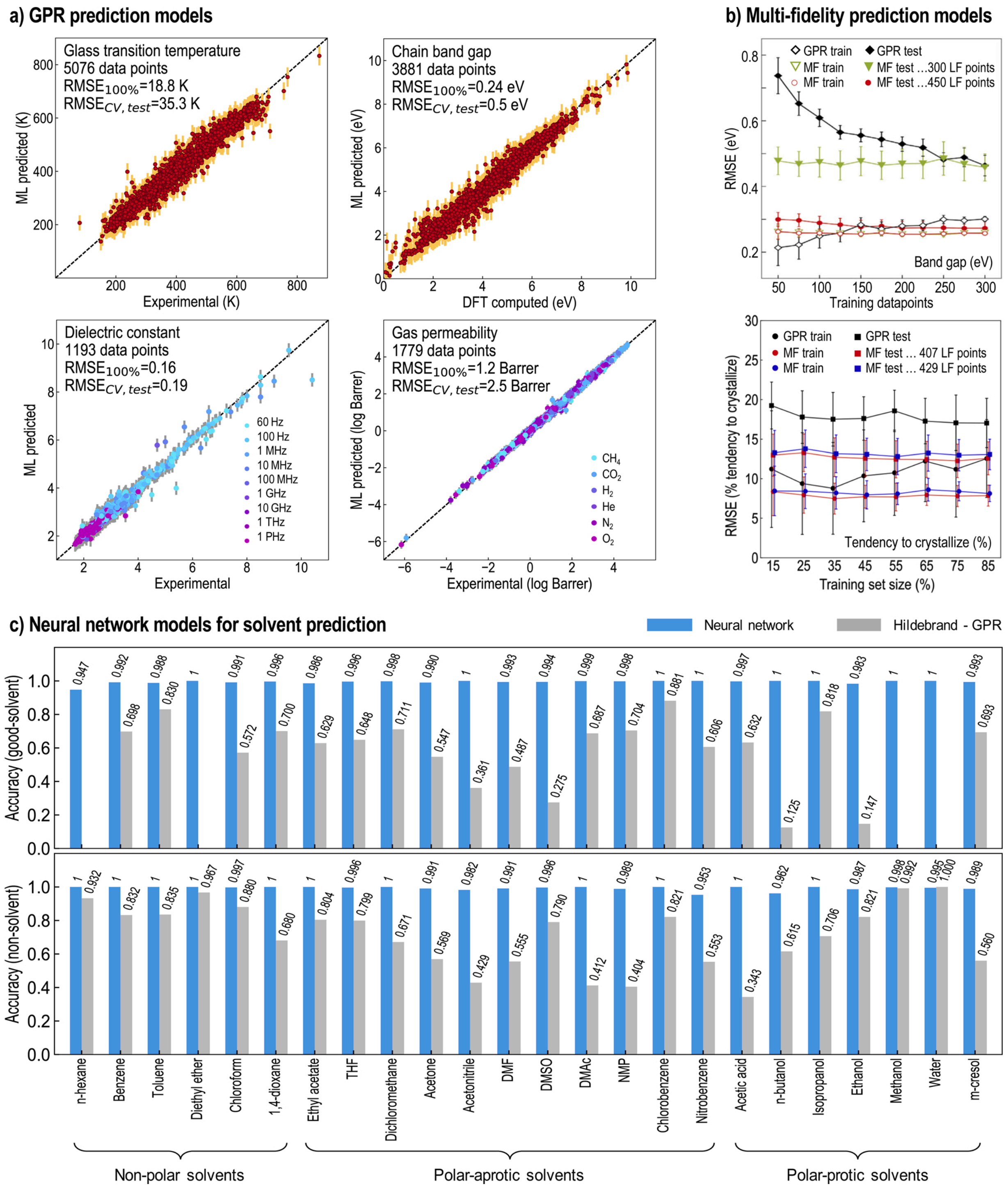


Fig. 5. (a) Parity plots of GPR predicted and true values of the glass transition temperature, band gap of sing-chain polymers, dielectric constant and gas permeability. In the case of the gas permeability model, 6 gases, i.e., O_2, N_2, CH_4, He, CO_2 and H_2, were considered and dielectric constant at 9 different frequencies was used in the dielectric constant model. CV-test RMSE is the average RMSE of the test subsets in 5 fold-CV [23]. Error bars represent GPR uncertainty. (b) A comparison of learning-curves for the GPR and multi-fidelity (MF) predicted band gap [48] and tendency to crystallize [106]. (c) Neural network-based solvent prediction accuracy of soluble (top) and insoluble (bottom) polymers for 24 solvents, including non-polar, polar-aprotic and polar-protic solvents, along with results from GPR models trained by Hildebrand parameters [49]. Figure (a), (b) and (c) are taken from Ref. [23], Ref. [48,106] and Ref. [49], respectively, with permission from AIP, ELSEVIER and ACS Publications.

using 107 high-fidelity data directly measured by experiments and 429 low-fidelity data estimated using a combination of experimental and group-contribution methods. In the latter, 382 hybrid and PBE computed band gap values were used as high- and low-fidelity data in the MF model. Fig. 5b) compares the learning performance of the MF models against single-fidelity GPR models trained on the respective high-fidelity polymer property data, i.e., the RMSE on the training and the test set as a function of the training size of the high-fidelity data. We note that MF models surpass the GPR model accuracy (trained on the high-fidelity data alone) at a much smaller fraction of the high-fidelity training data. This is mainly because the relatively large volume of the low-fidelity data, although somewhat inaccurate, allows the MF model to learn polymer property trends. These findings indicate that there are benefits to employing the MF approach, especially in situations wherein resource demanding high-fidelity experimental data can be combined with a large number of low-fidelity and inexpensive computational data.

While the first proof-of-principle examples are just beginning to appear, MF models could have a considerable impact in the field of polymer informatics. It is worth pointing out that the accuracy of MF models depends on the ability of the shared subset of high- and low-fidelity data to learn the latent space of the two fidelities. Further, there is a necessity to improve upon the MF scheme. For instance, several levels of fidelity hierarchies can be present simultaneously in the polymer property datasets. The number of the co-kriging model parameters can significantly increase in such scenarios, leading to expensive computational cost. Consequently, advanced MF learning algorithms should be developed to speed up the learning process, particularly when several levels of fidelities are present in the polymer data sets.

4.3. Deep neural networks

Conventional machine learning techniques described above provide good property prediction accuracy. However, these methods are computationally efficient for systems with small dataset size only. Given the surge in the available computational/experimental data in materials science, deep neural networks (NN) are being increasingly utilized in polymer informatics. Fig. 4c) presents the general architecture of NNs, in which molecular or polymer fingerprints form the input layer. The following hidden layers are constructed with a specific activation function, e.g., the parametrized rectified linear unit (PReLU). The final output layer of the NN consists of neuron(s) for target properties, also with a specific activation function depending on the problem at hand (classification or regression). The details of various types of NNs and their uses in materials science are well-reviewed in Refs. [119–121]. Below we discuss the initial attempts to apply NNs for polymer properties prediction [43,49,122–125].

The selection of suitable polymer-solvent pairs is a critical step for polymer synthesis. Chandrasekaran et al. have developed a deep neural network model for solvent prediction [49]. In this work, 4595 polymers and 24 solvents, forming a total of 11,958 polymer + good-solvent pairs and 8469 polymer + non-solvent pairs, were used to train a binary classification NN model (i.e., given a polymer-solvent pair the model predicts if it a good-solvent or non-solvent for that polymer). A multilayer perceptron with special architecture was used: the first part of the NN composed of two input branches, one for the polymer descriptors generated using hierarchical fragment-based fingerprint described in Section 3 and the other for the solvent descriptors represented by one-hot encoding. In the second part, polymer and solvent latent space were concatenated into a single merged latent vector. Fig. 5c) shows the neural network prediction accuracy of soluble (top) and insoluble (bottom) polymers for 24 solvents, including non-polar, polar-aprotic and polar-protic solvents. Performance results for the GPR models trained using Hildebrand parameters of about 100 polymers [118] are also compared. In general, the performance of the NN model greatly outperforms that of the GPR model, mainly due to the higher level of diversity in the training data. Further, the Hildebrand parameter is only an approximate empirical approach to distinguish good-solvent against non-solvents, based on the notion of "like dissolves like". This deep neural network-based framework provides a more general, accurate, and efficient way to predict good-solvents vs. non-solvents for new polymers.

Additionally, researchers have applied NNs to build prediction models for glass transition temperature [122–124], polymer permeability to gases [91] and thermal conductivity (κ) of polymers [43,85,125]. For the glass transition temperature and polymer permeability, small datasets ($<= 150$ polymers) were applied to train NNs, raising concerns on the generality of obtained models for new cases. In the case of κ, Zhu et al. have used classical MD computed κ values of single-chain polymers and molecular fingerprints to train the NNs models. One potential concern is that the computational uncertainty, arising from the model difference between the adopted single chain and realistic polymers, may introduce additional noise in the ML models [85]. An alternative approach is to directly use experimental κ to train the model, although, only sparse data is available. To overcome this issue, Wu et al. [43] has utilized the transfer learning approach to learn the κ of polymers (58 data points), via training other properties of polymers with large data size (e.g., melting temperature, glass transition temperature and heat capacities). The performance of the obtained model is better than those trained only on the thermal conductivity data, because the shared features between κ and other properties and large training dataset are considered in the transfer learning algorithm. However, there are still challenges, as discussed in Section 7.

