ACS POLYMERS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/polymerau

Review

Data-Driven Methods for Accelerating Polymer Design

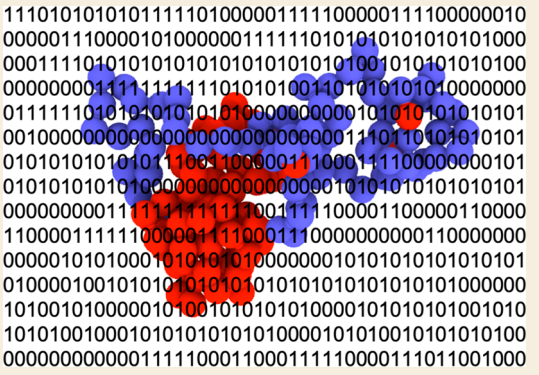
Tarak K. Patra*

Cite This: *ACS Polym. Au* 2022, 2, 8−26

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Optimal design of polymers is a challenging task due to their enormous chemical and configurational space. Recent advances in computations, machine learning, and increasing trends in data and software availability can potentially address this problem and accelerate the molecular-scale design of polymers. Here, the central problem of polymer design is reviewed, and the general ideas of data-driven methods and their working principles in the context of polymer design are discussed. This Review provides a historical perspective and a summary of current trends and outlines future scopes of data-driven methods for polymer research. A few representative case studies on the use of such data-driven methods for discovering new polymers with exceptional properties are presented. Moreover, attempts are made to highlight how data-driven strategies aid in establishing new correlations and advancing the fundamental understanding of polymers. This Review posits that the combination of machine learning, rapid computational characterization of polymers, and availability of large open-sourced homogeneous data will transform polymer research and development over the coming decades. It is hoped that this Review will serve as a useful reference to researchers who wish to develop and deploy data-driven methods for polymer research and education.



KEYWORDS: *Polymer Design, Machine Learning, AI, One-Hot Encoding, Polymer Genome*

1. INTRODUCTION

Polymers are highly correlated many-body systems with complex structures and dynamics spanning a wide range of length and time scales. Their relaxation processes involve complex phenomena such as vitrification, jamming, gelation, and semicrystallization, which are highly process-dependent and for which no comprehensive theoretical framework and understanding exist. These phenomena are strongly influenced by the chemical details of a polymer's building blocks. Optimal use of polymers in future technologies such as electronics, medicine, and energy devices demands a deeper understanding of the connections between molecular chemistry and materials processing while establishing strategies for their rapid and rational design.[1−3] Advancement in experimental, theoretical, and computational polymers research and their close integration can potentially improve the current understanding of polymers and accelerate their design by generating a large volume of data. However, rapid production of polymer structure−property data and their utilization for new material development require synergies among several fields of STEM (Science Technology Engineering Mathematics) including data science, high-performance computing, machine learning, numerical optimization, automation science, and materials informatics along with polymer physics, chemistry, and processing. There are several challenges in integrating these disciplines including translating polymer chemistry into machine readable fingerprints, standardization of materials data formats, data sharing, mining and learning from data that are explainable and interpretable, developing transferable predictive models from data, and reverting fingerprints into chemical formulas.[4−9] Addressing these challenges is vital for moving from traditional trial-and-error polymer development toward rational data-driven polymer design.

Unlike biomolecules where informatics-based design has been accelerated by the availability of large, open, and homogeneous structure−property relation databases such as the Protein Data Bank,[10,11] similar data banks in synthetic polymers are sparse, heterogeneous, and often outright unavailable.[12,13] Moreover, due to a general lack of rapid, parallelizable techniques for measuring polymer properties, building such databases de novo for a specific design problem is often challenging and resource-prohibitive. Molecular simulations, machine learning (ML), and advanced optimization methods can potentially address these challenges and accelerate polymer design. Such data-driven materials design strategies are on the rise, and they have shown

Received: September 20, 2021
Revised: December 6, 2021
Accepted: December 6, 2021
Published: December 28, 2021

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

success in many materials design problems.[14−28] Toward this end, in recent times, large materials databases have been created to accelerate this data-driven paradigm of material research. A list of publicly available databases can be seen elsewhere.[26,29] Although such data-driven approaches were initially started for inorganic materials and small organic molecules,[7,30−35] they have subsequently gained popularity and potential in polymer and soft matter research.[20,36−41] Many databases are now available for synthetic polymeric materials as well.[42−46] These databases contain mostly computationally estimated structure−property data and a lesser amount of experimentally measured data. This is because the high-throughput large-scale data generation via experimentation is still a substantial challenge. However, computational methods are now routinely used to generate large-scale materials data. In particular, molecular modeling and simulation of polymers have now grown into a matured field and provide efficient methods for the reliable estimation of polymer properties.[47−63] These methods are now implemented in open-sourced software that are highly scalable with system size in a HPC (High Performance Computing) environment. Some of the popular large-scale classical polymer simulation packages are LAMMPS,[64,65] GROMACS,[66,67,67] NAMD,[68,69] DL_POLY,[70] DL_MONTE,[71] and Cassandra.[72,73] These classical simulation packages can be utilized for reliable estimation of a wide range of equilibrium and steady-state properties including structure factor, radius of gyration, phase equilibria, viscosity, thermal conductivity, surface tension, ion diffusion, gas permeation, and many other properties of polymeric materials. Quantum level simulations can be performed for band gap, dielectric constant, refractive index, and other properties relevant to the electronic scale. VASP,[74,75] QUANTUM ESPRESSO,[76−78] and Gaussian[79] are very commonly used packages for quantum chemical simulations. These packages have been used regularly to perform large-scale molecular simulations in advanced computing hardware with a large number of CPU (central processing unit) cores in combination with powerful GPUs (graphics processing units) in MPI (message passing interface) architectures for rapid characterization of large number polymer systems simultaneously. The large amounts of simulated data can be utilized to downselect a few promising candidates for a target application and direct the experimental investigations to a very narrow region of the search space. Hence, this increasing trend in the availability of polymer data and rapid computation of polymer properties are poised to shorten the materials development time scale, which is typically 15−25 years.[80,81]

However, the optimal design of polymers cannot be addressed solely by building large amounts of computational or experimental databases. The primary bottleneck in the optimal design of a polymer is its astronomically large combinatorial sequence space. For instance, a linear copolymer chain with an n number of possible monomers and m type of chemical moieties will have $m^n/2$ sequences. The denominator is to eliminate the double-counting of a polymer, as a sequence and its reverse sequence represent one copolymer. Even for an AB type copolymer, i.e., two types of chemical moieties with a chain length of 50, the total combination is 2^{49}, which is over 10^{15}. In practice, many copolymers possess more than two chemical moieties and several hundreds of monomer units. Given that such an enormous sequence space needs to be explored to identify the best candidate for a given application, it is highly desirable to minimize the number of property measurements (computer simulations or experiments) to complete a design task within a reasonable time. ML and advanced optimization techniques can play an important role in reducing the number of property measurements in a design cycle.[82,83] Pre-existing and open-source data (both experimental and simulation data) can serve as a starting point for such rational exploration of polymer search space.[39,84−86] More importantly, global optimization techniques and ML models can guide the data generation toward unknown regions of a polymer's physicochemical space and, thus, help in avoiding repeated sampling in the same region of the space. How much data are required to identify an optimal polymer and how to choose an efficient algorithm for a specific polymer design problem are important open questions in data-driven polymer research.

This Review surveys the multiscale featurization of polymers, summarizes ML methods for predicting polymer properties, and examines various ways of integrating first-principle methods, ML, and optimization methods for polymer design problems. I note that DFT, MD, and other physics-based theoretical computations and experimental measurements of polymer properties are referred as first-principle methods in this Review. A few representative case studies on the application of data-driven methods in polymer design problems are discussed, and their key findings are reported. I note that Ferguson has reported an excellent review article which can be referenced for an introduction to machine learning methods, particularly unsupervised methods, that are suitable for soft matter research.[87] Here, I particularly focus on the supervised methods and other data-driven strategies that are suitable for the design of polymers. Furthermore, Audus and de Pablo have reported a viewpoint on some of the challenges and opportunities in polymers informatics.[88] Jackson, Webb and de Pablo have also reported a review of recent advances in machine learning toward multiscale soft materials design.[37] Ramprasad and co-workers have reviewed the roadmap for rational polymer design.[38,89] Along these lines, there are other excellent review articles that focus on the applications of ML and data science in various polymers and soft matter research.[90−93] This review aims to provide complementary perspectives of data-driven polymer research and highlights some of the recent works on data-driven synthetic polymers design that have not been covered in previous review articles. It specifically discusses the application of data-driven methods for single-chain polymer design, polymer-membrane, polymer compatibilizer, polymer dielectrics and heat conducting polymer research.

The flow of the article is organized as a roadmap for a polymer design study. First, the fingerprinting of a polymer and defining its design space are discussed in section 2. Then, the data generation and ML model development procedures are described in section 3. Section 4 presents the polymer design workflows that integrate optimization algorithms, ML models, and/or first-principle methods. Five representative case studies are presented in section 5. In each of these sections, current practices, limitations, and yet-to-be-solved challenges are discussed and some interesting and important works that have attempted to tackle these challenges are cited. Moreover, future research opportunities are highlighted in section 6. I hope that this Review will be useful to make an informed-decision to set up a new polymer design study from the plethora of tools that are offered by data science and computational techniques and to stimulate methodological advancement for efficient data-driven polymer design.

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

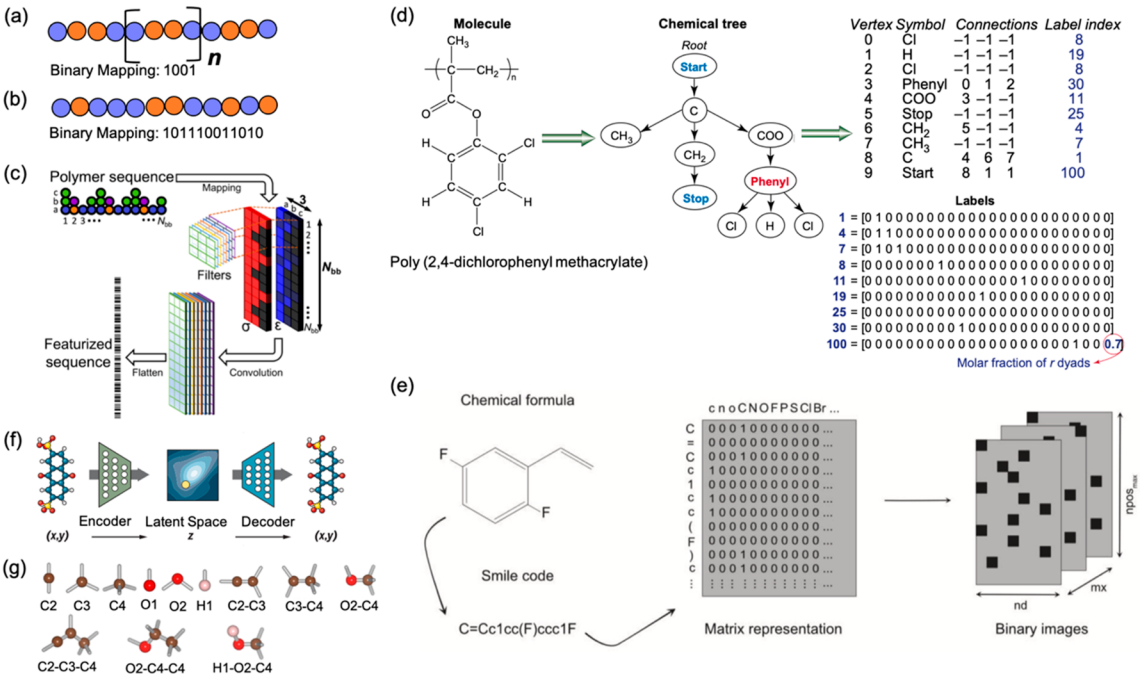


Figure 1. Feature space representation of polymers. (a,b) Binary string representation of a copolymer with two chemical moieties. (c) Property coloring featurization of a coarse-grained polymer. Adapted with permission from ref 39. Copyright 2020 The Authors. (d) Poly(2,4-dichlorophenyl methacrylate) structure, its chemical tree, numerical labels and connection table. Adapted with permission from ref 95. Copyright 2007 Elsevier. (e) Converting chemical formula into a binary image for CNN model development. Adapted with permission from ref 97. Copyright 20202 Elsevier. (f) Schematic representation of an autoencoder that maps the discrete molecular representation to a continuous latent space. Adapted with permission from ref 35. Copyright 2018 AAAS. (g) Molecular motifs of several types: C2, C3, C4, O1, O2 and H1 are atom types, C2−C3, C3−C4 and O2−C4 are bond types, and C2−C3−C4, O2−C4−C4 and H1−O2−C4 are two bonds catenations. Adapted with permission from ref 14. Copyright 2015 American Physical Society.

