

Potentials and challenges of polymer informatics: exploiting machine learning for polymer design

Stephen Wu,^{*,†,‡} Hironao Yamada,[†] Yoshihiro Hayashi,[†] Massimiliano Zamengo,[§]
and Ryo Yoshida^{†,‡,||}

[†]*The Institute of Statistical Mathematics, Research Organization of Information and
Systems, Tachikawa, Tokyo 190-8562, Japan*

[‡]*The Graduate University for Advanced Studies, SOKENDAI, Tachikawa, Tokyo 190-8562,
Japan*

[¶]*School of Pharmacy, Tokyo University of Pharmacy and Life Sciences, Hachioji, Tokyo
192-0392, Japan*

[§]*School of Materials and Chemical Technology, Tokyo Institute of Technology, Meguro,
Tokyo 152-8550, Japan*

^{||}*National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan*

E-mail: stewu@ism.ac.jp

Abstract

There has been rapidly growing demand of polymeric materials coming from different aspects of modern life because of the highly diverse physical and chemical properties of polymers. Polymer informatics is an interdisciplinary research field of polymer science, computer science, information science and machine learning that serves as a platform to exploit existing polymer data for efficient design of functional polymers. Despite many potential benefits of employing a data-driven approach to polymer design, there has been notable challenges of the development of polymer informatics

attributed to the complex hierarchical structures of polymers, such as the lack of open databases and unified structural representation. In this study, we review and discuss the applications of machine learning on different aspects of the polymer design process through four perspectives: polymer databases, representation (descriptor) of polymers, predictive models for polymer properties, and polymer design strategy. We hope that this paper can serve as an entry point for researchers interested in the field of polymer informatics.

Introduction

Polymers are one of the most important classes of material in modern society, as its applications range from the plastic bags and bottles used in daily life to a variety of electronics, and even structural components in the aerospace industry. A polymer is a material made of a collection of chains that are built by connecting many repeated units, called monomers. These chains can form diverse structures that contribute to the highly diverse physical and chemical properties of different types of polymers. Some polymers can be consisting of more than one type of monomer to form even more complicated topological structures across different length scales. The research field of polymer science and engineering emerged to understand, control, and design novel polymers that can be used to satisfy the rapidly growing demand on highly functional materials coming from different aspects of modern life. While polymers can be categorized into natural or synthetic polymers, we will focus our discussion on the latter in this paper.

Following the mainstream of materials science, polymer science has gone through multiple major paradigm shifts. The early studies of polymers flourished in the first half of the 20th century was closely associated with H. Staudinger who received the Nobel Prize in 1953.¹ Initially, discovery of polymers was mainly based on an empirical approach, i.e., relying on many trial-and-error experiments. Accumulation of experimental experiences has led to developments of theoretical and simple statistical models for guiding the design of new

polymers. These include many important work by the 1974 Nobel prize recipient P. J. Flory, the group contribution method and so on.^{2–4} Following the rapid advances of computing power in the last few decades, computational methods has become one of the main tools to study properties of polymers.⁵ Recent developments of simulation techniques for various types of polymers opened up opportunities to computationally study polymers across different length scales.^{6,7} While generating experimental data of polymers is often costly and time-consuming, modern supercomputers provide new opportunities for building larger databases of polymers computationally. With the drastic expansion of data size in science, a data-driven approach for scientific discovery is said to be the 4th paradigm of science. Polymer informatics is an interdisciplinary research field of polymer science, computer science, information science and machine learning that serves as a platform to mine the precious polymer data for new knowledge. Yet, there has been notable challenges of the development of polymer informatics attributed to the complex hierarchical structures of polymers.^{8,9}

Design of a polymer can be broken down into three parts corresponding to the three steps in the typical production process of polymers: design of monomers (polymerization), microstructures (crystallization), and material processing (manufacturing) (see Figure 1). Monomers, the building blocks of polymers, contribute to the foundation of potential properties of the eventually produced polymers. While molecular size is one of the important factors that influences the properties of an organic material, the “size effect” of polymers is not directly correlated with the size of the monomers because a large collection of small monomers (e.g., ethylene) can also build long polymer chains the same way large monomers do. Instead, various metrics based on the molecular weight distribution (MWD) of a polymer are often used as a reference to relate the “size of polymer” to polymer properties. Different polymers built from the same monomer can have different MWD by controlling the polymerization process, which can lead to significantly different physical and chemical properties.^{10–12} Furthermore, the collection of polymer chains can form very different crystal structures through a variety of crystallization processes, affecting the material properties of

the resulting polymers. For example, Yi et al. controlled the crystallinity and orientation of poly(3-hexylthiophene) molecules to optimize the performance of solar cells.¹³ Finally, the same type of polymer in the microscopic scale can undergo different manufacturing process, such as stretching or mixing additives, to further enhance or alter its properties in order to fulfil specific needs from a broad range of applications.¹⁴ Ideally, the design space of polymers covers all parameters involved in the three production steps, for example, the molecule space of a single or multiple monomer(s) (namely homopolymers or copolymers, respectively), temperature and types of polymerization process, additives or fillers, molding methods, etc. In practice, we often focus on a subset of the parameters while keeping other fixed to reduce the enormous search space. For example, Wu et al. focused on the design of monomer that has a high probability of making liquid crystal polymers with high thermal conductivity after compressed to a polymer thin film.¹⁵