4.4. Polymer genome online platform

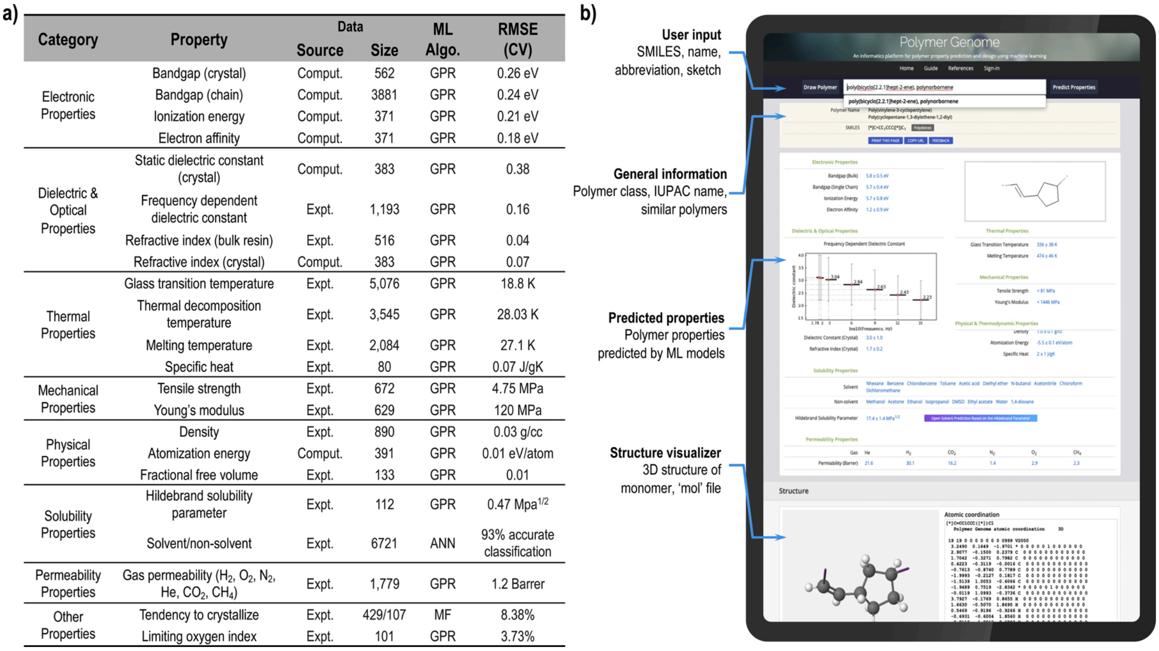
Significant efforts are also being made to provide easy access to the aforementioned polymer prediction models. In this regard, the Polymer Genome online platform (www.polymergenome.org) has been developed to instantly provide property predictions for polymers. As summarized in Fig. 6a), various polymer property prediction models have been implemented, including electronic, dielectric, thermal, mechanical and other important properties. The source and size of the training data, the applied algorithms and the expected errors ($RMSE_{CV}$) for each of the property models are also provided. Fig. 6b) shows a typical output of Polymer Genome, taking the example of Polynorbornene. The SMILES ([*]C=CC1CCC([*]C1)), which forms the repeat unit of Polynorbornene, is provided as the input to Polymer Genome, where [*] denotes the connection points of the repeat units. ML predicted properties of this polymer are shown in a tabular format. The 3D visualization of the structure with atomic coordinates is also provided at the bottom of the page. More detailed information is available in Ref. [22,23].

5. Polymer design algorithms

Once the polymer surrogate models are trained, they can be utilized to accelerate the polymer discovery process. Below we outline two distinct strategies for this. While the first relies on screening candidates that meet target property requirements based on predictions for a predetermined candidate pool, the other utilizes genetic and generative models to directly produce desirable candidates.

5.1. Enumeration

One of the dominant applications of machine learning techniques is to significantly accelerate the rational design and discovery of new materials by efficiently searching a pre-determined chemical space. The previously discussed ML models are used to predict the properties of a large pool of candidate polymers enumerated based on some physically or chemically motivated scheme, followed by a down-selection procedure based on certain screening criteria. The end result is a rank-ordered list of promising candidates for the target application. The initially enumerated candidates may be previously synthesized polymers, or hypothetical polymers made by human experts or machine (e.g., genetic algorithm). Fig. 7a) shows the trends, in a form similar to "Ashby plots", of various ML predicted properties (such as glass transition temperature, band gap, dielectric constant (at THz) and density) for ten-of-thousands of known/synthesizable polymers. These synthesized polymers have been manually accumulated from various resources, as discussed in Section 2. Depending on the property requirements for specific applications, different combinations of properties can be selected. For instance, polymers that are tolerant to extreme temperatures require large band gap, high glass transition temperature and dielectric constant, whereas polymers electrolytes used in Li-ion



Category	Property	Data Source	Data Size	ML Algo.	RMSE (CV)
Electronic Properties	Bandgap (crystal)	Comput.	562	GPR	0.26 eV
	Bandgap (chain)	Comput.	3881	GPR	0.24 eV
	Ionization energy	Comput.	371	GPR	0.21 eV
	Electron affinity	Comput.	371	GPR	0.18 eV
Dielectric & Optical Properties	Static dielectric constant (crystal)	Comput.	383	GPR	0.38
	Frequency dependent dielectric constant	Expt.	1,193	GPR	0.16
	Refractive index (bulk resin)	Expt.	516	GPR	0.04
	Refractive index (crystal)	Comput.	383	GPR	0.07
Thermal Properties	Glass transition temperature	Expt.	5,076	GPR	18.8 K
	Thermal decomposition temperature	Expt.	3,545	GPR	28.03 K
	Melting temperature	Expt.	2,084	GPR	27.1 K
	Specific heat	Expt.	80	GPR	0.07 J/gK
Mechanical Properties	Tensile strength	Expt.	672	GPR	4.75 MPa
	Young's modulus	Expt.	629	GPR	120 MPa
Physical Properties	Density	Expt.	890	GPR	0.03 g/cc
	Atomization energy	Comput.	391	GPR	0.01 eV/atom
	Fractional free volume	Expt.	133	GPR	0.01
Solubility Properties	Hildebrand solubility parameter	Expt.	112	GPR	0.47 Mpa$^{1/2}$
	Solvent/non-solvent	Expt.	6721	ANN	93% accurate classification
Permeability Properties	Gas permeability (H_2, O_2, N_2, He, CO_2, CH_4)	Expt.	1,779	GPR	1.2 Barrer
Other Properties	Tendency to crystallize	Expt.	429/107	MF	8.38%
	Limiting oxygen index	Expt.	101	GPR	3.73%

Polymer Genome
An informatics platform for polymer property prediction and design using machine learning
Home Guide References Sign-in
Draw Polymer | poly(bicyclo[2.2.1]hept-2-ene), polynorbornene | Predict Properties
poly(bicyclo[2.2.1]hept-2-ene), polynorbornene

Polymer Name: Poly(vinylene-3-cyclopentylene)
Poly(cyclopentane-1,3-diylethene-1,2-diyl)
SMILES: [*]C=CC$_1$CCC([*])C$_1$ Polydienes
PRINT THIS PAGE | COPY URL | FEEDBACK

Electronic Properties
Bandgap (Bulk): 5.8 ± 0.5 eV
Bandgap (Single Chain): 5.7 ± 0.4 eV
Ionization Energy: 5.7 ± 0.8 eV
Electron Affinity: 1.2 ± 0.9 eV

Dielectric & Optical Properties
Frequency Dependent Dielectric Constant
Dielectric Constant (Crystal): 3.0 ± 1.0
Refractive Index (Crystal): 1.7 ± 0.2

Thermal Properties
Glass Transition Temperature: 336 ± 38 K
Melting Temperature: 474 ± 46 K

Mechanical Properties
Tensile Strength: < 81 MPa
Young's Modulus: < 1446 MPa

Physical & Thermodynamic Properties
Density: 1.0 ± 0.1 g/cc
Atomization Energy: -5.5 ± 0.1 eV/atom
Specific Heat: 2 ± 1 J/gK

Solubility Properties
Solvent: Nhexane Benzene Chlorobenzene Toluene Acetic acid Diethyl ether N-butanol Acetonitrile Chloroform Dichloromethane
Non-solvent: Methanol Acetone Ethanol Isopropanol DMSO Ethyl acetate Water 1,4-dioxane
Hildebrand Solubility Parameter: 17.4 ± 1.4 MPa$^{1/2}$ Open Solvent Prediction Based on the Hildebrand Parameter

Permeability Properties

Gas	He	H_2	CO_2	N_2	O_2	CH_4
Permiability (Barrer)	21.6	30.1	16.2	1.4	2.9	2.3

Structure

Atomic coordination

```
[*]C=CC1CCC([*])C1
 Polymer Genome atomic coordination     3D

19 19 0 0 0 0 0 0 0999 V2000
 3.2490  0.1649 -1.9701 * 0 0 0 0 0 1 0 0 0 0 0 0
 2.9077 -0.1500  0.2379 C  0 0 0 0 0 0 0 0 0 0 0 0
 1.7042 -0.3271  0.7982 C  0 0 0 0 0 0 0 0 0 0 0 0
 0.4223 -0.3119 -0.0016 C  0 0 0 0 0 0 0 0 0 0 0 0
-0.7613 -0.8740  0.7789 C  0 0 0 0 0 0 0 0 0 0 0 0
-1.9993 -0.2127  0.1817 C  0 0 0 0 0 0 0 0 0 0 0 0
-1.5139  1.0053 -0.6066 C  0 0 0 0 0 0 0 0 0 0 0 0
-1.9489  0.7519 -2.8342 * 0 0 0 0 0 1 0 0 0 0 0 0
-0.0119  1.0993 -0.3736 C  0 0 0 0 0 0 0 0 0 0 0 0
 3.7927 -0.1769  0.8655 H  0 0 0 0 0 0 0 0 0 0 0 0
 1.6630 -0.5070  1.8690 H  0 0 0 0 0 0 0 0 0 0 0 0
 0.5469 -0.9196 -0.9266 H  0 0 0 0 0 0 0 0 0 0 0 0
```

Fig. 6. (a) Summary of various property prediction models implemented in the Polymer Genome online platform (www.polymergenome.org). RMSE(CV) denotes the average RMSE errors on the test subset during the 5 fold-CV. (b) Overview of the Polymer Genome platform. Polynorbornene is used as an example of user input to show the obtained ML predicted properties and 3D structure visualization. Figure (a) is taken from Ref. [23] with permission from AIP Publications.

batteries require desired electron affinity, band gap, and ionization energy. Polymer membranes, on the other hand, require suitable gas permeability-selectivity pairs.