2. POLYMER FEATURIZATION AND DESIGN SPACE

Optimization algorithms, ML-model development, storing, sorting, and accessing large volumes of structure−property data need to represent polymers in a machine-readable numerical representation.[94] Significant efforts have been dedicated to establish and validate such numerical representation of molecules for machine learning and computer aided design. These numerical representations are known as fingerprints or descriptors or features. There are several classes of mapping reported, and they are typically decided based on the scale of representation. For instance, for a sequence design problem, a very common strategy is to represent the sequence of chemical moieties as a binary number. In several other applications, polymer structures are represented as a graph or image that describes each and every atom of a molecule or a polymer subunit explicitly. Often, a polymer subunit is represented by a chemical tree, where each node is an atom or a group of atoms. Such representation is useful for developing a recursive neural network model for predicting polymer properties.[95] In this regard, one of the popular descriptors of macromolecules is SMILES[96] (Simplified Molecular Input Line Entry System) string representation. The SMILES string representations are heavily used for the chemical space exploration of polymers.[90] The SMILES string representation of a restricted chemical space can be converted into a two-dimensional binary matrix representation that can also serve as an input to machine learning model for predicting polymer properties.[97,98] Similarly, deep learning methods can be employed for autonomous extraction of descriptors.[99,100] This process of utilizing domain knowledge for creating features that will be inputted into a computer algorithm is known as feature engineering in ML literature. As features impact the performance of an ML model, efficient featurization of polymers are very important for data-driven polymer design. Also, the features must be compatible as inputs of an ML model. For example, a numerical representation of a polymer is required for building an artificial neural network (ANN) model, while the convolutional neural network (CNN) needs an image-based representation of the polymer as its input. The ranges of these variables define the feature space of a given polymer. The number of descriptors or the dimension of the feature space of a polymer depends on the nature of exploration. Typically, a design problem starts with defining a search space, which is usually a subspace of the feature space, and restricting the search in a specific region of physicochemical space based on our interests, a priori understanding and feasibility within limited resources. Here, I describe very common and popular featurization schemes that have recently been used for synthetic polymer design at sequence and subunit levels. Featurization at subunit levels is preferred for homopolymers where all the repeating subunits are chemically identical. On the other hand, sequence level featurization is essential for copolymers where multiple chemical moieties are sequentially connected in polymer topologies.

2.1. One-Hot Encoding (OHE) of Polymer Sequence

In machine learning and data science literature, converting text and other categorical data into a numerical representation is commonly called one-hot encoding. One-hot encoding yields a numerical data structure that can be used as the input of an ML model.[101,102] The central part of feature engineering for ML model/algorithm development is OHE. A simple example of OHE is the mapping of a long copolymer with two chemical

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

moieties to a binary number in which "0" and "1" represent two different moieties.[103−106] The length of the binary string is determined by the nature of the polymer. In the case of polymers that consist of a repeating subunit, a binary number equivalent to the repeating subunit will serve as a OHE representation for ML model development, as shown in Figure 1a. In this case, the design space is limited to the repeating unit. The number of candidate polymers will be the total number of permutations within the size of the repeating unit. However, for stochastic sequences, the entire polymer sequence is mapped to a binary string as shown in Figure 1b; and the number of possible candidates grow exponentially with the polymer chain length. Polymers with more than two chemical moieties and their chemical connectivity including branching can be encoded into such binary representation.[39,84]

2.2. Property Coloring

As an alternative to OHE of a polymer sequence, de Pablo and co-workers have proposed a flexible featurization approach, viz., property coloring that encodes a polymer in an image as shown in Figure 1c.[39] The advantage of property coloring is that it accounts for local physical and chemical environments along with the sequence. A polymer is initially mapped to a three-dimensional array. These three dimensions correspond to the number of beads/chemical moieties in a monomer, properties of the beads such as their size and interaction strength, and sequence of monomers in the polymer chain. This array of information is passed through filters to produce a convoluted image. This image of the polymer can be used to build a CNN. This featurization is found to be very efficient for ML model development for coarse-grained polymer systems.[39]

2.3. Chemical Tree

Monomer and/or repeat unit level descriptions can be directly represented by one of the above binary schemes. However, atomic level description and subsequent design of chemical composition of a polymer subunit can be conducted by forming a chemical tree as shown in Figure 1d. The node of the tree is labeled with the appropriate chemical groups. The molecular graph is made by splitting the compound into atomic groups and placing them in the vertexes. Solaro and co-workers have proposed a "1-of-*n*" coding scheme that labels the chemical symbols of the atomic groups by a numerical vector.[95] The chemical information is then stored in a connection table that records the connections among the vertexes and their subtrees. This tree representation of chemical structure is flexible and allows defining molecular fragmentation at the desired level. The tree representation of the unit of a homopolymer is shown to be very efficient for building recursive neural network model (RNN) for predicting polymer properties.[95,107,108] This monomer level fingerprinting can be implemented with the help of open-sourced tools such as RDKit.[109]

2.4. One-Hot Encoding (OHE) of SMILES Strings

The SMILES code is a line notation that encodes a molecular structure into a string of characters. For small molecules and periodic polymers, SMILES serves as an effective descriptive code for text-based ML model development. These SMILES representations can be converted into vector representation for efficient ML algorithm development. For example, Miccio and Schwartz have used an OHE scheme to transform SMILES string into a binary image.[97] The binary image is constructed with the help of a dictionary that includes all possible SMILES characters. As shown in Figure 1e, this encoding starts with establishing a binary matrix of size nd × $npos_{max}$. Here, nd is the number of characters in the dictionary and $npos_{max}$ is the number of characters in a SMILES string of a candidate polymer. All the columns of a given row are populated with "0" except the one where the library character and SMILES string character coincide. Such points of coincidence are assigned a value of "1". Such a binary matrix is converted into a binary image where "0" and "1" correspond to gray and black colors, respectively. This way, all the candidate polymers in a data set can be transformed into binary images, which can serve as input to a CNN model. I note that the SMILES string representation is further extended to bigSMILES representation that is more suitable for nonperiodic and stochastic polymers.[110] The bigSMILES representation provides additional bonding descriptors that define how different monomer units are connected to form a long polymer. However, ML model developments based on bigSMILES representation, and its efficacy are not explored intensively.

2.5. Autoencoding

A discrete molecular representation such as SMILES strings can be converted into continuous variables in a latent space using an autoencoder.[35,100,111] An autoencoder is a special class of neural network that compresses high dimensional data to a lower dimensional representation. As shown in Figure 1f, it consists of an encoder and a decoder. While the encoder maps the input to a lower dimensional representation, the decoder converts the lower dimensional representation to its actual representation. The methodological details of building an autoencoder can be found in our recent work.[99] The lower dimensional representation is known as latent space, and it can be used as an input to a predictive model. This continuous representation of molecules provides an open-ended space and is very efficient for gradient-based optimization and exploration of chemical space.[100]

2.6. Motif-Based Fingerprinting

Ramprasad and co-workers have introduced a motif-based hierarchical fingerprinting method to produce numerical representation of small molecules.[14] The motif-based fingerprinting is developed based on the quantity of a specific atom present in a molecule, its principles of chemical bond formation and coordination with other atoms in the molecule. These chemical principles can be utilized to constitute all possible motifs of a molecule. For example, an oxygen atom can form a bond with either one or two other atoms, and a carbon can form a bond with either two, three, or four other atoms. Therefore, the possible oxygen type motifs are O2 and O1; and carbon type motifs are C2, C3, and C4. Similarly, different chemically possible bonds are considered as bond type motifs, and chemically possible catenations of bonds are considered as angle type motifs. A schematic representation of several motifs is shown in Figure 1g. Now, the chemical compositions and all the possible motifs of a molecule can be converted into fingerprints of different orders. The zeroth order fingerprints of a molecule are simply the fractions of different atomic species present in it. Likewise, first, second, and third order fingerprints are the number of specific atom type, bond type, and angle type motifs in the molecule, respectively, normalized by total number of atoms in the molecule. For obvious reasons, these fingerprints should satisfy several constraints arising from their definition and chemistry. Readers can find all the constrained relations and more details of the motif-based fingerprinting elsewhere.[14] This motif-based fingerprinting concept is further extend for long

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

polymeric systems wherein a building blocks like CH2 and C6H4 can be considered as motifs instead of atoms.[94] In this case, zeroth order fingerprints are the number of different types of building blocks, normalized by the total number of building blocks in a polymer. Subsequently, higher order fingerprints can be generated by counting the number of pairs, triplets, quadruplets and longer segments of building blocks in a polymer. This higher order fingerprints are essential to capture the impact of sequences on polymer properties.[112]

The above featurization schemes are primarily adopted for synthetic polymers. However, for biomolecules, BioVect is a popular featurization scheme.[113] Within the BioVec scheme, gene sequences are represented by GeneVec and protein sequences are represented by ProtVec. For biomolecule-specific featurization, readers are directed to the work of Asgari and Mofrad.[113] Now, given this wide varieties of featurization, how should one decide a featurization scheme for a given problem? How does one select features when the structure−property correlations are not known but the data are available? How many features are sufficient to build an ML model? These are the important considerations for data-driven structure−property model development. Although there is no specific agreed-upon strategy for polymer feature engineering, typically features are decided based on their ability to capture hidden dependency relationships in the data and, most importantly, improve the prediction quality of an ML model. For example, for a polymer compatibilizer, the sequence of monomer is more important than the exact chemical details of a monomer and, therefore, features that uniquely describe the sequence of a given polymer might be sufficient to build a model for predicting interfacial energy of a compatibilizer system. On the other hand, for polymer dielectrics, the chemical composition of a monomer/subunit determines the dielectric constant and band gap of the materials. Therefore, the features should represent the chemical constituents of a polymer segment. It is also important that feature vectors should include the details of chain size along with subunit level chemical composition of a homopolymer if its properties are size dependent. More details of these two design studies are discussed in section 5. Overall, the feature engineering requires domain knowledge and some extent of trial-and-error exercise to identify physically meaningful variables or features that can be correlated to target properties via machine learning.[114−118]

3. PREDICTIVE MACHINE LEARNING MODELS

Numerical features of a polymer can be used to build predictive ML model. It requires leveling a set of features that define a polymer structure by its properties, and a collection of feature-property data is utilized to train and build an ML model. Such an ML model serves as a cheaper, albeit low fidelity, surrogate for the high-fidelity first-principle-based simulations and experiments that are expensive.[89,119−121] In ML-evaluated polymer design, the properties of a large number of candidate polymers are not directly measured using first-principle methods, thereby reducing the cost of an exhaustive search of chemical and sequence space of a polymer. Pre-existing data that are labeled by their property values are used for the training and development of these predictive models. These are called supervised machine learning as the data are labeled by their properties. There are several supervised machine learning methods that can be deployed for building predictive models for polymers. Kernel ridge regression (KRR), support vector machine (SVM), Gaussian process regression (GPR), ANN, and random forest regression are a few common methods that have found application in materials data science.[122−125] Within the ANN framework, several types of surrogate models can be developed for materials application such as CNN,[97] RNN,[39] and generative adversarial network (GAN).[126] These networks typically consist of multiple layers of neurons, and they are commonly known as deep learning (DL) models. The past few years have witnessed the surge of generative neural network models as an attractive strategy for molecular property prediction.[33,35,127−130] A generative model aims to capture the distribution of data, both structures and properties of a material, and relate them in a nonlinear way.[35,126] The most interesting aspect of a generative model is that it represents molecules in a continuation latent space.