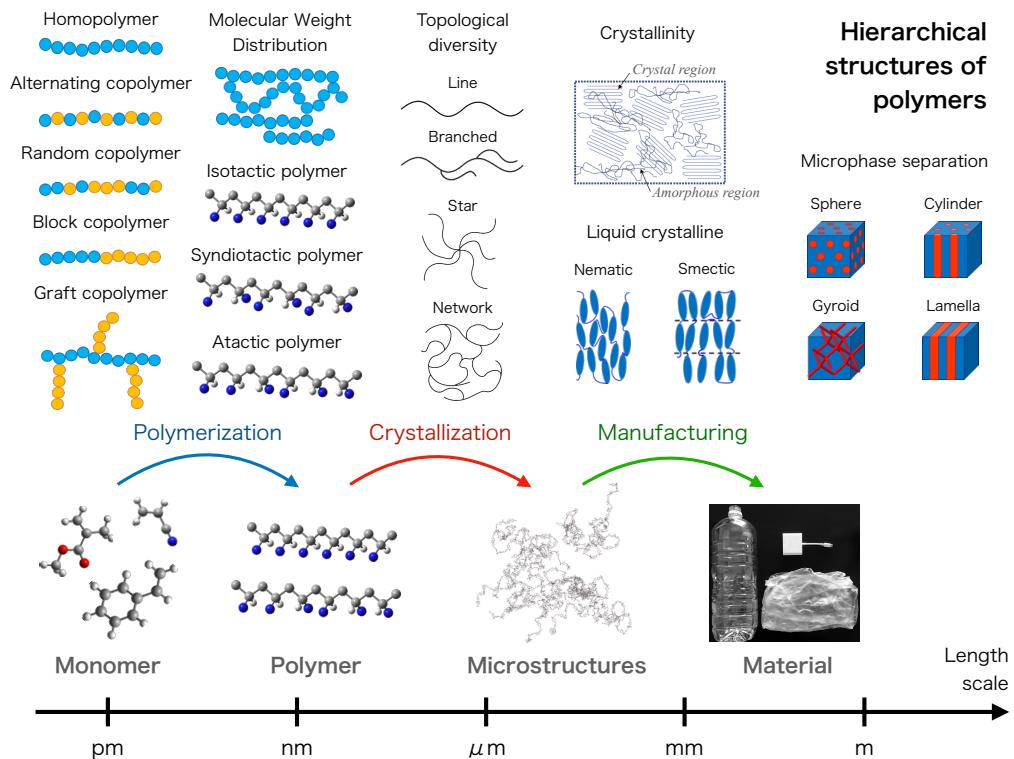


Figure 1: Overview of polymer design across different length scales.

A machine learning approach of polymers design on such a large design space requires a

very large data set, in terms of both quantity and diversity. Unfortunately, openly available large polymer database is still limited⁸ and historical data is often highly biased to a few types of polymers or polymer properties. For example, in PoLyInfo,^{16,17} which is one of the largest database of polymers, around 30% of data related to thermal properties are covered by only 10 different polymers and over 40% of data is glass transition temperature (see Figure 2). In order to achieve a fully data-driven process of polymer design, continuous efforts have been made to bridge the demand of machine learning technologies and the current state of polymer informatics. In this paper, we review and discuss the applications of machine learning on different aspects of the polymer design process through four perspectives: polymer databases, representation (descriptor) of polymers, predictive models for polymer properties, and polymer design strategy. Illustrative examples are also given using the open-source materials informatics software, XenonPy.¹⁸ We hope that this paper can serve as an entry point for researchers interested in the field of polymer informatics, who may be coming from any of the scientific fields covered by polymer informatics.

Machine learning in polymer informatics

The goal of machine learning is to develop computer algorithms that can automatically improve their ability to solve a target problem by extracting information from past experience (training data). A basic implementation of this idea is to build a mapping f from an input x to an output y when given some relevant data D . Here, x is the representation (or descriptor) of a problem of interest, which we will discuss in detail in the next section. Depending on the choice of f , y and D , machine learning is often categorized into:

- **Supervised learning** Directly building f that maps x to the desired output of interest y by learning from many examples of pairings between different x and y in D . Such data that contains pairs of x and y is called labelled data. Supervised machine learning is often subcategorized into regression or classification, where y is

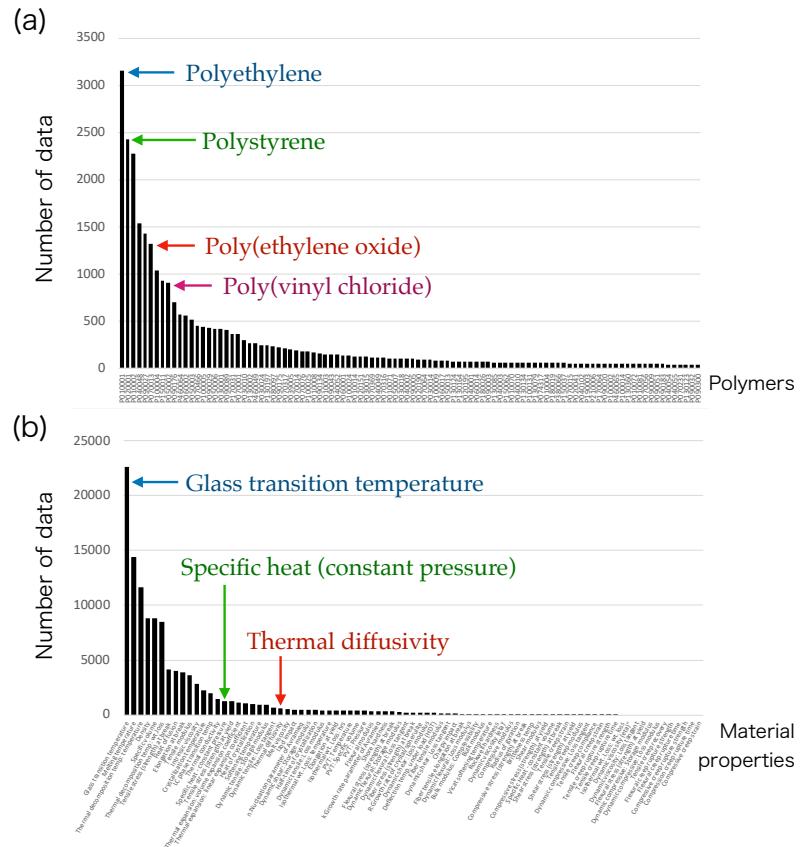


Figure 2: Statistics of 54151 data entries of 83 different polymer properties related to thermal properties recorded in PoLyInfo (extracted on April 2016). (a) Histogram of the number of data for the top 100 polymers with the most data is plotted in descending order. (b) Histogram of the number of data for 83 different polymer properties is plotted in descending order.

either a continuous variable or a discrete variable, respectively. An example would be building a model to predict the glass transition temperature of homopolymers (y) from composition and bonding information of the corresponding monomers (x).^{19,20}

- **Unsupervised learning** Learning the underlying structure of x using unlabelled data, i.e., D has no information about the output of interest y . Typical techniques of unsupervised machine learning are clustering and dimension reduction, where x is mapped to some categorical labels or lower dimensional space, respectively. For example, Xu et al. used these techniques to study phase transitions of polymer configurations.²¹ Note that the learned mapping is not necessarily directly correlated with the targeted y as such information is not included in D .
- **Reinforcement learning** Learning a strategy to achieve a certain goal in an interactive environment. Closely related to the problem of experimental design, the goal here is to learn f that maps the current state x to a possible action y . Successful training of f through repeated engagement to the system can take the system closer to the final goal. Typically, a reward function is defined to quantify the progress of the system and the size of D increases continuously during the interactive engagement process. An example would be developing optimal strategy to control MWD of a class of polymers.²²

Depending on the nature of application, available data and computing resources, it is important for researchers to frame their problems under an appropriate class of machine learning, to pick a suitable learning algorithm, and to represent the materials of interest using an effective descriptor. We attempt to provide useful hints from existing literature and our own experience in the following sections.