Following this forward design pipeline, Mannodi Kanakkithodi and co-workers identified promising polymer dielectrics with desired dielectric constant and band gap from a series of human-designed hypothetical polymers, made up of 4, 6 and 8 building blocks (e.g., $–CH_2–$, –CO–) [51]. Likewise, Chen et al. proposed five representative polymer candidates satisfying high glass transition temperature and required dielectric constant for high temperature, energy density capacitors and microelectronic devices from a pool of synthesized polymers [52]. Additionally, Wu et al., on the other hand, used a surrogate thermal conductivity model based on transfer learning to screen promising candidates with target glass transition and melting temperatures, resulting in the synthesis of polymers with thermal conductivities of 0.18–0.41 W/mK [43]. Another successful example is from Kumar [42], wherein two polymer membranes with excellent CO_2/CH_4 separation performance were discovered from over 11,000 homopolymers, guided by GPR based gas permeability prediction models. These findings strongly advocate the success of machine learning assisted forward design approach to discover polymer candidates for specific applications.

5.2. Sequential (Active) learning

The polymer design algorithms discussed above provide a subset of promising polymer candidates with tailored properties for further validation via experimental synthesis or high-fidelity computations. However, these models are passive, with inherent errors in the property predictions owing to the limitations, such as bias or limited size of the training data. Thus, how one can dynamically and efficiently optimize polymers for the next experiment (or computation) is an important problem in materials discovery. It is far from trivial to select optimal candidates based purely on human intuition. In recent years, active-learning algorithms that exploit Bayesian optimization frameworks have been developed to effectively guide experiments or high-throughput computations for materials design, e.g., optimizing GaN LED structures, $BaTiO_3$ based piezoelectrics, and other inorganic materials for thermoelectric and electronic devices [127].

As illustrated in Fig. 7b), active learning algorithms consist of three important components: (1) a surrogate model for the target property prediction; (2) an acquisition function to select the optimal point for the next experiment; (3) addition of the newly performed experiment to the knowledge dataset [126,127]. The surrogate models in part (1) can be trained using various algorithms introduced in Section 4. To provide both prediction and uncertainty values of the target property, Gaussian process-based algorithms are common approaches used in active learning. There are other methods, such as support vector regression or decision trees in combination with bootstrapping methods, that estimate both the target property and its uncertainty. In part (2), the user can search unlabeled data by either using the prediction uncertainties (called exploration), or by maximizing the target prediction values (called exploitation), or by balancing between exploration and exploitation. In the last part, the newly generated data from the new experiment is supplemented into the knowledge dataset to retrain the surrogate model. The whole pipeline is repeated until the target candidate is achieved.

In the polymer domain, Kim et al. benchmarked the use of active learning to efficiently search polymers with $T_g > 450$ K [126]. Fig. 7b) illustrates the average number of experiments required to discover 1–10 polymers with a glass transition temperature of above 450 K, starting from an initial training dataset of 5 polymers. The error bars denote the standard deviation across the 50 different runs. We note that on average 30, 46, 98 and 234 experiments were required to discover 10 high-glass transition temperature polymers using acquisition function definitions based on balanced exploitation and exploration, exploitation, exploration and random approaches, respectively. These findings indicate that the balanced exploitation and exploration method showed the best performance in terms of discovering promising polymer candidates. Additionally, Huan et al. have applied the active learning to automatically select polymer candidates for high-throughput DFT computations and find polymer dielectrics with a large band gap. More details are shown in Section 5. It is evident that the integration of active-learning within the materials discovery pipeline can guide materials design and dataset expansion in an efficient and targeted fashion.

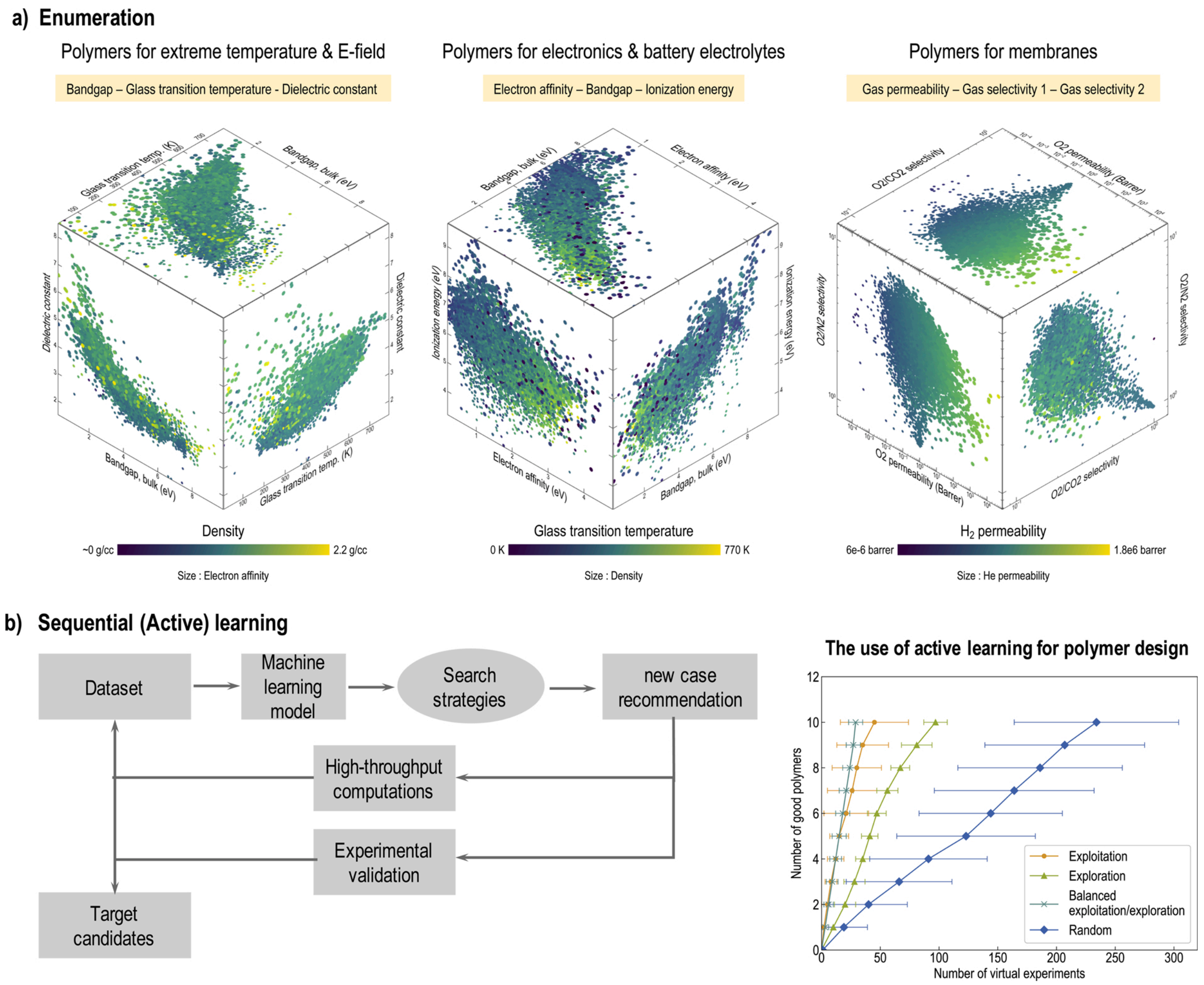


Fig. 7. (a) ML predicted properties of ten-of-thousands of enumerated known/synthesizable polymers. Different combinations of properties can be selected to screen polymers for the specific application, for example, large band gap, high glass transition temperature and dielectric constant for polymers tolerant to extreme temperature and electric field. (b) Sequential (active) learning workflow (left) and its use for polymer design (right): number of experiments required (on average) to discover 1–10 polymers with glass transition temperature greater than 450 K when starting with an initial dataset size of five polymers. The average is calculated using 50 different runs and the standard deviation is denoted by the error bar. This figure is taken from Ref. [126] with permission from Cambridge University Press.

5.3. *Evolutionary strategies*

Another approach to polymer discovery is the “inverting the prediction pipeline”. Contrary to the enumeration approach that relies on virtual screening of polymers from a pre-defined candidate set using instantaneous property prediction from surrogate models, inversion problems focus on directly generating polymers that satisfy given property objectives, making it a more general approach to materials discovery. Two approaches for direct materials design have emerged: first, the use of evolutionary methods, such as the genetic algorithm (GA) [128,130], and second, generative ML approaches, such as variational autoencoders (VAE) [131] and generative adversarial networks (GAN) [132]. We describe the evolutionary approaches here, while generative methods are discussed in Section 5.4.

GA is based on the principle of natural selection. The inherent structure of a polymer makes its treatment using GA straightforward-a polymer can be thought of as a sequence of chemical building blocks (e.g., CH_2, C_6H_6, or blocks B_{12}, B_{13} in Fig. 8a) connected to each other by covalent bonds (analogous to DNA base pairs), and the properties of a polymer are functions of the sequence of constituent chemical building blocks (analogous to how oculocutaneous albinism II (OCA2) gene sequence mostly dictates human eye color). Within the GA approach a series of *crossover*, *mutation* and *selection* operations are applied to discover new candidate polymers with desired properties. It starts with a random generation of (say, 100) polymers, whose chemical building blocks are modified using crossover-pruning and mixing of monomer building blocks-and mutation-random alterations to monomer building blocks-operations to obtain a large set of offspring polymers, as illustrated in Fig. 8a). Next, the top offspring candidates with desired properties are selected based on their user-defined objective score to form the next generation of polymers. This GA cycle is repeated until a sufficient number of candidates with desired properties are obtained, as in Fig. 8b). Besides polymer discovery, GA has also been utilized to solve other problems in materials science, such as developing functional forms of interatomic potentials [133], and discovering hidden material property relations [134].