Here, some common principles of ML model development are briefly discussed and references are provided for further reading and implementation. The choice of regression model for a given problem is very important, and there are no well-established guidelines available for this purpose. Normally, one selects a method depending on the kind of available data and the type of model that best fits the data. The accuracy, interpretability, scalability, and complexity vary across these methods and across systems. A common tendency is to try a few of the numerous models and select the best one in terms of accuracy and efficiency for a given problem. A standard practice in ML model development includes dividing the structure−property data into two sets: training set and test set. The training set is used to build the model, and the test set is to examine the performance of the model for unknown data. Appropriate care is taken to avoid overfitting and underfitting during the model development. Readers can refer the work of Wang et al., who have recently discussed some of the best practices for ML model developments.[131] The mathematical foundation of these regression methods can be found elsewhere.[132,133] These methods are implemented as library functions in application programming interfaces (APIs) such as scikit-learn,[134,135] Keras,[136] and PyTorch,[137] which can be easily used for predictive model development. Moreover, two web sites, https://machinelearningmastery.com/ and https://towardsdatascience.com/, that host a large number of blogs on various critical aspects of ML models selection and development are available.

4. DESIGN WORKFLOWS

First-principle methods or predictive ML models or their combination can be utilized for polymer design. Historically, all the data-driven methods that are used for screening polymer configurational space and polymer design can be categorized into the following four classes.

(I) **Edisonian Design.** High throughput computations or experiments are used to calculate the property of a large number of candidate materials. The top few candidates based on the databases are selected for further investigation and deployment in a target application. This is a typical trial-and-error Edisonian approach where candidate structures are selected either randomly or based on an intuitive understanding of a polymer's configurational and chemical spaces.

(II) **ML-Evaluated Design.** A pre-existing structure−property data set is used to build an ML model. The ML model is then used to predict the property of a large number of candidate materials. The top candidates identified based on the ML prediction are selected for first-principle-based study and further analysis. Contrary to the first method, this strategy

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

screens a material's search space based on the prediction of an ML model. But the primary bottleneck of such ML-evaluated screening is that the ML methods are better suited for interpolation. Its accuracy tends to decline in search of extremal properties that fall outside the known range of property.[103]

(III) **First-Principle-Based Inverse Design.** Inverse design is a promising approach that identifies an optimal set of parameters for a target property. This requires integrating an optimization algorithm with a first-principle-based computational or experimental measurement of the properties of candidate polymers. The optimization algorithm iteratively produces new candidates whose properties are measured/estimated based on computations/experiments. This process continues until predefined stopping criteria are reached. This is perhaps the best and robust strategy of polymer design. However, on-the-fly characterization/computation of properties for a large number of candidates can be time-consuming.

(IV) **ML-Evaluated Inverse Design.** The most time-consuming part of the first-principle-based inverse design is the on-the-fly direct measurement of polymer properties within a given design cycle. There have been many recent attempts to address this issue where inverse designs are performed based on the property prediction made by an ML model. An ML model is built with pre-existing data, and then it is integrated with an optimization algorithm for ML-evaluated screening of the search space. This is certainly a promising method to speed up the design cycle with a caveat of training-target mismatch. As mentioned earlier, the ML methods are interpolative in nature, and, therefore, the success of ML-evaluated inverse design relies on the ability of ML models to predict properties that are outside the range of their training data.

Categories I and II solve the forward problem of polymer characterization. However, categories III and IV combine the polymer characterization with a strategy for minimizing the number of candidate polymers assessed en route to the ultimate target polymer and direct the search toward target/optimal values. This strategy is an "inverse" of the forward problem, wherein property values of the evaluated candidates are analyzed to decide and select the next set of candidate polymers to be characterized. Thus, categories III and IV are commonly known as inverse design methods. The core of these two inverse design methods is an optimization algorithm. In the case of first-principle-based inverse design, the objective function of the optimization method is measured/calculated via experiment or molecular simulation or any other physics-based method. On the other hand, the ML-evaluated inverse design uses an ML predictive model to estimate the objective function. The first-principle-based inverse design method does not need a priori data, and it generates structure−property data on the fly. On the other hand, the ML-evaluated inverse design needs structure−property data before running the design cycle to build the predictive model. How one should generate/sample data for building such an ML model, which will be utilized later for predicting the objective function of an optimization algorithm, is an interesting and open question.

General purpose "black-box" optimization algorithms are always in demand for any design and optimization problem. However, a number of "no free lunch" theorems suggest that any superior performance of an optimization algorithm over one class of problems is offset by its performance over another class.[138] A detailed discussion on the connection between effective optimization algorithms and the problems they are solving can be found in the work of Wolpert and Macready.[138]

Although there is no generic prescription on how to choose an optimization algorithm for a specific problem, Bayesian optimization (BO), genetic algorithm (GA), and Monte Carlo tree search (MCTS) are most commonly used in materials design problems. Their performance and efficiency vary significantly in a given materials design problem.[139−141] Here, the underlying principles of these algorithms are discussed, and their workflows in the context of polymer design are outlined.

4.1. Genetic Algorithm

A GA is a metaheuristic global optimization method that aims to mimic the process of natural selection to optimize a system's properties. It is widely used as a versatile strategy for material property optimization and design.[142] It has been increasingly adopted in the field of polymers[143−147] and other materials design problems[148−151] as well. The algorithm begins with a set of initial candidate polymers. The objective functions for all the candidates are either quantified using a first-principle-based method or predicted by an ML model. Afterward, the algorithm iteratively selects new candidates based on genetic operators such as elitism, selection, crossover, and mutation.[152−154] The evolutionary process usually continues until some criterion for convergence is satisfied or it is terminated because of a maximal time constraint. The workflow of a GA for a polymer design problem is shown in Figure 2. There are open-sourced GA codes

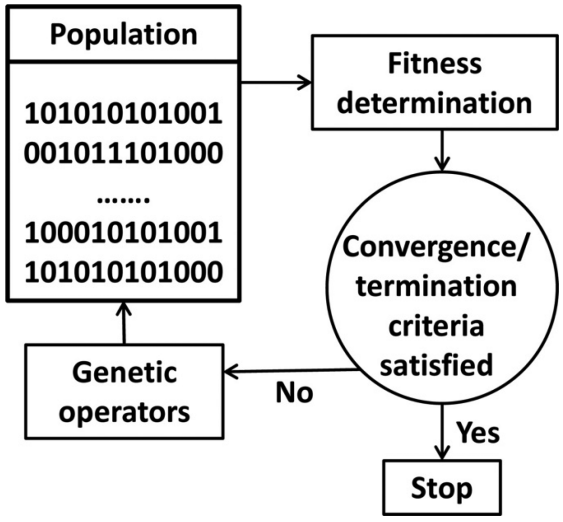


Figure 2. Schematic representation of a genetic algorithm. The population consist of candidate polymers in its feature representation. Adapted with permission from ref 103. Copyright 2017 American Chemical Society.

that can be utilized for polymer design. For example, Henning et al. has developed an open source code, viz., GASP that interfaces genetic algorithms with molecular simulation packages such as LAMMPS, VASP, and GULP.[155,156] Although GASP is primarily used for structure and phase diagram prediction of inorganic materials, it can be easily extended for polymer design.

4.2. Monte Carlo Tree Search (MCTS)

MCTS is a powerful global optimization method and is very popular in computer gaming algorithms such as Alpha Go, Bridge, Poker, and many other video games.[157,158] It has recently been adopted for materials design problems.[141,159−161] It integrates a tree search algorithm with reinforcement learning.[162,163] The algorithm begins with building a shallow tree of nodes, where each node represents a point in the search space. It

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

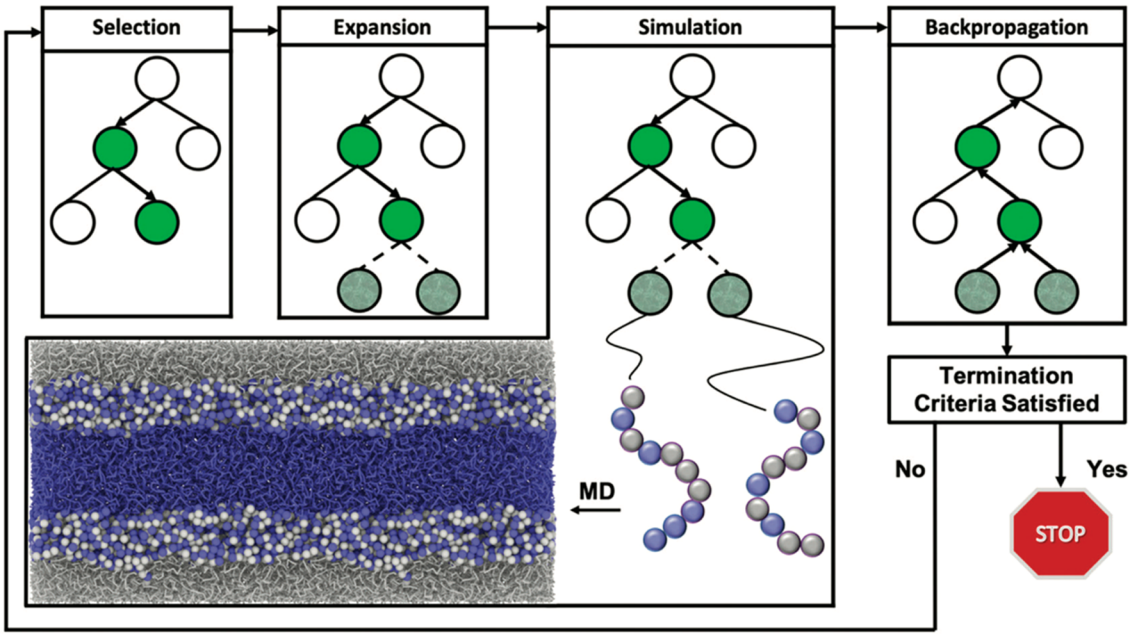


Figure 3. MD-MCTS workflow for a polymer design. All the nodes of a tree are the points in the search space. In a polymer design problem, each node represents a candidate polymer. Four steps—selection, expansion, simulation, and backpropagation are iteratively continued until a predefined termination criterion is satisfied. In the simulation step, the objective functions of the candidate structures are determined via MD simulations. Reproduced from ref 164 with permission from the Royal Society of Chemistry.

subsequently generates downstream pathways by a rollout procedure. The algorithm simultaneously explores many different pathways to reach the optimal point and exploits a single pathway that has the greatest estimated value of the search function. This simultaneous exploration and exploitation and an appropriate trade-off principle between them make the algorithm very efficient in identifying the global optimal point in a given function. It rapidly surmounts metastable and suboptimal points in a search space by growing other branches of the tree utilizing the trade-off mechanism between exploration and exploitation. We have recently integrated molecular dynamics simulation and Monte Carlo Tree Search for copolymer design.[164] A schematic representation of the workflow is shown in Figure 3. The algorithm is implemented in open-sourced codes such as MDTS,[141] which can be utilized for inverse design of polymers.