Database

One of the most important components in machine learning is data. The quality and quantity of available data determine the scope of solvable problems in an application. A rule of thumb in machine learning is that purely data-driven models are not reliable when extrapolating. In other words, it is dangerous to trust the prediction of a model on a material that is not similar to the ones in the training data. Therefore, high quality large database for a diverse set of polymers is always of high demand in polymer informatics. Table 1 shows a list of online databases that contain polymer data. There also exists large amount of publications on polymer technology or database that collects recipes of polymer synthesis (e.g., NIST Synthetic Polymer MALDI Recipes Database²³). These types of information require extra processing effort before being useful for polymer informatics.

Comparing to other research fields (e.g., image recognition) that benefit from modern machine learning technology, deep learning specifically, the number and size of open databases in polymer informatics are significantly smaller. We have accumulated a large amount of polymer data throughout the history of polymer science, but many historical data recorded in handbooks or publications are not well-organized, and a lot of the industry-owned data are not openly available. These issues have become the bottleneck for the development of polymer informatics.⁸ With the advancements of computational simulation technologies and supercomputers, we expect an increasing interest and opportunity to build large scale computational databases of polymers. Meanwhile, technology of high-throughput experiments²⁸ and the use of robotics combined with artificial intelligence²⁹ provide new opportunities to build experimental database of polymers efficiently. Making these databases open for research purposes will be the key to the success of polymer informatics.

Table 1: List of online polymer databases. Numbers are extracted on September 25, 2020.

Name (link)	Descriptions
PoLyInfo (polymer.nims.go.jp)	Polymer database supported by National Institute for Materials Science (NIMS) where data is mainly extracted from academic literature (covering 18,044 literature data). The database includes 367,711 property data points of various kinds of polymers built from 18,015 different monomers. ^{16,17}
Polymer Genome - Khazana (khazana.gatech.edu)	An open online platform that stores computational and experimental data from 24 publications. The database includes property data of 1,412 different polymers/organic materials and 2,657 different inorganic materials. ^{20,24}
Polymer Property Predictor and Database (pppdb.uchicago.edu)	Online polymer database maintained by CHiMaD that includes 263 and 212 data entries of Flory-Huggins χ value and glass transition temperature, respectively, extracted from the literature.
NanoMine (materialsmin.org/nm#)	An open platform for data sharing that includes images of polymer microstructures and property data of polymers. ^{25,26}
Cambridge Structural Database (www.ccdc.cam.ac.uk/structures)	Crystal structure database of organic and inorganic materials that includes more than 1,000,000 structures, where around 11% is polymeric.
CROW (polymerdatabase.com)	An online data source that includes thermo-physical data of polymers. The source is either experimental data from the literature and/or calculated values from similarity analysis or quantitative structure property relationships.
Polymers: A Property Database (poly.chemnetbase.com)	Online database of various polymer properties used to support the book <i>Polymers: A Property Database</i> by Wiley. ²⁷
Citrination (citrination.com)	Materials informatics platform that includes publicly available data of mechanical properties and solid surface energy of polymers.
CAMPUS (www.campusplastics.com)	Material property database of 9,236 commercial polymer grades.
Identify (www.netzsch-thermal-analysis.com)	Commercial software and database that includes differential scanning calorimetry curves for more than 600 commercial polymers.

Descriptor

The fundamental purpose of a material descriptor is to uniquely encode materials in a compact form in order to allow for efficient machine learning. This is particularly difficult for polymers due to their hierarchical structures.³⁰ An example of unique representation of polymers would be the Polymer Markup Language, which is designed to include complete information of a polymer ranging from compositional information to all the processing parameters.³¹ While this representation may be suitable for building polymer databases, it is not compact enough to be used as input of a model for machine learning purposes. In fact, a good material descriptor should consider the tradeoff between ability to uniquely represent a material, easiness to obtain or calculate, and sensitivity to targeted application.³² In other words, while possible, it is unlikely that a single descriptor can be used for all polymer applications. For example, to search for an efficient polyimide for controlled inkjet deposition, descriptor needs to be sensitive to specific microstructures formed by the polymer chains.³³ On the other hand, designing polymers with certain phase transition properties will require a descriptor that identifies key atomic-level structures. Different level of “fineness” of descriptor should be selected depending on the physical and chemical properties of interest to the target problem.³⁴ One example that attempted to capture such hierarchy of descriptor in a Python package is the Materials knowledge systems.³⁵ Here, we introduce and compare a few descriptors that could be useful in polymer informatics.

Processing conditions of polymers, such as temperature, additives and solvents, can be directly included in a descriptor. Representations of the microstructure and molecular structure of polymer chains are less intuitive. Certain machine learning models in deep learning allow direct use of images as input, but often require a significant amount of data to train the models.³⁶ Persistant homology is a technique to extract statistical features from topological structures,³⁷ but its application in polymer informatics is yet to be investigated. Graphical kernel is another option to numerically encode the polymer structures that can be represented in a simple graph.³⁸ However, finding an efficient graph representation for

the complex polymer structures is challenging, especially because different types of polymer chains can form structures in different length scales. Many of the descriptors commonly used in polymer informatics focus on capturing the molecular structure of monomers composing the polymer chains.

Material descriptors or fingerprints developed for small organic molecules are directly applicable to represent polymers based on their monomers. Modern machine learning algorithms also allow direct use of molecular graph or simplified molecular-input line-entry system (SMILES)³⁹ strings as input of a model.⁴⁰ Picking a descriptor suitable for the problem of interest may significantly improves the efficiency of machine learning.³⁴ However, such representation omits information about the chemical bonds between the monomers. An alternative is to consider oligomer consisting of n monomers, but there is no clear answer to how to pick n . The larger n is, the more representative the oligomer would be to the polymer. Nevertheless, certain types of descriptors, such as the physical descriptors, will take significantly longer time to calculate for a larger molecule. Furthermore, some fingerprints could be biased toward polymers with smaller or larger monomers depending on the choice of n . For example, a fingerprint that counts if there are more than 3 benzene rings in a molecule will not be able to distinguish between polystyrene and poly(bisphenol A carbonate) when $n > 3$. Figure 3 shows how different fingerprints may have different convergence behavior with respect to n . Wu et al. proposed to calculate descriptor of polymers using an infinitely long chain of their corresponding monomers.¹⁹ However, the bias issue for some fingerprints remains unsolved.