A critical component of the GA design scheme is the evaluation of the objective function during the selection stage. This step has been a major bottleneck for polymer discovery since property estimation through experiments or computations is very expensive and time-consuming [135]. However, with the recent development of cheap and reliable polymer property models (Section 5.1), the objective function can now be computed in a fraction of a second. This allows the GA process to truly explore a very rich chemical polymer space, going well-beyond the pre-defined candidate sets. Furthermore, by setting-up a property weighted objective function, polymers that simultaneously satisfy

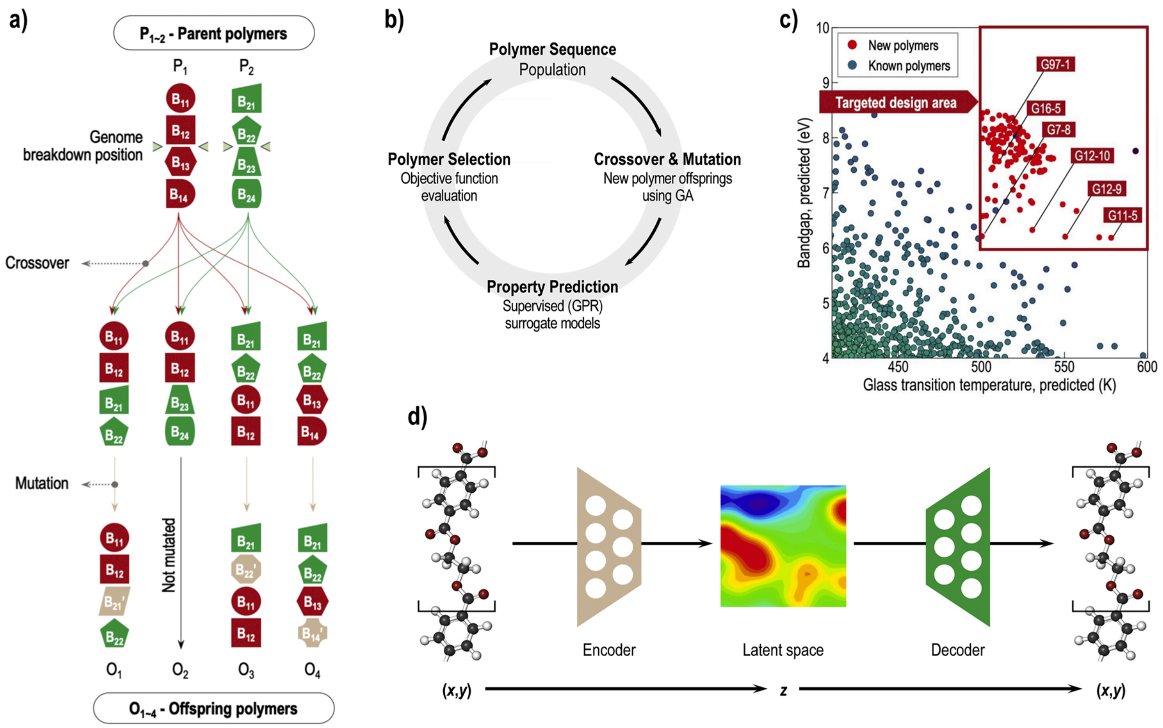


Fig. 8. Polymer inverse design using machine learning. Use of evolutionary algorithms, such as GA, for design of polymers with target properties; (a) basic operations of crossover and mutation to generate polymer offsprings, taken from Ref. [128] with permission from ELSEVIER Publications; (b) the different stages of the iterative evolutionary process, involving population of new candidates, evaluation of fitness function using property-prediction models, and selection of the next generation with best fitness, taken from Ref. [129] with permission from ACS Publications; (c) results for an exemplary polymer design problem of high glass transition temperature and large band gap, taken from Ref. [128] with permission from ELSEVIER Publications; (d) Use of variational autoencoders (VAE) for polymer design. The latent space is searched to find polymers with desired properties, which are 'generated' using the decoder mapping.

multiple property criteria can be targeted.

Kim and co-workers used GA to design polymers with high glass transition temperature and large band gap, which are useful for high-energy capacitors because of their stability at both high temperatures and electric fields [128]. Notably, this is a difficult design problem, with only 4 out of thousands of known polymers displaying glass transition temperature $>$ 500 K and band gap $>$ 6 eV. Two interesting aspects of their design process was the choice of the chemical building blocks, and the underlying property prediction models. The former consisted of a comprehensive list of 3045 chemical blocks, extracted from ~12,000 synthetically known polymers using the concept of BRICS—similar to the hypothetical polymer design work in Section 2. Each block had 1–4 connection points that were used to form/break bonds with other chemical blocks during the crossover and mutation operations. For the latter glass transition temperature and band gap prediction models, they used two independent GPR surrogate models based on a hierarchical polymer fingerprinting scheme (Section 3) that were trained on an experimental and DFT computed dataset of 5072 and 562 polymers, respectively. During 100 generations of the GA cycle, they successfully designed 132 new polymers that meet the target properties, as opposed to only 4 previously known cases (see Fig. 8c)). Furthermore, their analysis of the identified polymer candidates revealed insights about the key fragments that promote high glass transition temperature and large band gap in polymers, such as the presence of terminal difluorocarbon or trifluoromethyl, saturated 5-or 6-membered rings, oxolane, etc. These findings are compatible with known chemistry. For instance, fluorine atoms induce large band gap through the formation of lower (higher) C–F sigma bonding (anti-bonding) orbitals. A similar approach has been utilized to design polymers with high dielectric constant, although it considered a relatively small number of possible chemical building blocks [51].

In a different study, Pilania et al. used GA to design bio-advantaged (biosynthesizable and biodegradable) Polyhydroxyalkanoate (PHA)-based polymers with desired glass transition temperature values [129]. A machine learning model trained on an experimentally-measured and carefully-curated glass transition temperature values for a wide range of PHA homo- and co-polymers were combined with a GA-based search and optimization routine to explore a much wider chemical space formed by multi-component polymer chemistries, beyond co-polymers. Furthermore, by explicitly integrating the prediction uncertainties and number of polymer components in the GA objective function, they were able to focus their search on polymers containing desired number of components (ternary, quaternary, etc.) where the confidence level in the machine-learned glass transition temperature predictions were higher than a pre-specified cutoff value.

5.4. Generative models

Based on the concept of unsupervised learning, VAE and GAN offer a different route for targeting inverse polymer design. They learn a mapping from a continuous latent space to the polymer space, using which new candidates with desired properties are generated after solving the optimization problem in the more amenable latent space. For example, in the case of VAE, the encoder unit learns to represent a polymer in a high-dimensional (say, 100–200) continuous (latent) space, while the decoder unit learns to map back a vector in the latent space to a valid polymer, as shown in Fig. 8d). Both mappings are important from a materials design perspective: the encoder provides a fingerprinting scheme that can be exploited by different "forward models", while the decoder provides a pathway to systematically search polymers in a proxy latent space using different optimization schemes, and later generate the desired polymer candidate associated with the optimal latent point. Although the VAE and GAN approaches have received attention for molecule or drug discovery [136–140], they are

Table 2
Desired properties of polymer candidates for various applications

Applications	Representative desired polymer properties
Capacitors (polymer dielectrics)	Large band gap, high charge injection barriers, high glass transition temperature
Li-ion batteries (polymer electrolytes)	Large electrochemical stability window, high ionic conductivity, high Li-ion transference and mechanical strength
Polymer membrane	High permeability and selectivity for gas pairs
Electronic devices (conducting polymers)	High electrical conductivity

only beginning to be exploited for extended systems such as polymers.

A key challenge in developing such a generative model is that the decoder unit should map a continuous latent space to a discrete and structured material space, which should represent a valid candidate material as dictated by chemistry. For example, in case of polymers, the decoder output should necessarily have two chain ends, or the involved C and O atoms should display a valency of 4 and 2, respectively. This goal of enforcing the decoder to output valid polymers was recently achieved by Batra et al. [44] using syntax-directed VAE. The strategy involves representing the polymers using their SMILES representation, and then imposing the decoder to obey both the syntactic and semantic constraints associated with the class of polymers; syntactic refers to the grammatical rules inherent to the SMILES language, while semantic refers to the contextual constraints driven by polymer chemistry. The inclusion of explicit syntax and semantics in the VAE model improves the quality of the learned latent space. It also leads to a high occurrence of valid polymer SMILES upon decoding, making the process of discovery efficient.

Batra et al. coupled the unsupervised syntax-directed VAE with the supervised GPR method to discover polymers with high glass transition temperature and large band gap [44]. They used the encoder unit to fingerprint the polymers, which were then mapped to the respective glass transition temperature and band gap values using GPR. To train the VAE model they had to overcome a crucial data sparsity challenge: to-date the total number of chemically diverse polymers synthesized is $\sim$12,000, while a VAE model usually requires $>$100,000 points for its training. They used retro-synthetic ideas to generate a representative hypothetical dataset of $\sim$250,000 polymers, constructed from the previously discussed set of 3045 chemical building blocks. For the discovery of polymers with target properties, they first encode a few known polymers that satisfy the given design criteria to find regions in the latent space where desirable polymers are expected to be present. Linear interpolations within these preferred regions of the latent space are then used to select latent points for which GPR property predictions meet the desired goals. Finally, the decoder is used to obtain the polymer SMILES associated with such selected latent vectors. Several hundreds of new polymers that meet the target property objectives were generated using this process. We anticipate that the concepts of transfer learning, multi-task learning and semi-supervised learning will advance the use of generative models for polymer discovery.

The following comparisons between the GA and the generative techniques for inverse design can be made. First, GA is relatively easy to interpret and entails little efforts to tune the involved parameters (mutation chance, initial population, etc.). In contrast, VAE models being based on NNs are almost impossible to interpret and often entail hefty parameter tuning efforts. Second, the space explored by GA is somewhat constrained by the polymer building blocks, but the SMILES based polymer representation allows VAE to explore a much wider chemical space in a truly unconstrained manner. Lastly, prior knowledge can be easily incorporated in GA, for instance, by biasing the initial population and/or the mutation operation towards favorable building blocks. However, more comparative studies would be needed in the future to establish methods that are appropriate under different scenarios.

6. Application examples

Polymers are useful in a range of applications. To be a good candidate for any specific application, they must meet multiple desired property requirements, as summarized in Table 2 for selected applications. Below, we comment on a few such applications, with an emphasis on key properties relevant for those applications which may be used to formulate screening criteria (also captured in Table 2).