4.3. Bayesian Optimization and Active Learning

Bayesian optimization (BO) is a sequential design strategy that promises greater automation and, thus, is gaining popularity for autonomous designs.[165] It has been successfully implemented for many materials design problems.[139,166−169] The two key ingredients of a BO are a surrogate model and an acquisition function. The surrogate model is an ML model that is built upon past observations/data, and it predicts the properties for a given new structure. On the other hand, the acquisition function assigns scores to each new candidate structure according to the utility of measuring their properties via a first-principle method. The acquisition function uses the surrogate model to decide the score of a candidate structure. Any of the above regression algorithms, discussed in section 3, can be used as a surrogate model within the framework of a BO. This strategy is also known as active learning or adaptive learning, as it actively searches for new candidates in the design space and progressively rebuilds the surrogate model with increasing amounts of training data. There are many active learning strategies used in materials designs in recent times within the framework of BO. These methods are primarily varied on their choices of regression algorithms and strategies of selecting new candidates. The combination of selector and regressor determines the efficiency and performance of a BO run. As shown in Figure 4, a generic

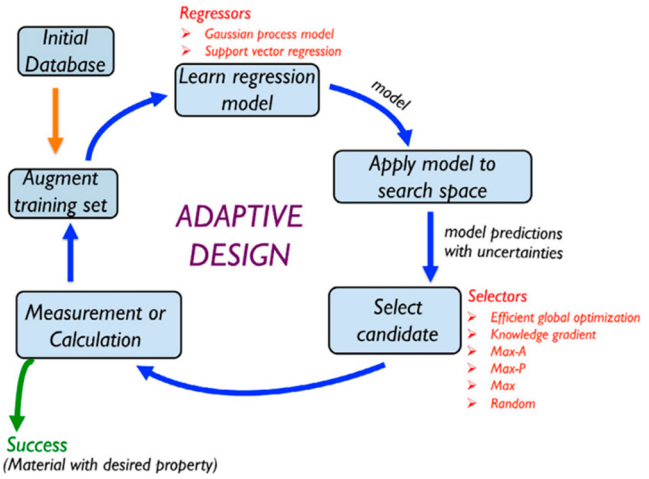


Figure 4. Workflow of a BO-based adaptive design scheme. Here, Measurement or Calculation indicates estimation of the material property via first-principle methods. The combination of regressor and selector is vital for the efficiency of the design scheme. Adapted with permission from ref 124. CC BY 4.0.

design cycle within the BO framework consists of following steps. (1) Select a prior for the design space based on the initial database. (2) Estimate the posterior given the prior and current data. (3) Deploy the posterior to determine the next candidates to evaluate according to the acquisition function. (4) Conduct first-principle calculations/measurements to obtain the new data. Steps 2 to 4 are repeated iteratively to explore the design

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

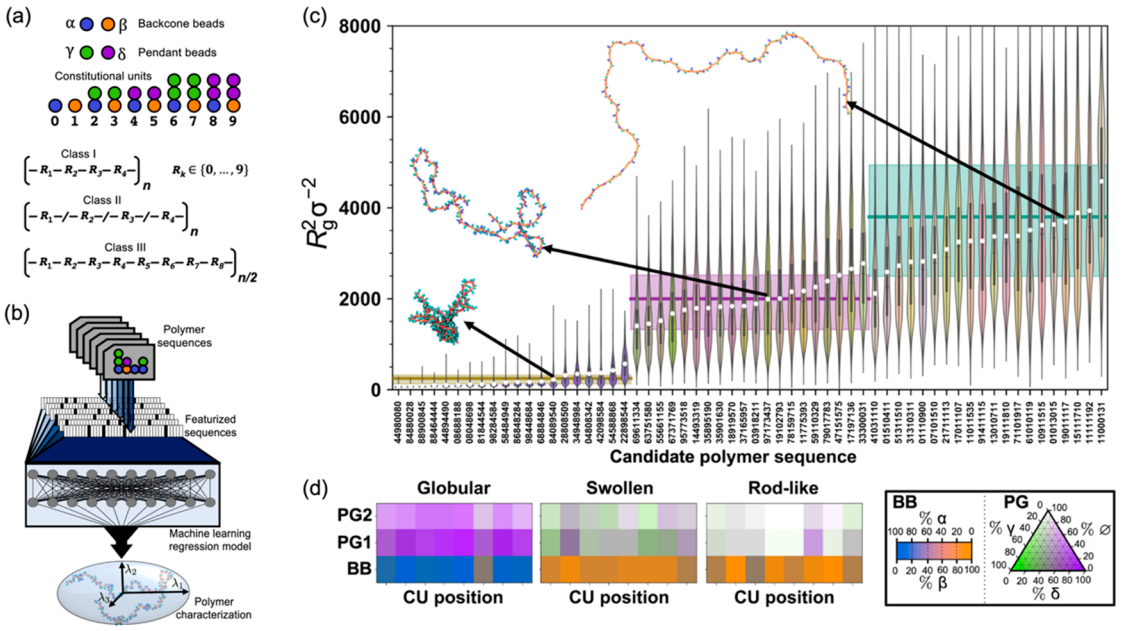


Figure 5. Single chain polymer design. The chemical moieties, constitutional units, and polymer topologies are schematically shown in (a). Polymer sequences are converted into featured sequences that serve as the input to the neural network model as shown in (b). The neural network model is used to screen the sequence space and identifies three target structures, globule, swollen, and rodlike aggregates, whose radius of gyration is shown in (c) for several candidate polymers. The violin plot for each sequence implies the distribution of values underlying the mean, with a notch at the median value. Here, the bar is extending from 25th to the 75th percentile values. The color of each violin is commensurate with the average composition of the sequences. The three target values are indicated by three horizontal lines, and shaded regions across the lines indicate the average spread between the 25th and 75th percentiles for class-I polymers of similar size. The average color composition of the sequences for each target structure is shown in (d). Here, BB, PG1, and PG2 denote backbone bead, first pendent group, and second pendent group, respectively. The color contributions for all the four types of bead are shown in the right side boxed legend wherein ⌀ indicates the absence of a pendent bead. Adapted with permission from ref 39. Copyright 2020 The Authors.

space of a material until convergence criteria are achieved. COMBO[170] and GPyOpt[171] are two well-known open-sourced codes that can be employed for materials design within the BO framework.

There are several other optimization techniques that are successfully used for polymer design problems. Notable examples are Particle Swarm Optimization[172,173] and tree-structured Parzen estimator (TPE) algorithm.[39]

5. REPRESENTATIVE EXAMPLES OF DATA-DRIVEN DESIGN

5.1. Targeted Polymer Sequence Design

Precise control of the sequence of monomers in a copolymer is an attractive goal of polymer engineering with a wide range of potential applications.[174,175] The compactness, rigidity, and stability of a single chain polymer nanoparticle are strongly influenced by its sequence of monomers.[176] CGMD simulation, advanced optimization, and ML have been used to understand sequence−structure relations of single molecule polymers and design sequences for target structures.[177−180] Here, I highlight one case study wherein de Pablo and co-workers have combined coarse-grained polymer genome, deep neural network model and sequential model-based optimization technique (SMBO)[181] for predicting the sequence of a polymer that produce a target structure.[39] This particular study has focused on designing very generic coarse-grained polymers with target structures. The model polymer is made of four coarse-grained beads named as α, β, γ, and δ, as shown in Figure 5a. These beads are the representation of different chemical moieties with varied solvophobicity. This is achieved by distinctly assigning interaction parameters (ϵ) to different beads. A high value of ϵ represents a stronger attractive force. The backbone of the polymer is formed by the α and β type moieties, while γ and δ constitute the side chains of the polymer. Within this chemical space, one can build 10 possible unique constitutional units (CUs) as shown in Figure 5a. The authors have studied three classes of polymers that are formed by connecting these CUs using two-, three-, and four-body intramolecular potentials. These three classes of polymers are distinguished by the number of CUs and their specific arrangement to build a polymer. As shown in Figure 5a, class-I and class-II polymers are made of four CUs, while class-III may contain a maximum of eight CUs. The class-I and class-III polymers have four and eight constitutional repeat units (CRUs). However, class-II polymers are stochastically generated sequences and cannot be defined by CRU. There are 1540 unique polymers that are possible in the class-I category. Number of unique polymers in class-II category is astronomically large. Within this framework, the authors have conducted implicit solvent CGMD simulations of large number of single-chain polymers that are composed of 400 CUs with varied sequences. The CGMD trajectories are utilized to calculate the radius of gyration of all the polymer sequences. The sequence-radius of gyration (R_g) data constitutes the polymer genome that is used for ML model development. The sequences are mapped to two different feature spaces by one-hot encoding (OHE) and property coloring for ML purposes. Both the features are shown to be efficient for modeling sequence-R_g correlation. The ML model development workflow is schematically shown in Figure 5b. The model is built upon class-1 polymer data, and it has predicted the properties of both class-I and class-II with high fidelity. However, the performance is