Polymer chain can be separated into a backbone and its side chains. Distinguishing the two components is theoretically important to predict polymer properties. However, it is not clear how to efficiently encode such information into a general descriptor for different classes of polymers. Similarly, developing descriptors for polymers consisting of more than one type of monomer is still an open challenge. For alternating copolymer, we can simply consider the combined repeating unit of multiple monomers as a “metamonomer”, but the molecule

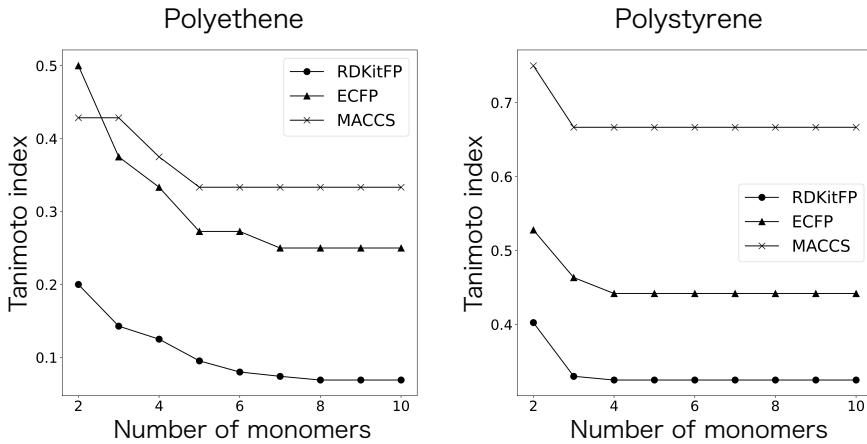


Figure 3: Convergence of the change of fingerprints for two polymers as a function of the number of monomers in the oligomer. The change is measured by Tanimoto index between the fingerprints of the monomer and the oligomer (n from 2 to 10), which represents the similarity of the two fingerprints. Three RDKit fingerprints implemented in XenonPy are used: *RDKitFP* denotes the standard fingerprint, *ECFP* denotes the Morgan fingerprint, and *MACCS* denotes the MACCS keys.

may become too big that conventional descriptors or fingerprints are not efficient anymore. Efficient descriptor for block or graft copolymers is yet to be found. Embedding methods using neural network can be a potential option, but problems such as invariance to the input order of monomers or generality to different types of chemical reactions used to form the polymers using multiple monomers remain to have no clear solution.

Predictive models

One of the key components in polymer design is the prediction of polymer properties relevant to a target application. Some of the important properties often considered in the polymer industry include transparency, glass transition temperature, toughness, etc.⁴¹ The ability to predict properties of different classes of polymers opens up new computational design approaches for polymers, such as high-throughput screening and inverse design, which will be discussed in the next section. One approach to predict polymer properties is building empirical equations based on physical descriptors and some basic properties of polymers.

An effort to collect such equations in Python has been made by the *thermo* package.⁴² This approach can be very powerful, but could be difficult to use when predicting properties for a new design with not much knowledge available yet. Another approach that is more suitable for design purpose, called the quantitative structure–property relationship (QSPR) modeling, aims at mapping structural descriptors of a material to its property. In particular, models that rely only on 2D structure information of the polymer is preferred as it avoids intensive geometry optimization for the polymer molecules. To date, around 300 articles can be found on the web of science database when searching with keywords QSPR and polymer.

Group contribution method is one of the earliest data-driven approaches to predict complex properties of different polymers.³ The basic idea is that certain groups of bonded atoms within different molecules may have common effects on a specific material property. Linear regression is used to model the interaction between groups when data is limited, but higher order models can also be used. Machine learning models using fingerprints as input can be seen as an extension to this idea, since the fingerprints are collections of rules that check the existence of different structural groups. Some commonly used models include elastic net, support vector machine, random forest, Gaussian process, neural network, etc. Table 2 summarizes some recent applications of these algorithms in polymer informatics.

Machine learning models are inherently interpolative, i.e., their predictions are reliable only around the domain close to the training data. The range of properties prediction for certain types of polymers with a reasonable accuracy governs the potential search space one can cover, i.e., the performance of the final design. In other words, the training data determines the feasible design space of the target application. The concept of applicability domain (AD) developed in Cheminformatics is used to quantify the reliable region of a QSPR model.⁴⁹ The concept of uncertainty in statistics is also a popular metric believed to be strongly correlated with the validity of a prediction.⁵⁰ Figure 4 shows how different models fail to predict different materials properties when extrapolating from the given data. One idea to tackle this issue is called transfer learning, that is to exploit information learned from

Table 2: Examples of QSPR for different polymer properties using machine learning technologies. Corresponding references are cited in the property column. For properties, ΔE denotes atomization energy, ϵ_{gap} denotes bandgap, κ denotes dielectric constant, ρ denotes density, HOMO denotes highest occupied molecular orbital, LUMO denotes lowest unoccupied molecular orbital, ϵ_{opt} denotes optical gap, η denotes refractive index, δ denotes solubility parameter, T_g denotes glass transition temperature, E_g denotes glass modulus, E_r denotes rubber modulus, and $\tan\delta_{max}$ denotes peak height of viscoelastic loss tangent. For descriptors, *Mix* denotes a mix of various descriptors specified by Kim et al.,²⁰ *ICD* denotes the infinite chain descriptors,¹⁹ *Str* denotes customized strings by Jørgensen et al.,⁴³ *D&P* denotes a combination of the Dragon⁴⁴ and PaDEL⁴⁵ descriptors, and *Img* denotes direct use of 2D microstructure images. For models, *GP* denotes Gaussian process, *SVM* denotes support vector machine, *PLS* denotes partial least squares regression, *VAE* denotes using the best regression model based on the hidden layer of a variational autoencoder as described by Jørgensen et al.,⁴³ and *CNN* denotes a multi-task learning convolutional neural network. For test method to calculate root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2) of the QSPR models, *CV-5* denotes a 5-fold cross validation, *Split-X* denotes a X% random splitting of test data from the full data set, and *Select-27* denotes manually picking 27 data points as test data. (* Mean absolute percentage error is measured for these studies)