6.1. Polymer dielectrics design for high energy density capacitors

Polymer-based dielectric capacitors are widely used in energy storage devices [5,4,141–145]. Given the increasing needs of high energy density capacitors, the development of polymer informatics can significantly facilitate the discovery of novel polymer dielectrics [25,29,98,99,143]. Typically, good polymer dielectrics for high energy density capacitors need to satisfy several property requirements, e.g., high dielectric constant and high breakdown strength (which is positively correlated with band gap and charge injection barriers of metal/polymer interfaces [146]). Further, polymers with high glass transition temperature are desired for high-temperature capacitors to enhance the thermal stability at extreme temperature [144,145]. Thus, the criteria of high glass transition temperature and ε, large band gap and high charge injection barriers can be utilized, in combination with machine learning, to screen polymer candidates tailored to extreme high-temperature and electric field. For instance, several representative dielectrics films with high dielectric constant and band gap, have been successfully designed and synthesized using computation- and data-driven strategies [25,99]. Additionally, many representative polymer dielectrics are being proposed for high-temperature capacitors by either screening known/hypothetical polymers using the enumeration method [52,146] or using the generative models, such as GA [128] and VAE [44], as described in Section 5.

6.2. Polymer membrane design for gas separation

Polymers are also promising candidates for gas separation due to their high surface area [42,147,148]. A typical class of polymers called polymers of intrinsic microporosity, has attracted great attention since the early 1990s [147]. The present polymer membranes suffer from low selectivity and physical aging, calling for the exploration of novel polymeric membranes. However, it is non-trivial to find promising polymer membranes with a combination of high permeability and selectivity (or above the upper bound of “Robeson plots” [149]) for different gas pairs, e.g., O_2/N_2. Some initial attempts have been performed to speed up the polymer membrane search using data-driven approaches, for instance, building gas permeability prediction models [111] (see Section 4) and identifying polymer membrane candidates for CO_2/CH_4 separation using the enumeration method [42].

6.3. Polymer electrolytes design for Li-ion batteries

Rechargeable Li-ion batteries have been widely adopted in many applications from micro-electronics to aerospace. Motivated by their commercial need, the development of novel and safer solid polymer electrolyte materials has caught ever-increasing attention [10, 150–152]. To optimize the performance of Li-ion batteries, the polymer electrolytes should have a wide electrochemical stability window, high ionic conductivity and Li-ion transference, and low glass transition temperature. Since it is time-consuming to search optimal electrolytes using experiments, data-driven aided polymer design strategies provide a great opportunity. For example, we previously noted that Wang et al. designed novel polymer electrolytes with high ionic conductivity using machine learning aided coarse-grained molecular dynamics simulations [37]. Additionally, the property prediction models (Section 4) and the design algorithms (Section 5) discussed above are powerful methods to screen/design polymer electrolytes satisfying multiple property requirements.

6.4. *Conducting polymers design for electronic applications*

Although polymers are usually insulators, there is a class of intrinsically conducting polymers used in electronic devices, such as light-emitting diodes, field-effect transistors and organic solar cells [6,153, 154]. Molecular doping is often used to further increase the conductivity of polymers [153], but it slows down the discovery of optimal polymer-dopant pairs with high conductivity because of the complex nature of the electron transfer mechanisms, dopants and polymers interactions, and processing conditions. This situation can be improved using polymer informatics, e.g., developing conductivity prediction models and screening optimal polymer-dopants pairs using the enumeration method. It is supported by the discovery of several high-performing donor/acceptor pairs for organic solar cells using random forest and boosted regression trees based property prediction models [53].

6.5. *Biodegradable and depolymerizable polymers discovery*

Bioplastics, such as those derived from plants and bacteria, are rich in highly oxygenated molecules. They can be utilized in the production of monomers capable of facile conversion to plastic materials that are easily degradable in the environment [155]. However, to fully harness the power of these nonconventional biosynthesis routes, it needs to establish structure-property relationships to identify desired application-specific optimal chemistries. To understand this problem better, Pilania et al. have proposed a machine learning route to learn structure-property mappings in PHA-based polymers from polymer data [129]. Moreover, it is critical to discover new biodegradable polymer candidates with high biodegradability. Because low crystallinity, melting temperature and glass transition temperature lead to large amorphous regions and favor biodegradation, ration-design of biodegradable polymers satisfying these properties using data-driven methods can be an important research topic. For instance, some new biodegradable polymers with desired glass transition temperature have been designed recently using GA [129] (see Section 5).

Additionally, depolymerizable polymers are playing an increasingly important role in practical applications, especially in drug delivery, recyclable plastics, self-healing and recyclable coating materials [156, 157]. Such great interest is motivated by the fact that depolymerizable polymers, upon exposure to particular stimuli, can be triggered to rapidly depolymerize into monomers at moderate to relatively low temperatures. As a result, polymers with low ceiling temperatures are desirable, where ceiling temperature is the temperature at which the polymerization and depolymerization rates are in equilibrium. Because of the limited available number of known polymers with low ceiling temperatures, it is greatly desired to propose computational strategies to estimate the ceiling temperature. Further, the data-driven design tools involved in polymer informatics could be applied to rapidly screen such depolymerizable polymers.

7. Critical next steps

7.1. *Beyond homopolymers*

So far, many data-driven approaches have been limited to homopolymers. The space of co-polymers, polymer blends and polymers with additives/nanocomposites remains largely unexplored but has great practical significance. Brinson and co-workers have spent significant efforts to develop “NanoMine” for polymer nanocomposites analysis and design [158]. However, it is still non-trivial to treat these types of polymers, because of their complicated chemical and physical structures. Co-polymers consist of two/more monomer or basic building unit types, and could be branched or linear co-polymers (further classified as block, alternating and random co-polymers based on the structural arrangement of different monomers). Polymer blends are mixtures of two or more polymers, including homogeneous, immiscible and heterogeneous polymer blends. The ratios and structural arrangements of different monomers (or polymers) significantly impact properties of co-polymers and polymer blends, but only sparse data is available on this topic. Moreover, it is challenging to systematically and dynamically collect such data from various resources. Thus, advanced techniques need to be developed to collect, represent and learn data of more complex varieties of polymers.

7.2. *Sustainable data capture*

The core requirement for polymer informatics is a broad-based data acquisition and management infrastructure. In addition to the limited number of available polymer databases and polymer handbooks, a large amount of scientific data remains untapped in numerous scientific journals, including text, tables or figures. While the manual text excerption of such journals is very time consuming and laborious, machine learning-based NLP methods are more powerful and promising tools to expedite and automate this process. The application of NLP tools in material science is still in its infancy. More efforts are needed to incorporate materials or polymers domain knowledge into existing NLP algorithms (e.g., word2vec [159]) to train word-vectors (numerical vectors that represent distinct words) for scientific information retrieval. To achieve this goal, Named Entities Recognition (NER) is the most important step, i.e., tokenizing the words into scientific meanings (e.g., chemical species, synthesis conditions and characterization methods). ChemDataExtractor [160], ChemSpot [161] and ChemTagger [162] are available toolkits for extracting chemical information of materials from scientific articles, such as inorganics and molecules. Similar tools need to be developed for the polymer domain.

However, polymers pose additional challenges [21,26], as there is no standard or complete polymer name entity dictionaries. A collection of source-based, structure-based, traditional and abbreviation names are interchangeably used to name polymers [26]. For example, polyethylene is also called PE, poly(ethylene) and poly-(ethylene), but all these possible names should be treated as the same entity (in a process referred to as “normalization” by the NLP community). In addition to names, more efforts are required to assign polymer notations for specific categories, e.g., properties, synthesis recipes and characterization technologies. Therefore, it is of great importance to create unique and standard polymer related dictionaries in the future. Other important issues include building efficient toolkits to interpret monomer SMILES from polymer names, identifying structure (or polymer names)-property relationships in texts, and extracting valuable material property contained in images and tables.

7.3. *Polymer representation and learning*

As discussed in Section 3, molecular or polymer-based fingerprints can provide acceptable prediction accuracy for many polymer properties, e.g., glass transition temperature, band gap, dielectric constant and gas permeability. This is because the chemical structure of the monomers plays a dominant role in determining these properties. However, other important polymer properties, including crystallinity, mechanical properties (e.g., tensile strength) and solution behavior, strongly depend on their molecular weights, morphologies (linear or cross-link) and processing conditions (temperature, pressure and cooling-rates). Incorporating these descriptors in the fingerprint framework is critical to the creation of accurate, robust and universal property prediction models.

In addition to enriching the polymer fingerprint definition, more advanced neural networks algorithms can be applied for learning the latent knowledge, property prediction and polymer generations. For instance, the transfer-learning or multi-task learning approaches have great potential to deal with the sparse data issue in polymers. The former modifies the latent features learned using one source task to learn a different target task, while the latter trains multiple source tasks and the shared features used to learn a target task. The key concept common between these two methods is the learning of a shared (polymer) representation between related properties (or materials). These algorithms have been successfully applied in the domain of drug design and bioinformatics [163–165]. In polymers, as mentioned in Section 4, Yamada et al. have used the transfer-learning method to predict properties of polymers using pre-trained models of molecules and inorganic materials [125]. However, large and diverse datasets of related property (tasks) are essential for the success of such models, as only then there is a high chance of learning transferable features and achieving accurate predictions for the target task.

Another important topic is the use of graph neural networks (GNN) in polymer informatics. In contrast to traditional manually designed fragment-based ML models, GNN represents materials as graphs (typically, atoms as nodes and their bonds as edges) and automatically find their optimal fingerprint representation depending on the downstream learning task, leading to its wide applications for molecules. However, the use of GNN for polymers has been limited [104,124] owing to the difficulty in treating large-scale polymers using graphs. Further, polymers are made up of numerous repeat units, and the best way to treat connection points between repeat units in a graph is unclear. Using oligomers to replace polymers is a potential solution [104,124], however, its prediction capability needs to be tested. Additionally, ideas on graph generative methods for molecules, e.g., atoms- and substructure-based encoder-decoder methods [45] could be extended for polymers using GNN. Another interesting approach of graph-to-graph translation was recently put forth to optimize molecules with desired properties, by assembling one molecular graph with another of the target properties [46]. All of these techniques can be adapted for polymers, provided the following challenges are addressed. Many polymers have large-sized monomers (>50 atoms), making it difficult to correctly assemble potential fragments during the decoding process. Motifs-based methods can greatly increase the reconstruction and validation accuracy for polymer generation by using large-size motifs as building blocks [45]. However, other concerns remain, such as chemical or thermodynamic stability of generated polymers and their synthetic feasibility.