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

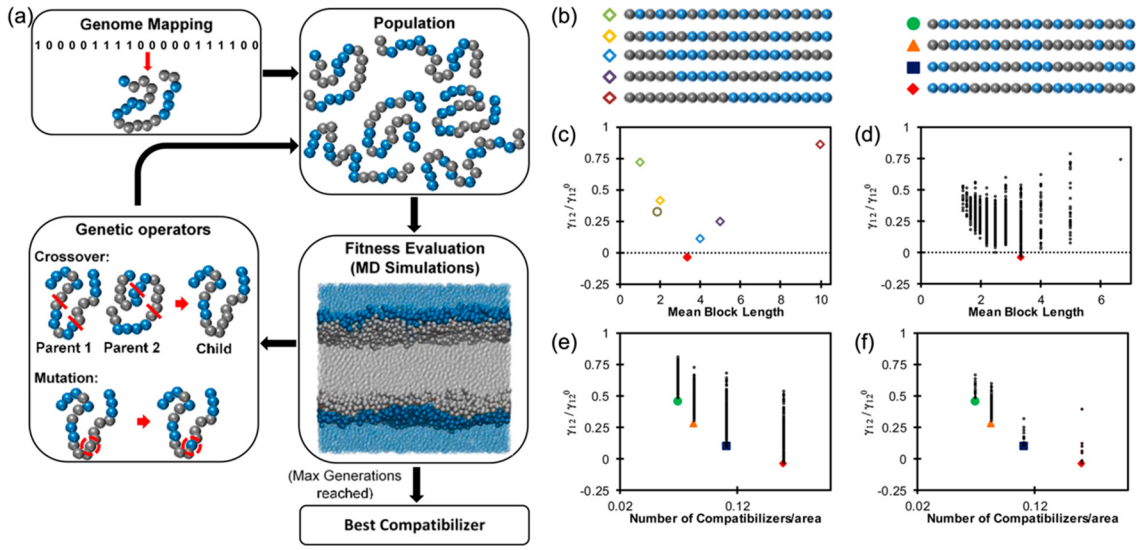


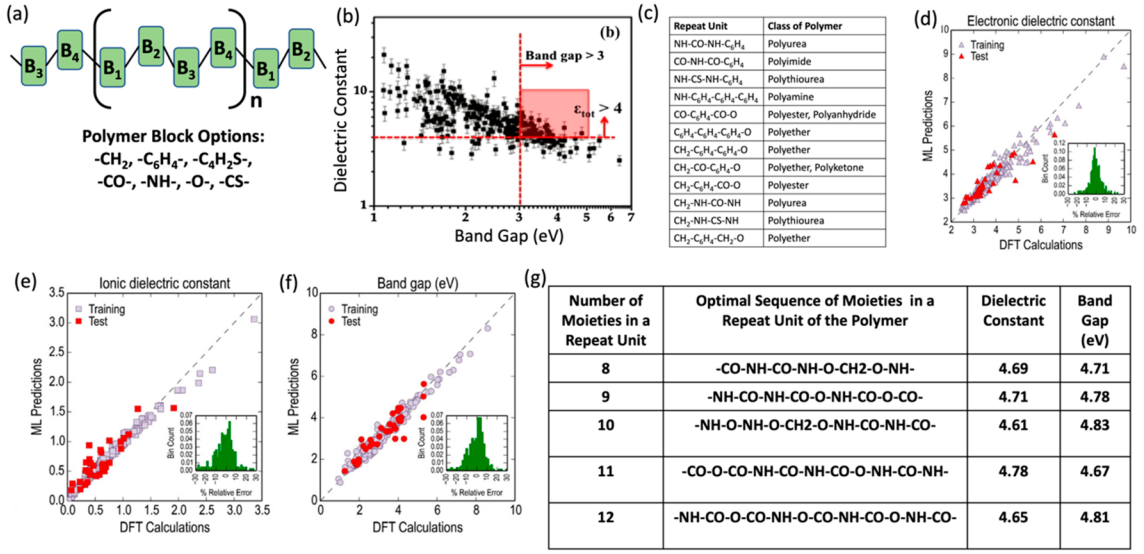
Figure 6. Inverse design of compatibilizer polymer. (a) Workflow of the molecular dynamics simulation-based genetic algorithms. Human design and GA identified compatibilizer polymers sequences are schematically shown in (b). The gray and blue beads represent two chemical moieties of different chemical affinity. The open symbol and filled symbols correspond to human design and GA identified sequences, respectively. The surface tension of the system for a compatibilizer concentration of 0.159 per unit interfacial area is plotted as a function of mean block length of a sequence for regular compatibilizer polymers along with GA optimized sequence in (c). The surface tensions of all the candidate sequences screened during a GA run for a compatibilizer concentration of 0.159 are plotted in (d) as a function of mean block length. The optimized candidate is shown as a red diamond. The surface tension of systems sharing the same mean block length as the optimized compatibilizer at each concentration studied is shown in (e). Similarly, the surface tension of systems sharing the same block length distribution as the optimized compatibilizer at each concentration studied are shown in (f). The surface tension value for compatibilizer systems γ_{12} in (c)−(f) are normalized by that of a polymer blend without compatibilizer polymers in the interface (γ_{12}^0). Reproduced from ref 104. Copyright 2017 American Chemical Society.

slightly inferior for class-II polymers. Moreover, the authors have used this ML model to design class-III polymers with three target structures viz., globule, swollen and rod-like aggregate. This is attained by setting a target value of the R_g during an optimization run that corresponds to a target structure based on a priori understanding of the structure-R_g correlation of the model polymer. The target gyration for globule, swollen and rod-like structures are $\langle R_g^2 \rangle \sigma^{-2}$ = 250, 2000 and 3800, respectively. Here, σ is the unit of length. The authors have employed a SMBO type of algorithm, viz., tree structure Parzen estimator (TPE) that generates a candidate sequence, compares its R_g value, as predicted by the ML model, with the target value, and proposes a new candidate based on historical performance. Although the exact R_g value is not achieved, the algorithm identified multiple sequences that are within the range of target R_g and structures. As shown in Figure 5c, the authors have identified 20 candidate polymer sequences for each of the three target structures using SMBO-TPE algorithm. The average sequence composition of these target structures is analyzed, and their characteristics are assessed by using a color scheme as shown in Figure 5d. The most important aspect of this work is that the ML model, which is built in one region of the sequence space, is transferable to other regions. This is an example of the ML-evaluated inverse design. The success of this approach is perhaps because both the regions of the sequence space—training and testing are mapped to the same region of the structure space. Nevertheless, the work points toward a new paradigm of ML-based inverse design and exploration outside the known range of properties.

5.2. Inverse Design of Polymer Compatibilizer

A compatibilizer is an additive that reduces the interfacial energy between two immiscible polymers and provides mechanical and thermal stability of their blends. It has widespread applications in emulsification,[182] coalescence and stabilization of polymer blends,[183,184] and barrier materials.[185] These compatibilizer polymers are made by covalently connecting repeat units that have different preferences for the two phases; and this particular architecture helps them to work like a bridge between the two phases. Diblock copolymers, where two polymer chains of different chemical affinities are linked by a covalent bond, have long been perceived as an effective compatibilizer.[186−188] However, many subsequent works indicate a finer level of architectural control of a copolymer, such as a random copolymer where two types of moieties are distributed randomly in the polymer architecture,[189−191] can provide improved compatibility. This opens up new possibilities for monomer level sequence control of a compatibilizer and provides a large design space of a copolymer compatibilizer. In a recent work, Simmons and co-workers have used a molecular simulation-based genetic algorithm for rational design of a AB type compatibilizer[104] by employing a generic bead−spring polymer of Kremer−Grest type to model the system, and reported several sequence defined polymers that outperform many periodic copolymers and random copolymers. This is an example of first-principle-based inverse design. The workflow of the MD-based genetic algorithm and key findings of their study[104] are summarized in Figure 6. The design workflow combines generic operations and MD simulations as shown in Figure 6a. The Figure 6b schematically shows optimal sequences identified by the MD-GA design scheme along with intuitive periodic copolymers that are commonly used for compatibilization applications. It clearly indicates that the optimized sequences are nonintuitive, nonperiodic, and highly irregular. A comparison of the interfacial tensions of the MD-GA identified sequence and periodic sequences are shown in Figure 6c for a specific compatibilizer concentration. The interfacial tension of all the

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26



Repeat Unit	Class of Polymer
NH-CO-NH-C_6H_4	Polyurea
CO-NH-CO-C_6H_4	Polyimide
NH-CS-NH-C_6H_4	Polythiourea
NH-C_6H_4-C_6H_4-C_6H_4	Polyamine
CO-C_6H_4-CO-O	Polyester, Polyanhydride
C_6H_4-C_6H_4-C_6H_4-O	Polyether
CH_2-C_6H_4-C_6H_4-O	Polyether
CH_2-CO-C_6H_4-O	Polyether, Polyketone
CH_2-C_6H_4-CO-O	Polyester
CH_2-NH-CO-NH	Polyurea
CH_2-NH-CS-NH	Polythiourea
CH_2-C_6H_4-CH_2-O	Polyether

Number of Moieties in a Repeat Unit	Optimal Sequence of Moieties in a Repeat Unit of the Polymer	Dielectric Constant	Band Gap (eV)
8	-CO-NH-CO-NH-O-CH2-O-NH-	4.69	4.71
9	-NH-CO-NH-CO-O-NH-CO-O-CO-	4.71	4.78
10	-NH-O-NH-O-CH2-O-NH-CO-NH-CO-	4.61	4.83
11	-CO-O-CO-NH-CO-NH-CO-O-NH-CO-NH-	4.78	4.67
12	-NH-CO-O-CO-NH-O-CO-NH-CO-O-NH-CO-	4.65	4.81

Figure 7. Rational design of polymer dielectrics. (a) Chemical and sequence space is schematically shown. (b) Total dielectric constant of polymers are plotted as a function of their band gap. The candidates with dielectric constant greater than 4 eV and band gap greater than 3 eV are highlighted by the red arrow lines. Adapted with permission from 194. Copyright 2013 Elsevier. (c) Some of the 4-block polymers that have high dielectric constant and high band gap are tabulated. The performance of the ML model in predicting electronic dielectric constant, ionic dielectric constant, and band gap is shown in (d)−(f), respectively, for training and test sets. Insets of (d)−(f) report the percentage of relative error in ML model prediction with respect to their actual DFT calculated values. Panels (d)−(f) are adapted with permission from ref 112. CC BY 4.0. (g) The optimal sequences of moieties that are identified by an evolutionary search are listed along with their dielectric constant and band gap, as reported in ref 112.

sequences that are screened during a MD-GA run are plotted as a function of their mean block lengths in Figure 6d. The data suggest that there is no generic correlation between surface tension and mean block length of a compatibilizer polymer. Interestingly, systems with same mean block lengths and/or same block length distributions exhibit a wide range of surface tension, as indicated in Figure 6e and f. The authors have also developed an analytical model for the optimal conformation of a compatibilizer polymer at the interface that suggests sequences that combine long and short blocks provide higher stability over periodic compatibilizer with a monodisperse block length distribution.

5.3. Rational Design of Polymer Dielectrics

The dielectric constant of a polymer determines its ability to polarize in the presence of an electric field. Polymers with very high and low dielectric constants are essential for electronic applications. Low dielectric polymers are used for insulations such as isolating signal carrying conductors and suppressing coupling between closely packed metal lines in an integrated circuit (IC). On the other hand, high dielectric polymers are demanding for semiconductor and high-density energy storage devices. Many of the well-known polymers have low dielectric constant and they are commonly used for insulating purposes. However, achieving high dielectric constant in polymeric materials is challenging.[192] There have been many recent efforts to design high dielectric polymer via data-driven strategies.[38,112,193−197] Ramprasad and co-workers have reported a rational design strategy wherein they have conducted density functional theory (DFT) and density functional perturbation theory (DFPT) calculations to estimate one-dimensional optimal structure, band gap and dielectric constant of 267 polymers.[193−195] These polymers' repeat unit has four chemical moieties chosen from seven building blocks as shown in Figure 7a. Based on chemical intuition, amenability of synthesis and obvious unstable combinations, the authors have excluded several sequences and selected 267 candidates that are experimentally realizable. The dielectric constant and bandgap of these polymers exhibit an inverse correlation as shown in Figure 7b. This inherent correlation is the primary bottleneck to develop polymer dielectrics for energy device applications that require high band gap as well as high dielectric constant. The authors have suggested materials that pose high dielectric constant and moderately high band gap for capacitive energy storage application, which can be drawn from this data set. With these applications in mind, the authors have selected top candidates from this pool of 267 candidates that have dielectric constant greater than 4 eV and band gap greater than 3 eV. These top candidates are tabulated in Figure 7c. The top three from this table were further studied and then successfully synthesized. This is a typical Edisonian design, and the research group subsequently proposed an ML-evaluated inverse design. For this purpose, DFT calculations are conducted for three-dimensional (3D) system of all the candidates.[112] Machine learning predictive models are developed based on the 3D structure data set. The performance of the ML model in predicting the electronic dielectric constant, ionic dielectric constant, and band gap are shown in Figure 7d−f. Although the ML model is based on four block polymer data, it has been shown to predict properties for polymers of other block sizes with high accuracy. This ML model is then used for inverse design of polymer dielectric via an evolutionary search with a target dielectric constant of 5 and band gap of 5 eV. The top candidates that are identified using the surrogate-model-based evolutionary search are tabulated in Figure 7g for five different sizes of the repeat unit. The dielectric constant and band gap of the EA identified candidates are very close to their target values.