Property	Data size	Descriptor	Model	Test method	RMSE	MAE	R^2	Unit
ΔE^{20}	392	Mix	GP	CV-5	0.01	0.01	0.999	eV/atom
ϵ_{gap}^{19}	155	ICD	SVM	Split-20	—	—	0.88	eV
ϵ_{gap}^{20}	382	Mix	GP	CV-5	0.3	0.23	0.971	eV
ϵ_{gap}^{43}	3,989	Str	VAE	CV-5	—	74	—	meV
κ^{19}	155	ICD	SVM	Split-20	—	—	0.96	—
κ^{20}	384	Mix	GP	CV-5	0.48	0.32	0.815	—
ρ^{20}	173	Mix	GP	CV-5	0.05	0.03	0.938	g/cm ³
HOMO ⁴³	3,989	Str	VAE	CV-5	—	66	—	meV
LUMO ⁴³	3,989	Str	VAE	CV-5	—	43	—	meV
ϵ_{opt}^{43}	3,989	Str	VAE	CV-5	—	70	—	meV
η^{20}	384	Mix	GP	CV-5	0.08	0.05	0.892	—
η^{46}	221	D&P	PLS	Split-30	—	0.004	0.899	—
η^{47}	527	Mix	GP	Select-27	0.05	—	0.88	—
δ^{20}	113	Mix	GP	CV-5	0.56	0.4	0.955	MPa ^{1/2}
T_g^{19}	270	ICD	SVM	Split-20	—	—	0.95	K
T_g^{20}	451	Mix	GP	CV-5	17.74	12.79	0.944	K
E_g^{48}	11,000	Img	CNN	Split-15	—	0.68	—	%*
E_r^{48}	11,000	Img	CNN	Split-15	—	3.12	—	%*
$\tan\delta_{max}^{48}$	11,000	Img	CNN	Split-15	—	3.58	—	%*

a relevant task for improving prediction of another task. Yamada et al. demonstrated the successful applications of transfer learning in different materials science problem, including polymers.⁵¹ Intuitive ideas for knowledge transfer include transferring from a global material space to a local domain, from a material property with rich data to a physically linked property with little data, or from computational data to experimental data. The latter idea is also called multi-fidelity learning, which has been successfully applied to predict crystallization tendency⁵² and bandgap⁵³ of polymers.

Polymer design

While there are increasing examples of machine-assisted polymer design, there has not been any report of end-to-end design example that covers every step from monomer design to manufacturing process. Instead, polymer informatics has been used to improve design efficiency within each step of the polymer design process. For example, Wu et al. discovered new homopolymers with high thermal conductivity that are validated experimentally¹⁵ and Li et al. developed an algorithm that discovers optimal strategy to experimentally control MWDs of various polymers.²² There are three types of design strategies based on machine learning technology: high-throughput screening, inverse design, and experimental design. We will discuss various efforts of applying these methods to polymer design in this section.

High-throughput screening

High-throughput screening aims at identifying potential candidates of interest by conducting computational or experimental tests on a large pool of candidates. In polymer informatics, a library of chemically or synthetically feasible polymer candidates is built and then computationally screened by predictive models relevant to the target material properties. When the search space is finite and tractable, the library is simply composed of the exhaustive list of candidates, such as selection of optimal processing method within the existing technologies.

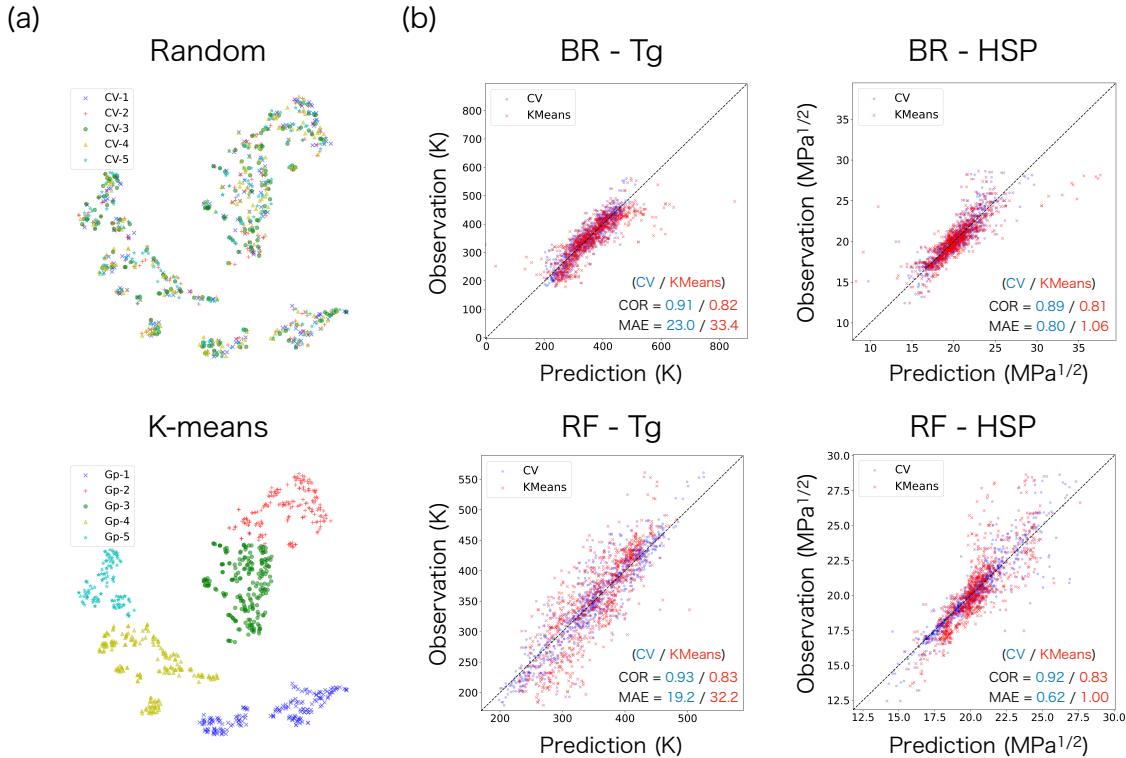


Figure 4: Demonstrating extrapolation power of machine learning models using data from Polymer Genome.²⁰ (a) Visualization of a 2D projection of the 200 physical descriptors in RDKit using t-distributed Stochastic Neighbor Embedding⁵⁴ with perplexity = 30. The data is further separated into five groups, either randomly or through K-means clustering⁵⁵ with number of cluster = 5. The groupings are used for cross validation to test the prediction performance on glass transition temperatures (Tg) and Hildebrand solubility parameter (HSP). (b) Plots of predictions versus observed data for Tg and HSP based on Bayesian ridge regression (BR) and random forest (RF). Crosses are results from cross validation based on K-means clustering, i.e., each set of test data is not similar to the training data (extrapolation). Circles are results from cross validation based on random picking. The extrapolation case has a worse prediction accuracy all four cases, where the failure patterns are different between BR and RF.