7.4. Polymer retro-synthesis planning

Even with the knowledge of which polymer to make for a given application (designed, for instance, using intuition, computation or machine learning), the realization of the polymer can still be very slow because of synthesis challenges. There are various uncertain factors, such as unavailability, toxicity or high cost of the raw materials or demanding technical steps. In the past, the synthesis pathways adopted for a target polymer have been heavily dependent on the domain knowledge and personal preferences of experimenters. Computer-assisted retro-synthesis techniques have been widely developed in the last several decades to identify a series of reaction pathways leading to the synthesis of a target product. In the domain of molecules, either template-based [54,166] or template-free [105,167,168] machine learning approaches have been built for product prediction and have achieved promising results. However, no such method has been developed yet for the case of polymers. Complications in the polymer synthesis processes, e.g., various polymerization mechanisms (such as addition, ring-opening and condensation polymerization), the selection of optimal monomers and solvent pairs, processing conditions (such as cooling rates or annealing temperatures) will need to be considered. Further, unlike molecules, there is no library of reaction templates for polymerization. Nevertheless, experimental polymer synthesis data is plentiful, which can be accumulated manually or using NLP methods, and processed appropriately to develop machine learning models for polymer synthesis and retro-synthesis planning.

7.5. Autonomous integration of experimental and computational workflows

As all the different pieces of AI-assisted chemical search, retro-synthesis planning and processing optimization come together, the idea of autonomous polymer synthesis and design is expected to become a reality. In fact, examples of autonomous robot researchers with the ability to synthesize drugs for tropical diseases [169], carbon nanotubes with targeted growth rates [170], layered superlattices [171], and even perform X-ray scattering measurements [172], have already been demonstrated recently. However, polymers owing to their structural, chemical and processing complexity pose unique challenges for autonomous design. For instance, the average molecular weight of a polymer, which predominantly dictates its properties, is highly sensitive to the processing time and conditions. Learning such complex relations, from the data alone, will be challenging for an autonomous researcher. The real-time/in-line characterization of polymers is also difficult owing to their complex semi-crystalline/amorphous structure, or due to the different degree of branching or stereochemical relationships. Nonetheless, the prowess of autonomous labs in terms of time and cost benefits, experimentation consistency, long hours of operation, and efficient and robust search of parameter spaces is expected to guide polymer discovery in the future.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

This work has benefited from the generous support by the Office of Naval Research, the Toyota Research Institute, the Department of Energy and the National Science Foundation on machine learning related research through several grants. G.P. acknowledges support from the Laboratory Directed Research and Development (LDRD) program of Los Alamos National Laboratory under the BioManIAC project # 20190001DR. C. Ku. acknowledges support from the Alexander von Humboldt Foundation. R.B acknowledges support by LDRD funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357, and the use of the Center for Nanoscale Materials, an Office of Science user facility, supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. Joseph Kern is gratefully acknowledged for a critical reading of the manuscript.

References

[1] A.J. Peacock, A. Calhoun, Polymer Chemistry: Properties and Application, Carl Hanser Verlag GmbH Co KG, 2012.
[2] P.C. Hiemenz, T.P. Lodge, Polymer Chemistry, CRC Press, 2007.
[3] C. Wong, Polymers for Electronic & Photonic Application, Elsevier, 2013.

[4] T.D. Huan, B. Steve, T. Gilbert, L. Christian, C. Miko, K. Sanat, R. Rampi, Prog. Mater. Sci. 83 (2016) 236–269.

[5] Q. Tan, P. Irwin, Y. Cao, IEEJ Trans. FM 126 (11) (2006) 1153–1159.

[6] A.C. Mayer, S.R. Scully, B.E. Hardin, M.W. Rowell, M.D. McGehee, Mater. Today 10 (11) (2007) 28–33.

[7] F.M. Haque, S.M. Grayson, Nat. Chem. (2020) 1–12.

[8] T. Leigh, P. Fernandez-Trillo, Nat. Rev. Chem. (2020) 1–20.

[9] K. Ghosal, B.D. Freeman, Polym. Adv. Technol. 5 (11) (1994) 673–697.

[10] C. Sequeira, D. Santos, Polymer Electrolytes: Fundamentals and Applications, Elsevier, Amsterdam, Netherlands, 2010.

[11] R. Geyer, J.R. Jambeck, K.L. Law, Sci. Adv. 3 (7) (2017) e1700782.

[12] M.I. Jordan, T.M. Mitchell, Science 349 (6245) (2015) 255–260.

[13] A. Gopnik, Sci. Am. 316 (6) (2017) 60–65.

[14] Y. Liu, T. Zhao, W. Ju, S. Shi, J. Materiomics 3 (3) (2017) 159–177.

[15] B. Meredig, Chem. Mater. 31 (23) (2019) 9579–9581, https://doi.org/10.1021/acs.chemmater.9b04078.

[16] J. Schmidt, M.R. Marques, S. Botti, M.A. Marques, NPJ Comput. Mater. 5 (1) (2019) 1–36.

[17] T.J. Oweida, A. Mahmood, M.D. Manning, S. Rigin, Y.G. Yingling, MRS Adv. 5 (7) (2020) 329–346, https://doi.org/10.1557/adv.2020.171.

[18] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, Sci. Data 6 (1) (2019) 203, https://doi.org/10.1038/s41597-019-0224-1.

[19] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, NPJ Comput. Mater. 3 (1) (2017) 54.

[20] D.J. Audus, J.J. de Pablo, ACS Macro Lett. 6 (10) (2017) 1078–1082, https://doi.org/10.1021/acsmacrolett.7b00228, pMID: 29201535.

[21] J.S. Peerless, N.J. Milliken, T.J. Oweida, M.D. Manning, Y.G. Yingling, Adv. Theory Simul. 2 (1) (2019) 1800129.

[22] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, J. Phys. Chem. C 122 (31) (2018) 17575–17585.

[23] D.H. Tran, K. Chiho, C. Lihua, C. Anand, B. Rohit, V. Shruti, K. Deepak, P. L. Jordan, G. Rishi, S. Pranav, J.L. Manav, S. Ramprasad, R. Madeline, Rampi, J. Appl. Phys. 128 (2020) 171104.

[24] A. Chandrasekaran, C. Kim, R. Ramprasad, Machine Learning Meets Quantum Physics, Springer, 2020, pp. 397–412.

[25] A. Mannodi-Kanakkithodi, A. Chandrasekaran, C. Kim, T.D. Huan, G. Pilania, V. Botu, R. Ramprasad, Mater. Today 21 (7) (2018) 785–796.

[26] N. Adams, P. Murray-Rust, Macromol. Rapid Commun. 29 (8) (2008) 615–632.

[27] J. Mark, Polymer Data Handbook, Oxford University Press, 1999.

[28] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, M. Yamazaki. 2011 International Conference on Emerging Intelligent Data and Web Technologies, IEEE, 2011, pp. 22–29.

[29] T.D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, Sci. Data 3 (2016) 160012, https://doi.org/10.1038/sdata.2016.12.

[30] L. Chen, R. Batra, R. Ranganathan, G. Sotzing, Y. Cao, R. Ramprasad, Chem. Mater. 30 (21) (2018) 7699–7706, https://doi.org/10.1021/acs.chemmater.8b02997.

[31] L. Chen, S. Venkatram, C. Kim, R. Batra, A. Chandrasekaran, R. Ramprasad, Chem. Mater. 31 (12) (2019) 4598–4604, https://doi.org/10.1021/acs.chemmater.9b01553.

[32] H.S. Kim, S.M. Huang, Y.G. Yingling, MRS Adv. 1 (25) (2016) 1883–1889, https://doi.org/10.1557/adv.2016.91.

[33] K.-H. Shen, L.M. Hall, Macromolecules 53 (10) (2020) 3655–3668, https://doi.org/10.1021/acs.macromol.0c00216.

[34] S. Zhu, N. Lempesis, P.J. in t Veld, G.C. Rutledge, Macromolecules 51 (22) (2018) 9306–9316, https://doi.org/10.1021/acs.macromol.8b01922.

[35] S. Mogurampelly, O. Borodin, V. Ganesan, Annu. Rev. Chem. Biomol. Eng. 7 (2016) 349–371.

[36] Y. Seo, K.-H. Shen, J.R. Brown, L.M. Hall, J. Am. Chem. Soc. 141 (46) (2019) 18455–18466, https://doi.org/10.1021/jacs.9b07227, pMID: 31674178.

[37] Y. Wang, T. Xie, A. France-Lanord, A. Berkley, J.A. Johnson, Y. Shao-Horn, J. C. Grossman, Chem. Mater. 32 (10) (2020) 4144–4151, https://doi.org/10.1021/acs.chemmater.9b04830.

[38] Y. An, S. Singh, K.K. Bejagam, S.A. Deshmukh, Macromolecules 52 (13) (2019) 4875–4887, https://doi.org/10.1021/acs.macromol.9b00615.

[39] Y. An, S.A. Deshmukh, Chem. Comm. 56 (65) (2020) 9312–9315.

[40] T.D. Huan, R. Ramprasad, J. Phys. Chem. Lett. 11 (15) (2020) 5823–5829, https://doi.org/10.1021/acs.jpclett.0c01553, pMID: 32609529.

[41] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, https://doi.org/10.1021/ci00057a005.

[42] J.W. Barnett, C.R. Bilchak, Y. Wang, B.C. Benicewicz, L.A. Murdock, T. Bereau, S. K. Kumar, Sci. Adv. 6 (20) (2020) eaaz4301.

[43] S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, et al., NPJ Comput. Mater. 5 (1) (2019) 1–11.

[44] R. Batra, H. Dai, H. Tran, L. Chen, C. Kim, G. Will, L. Song, R. Ramprasad, Polymers for extreme conditions designed using syntax-directed variational autoencoders, Chem. Mater. (2020). Article ASAP, 2020.