5.4. Inverse Design of Thermally Conductive Polymers

Polymers have emerged as promising materials for energy storage devices, semiconductors and many micro/nanofluidics devices due to their tunable properties and easy process-

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

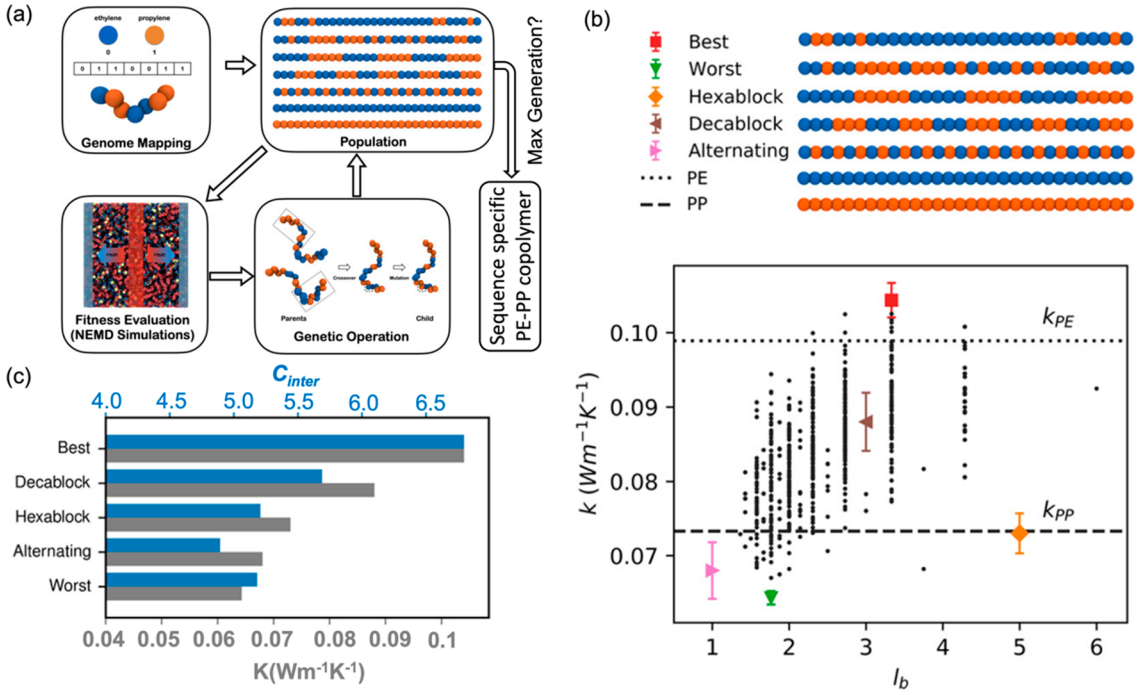


Figure 8. Evolutionary design of polymer thermal conductivity. (a) Polyethylene-polypropylene (PE-PP) copolymer genome mapping and the cycle of NEMD simulation-based genetic algorithm. (b) Thermal conductivity of all the candidate polymers that have appeared in the GA search are shown as a function of their mean block length (l_b) along with hexablock, decablock, and alternating block copolymers. The conductivities of two homopolymers, polyethylene (K_{PE}) and polypropylene (K_{PP}), are shown by the dashed lines. (c) The intermolecular coordination number and conductivity of the best and worst candidates along with decablock, hexablock, and alternating block copolymers are shown in a bar plot. Reproduced with permission from ref 105. Copyright 2021 American Chemical Society.

ability.[198−204] These devices produce a substantial amount of heat during their operation, and therefore, efficient heat dissipation is vital for their stability, durability, and sustainability. Polymers inherently show poor thermal conductivity and improving their thermal conductivity is essential for their successful deployment in above-mentioned applications. Hence, much effort has been directed toward understanding and tailoring the thermal conductivity of polymeric materials.[205−209] These studies have indicated that the conductivity of polymers intricately correlates to their alignment, chemical composition, blending with other compounds, and many other processing and environmental conditions. These complex correlations offer a large design space for thermally conducting polymers. Recently, Müller-Plathe and co-workers have combined nonequilibrium molecular dynamics (NEMD) and genetic algorithm for first-principle-based inverse design of a binary copolymer for optimal thermal conductivity.[105] As shown in Figure 8a, the authors linked NEMD and a genetic algorithm and evaluated the thermal conductivity of polyethylene-polypropylene (PE-PP) copolymers for ∼600 candidate sequences. A binary mapping is used to build the feature space of the PE-PP copolymer. The thermal conductivities of all the copolymer sequences that are screened during the design cycle are shown in Figure 8b along with regular block copolymers and two homopolymers. The data suggest that there is more than 70% variation in the thermal conductivity among all the candidate polymers. The optimal sequence for the highest thermal conductivity is evidentially nonperiodic and nonintuitive. The block distribution of this MD-GA identified sequence is very different from the intuitive human design polymers. A visual inspection indicates that the GA-optimized sequence contains a long block of PE in the middle region of the chain, and a small fraction of PP monomer near the two ends of the chain. The authors argue that this specific arrangement of blocks in the polymer topology balances the thermal energy transfer via bonded and nonbonded interactions and assists the candidate polymer to achieve maximum thermal efficiency. Moreover, as shown in Figure 8b, there is no generic correlation between thermal conductivity and mean block length of the copolymer. Interestingly, the thermal conductivity strongly correlates with the coordination number of many copolymer sequences as can be seen in Figure 8c. The polymer with the highest thermal conductivity has the highest coordination number. This coordination number, c_{inter}, is a measure of the number of molecular contacts in the system and usually is commensurate with strong nonbonded attractive forces. It suggests a very compact molecular packing in the GA-identified highest conductivity polymers in comparison to regular block copolymers.

5.5. Rational Design of Gas-Separation Polymer Membrane

Polymer membranes have long been used for separating gas mixtures such as the removal of CO_2 from natural gas, hydrogen recovery, and carbon capture from points of sources before it enters the atmosphere. The design of such polymer membranes for gas separation is based on empirical observation and not based on an exhaustive search of the chemical and conformational space of polymers. The target properties of gas separation membranes are the selectivity and permeability of a gas that are inversely correlated, and therefore, the development of polymers that offer high selectivity and permeability of a gas is challenging.[210−216] In a recent study, Kumar and co-workers have addressed this problem using a machine learning model for predicting the selectivity and permeability of several combinations of gases of a large number of polymer membranes.[84] The computational workflow of this ML-evaluated design study is

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

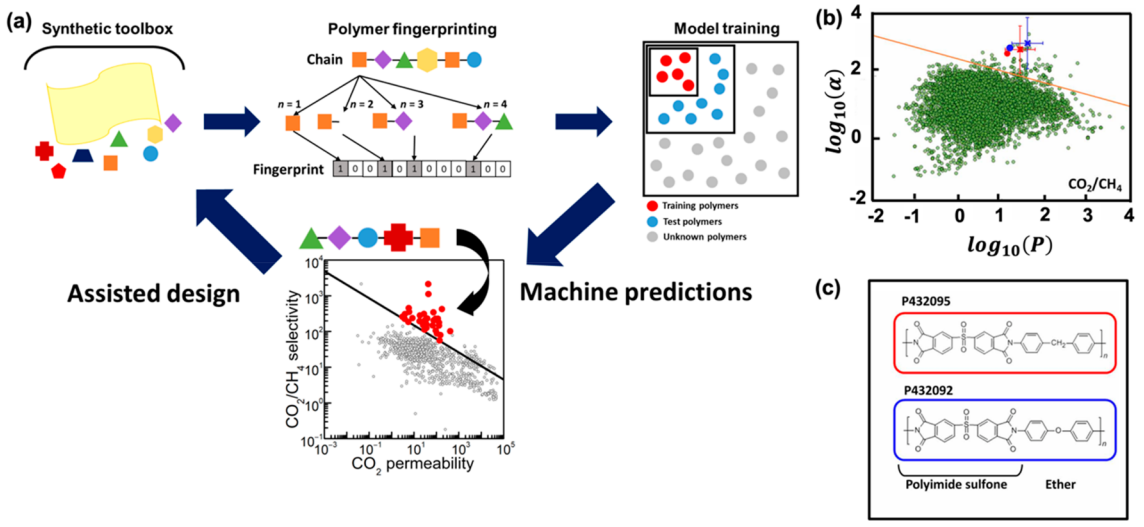


Figure 9. Data-driven design of a gas-separation polymer. The design workflow is shown schematically in (a). Here, the linear combination of chemical moieties are fingerprinted as a binary string that serves as an input to a machine learning model. The training data is randomly selected from literature database. The model is used to predict the permeability and selectivity of a large set of literature data. (b) ML-based Robeson plot for CO_2/CH_4. The red and blue colored cross legends represent the two top performing candidates predicted by the ML model. Their actual values are marked as red and blue colored circle legends. The repeat units of the two top performing candidate polymers are shown in (c). Adapted with permission from ref 84. Copyright 2020 The Authors.

Table 1. Summary of Five Case Studies[a]

applications (domain of interest)	type of model/featurization	origin of data	ML/ optimization	approximate data size	design category
single chain polymer structure[39]	coarse grained/property coloring	phenomenological CGMD	DNN/ SMBO	1540	ML-evaluated inverse design
polymer compatibilizer[188]	coarse-grain/binary mapping	phenomenological CGMD	GA	960	first-principle-based inverse design
thermally conducting polymer[207]	united atom/binary mapping	chemically Informed CGMD	GA	600	first-principle-based inverse design
polymer dielectric[108,192]	atomistic/motif-based fingerprints	DFT	KRR/GA	284	Edisonian design and ML-evaluated inverse design
polymer membrane[84]	binary mapping	experiments	GPR	500	ML-evaluated design

[a]Since polymer membrane design is based on the experimental data, it has no entry for "type of model" in column 2. Phenomenological CGMD simulation uses a bead−spring Kremer−Grest polymer model, while the chemically informed CGMD simulation uses united atom model within the framework of OPLS force field. Here, the data in column 4 are actual structure−property data calculated/measured using physics-based methods collected either during the optimization run or before the optimization run for the case of ML-based design. For the polymer dielectric study, KRR/GA is used for ML-evaluated inverse design.

shown in Figure 9a. An ML model, viz., GPR, is built using ∼250 data points. It shows a R^2 (coefficient of determination) score of 0.8 for the test set of ∼85 data points. The ML model is used to predict the gas separation behavior of 11 000 unknown polymer systems. The two topmost performing polymers based on the ML prediction are experimentally tested, and it appears that their actual selectivity and permeability are in close agreement with the ML prediction. The permeability and selectivity of the top two candidates along with others are shown in Figure 9b. The chemical units of the top two polymers are shown in Figure 9c. These candidate polymers are exceptional as they fall outside the known limit of the selectivity−permeability range as shown in the "Robeson plot". Further study is required to understand the gas separation mechanism and the structure of these two polymers and why they exhibit superior properties than others.

The above examples span a diverse range of data sources from phenomenological CG simulations to classical atomistic simulations to quantum simulations and experiments. Application, featurization, and ML architectures are selected according to the data and specific design goals. Table 1 summarizes different levels of development and applicability. Single chain polymer design utilizes implicit MD simulations within the Brownian dynamics framework. This is computationally faster than explicit MD simulations that include both the polymer and solvent particles. On the other hand, the compatibilizer design study conducts phenomenological CGMD simulations of polymers in a bulk melt-state. The nature of this problem requires simulation of a large number of particles, and therefore, data generation is computationally more expensive than that for the single molecular design study. But, the repeat unit of both of these polymeric systems is a Kuhn segment of a polymer.[217] On the other hand, a thermally conducting polymer design study is carried out for shorter length scales, for example, a united atom (UA) level feature of a polymer where each bead is represented by a small group of atoms; in this case study, they are either CH, CH_2, or CH_3. These UA level CG simulations are even computationally more expensive than the phenomenological CGMD simulations. Finally, the DFT simulations, where all atoms are explicitly considered, is computationally the most expensive and, therefore, the polymer dielectric design is restricted to shorter periodic polymers. These case studies indicate that the design space is decided based on the

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

computational cost. Therefore, targeting a larger design space while the property measurements are time and/or resource intensive is still a substantial challenge. The current trend in data sharing and online data repositories can potentially mitigate these challenges and help in expanding the design space.