One can also use a library from any open databases. Otherwise, a generative algorithm is needed to construct a pool of candidates of interest. For example, a conventional approach is to define a set of molecular fragments of interest and exhaustively combine different numbers of fragments up to an upper limit to form the candidate set. Fragment can be obtained from organic molecules which have more large open databases, such as GDB-17⁵⁶ or Pubchem.⁵⁷ To expand beyond the search space limited by a finite variety of fragments, Wu et al. implemented a language model proposed by Ikebata et al.⁵⁹ on polymers represented in SMILES.⁵⁸ Different deep neural networks are also used to learn the SMILES or graph representation of organic molecules.^{60–62}

Giving a library of candidates, one can use pretrained predictive models to computationally screen out candidates with target properties. There is a significant amount of literature that applies high-throughput screening to different polymer applications. Recent examples of high-throughput screening in polymer designs using machine learning models include searching of high refractive index polymers^{46,63,64} and screening optoelectronic properties of conjugated polymers.⁶⁵ This strategy of polymer design may sound inefficient to solve a single design problem, because a significantly large library is needed to increase the probability of finding candidates of interest. However, a well-developed library that contains diverse candidates can be reused for many different design problems. With a rich database of candidates, high-throughput screening is a very attractive tool in many industrial applications.

Inverse design

Another design strategy, referred as the inverse design, is to perform targeted search in a materials space guided by knowledge extracted from existing data. While predictive model is a mapping from an input x (polymer) to an output y (material property), the goal of inverse design is to map a targeted range of y to a sub-domain of x . This can be achieved by solving an optimization problem using sophisticated algorithms, such as genetic algorithms, or by sampling the set of x with high probability to be in the targeted range of y under a

Bayesian framework. Both approaches can be implemented in an iterative algorithm:

1. Start with a set of initial candidates.
2. Propose a new set of candidates based on the existing ones.
3. Evaluate likeliness of the candidates to have the desired material properties.
4. Repeat step 2 and 3 for fixed number of times or until convergence.

In step 2, a generative algorithm is needed to propose new candidates similar to the one in the high-throughput screening. In the case of optimization, the likeliness metric in step 3 is usually a loss function measuring the distance of the predicted properties for the candidates and the targeted properties. For sampling, the likeliness metric is usually correlated with the probability of observing the candidates conditional on the targeted properties.

The inverse design problem is inherently ill-posed, i.e., there are different types of materials that may have the target properties. It is important that an inverse design algorithm can produce a diverse set of candidates in order to maximize the probability of discovering novel functional materials. Typical optimization algorithm can only search for the optimal solution. Sampling methods can have a higher chance to provide diverse candidates, but can also be trapped in a local mode, especial when the search space is high dimensional. Many efforts have been made to address this issue, such as limiting the search space size, projecting the search space to a low dimensional space, or employing an annealing algorithm, etc. Table 3 shows some recent examples of successful inverse design of different polymer applications using machine learning technologies.

Experimental design

The previous two design strategies relies heavily on efficiency of the generative model and accuracy of the predictive model, which, in turn, relies on the quality and quantity of training data. Since many properties of polymers have only limited data, increasing data size with

Table 3: Examples of polymer inverse design using machine learning technologies.

Publication	Target property	Method
Mannodi-Kanakkithodi et al. (2016) ⁶⁶	Dielectric constant and bandgap	Optimization using genetic algorithm
Jørgensen et al. (2018) ⁴³	LUMO and optical gap energy	Gradient-based optimization on embedded space in deep neural network
Pilania et al. (2019) ⁶⁷	Glass transition temperature	Optimization using genetic algorithm
Kumar et al. (2019) ⁶⁸	Phase behavior	Optimization using particle swarm optimization
Wu et al. (2019) ¹⁵	Thermal conductivity	Sampling with sequential Monte Carlo
Schadler et al. (2020) ⁶⁹	Three different dielectric properties	Optimization using genetic algorithm
Wu et al. (2020) ⁵⁸	Dielectric constant and bandgap	Sampling with sequential Monte Carlo

extra experimental or computational tests is inevitable to ensure that the machine learning models can cover a large enough design space. Instead of randomly picking candidates to perform the new tests, one can employ a recursive design process, i.e., new test candidates are optimally selected and tested, and the results are added to the existing data set for improving the optimal candidate selection in the next round of tests. Such trial-and-error design process is what a chemist would do in practice.

The goal of experimental design is to minimize the amount of new experiments required to reach a design goal. There are two perspectives to the optimality of candidate selection: (1) exploit the knowledge embedded in the existing data to make the best guess of which candidates may satisfy the design goal, or (2) explore candidates with the least information from the existing data to infer their properties. In a typical experimental design algorithm, we define a utility function that balance the tradeoff between the two perspectives. Candidates that optimize the utility function are selected for the next round of tests. Bayesian optimization and reinforcement learning are two commonly used algorithms for experimental

design. The former considers candidates with a high prediction uncertainty as exploratory and optimize the utility function accordingly. The latter treats the problem as a game with reward when achieving the design goal, where an agent is trying to learn the best strategy (minimum number of experiments) to get the maximum reward (reaching the design goal). Table 4 shows some recent examples of experimental design applications on different stages of the hierarchical polymers design process.

Table 4: Examples of polymer experimental design using machine learning technologies.

Publication	Target	Search space	Method
Li et al. (2017) ⁷⁰	Median length, median diameter and quality of fibers	Five synthetic process parameters	Bayesian optimization
Li et al. (2018) ²²	Shape of MWD	Amount of five chemical reagents	Reinforcement learning
Wang et al. (2018) ⁷¹	Interphase properties (dielectric and viscoelastic)	Hyperparameters in two interphase models	Bayesian optimization
Minami et al. (2019) ⁷²	Glass transition temperature	Mixing ratios of three selected polymers	Bayesian optimization
Kim et al. (2019) ⁷³	Glass transition temperature	736 predefined candidates in database	Bayesian optimization

Experimental design algorithm is the ideal solution when we do not have a large enough polymer database to begin with. However, polymer experiments are costly and syntheses of new polymers are difficult. Computational simulations are becoming more and more accessible, yet calculations for new polymers often required labor-intensive tuning of model parameters. Automatic simulation, synthetic planning, and property measurement for polymers remain challenging and are continuously explored by researchers in polymer informatics.

Discussion

Polymer informatics is a promising tool for discovery of novel polymers. With a sufficiently large data set to support the use of modern machine learning technology, a data-driven

approach of polymer design will significantly improve the pace of making new functional polymers, satisfying the rapidly expanding demand on polymeric materials in modern society (e.g., deformable electronic devices⁷⁴). One of the most important elements of polymer informatics is the availability of large open databases. Building such databases for polymer is challenging because of many reasons: (1) difficulty of encoding the hierarchical structure of polymers for machine learning purposes, (2) inconsistent naming rules throughout the history of polymer science, (3) lack of data sharing due to many privately own industrial data, etc. Nonetheless, this is an essential step towards a fully data-driven design process. Many efforts have been made to build the backbone technologies necessary to exploit the potential benefit of polymer informatics, such as developing new descriptors to better capture the physical properties of polymers and new simulation methods to estimate polymer properties with higher accuracy in a shorter time. The true power of polymer informatics is to release polymer scientists from the low efficiency trial-and-error design process, thus, to free up more time for higher level design concepts and theoretical advancements. Such opportunity can be realized only if we work together to push for a more open community in polymer science, where everyone can benefit from the new paradigm of polymer design. An easy first step to take would be to engage in the field of polymer informatics and experience the new way of studying polymers ourselves.