[45] W. Jin, R. Barzilay, T. Jaakkola, Hierarchical Generation of Molecular Graphs Using Structural Motifs, 2020 (arXiv preprint), arXiv:2002.03230.

[46] W. Jin, K. Yang, R. Barzilay, T. Jaakkola, Learning Multimodal Graph-To-Graph Translation for Molecular Optimization, 2018. arXiv:1812.01070.

[47] R. Batra, G. Pilania, B.P. Uberuaga, R. Ramprasad, ACS Appl. Mater. Interfaces 11 (28) (2019) 24906–24918.

[48] A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T.D. Huan, R. Ramprasad, Comput. Mater. Sci. 172 (2019) 109286.

[49] A. Chandrasekaran, C. Kim, S. Venkatram, R. Ramprasad, Macromolecules 53 (12) (2020) 4764–4769, https://doi.org/10.1021/acs.macromol.0c00251.

[50] A. Mannodi-Kanakkithodi, G. Treich, T.D. Huan, R. Ma, M. Tefferi, Y. Cao, G. Sotzing, R. Ramprasad, Adv. Mater. 28 (2016) 6277–6291.

[51] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Sci. Rep. 6 (2016) 20952, https://doi.org/10.1038/srep20952.

[52] L. Chen, C. Kim, R. Batra, J.P. Lightstone, C. Wu, Z. Li, A.A. Deshmukh, Y. Wang, H.D. Tran, P. Vashishta, et al., NPJ Comput. Mater. 6 (1) (2020) 1–9.

[53] Y. Wu, J. Guo, R. Sun, J. Min, NPJ Comput. Mater. 6 (1) (2020) 1–8.

[54] C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, K.F. Jensen, ACS Cent. Sci. 3 (5) (2017) 434–443.

[55] C.W. Coley, D.A. Thomas, J.A. Lummiss, J.N. Jaworski, C.P. Breen, V. Schultz, T. Hart, J.S. Fishman, L. Rogers, H. Gao, et al., Science 365 (6453) (2019) eaax1566.

[56] J. Brandrup, E.H. Immergut, E.A. Grulke, A. Abe, D.R. Bloch, Polymer Handbook, vol. 7, Wiley New York etc., 1989.

[57] G. Wypych, Handbook of Polymers, Elsevier, 2016.

[58] D. van Krevelen, K. te Nijenhuis, Properties of Polymers: Their Correlation With Chemical Structure; Their Numerical Estimation and Prediction From Additive Group Contributions, Elsevier Science, 2009.

[59] J.E. Mark, Polymer Data Handbook, 2009.

[60] http://www.polymerdatabase.com, howpublished = http://www.polymerdatabase.com.

[61] B. Ellis, R. Smith, Polymers: A Property Database, CRC Press, 2008.

[62] https://www.campusplastics.com, howpublished = https://www.campusplastics.com.

[63] J. Pionteck, M. Pyda, Polymer Solids and Polymer Melts, Part 2, Thermodynamic Properties-PVT-Data and Thermal Properties-Landolt-Boernstein-Polymer, 2014.

[64] https://pppdb.uchicago.edu, howpublished = https://pppdb.uchicago.edu.

[65] https://khazana.gatech.edu, howpublished = https://khazana.gatech.edu.

[66] D. Braun, H. Cherdron, M. Rehahn, H. Ritter, B. Voit, Polymer Synthesis: Theory and Practice: Fundamentals, Methods, Experiments, Springer Science & Business Media, 2012.

[67] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Nature 571 (7763) (2019) 95–98.

[68] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Chem. Mater. 29 (21) (2017) 9436–9444, https://doi.org/10.1021/acs.chemmater.7b03500.

[69] P. Yi, C.R. Locker, G.C. Rutledge, Macromolecules 46 (11) (2013) 4723–4733.

[70] M. Andreev, G.C. Rutledge, J. Rheol. 64 (1) (2020) 213–222.

[71] A.E. Marlowe, A. Singh, Y.G. Yingling, Mater. Sci. Eng. C 32 (8) (2012) 2583–2588.

[72] K.C. Daoulas, M. M&ldquo;uller, J.J. De Pablo, P.F. Nealey, G.D. Smith, Soft Matter 2 (7) (2006) 573–583.

[73] L. Chen, T.D. Huan, R. Ramprasad, Sci. Rep. 7 (2017) 6128.

[74] K. Chenoweth, S. Cheung, A.C. van Duin, W.A. Goddard, E.M. Kober, J. Am. Chem. Soc. 127 (19) (2005) 7192–7202.

[75] A. Vashisth, C. Ashraf, W. Zhang, C.E. Bakis, A.C. Van Duin, J. Phys. Chem. A 122 (32) (2018) 6633–6642.

[76] A. Vashisth, C. Ashraf, C.E. Bakis, A.C. van Duin, Polymer 158 (2018) 354–363.

[77] S. Fukushima, S. Tiwari, H. Kumazoe, R.K. Kalia, A. Nakano, F. Shimojo, P. Vashishta, AIP Adv. 9 (4) (2019) 045022.

[78] A. Vasilev, T. Lorenz, C. Breitkopf, Polymers 12 (5) (2020) 1081.

[79] S. Shenogin, A. Bodapati, P. Keblinski, A.J. McGaughey, J. Appl. Phys. 105 (3) (2009) 034906.

[80] D.V. Guseva, V.Y. Rudyak, P.V. Komarov, B.A. Bulgakov, A.V. Babkin, A. V. Chertovich, Polymers 10 (7) (2018) 792.

[81] I.-C. Yeh, J.W. Andzelm, G.C. Rutledge, Macromolecules 48 (12) (2015) 4228–4239.

[82] C.-L. Pai, M.C. Boyce, G.C. Rutledge, Polymer 52 (10) (2011) 2295–2301.

[83] D. Kamal, A. Chandrasekaran, R. Batra, R. Ramprasad, Mach. Learn.: Sci. Technol. 1 (2) (2020) 025003, https://doi.org/10.1088/2632-2153/ab5929.

[84] E.A. Algaer, F. M&rdquo;uller-Plathe, Soft Mater. 10 (1-3) (2012) 42–80.

[85] M.-X. Zhu, H.-G. Song, Q.-C. Yu, J.-M. Chen, H.-Y. Zhang, Int. J. Heat Mass Transf. 162 (2020) 120381, https://doi.org/10.1016/j.ijheatmasstransfer.2020.120381.

[86] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, Chem. Med. Chem. 3 (10) (2008) 1503–1507, https://doi.org/10.1002/cmdc.200800178.

[87] Rdkit, open source toolkit for cheminformatics.

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

[89] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Chem. Rev. 112 (5) (2012) 2889–2919, https://doi.org/10.1021/cr200066h, pMID: 22251444.

[90] X. Yu, X. Wang, X. Li, J. Gao, H. Wang, Macromol. Theory Simul. 15 (1) (2006) 94–99.

[91] H. Hasnaoui, M. Krea, D. Roizard, J. Membr. Sci. 541 (2017) 541–549, https://doi.org/10.1016/j.memsci.2017.07.031.

[92] T.-S. Lin, C.W. Coley, H. Mochigase, H.K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J.A. Johnson, J.A. Kalow, K.F. Jensen, B.D. Olsen, ACS Cent. Sci. 5 (9) (2019) 1523–1531, https://doi.org/10.1021/acscentsci.9b00476.

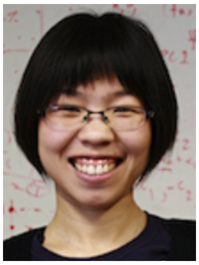
[93] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50 (5) (2010) 742–754.

[94] M.H. Segler, T. Kogej, C. Tyrchan, M.P. Waller, ACS Cent. Sci. 4 (1) (2018) 120–131.

[95] G.B. Goh, N.O. Hodas, C. Siegel, A. Vishnu, Smiles2vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties, 2017 (arXiv preprint), arXiv:1712.02034.

[96] C. Wang, G. Pilania, S. Boggs, S. Kumar, C. Breneman, R. Ramprasad, Polymer 55 (4) (2014) 979–988.