6. CONCLUSIONS AND OUTLOOK

Big data, distributed computing and machine learning are now playing a crucial role in materials research, and complementing the traditional theoretical and experimental methods.[22,26,218−220] World-wide efforts such as Materials Genome Initiatives[221−223] and FAIR Data[224−226] aim to integrate these paradigms to shorten the time scale of materials development. Here, the underlying principles of data-driven methods, their limitations and applications in polymer research are reviewed. I specifically focus on how data-driven methods can accelerate polymer design and aid in establishing new structure−property correlations. It has been found that a large body of reported works screens the polymer search space using ML models that are built upon preexisting experimental or simulation data. These ML models are subsequently used to predict the property of many unknown polymers. Although ML-models are of lower fidelity in comparison to first-principle methods, they can give a good estimation of materials properties when the test data are within the range or nearby region of their training data. They are, therefore, very suitable and successful for interpolation tasks. In recent times, there have been many attempts to integrate first-principle methods and advanced optimization methods for the inverse design of polymers. Such first-principle-based inverse design studies have identified new polymers as well as generated large amounts of data that are utilized to establish new correlations. Advancements in computing architecture and software and rapid progress in multiscale polymer simulations are enabling such a large-scale exploration of the chemical and sequence space of polymers. There have been cases where such computationally designed polymers are synthesized and characterized experimentally. Although high throughput parallel characterization and inverse design of polymers via computation are now being done routinely, such parallel synthesis and characterization of a large number of candidate polymers via experiment are rare. This requires overcoming challenges associated with the development of high throughput instruments and integrating them with optimization and ML algorithms. Nevertheless, high-throughput experimental setups[227] and robot-assisted synthesis and design approaches[228,229] are showing early success and will be very important for future data-driven polymer research.

In terms of algorithms and workflow developments, there are many challenges that need to be addressed. Some of the current challenges and future research directions are outlined below.

6.1. Polymer Feature Engineering

Efficient and unique descriptors of polymers for machine representation and mathematical operation is central to the success of ML assisted polymer design. One-hot vector and property coloring are shown to be effective for CG level polymer representation. Similarly, autoencoders have been used for small organic molecules. More work is required for testing and validating these fingerprinting schemes for long chain polymers and atomic level model systems. Unique and efficient descriptors are always on-demand for easily identifying target polymers. It can also help build new correlations and advance the current understanding of structure−property correlations of polymers. Graph and other hierarchical representations of polymers, especially stochastic sequences, are an area that requires further study.

6.2. Experimentally Realizable Polymer Design

One of the major issues in the computational design of polymers is to direct the search toward experimentally realizable candidate materials. Most of the reported works do not account for synthetic limitations. This requires a deeper understanding of the polymer chemistry and synthesis to develop appropriate constrained parameters in setting up the optimization problem.

6.3. Multiobjective Design

The majority of the previous attempts were focused on the optimization of one single property. However, successful utilization of designed polymers in industries requires a combination of properties that are often antithetical. For example, ion conductivity and mechanical properties of ion containing polymers are inversely correlated. Similarly, dielectric constant/band gap and gas permeability/selectivity are commonly known inverse correlations in polymeric materials. Future design efforts should be directed in the search of materials that can potentially expand the boundary formed by these inverse correlations. In multiobjective optimizations, a set of solutions is targeted that are commonly known as Pareto optimal frontier, instead of finding global optimal points of individual parameters.[230−232] In this direction, Yoo et al. have recently proposed a Pareto active learning procedure for multiobjective design of polymer.[233]

6.4. Simultaneous Exploration of Sequence and Chemical Space

Large scale exploration of chemical space and sequence space within a single design study is still a substantial challenge even using all the modern computational tools. Most of the published works have attempted to explore one of these two spaces to a great extent while limiting the other in a narrow range. It is anticipated that the increasing trends in computational speed, open-sourced codes and data, advancement in polymer feature engineering, and efficient computation workflows will enable large-scale exploration of the chemical space and sequence space simultaneously.

6.5. Training Target Mismatch

Many critical design problems target properties of materials that are outside the known range of their values. Machine learning models are inherently interpolative and tend to be less reliable in the search of extremal candidates. Often, design algorithms produce new molecules that lie in the "dead region" of the configuration space that is far away from data that are used to build an ML model.[100] Within the framework of GA-based materials design, we have proposed the NBGA (neural network biased genetic algorithm) to mitigate these challenges.[103] Griffiths and Hernández-Lobato have proposed a constrained Bayesian optimization approach to tackle the training-target mismatch within the framework of generative model development.[166] More research is essential for a holistic understanding of this training-target mismatch and developing more efficient strategies to tackle this problem.

Overall, with the gradual increment in the volume of open-sourced structure−property data and software, continuous improvement in ML algorithms, and their efficient integration with first-principle-based methods, data-driven methods are expected to become more reliable and easily accessible tools for polymer research.

https://doi.org/10.1021/acspolymersau.1c00035
ACS Polym. Au 2022, 2, 8−26

AUTHOR INFORMATION

Corresponding Author

Tarak K. Patra − *Department of Chemical Engineering, Center for Atomistic Modeling and Materials Design and Center for Carbon Capture Utilization and Storage, Indian Institute of Technology Madras, Chennai, TN 600036, India;* orcid.org/0000-0002-6002-0922; Email: tpatra@iitm.ac.in

Complete contact information is available at: https://pubs.acs.org/10.1021/acspolymersau.1c00035

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

The author acknowledges financial support from National Supercomputing Mission (NSM) and Science and Engineering Research Board (SERB). The author is grateful to Praneeth Srivanth Ramesh, Himanshu and Vibhu Vardhan Singh for their input in the manuscript preparation.

REFERENCES

(1) Allcock, H. R. Rational Design and Synthesis of New Polymeric Material. *Science* **1992**, *255* (5048), 1106−1112.

(2) Rubinstein, M.; Colby, R. H. *Polymer Physics*; Oxford University Press, 2003.

(3) Hiemenz, P. C.; Lodge, T. P. *Polymer Chemistry*; CRC Press, 2007.

(4) Draxl, C.; Scheffler, M. The NOMAD Laboratory: From Data Sharing to Artificial Intelligence. *J. Phys. Mater.* **2019**, *2* (3), 036001.

(5) Lipton, Z. C. The Mythos of Model Interpretability. *ArXiv (Machine Learning)*, March 6, **2017**, 160603490, ver. 3.

(6) Montavon, G.; Samek, W.; Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process.* **2018**, *73*, 1−15.

(7) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful Classification of Crystal Structures Using Deep Learning. *Nat. Commun.* **2018**, *9* (1), 2775.

(8) Ribeiro, M. T.; Singh, S.; Guestrin, C. Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv (Machine Learning)*, August 9, **2016**, 160204938, ver. 3.

(9) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *ArXiv (Artificial Intelligence)*, November 25, **2017**, 170507874, ver. 2.

(10) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(11) wwPDB consortium; Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. D.; Christie, C.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Liang, Y.; Lowe, R.; Peisach, E.; Periskova, I.; Randle, C.; Rose, A.; Sekharan, M.; Shao, C.; Tao, Y.-P.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Young, J.; Zardecki, C.; Zhuravleva, M.; Kurisu, G.; Nakamura, H.; Kengaku, Y.; Cho, H.; Sato, J.; Kim, J. Y.; Ikegawa, Y.; Nakagawa, A.; Yamashita, R.; Kudou, T.; Bekker, G.-J.; Suzuki, H.; Iwata, T.; Yokochi, M.; Kobayashi, N.; Fujiwara, T.; Velankar, S.; Kleywegt, G. J.; Anyango, S.; Armstrong, D. R.; Berrisford, J. M.; Conroy, M. J.; Dana, J. M.; Deshpande, M.; Gane, P.; Gáborová, R.; Gupta, D.; Gutmanas, A.; Koča, J.; Mak, L.; Mir, S.; Mukhopadhyay, A.; Nadzirin, N.; Nair, S.; Patwardhan, A.; Paysan-Lafosse, T.; Pravda, L.; Salih, O.; Sehnal, D.; Varadi, M.; Vařeková, R.; Markley, J. L.; Hoch, J. C.; Romero, P. R.; Baskaran, K.; Maziuk, D.; Ulrich, E. L.; Wedell, J. R.; Yao, H.; Livny, M.; Ioannidis, Y. E. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47* (D1), D520−D528.

(12) Bereau, T.; Andrienko, D.; Kremer, K. Research Update: Computational Materials Discovery in Soft Matter. *APL Mater.* **2016**, *4* (5), 053101.

(13) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889−2919.

(14) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92* (1), 014106.

(15) Hautier, G.; Jain, A.; Ong, S. P. From the Computer to the Laboratory: Materials Discovery and Design Using First-Principles Calculations. *J. Mater. Sci.* **2012**, *47* (21), 7317−7340.

(16) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9* (8), 3404−3419.

(17) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89* (9), 094104.

(18) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13* (6), 579−633.

(19) Amis, E. J. Combinatorial Materials Science: Reaching beyond Discovery. *Nat. Mater.* **2004**, *3* (2), 83−85.

(20) Breneman, C. M.; Brinson, L. C.; Schadler, L. S.; Natarajan, B.; Krein, M.; Wu, K.; Morkowchuk, L.; Li, Y.; Deng, H.; Xu, H. Stalking the Materials Genome: A Data-Driven Approach to the Virtual Design of Nanostructured Polymers. *Adv. Funct. Mater.* **2013**, *23* (46), 5746−5752.

(21) Parish, E. J.; Duraisamy, K. A Paradigm for Data-Driven Predictive Modeling Using Field Inversion and Machine Learning. *J. Comput. Phys.* **2016**, *305*, 758−774.

(22) 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* (6), 1017−1026.

(23) Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* **2016**, *4* (5), 053208.

(24) Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. Big-Deep-Smart Data in Imaging for Guiding Materials Design. *Nat. Mater.* **2015**, *14* (10), 973−980.

(25) Rajan, K. Materials Informatics: The Materials "Gene" and Big Data. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 153−169.

(26) Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv. Sci.* **2019**, *6* (21), 1900808.

(27) Choudhary, K.; Garrity, K. F.; Reid, A. C. E.; DeCost, B.; Biacchi, A. J.; Hight Walker, A. R.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A.; Davydov, A.; Jiang, J.; Pachter, R.; Cheon, G.; Reed, E.; Agrawal, A.; Qian, X.; Sharma, V.; Zhuang, H.; Kalinin, S. V.; Sumpter, B. G.; Pilania, G.; Acar, P.; Mandal, S.; Haule, K.; Vanderbilt, D.; Rabe, K.; Tavazza, F. The Joint Automated Repository for Various Integrated Simulations (JARVIS) for Data-Driven Materials Design. *Npj Comput. Mater.* **2020**, *6* (1), 1−13.

(28) Morgan, D.; Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 71−103.

(29) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.

(30) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12* (3), 191−201.