Acknowledgement

This work was supported in part by the “Materials Research by Information Integration” Initiative (MI2I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). S.W. gratefully acknowledges financial support from JSPS KAKENHI Grant Number JP18K18017. R.Y. acknowledges financial support from a Grant-in-Aid for Scientific Research (B) 15H02672 and a Grant-in-Aid for Scientific Research (A) 19H01132 from the Japan Society for the Promotion of Science (JSPS).

References

- (1) Feldman, D. Polymer History. *Designed Monomers and Polymers* **2008**, *11*, 1–15.
- (2) Flory, P. J. *Statistical Mechanics of Chain Molecules*; John Wiley and Sons, New York, 1969.
- (3) van Krevelen, D. W.; te Nijenhuis, K. *Properties of Polymers: their Correlation with Chemical Structure; their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions*, 4th ed.; Elsevier, Amsterdam, 2009.
- (4) Bicerano, J. *Prediction of Polymer Properties*; Marcel Dekker, New York, 2002.
- (5) Saha, S.; Bhowmick, A. K. An Insight into molecular structure and properties of flexible amorphous polymers: A molecular dynamics simulation approach. *Journal of Applied Polymer Science* **2019**, *136*, 47457.
- (6) Steinhauser, M.; Hiermaier, S. A Review of Computational Methods in Materials Science: Examples from Shock-Wave and Polymer Physics. *International Journal of Molecular Sciences* **2009**, *10*, 5135–5216.
- (7) Gartner, T. E.; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* **2019**, *52*, 755–786.
- (8) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Letters* **2017**, *6*, 1078–1082.
- (9) Kumar, J. N.; Li, Q.; Jun, Y. Challenges and opportunities of polymer design with machine learning and high throughput experimentation. *MRS Communications* **2019**, *9*, 537–544.

- (10) Imrie, C. T.; Karasz, F. E.; Attard, G. S. The Effect of Molecular Weight on the Thermal Properties of Polystyrene-Based Sidechain Liquid-Crystalline Polymers. *Journal of Macromolecular Science—Pure and Applied Chemistry* **1994**, *31*, 1221–1232.
- (11) Nunes, R. W.; Martin, J. R.; Johnson, J. F. Influence of molecular weight and molecular weight distribution on mechanical properties of polymers. *Polymer Engineering & Science* **1982**, *22*, 205–228.
- (12) Fetters, L. J.; Lohse, D. J.; Richter, D.; Witten, T. A.; Zirkel, A. Connection between Polymer Molecular Weight, Density, Chain Dimensions, and Melt Viscoelastic Properties. *Macromolecules* **1994**, *27*, 4639–4647.
- (13) Yi, A.; Chae, S.; Hong, S.; Lee, H. H.; Kim, H. J. Manipulating the crystal structure of a conjugated polymer for efficient sequentially processed organic solar cells. *Nanoscale* **2018**, *10*, 21052–21061.
- (14) Pascu, M.; Vasile, C. *Practical Guide to Polyethylene*; Smithers Rapra Publishing, 2005.
- (15) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Computational Materials* **2019**, *5*, 66.
- (16) National Institute for Materials Science, PoLyInfo. http://polymer.nims.go.jp/index_en.html **2011**, Last checked: September 25, 2020.
- (17) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: polymer database for polymeric materials design. 2011 Int. Conf. on Emerg. Intell. Data Web Technol. Tirana, Albania, 2011; pp 22–29.
- (18) Liu, C.; Wu, S.; Yoshida, R. XenonPy. <https://xenonpy.readthedocs.io/en/latest/> **2016**, Last checked: September 25, 2020.

- (19) Wu, K.; Sukumar, N.; Lanzillo, N. A.; Wang, C.; “Rampi” Ramprasad, R.; Ma, R.; Baldwin, A. F.; Sotzing, G.; Breneman, C. Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. *Journal of Polymer Science Part B: Polymer Physics* **2016**, *54*, 2082–2091.
- (20) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *The Journal of Physical Chemistry C* **2018**, *122*, 17575–17585.
- (21) Xu, X.; Wei, Q.; Li, H.; Wang, Y.; Chen, Y.; Jiang, Y. Recognition of polymer configurations by unsupervised learning. *Physical Review E* **2019**, *99*, 043307.
- (22) Li, H.; Collins, C. R.; Ribelli, T. G.; Matyjaszewski, K.; Gordon, G. J.; Kowalewski, T.; Yaron, D. J. Tuning the molecular weight distribution from atom transfer radical polymerization using deep reinforcement learning. *Molecular Systems Design & Engineering* **2018**, *3*, 496–508.
- (23) NIST, Synthetic Polymer MALDI Recipes Database. <https://maldi.nist.gov/> **2014**, *Last checked: September 25, 2020*.
- (24) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Scientific Data* **2016**, *3*, 160012.
- (25) Zhao, H.; Li, X.; Zhang, Y.; Schadler, L. S.; Chen, W.; Brinson, L. C. Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design. *APL Materials* **2016**, *4*, 053204.
- (26) Zhao, H.; Wang, Y.; Lin, A.; Hu, B.; Yan, R.; McCusker, J.; Chen, W.; McGuinness, D. L.; Schadler, L.; Brinson, L. C. NanoMine schema: An extensible data representation for polymer nanocomposites. *APL Materials* **2018**, *6*, 111108.

- (27) Ellis, B.; Smith, R. *Polymers: A Property Database*, 2nd Edition; CRC Press, 2020.
- (28) Oliver, S.; Zhao, L.; Gormley, A. J.; Chapman, R.; Boyer, C. Living in the Fast Lane—High Throughput Controlled/Living Radical Polymerization. *Macromolecules* **2019**, *52*, 3–23.
- (29) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A mobile robotic chemist. *Nature* **2020**, *583*, 237–241.
- (30) Baer, E.; Hiiltner, A.; Keith, H. D. Hierarchical structure in polymeric materials. *Science* **1987**, *235*, 1015–1022.
- (31) Adams, N.; Winter, J.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *Journal of Chemical Information and Modeling* **2008**, *48*, 2118–2128.
- (32) Zhou, T.; Song, Z.; Sundmacher, K. Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design. *Engineering* **2019**, *5*, 1017–1026.
- (33) Hart, L. R.; Harries, J. L.; Greenland, B. W.; Colquhoun, H. M.; Hayes, W. Molecular design of a discrete chain-folding polyimide for controlled inkjet deposition of supramolecular polymers. *Polymer Chemistry* **2015**, *6*, 7342–7352.
- (34) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017**, *3*, 54.
- (35) Brough, D. B.; Wheeler, D.; Kalidindi, S. R. Materials Knowledge Systems in Python—a Data Science Framework for Accelerated Development of Hierarchical Materials. *Integrating Materials and Manufacturing Innovation* **2017**, *6*, 36–53.