[97] K. Wu, N. Sukumar, N. Lanzillo, C. Wang, R. "Rampi" Ramprasad, R. Ma, A. Baldwin, G. Sotzing, C. Breneman, J. Polym. Sci. Pol. Phys. 54 (20) (2016) 2082–2091.
[98] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Sci. Rep. 6 (2016).
[99] A. Mannodi-Kanakkithodi, G.M. Treich, T.D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing, R. Ramprasad, Adv. Mater. 28 (30) (2016) 6277–6291.
[100] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Advances in Neural Information Processing Systems, 2015, pp. 2224–2232.
[101] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, J. Comput. Aided Mol. Des. 30 (8) (2016) 595–608.
[102] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph Neural Networks: A Review of Methods and Applications, arXiv:1812.08434 (arXiv preprint).
[103] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S.M. Lin, W. Zhang, P. Zhang, H. Sun, Bioinformatics 36 (4) (2019) 1241–1251, https://doi.org/10.1093/bioinformatics/btz718. https://academic.oup.com/bioinformatics/article-pdf/36/4/1241/32527566/btz718.pdf.
[104] M. Zeng, J.N. Kumar, Z. Zeng, R. Savitha, V.R. Chandrasekhar, K. Hippalgaonkar, Graph Convolutional Neural Networks for Polymers Property Prediction, 2018 (arXiv preprint), arXiv:1811.06231.
[105] W. Jin, C. Coley, R. Barzilay, T. Jaakkola, Advances in Neural Information Processing Systems, 2017, pp. 2607–2616.
[106] S. Venkatram, R. Batra, L. Chen, C. Kim, M. Shelton, R. Ramprasad, J. Phys. Chem. B 124 (28) (2020) 6046–6054, https://doi.org/10.1021/acs.jpcb.0c01865, pMID: 32539396.
[107] F. Jabeen, M. Chen, B. Rasulev, M. Ossowski, P. Boudjouk, Comput. Mater. Sci. 137 (2017) 215–224.
[108] V. Venkatraman, B.K. Alsberg, Polymers 10 (1) (2018) 103.
[109] X. Yu, Fibers Polym. 11 (5) (2010) 757–766.
[110] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Sci. Rep. 3 (2013) 2810.
[111] G. Zhu, C. Kim, A. Chandrasekarn, J.D. Everett, R. Ramprasad, R.P. Lively, J. Polym. Eng. 1 (2020) (ahead-of-print).
[112] G. Pilania, P.V. Balachandran, J.E. Gubernatis, T. Lookman, Synthesis Lectures on Materials and Optics, 1(1, Morgan & Claypool Publishers, 2020, pp. 1–188.
[113] P. Zaspel, B. Huang, H. Harbrecht, O.A. von Lilienfeld, J. Chem. Theory Comput. 15 (3) (2018) 1546–1559, publisher: ACS Publications.
[114] G. Pilania, J.E. Gubernatis, T. Lookman, Comput. Mater. Sci. 129 (2017) 156–163, publisher: Elsevier.
[115] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Phys. Rev. B 93 (11) (2016) 115104, publisher: APS.
[116] R. Ramakrishnan, P.O. Dral, M. Rupp, O.A. von Lilienfeld, J. Chem. Theory Comput. 11 (5) (2015) 2087–2096, publisher: ACS Publications.
[117] M.C. Kennedy, A. O'Hagan, Biometrika 87 (1) (2000) 1–13, publisher: Oxford University Press.
[118] S. Venkatram, C. Kim, A. Chandrasekaran, R. Ramprasad, J. Chem. Inf. Model. 59 (10) (2019) 4188–4194, https://doi.org/10.1021/acs.jcim.9b00656.
[119] Y. LeCun, Y. Bengio, G. Hinton, Nature 521 (7553) (2015) 436–444.
[120] A. Agrawal, A. Choudhary, MRS Commun. 9 (3) (2019) 779–792.
[121] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, vol. 1, MIT Press, Cambridge, 2016.
[122] W. Liu, C. Cao, Colloid Polym. Sci. 287 (7) (2009) 811–818.
[123] X. Chen, L. Sztandera, H.M. Cartwright, Int. J. Intell. Syst. 23 (1) (2008) 22–32.
[124] B.G. Sumpter, D.W. Noid, Macromol. Theory Simul. 3 (2) (1994) 363–378.
[125] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, ACS Cent. Sci. 5 (10) (2019) 1717–1730, https://doi.org/10.1021/acscentsci.9b00804.
[126] C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, MRS Commun. 9 (3) (2019) 860–866, https://doi.org/10.1557/mrc.2019.78.
[127] T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, NPJ Comput. Mater. 5 (1) (2019) 1–17.
[128] C. Kim, R. Batra, L. Chen, H. Tran, R. Ramprasad, Comput. Mater. Sci. 186 (2020) 110067.
[129] G. Pilania, C.N. Iverson, T. Lookman, B.L. Marrone, J. Chem. Inf. Model. 59 (12) (2019) 5013–5025, publisher: ACS Publications.
[130] T.D. Huan, A. Mannodi-Kanakkithodi, R. Ramprasad, Phys. Rev. B 92 (2015) 014106.
[131] D.P. Kingma, M. Welling. Second International Conference on Learning Representations, ICLR, vol. 19, 2014.
[132] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[133] A. Hernandez, A. Balasubramanian, F. Yuan, S.A. Mason, T. Mueller, NPJ Comput. Mater. 5 (1) (2019) 1–11.
[134] A.H. Gandomi, S. Sajedi, B. Kiani, Q. Huang, Autom. Constr. 70 (2016) 89–97.
[135] C.D. Fjell, H. Jenssen, W.A. Cheung, R.E. Hancock, A. Cherkasov, Chem. Biol. Drug Des. 77 (1) (2011) 48–56.
[136] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R. P. Adams, A. Aspuru-Guzik, ACS Cent. Sci. 4 (2) (2018) 268–276, https://doi.org/10.1021/acscentsci.7b00572, pMID: 29532027.
[137] M.J. Kusner, B. Paige, J.M. Hernández-Lobato, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, International Convention Centre, Sydney, Australia, 2017, pp. 1945–1954.
[138] H. Dai, Y. Tian, B. Dai, S. Skiena, L. Song, Syntax-Directed Variational Autoencoder for Structured Data, arXiv:1802.08786.
[139] J. You, B. Liu, R. Ying, V. Pande, J. Leskovec. Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc, Red Hook, NY, USA, 2018, pp. 6412–6422.
[140] N. De Cao, T. Kipf, Molgan: An Implicit Generative Model for Small Molecular Graphs, arXiv:1805.11973.
[141] B. Chu, X. Zhou, K. Ren, B. Neese, M. Lin, Q. Wang, F. Bauer, Q.M. Zhang, Science 313 (5785) (2006) 334–336.
[142] Q. Li, L. Chen, M.R. Gadinski, S. Zhang, G. Zhang, H.U. Li, E. Iagodkine, A. Haque, L.-Q. Chen, T.N. Jackson, et al., Nature 523 (7562) (2015) 576.
[143] V. Sharma, C. Wang, R.G. Lorenzini, R. Ma, Q. Zhu, D.W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G.A. Sotzing, Nat. Commun. 5 (2014) 4845.
[144] J.S. Ho, S.G. Greenbaum, ACS Appl. Mater. Interfaces 10 (35) (2018) 29189–29218.
[145] C. Wu, A. Deshmukh, Z. Li, L. Chen, A. Alamri, Y. Wang, R. Ramprasad, G. A. Sotzing, Y. Cao, Adv. Mater. (2020). Accepted.
[146] D. Kamal, Y. Wang, H.D. Tran, L. Chen, Z. Li, C. Wu, S. Nasreen, Y. Cao, R. Ramprasad, ACS Appl. Mater. Interfaces 12 (33) (2020) 37182–37187, https://doi.org/10.1021/acsami.0c09555, pMID: 32705867.
[147] Z.-X. Low, P.M. Budd, N.B. McKeown, D.A. Patterson, Chem. Rev. 118 (12) (2018) 5871–5911.
[148] M.L. Jue, R.P. Lively, React. Funct. Polym. 86 (2015) 88–110.
[149] L.M. Robeson, J. Membr. Sci. 320 (1–2) (2008) 390–400.
[150] R. Agrawal, G. Pandey, J. Phys. D: Appl. Phys. 41 (22) (2008) 223001.
[151] J. Mindemark, B. Sun, E. T&rdquo;orm&ldquo;a, D. Brandell, J. Power Sources 298 (2015) 166–170.
[152] A.M. Stephan, K. Nahm, Polymer 47 (16) (2006) 5952–5964, https://doi.org/10.1016/j.polymer.2006.05.069.
[153] A.I. Hofmann, R. Kroon, L. Yu, C. M&rdquo;uller, J. Mater. Chem. C 6 (26) (2018) 6905–6910.
[154] J. Brebels, J.V. Manca, L. Lutsen, D. Vanderzande, W. Maes, J. Mater. Chem. A 5 (46) (2017) 24037–24050.
[155] A.-C. Albertsson, M. Hakkarainen, Science 358 (6365) (2017) 872–873, publisher: American Association for the Advancement of Science.
[156] J.A. Kaitz, O.P. Lee, J.S. Moore, MRS Commun. 5 (2) (2015) 191–204.
[157] A.M. DiLauro, J.S. Robbins, S.T. Phillips, Macromolecules 46 (8) (2013) 2963–2968.
[158] H. Zhao, X. Li, Y. Zhang, L.S. Schadler, W. Chen, L.C. Brinson, APL Mater. 4 (5) (2016) 053204.
[159] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013 (arXiv preprint), arXiv:1301.3781.
[160] M.C. Swain, J.M. Cole, J. Chem. Inf. Model. 56 (10) (2016) 1894–1904.
[161] T. Rocktäschel, M. Weidlich, U. Leser, Bioinformatics 28 (12) (2012) 1633–1640.
[162] L. Hawizy, D.M. Jessop, N. Adams, P. Murray-Rust, J. Cheminform. 3 (1) (2011) 17.
[163] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, Massively Multitask Networks for Drug Discovery (Icml), 2015.
[164] B. Ramsundar, B. Liu, Z. Wu, A. Verras, M. Tudor, R.P. Sheridan, V. Pande, J. Chem. Inf. Model. 57 (8) (2017) 2068–2076, https://doi.org/10.1021/acs.jcim.7b00146.
[165] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, V. Svetnik, J. Chem. Inf. Model. 55 (2) (2015) 263–274, https://doi.org/10.1021/ci500747n.
[166] M.H. Segler, M. Preuss, M.P. Waller, Nature 555 (7698) (2018) 604–610.
[167] C.W. Coley, W. Jin, L. Rogers, T.F. Jamison, T.S. Jaakkola, W.H. Green, R. Barzilay, K.F. Jensen, Chem. Sci. 10 (2019) 370–377, https://doi.org/10.1039/C8SC04228D.
[168] H. Dai, C. Li, C. Coley, B. Dai, L. Song, Advances in Neural Information Processing Systems 32, Curran Associates, Inc, 2019, pp. 8872–8882.
[169] K. Williams, E. Bilsland, A. Sparkes, W. Aubrey, M. Young, L.N. Soldatova, K. De Grave, J. Ramon, M. De Clare, W. Sirawaraporn, et al., J.R. Soc. Interface 12 (104) (2015) 20141289.
[170] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, NPJ Comput. Mater. 2 (1) (2016) 1–6.
[171] S. Masubuchi, M. Morimoto, S. Morikawa, M. Onodera, Y. Asakawa, K. Watanabe, T. Taniguchi, T. Machida, Nat. Commun. 9 (1) (2018) 1–12.
[172] M.M. Noack, K.G. Yager, M. Fukuto, G.S. Doerk, R. Li, J.A. Sethian, Sci. Rep. 9 (1) (2019) 1–19.



Lihua Chen is presently a research scientist in the School of Materials Science and Engineering, Georgia Institute of Technology. She received her Ph.D. degree in the Materials Science program at the University of Connecticut from September 2012 to October 2017. Her doctoral thesis work focused on unraveling the electronic structure of polyethylene using first-principles theory. From 2018 to 2020, she was a postdoc fellow at the Georgia Institute of Technology. Her current work is applying computational modeling and machine learning techniques to accelerate polymer discovery for various applications.