- (36) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *In Advances in Neural Information Processing Systems 25* **2012**, 1097–1105.
- (37) Buchet, M.; Hiraoka, Y.; Obayashi, I. In *Nanoinformatics*; Tanaka, I., Ed.; Springer Singapore: Singapore, 2018; pp 75–95.
- (38) Vishwanathan, S.; Schraudolph, N. N.; Kondor, R.; Borgwardt, K. M. Graph Kernels. *Journal of Machine Learning Research* **2010**, *11*, 1201–1242.
- (39) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- (40) Miccio, L. A.; Schwartz, G. A. From chemical structure to quantitative polymer properties prediction through convolutional neural networks. *Polymer* **2020**, *193*, 122341.
- (41) Willbourn, A. H. Molecular design of polymers. *Polymer* **1976**, *17*, 965–976.
- (42) Caleb Bell and Contributors, thermo: Chemical properties component of Chemical Engineering Design Library (ChEDL). <https://github.com/CalebBell/thermo> **2016–2020**, *Last checked: September 25, 2020*.
- (43) Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine learning-based screening of complex molecules for polymer solar cells. *The Journal of Chemical Physics* **2018**, *148*, 241735.
- (44) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH Communications in Mathematical and in Computer Chemistry* **2006**, *56*, 237–248.
- (45) Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32*, 1466–1474.

- (46) Khan, P. M.; Rasulev, B.; Roy, K. QSPR Modeling of the Refractive Index for Diverse Polymers Using 2D Descriptors. *ACS Omega* **2018**, *3*, 13374–13386.
- (47) Lightstone, J. P.; Chen, L.; Kim, C.; Batra, R.; Ramprasad, R. Refractive index prediction models for polymers using machine learning. *Journal of Applied Physics* **2020**, *127*, 215105.
- (48) Wang, Y.; Zhang, M.; Lin, A.; Iyer, A.; Prasad, A. S.; Li, X.; Zhang, Y.; Schadler, L. S.; Chen, W.; Brinson, L. C. Mining structure–property relationships in polymer nanocomposites using data driven finite element analysis and multi-task convolutional neural networks. *Molecular Systems Design & Engineering* **2020**, *5*, 962–975.
- (49) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1912–1928.
- (50) Chatfield, C. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **1995**, *158*, 419–466.
- (51) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Central Science* **2019**, *5*, 1717–1730.
- (52) Venkatram, S.; Batra, R.; Chen, L.; Kim, C.; Shelton, M.; Ramprasad, R. Predicting Crystallization Tendency of Polymers Using Multifidelity Information Fusion and Machine Learning. *The Journal of Physical Chemistry B* **2020**, *124*, 6046–6054.
- (53) Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Computational Materials Science* **2020**, *172*, 109286.

- (54) van der Maaten, L. J. P.; Hinton, G. E. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
- (55) Lloyd, S. P. Least squares quantization in PCM. *Information Theory, IEEE Transactions* **1982**, *28*, 129–137.
- (56) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.
- (57) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Research* **2016**, *44*, D1202–1213.
- (58) Wu, S.; Lambard, G.; Liu, C.; Yamada, H.; Yoshida, R. iQSPR in XenonPy: A Bayesian Molecular Design Algorithm. *Molecular Informatics* **2020**, *39*, 1900107.
- (59) Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian molecular design with a chemical language model. *Journal of Computer-Aided Molecular Design* **2017**, *31*.
- (60) Cao, N. D.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *ArXiv* **2018**, *abs/1805.11973*.
- (61) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USA* **2018**, 6412–6422.
- (62) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **2018**, *4*, eaap7885.

- (63) Jabeen, F.; Chen, M.; Rasulev, B.; Ossowski, M.; Boudjouk, P. Refractive indices of diverse data set of polymers: A computational QSPR based study. *Computational Materials Science* **2017**, *137*, 215–224.
- (64) Afzal, M. A. F.; Haghighatlari, M.; Ganesh, S. P.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *The Journal of Physical Chemistry C* **2019**, *123*, 14610–14618.
- (65) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers. *Journal of Chemical Information and Modeling* **2018**, *58*, 2450–2459.
- (66) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Scientific Reports* **2016**, *6*, 20952.
- (67) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *Journal of Chemical Information and Modeling* **2019**, *59*, 5013–5025.
- (68) Kumar, J. N.; Li, Q.; Tang, K. Y. T.; Buonassisi, T.; Gonzalez-Oyarce, A. L.; Ye, J. Machine learning enables polymer cloud-point engineering via inverse design. *npj Computational Materials* **2019**, *5*, 73.
- (69) Schadler, L. S.; Chen, W.; Brinson, L. C.; Sundararaman, R.; Gupta, P.; Prabhune, P.; Iyer, A.; Wang, Y.; Shandilya, A. A perspective on the data-driven design of polymer nanodielectrics. *Journal of Physics D: Applied Physics* **2020**, *53*, 333001.
- (70) Li, C.; Rubín de Celis Leal, D.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.;

- Height, M.; Venkatesh, S. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Scientific Reports* **2017**, *7*, 5683.
- (71) Wang, Y.; Zhang, Y.; Zhao, H.; Li, X.; Huang, Y.; Schadler, L. S.; Chen, W.; Brinson, L. C. Identifying interphase properties in polymer nanocomposites using adaptive optimization. *Composites Science and Technology* **2018**, *162*, 146–155.
- (72) Minami, T.; Kawata, M.; Fujita, T.; Murofushi, K.; Uchida, H.; Omori, K.; Okuno, Y. Prediction of repeat unit of optimal polymer by Bayesian optimization. *MRS Advances* **2019**, *4*, 1125–1130.
- (73) Kim, C.; Chandrasekaran, A.; Jha, A.; Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Communications* **2019**, *9*, 860–866.
- (74) McBride, M.; Liu, A.; Reichmanis, E.; Grover, M. A. Toward data-enabled process optimization of deformable electronic polymer-based devices. *Current Opinion in Chemical Engineering* **2020**, *27*, 72–80.