(31) Montoya, J. H.; Winther, K. T.; Flores, R. A.; Bligaard, T.; Hummelshøj, J. S.; Aykol, M. Autonomous Intelligent Agents for Accelerated Materials Discovery. *Chem. Sci.* **2020**, *11* (32), 8517−8532.
(32) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10* (8), 1903242.
(33) Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849−860.
(34) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal-Organic Frameworks. *Mol. Syst. Des. Eng.* **2019**, *4* (1), 162−174.
(35) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(36) Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation. *J. Phys. Chem. B* **2020**, *124* (19), 3873−3891.
(37) Jackson, N. E.; Webb, M. A.; de Pablo, J. J. Recent Advances in Machine Learning towards Multiscale Soft Materials Design. *Curr. Opin. Chem. Eng.* **2019**, *23*, 106−114.
(38) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the Polymer Genome: A Roadmap for Rational Polymer Dielectrics Design and Beyond. *Mater. Today* **2018**, *21* (7), 785−796.
(39) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted Sequence Design within the Coarse-Grained Polymer Genome. *Sci. Adv.* **2020**, *6* (43), No. eabc6216.
(40) 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. *J. Phys. Chem. C* **2019**, *123* (23), 14610−14618.
(41) Ferguson, A. L.; Ranganathan, R. 100th Anniversary of Macromolecular Science Viewpoint: Data-Driven Protein Design. *ACS Macro Lett.* **2021**, *10* (3), 327−340.
(42) *Polymer Property Predictor and Database*. https://pppdb.uchicago.edu/ (accessed 2021-08-12).
(43) *CROW*. https://polymerdatabase.com/index.html (accessed 2021-08-12).
(44) *Polymers: A Property Database 2020*. https://poly.chemnetbase.com/faces/polymers/PolymerSearch.xhtml (accessed 2021-08-12).
(45) *Polymer Genome*. https://www.polymergenome.org (accessed 2021-08-12).
(46) Sha, W.; Li, Y.; Tang, S.; Tian, J.; Zhao, Y.; Guo, Y.; Zhang, W.; Zhang, X.; Lu, S.; Cao, Y.-C.; Cheng, S. Machine Learning in Polymer Informatics. *InfoMat* **2021**, *3* (4), 353−361.
(47) Kotelyanskii, M.; Theodorou, D. N. *Simulation Methods for Polymers*; CRC Press, 2004.
(48) Gartner, T. E.; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* **2019**, *52* (3), 755−786.
(49) Taylor, P. A.; Jayaraman, A. Molecular Modeling and Simulations of Peptide-Polymer Conjugates. *Annu. Rev. Chem. Biomol. Eng.* **2020**, *11* (1), 257−276.
(50) Carbone, P.; Ali Karimi-Varzaneh, H.; Muller-Plathe, F. Fine-Graining without Coarse-Graining: An Easy and Fast Way to Equilibrate Dense Polymer Melts. *Faraday Discuss.* **2010**, *144* (0), 25−42.
(51) Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3* (9), 754−769.
(52) Karimi-Varzaneh, H. A.; van der Vegt, N. F. A.; Muller-Plathe, F.; Carbone, P. How Good Are Coarse-Grained Polymer Models? A Comparison for Atactic Polystyrene. *ChemPhysChem* **2012**, *13* (15), 3428−3439.
(53) Milano, G.; Müller-Plathe, F. Mapping Atomistic Simulations to Mesoscopic Models: A Systematic Coarse-Graining Procedure for Vinyl Polymer Chains. *J. Phys. Chem. B* **2005**, *109* (39), 18609−18619.
(54) Qian, H.-J.; Carbone, P.; Chen, X.; Karimi-Varzaneh, H. A.; Liew, C. C.; Müller-Plathe, F. Temperature-Transferable Coarse-Grained Potentials for Ethylbenzene, Polystyrene, and Their Mixtures. *Macromolecules* **2008**, *41* (24), 9919−9929.
(55) Shen, K.-H.; Fan, M.; Hall, L. M. Molecular Dynamics Simulations of Ion-Containing Polymers Using Generic Coarse-Grained Models. *Macromolecules* **2021**, *54* (5), 2031−2052.
(56) Jayaraman, A. 100th Anniversary of Macromolecular Science Viewpoint: Modeling and Simulation of Macromolecules with Hydrogen Bonds: Challenges, Successes, and Opportunities. *ACS Macro Lett.* **2020**, *9* (5), 656−665.
(57) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys.: Condens. Matter* **2004**, *16* (15), R481−R512.
(58) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models. *J. Chem. Phys.* **2008**, *128* (24), 244114.
(59) Lu, L.; Voth, G. A. The Multiscale Coarse-Graining Method. In *Advances in Chemical Physics*; Rice, S. A., Dinner, A. R., Eds.; John Wiley & Sons, Inc., 2012; pp 47−81.
(60) Jackson, N. E.; Bowen, A. S.; de Pablo, J. J. Efficient Multiscale Optoelectronic Prediction for Conjugated Polymers. *Macromolecules* **2020**, *53* (1), 482−490.
(61) Lyubimov, I.; Wessels, M. G.; Jayaraman, A. Molecular Dynamics Simulation and PRISM Theory Study of Assembly in Solutions of Amphiphilic Bottlebrush Block Copolymers. *Macromolecules* **2018**, *51* (19), 7586−7599.
(62) Joshi, S. Y.; Deshmukh, S. A. A Review of Advancements in Coarse-Grained Molecular Dynamics Simulations. *Mol. Simul.* **2021**, *47* (10−11), 786−803.
(63) Dhamankar, S.; Webb, M. A. Chemically Specific Coarse-Graining of Polymers: Methods and Prospects. *J. Polym. Sci.* **2021**, *59*, 2613.
(64) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1−19.
(65) *LAMMPS Molecular Dynamics Simulator*. https://www.lammps.org/ (accessed 2021-09 -13).
(66) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701−1718.
(67) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435−447.
(68) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, *153* (4), 044130.
(69) *NAMD - Scalable Molecular Dynamics*. https://www.ks.uiuc.edu/Research/namd/ (accessed 2021-09-13).
(70) Smith, W.; Yong, C. W.; Rodger, P. M. DL_POLY: Application to Molecular Simulation. *Mol. Simul.* **2002**, *28* (5), 385−471.
(71) Brukhno, A. V.; Grant, J.; Underwood, T. L.; Stratford, K.; Parker, S. C.; Purton, J. A.; Wilding, N. B. DL_MONTE: A Multipurpose Code for Monte Carlo Simulation. *Mol. Simul.* **2021**, *47* (2−3), 131−151.
(72) Shah, J. K.; Marin-Rimoldi, E.; Mullen, R. G.; Keene, B. P.; Khan, S.; Paluch, A. S.; Rai, N.; Romanielo, L. L.; Rosch, T. W.; Yoo, B.; Maginn, E. J. Cassandra: An Open Source Monte Carlo Package for Molecular Simulation. *J. Comput. Chem.* **2017**, *38* (19), 1727−1739.
(73) *Cassandra*. https://cassandra.nd.edu/ (accessed 2021-09-13).

(74) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6* (1), 15−50.
(75) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47* (1), 558−561.
(76) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L; Cococcioni, M.; Dabo, I.; Dal Corso, A.; de Gironcoli, S.; Fabris, S.; Fratesi, G.; Gebauer, R.; Gerstmann, U.; Gougoussis, C.; Kokalj, A.; Lazzeri, M.; Martin-Samos, L.; Marzari, N.; Mauri, F.; Mazzarello, R.; Paolini, S.; Pasquarello, A.; Paulatto, L.; Sbraccia, C.; Scandolo, S.; Sclauzero, G.; Seitsonen, A. P; Smogunov, A.; Umari, P.; Wentzcovitch, R. M QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials. *J. Phys.: Condens. Matter* **2009**, *21* (39), 395502.
(77) Giannozzi, P.; Baseggio, O.; Bonfà, P.; Brunato, D.; Car, R.; Carnimeo, I.; Cavazzoni, C.; de Gironcoli, S.; Delugas, P.; Ferrari Ruffino, F.; Ferretti, A.; Marzari, N.; Timrov, I.; Urru, A.; Baroni, S. Quantum ESPRESSO toward the Exascale. *J. Chem. Phys.* **2020**, *152* (15), 154105.
(78) *QUANTUMESPRESSO*. http://www.quantum-espresso.org/ (accessed 2021-09-13).
(79) Gaussian.com. *Expanding the limits of computational chemistry*. https://gaussian.com/ (accessed 2021-09-13).
(80) Correa-Baena, J.-P.; Hippalgaonkar, K.; van Duren, J.; Jaffer, S.; Chandrasekhar, V. R.; Stevanovic, V.; Wadia, C.; Guha, S.; Buonassisi, T. Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule* **2018**, *2* (8), 1410−1420.
(81) Maine, E.; Garnsey, E. Commercializing Generic Technology: The Case of Advanced Materials Ventures. *Res. Policy* **2006**, *35* (3), 375−393.
(82) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142* (48), 20273−20287.
(83) Pilania, G. Machine Learning in Materials Science: From Explainable Predictions to Autonomous Design. *Comput. Mater. Sci.* **2021**, *193*, 110360.
(84) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv.* **2020**, *6* (20), No. eaaz4301.
(85) Wu, S.; Kondo, Y.; Kakimoto, M.; 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 Comput. Mater.* **2019**, *5* (1), 1−11.
(86) Wang, Y.; Xie, T.; France-Lanord, A.; Berkley, A.; Johnson, J. A.; Shao-Horn, Y.; Grossman, J. C. Toward Designing Highly Conductive Polymer Electrolytes by Machine Learning Assisted Coarse-Grained Molecular Dynamics. *Chem. Mater.* **2020**, *32* (10), 4144−4151.
(87) Ferguson, A. L. Machine Learning and Data Science in Soft Materials Engineering. *J. Phys.: Condens. Matter* **2018**, *30* (4), 043002.
(88) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6* (10), 1078−1082.
(89) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122* (31), 17575−17585.
(90) Sattari, K.; Xie, Y.; Lin, J. Data-Driven Algorithms for Inverse Design of Polymers. *Soft Matter* **2021**, *17*, 7607.
(91) Upadhya, R.; Kosuri, S.; Tamasi, M.; Meyer, T. A.; Atta, S.; Webb, M. A.; Gormley, A. J. Automation and Data-Driven Design of Polymer Therapeutics. *Adv. Drug Delivery Rev.* **2021**, *171*, 1−28.
(92) Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U. F.; Tang, S.; Bi, J.; Chen, W.; Li, Y. Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges. *Polymers* **2020**, *12* (1), 163.
(93) Bereau, T. Computational Compound Screening of Biomolecules and Soft Materials by Molecular Simulations. *Modell. Simul. Mater. Sci. Eng.* **2021**, *29* (2), 023001.
(94) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3* (1), 2810.
(95) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Starita, A.; Tiné, M. R. Prediction of the Glass Transition Temperature of (Meth)Acrylic Polymers Containing Phenyl Groups by Recursive Neural Network. *Polymer* **2007**, *48* (24), 7121−7129.
(96) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(97) Miccio, L. A.; Schwartz, G. A. From Chemical Structure to Quantitative Polymer Properties Prediction through Convolutional Neural Networks. *Polymer* **2020**, *193*, 122341.
(98) 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. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013−5025.
(99) Bhattacharya, D.; Patra, T. K. DPOLY: Deep Learning of Polymer Phases and Phase Transition. *Macromolecules* **2021**, *54*, 3065.
(100) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(101) Alkharusi, H. Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *Int. J. Educ.* **2012**, *4* (2), 202−210.
(102) Yadav, D. Categorical encoding using Label-Encoding and One-Hot-Encoder. *Medium*. https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd (accessed 2021-09-14).
(103) Patra, T. K.; Meenakshisundaram, V.; Hung, J.-H.; Simmons, D. S. Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn. *ACS Comb. Sci.* **2017**, *19* (2), 96−107.
(104) Meenakshisundaram, V.; Hung, J.-H.; Patra, T. K.; Simmons, D. S. Designing Sequence-Specific Copolymer Compatibilizers Using a Molecular-Dynamics-Simulation-Based Genetic Algorithm. *Macromolecules* **2017**, *50* (3), 1155−1166.
(105) Zhou, T.; Wu, Z.; Chilukoti, H. K.; Müller-Plathe, F. Sequence-Engineering Polyethylene-Polypropylene Copolymers with High Thermal Conductivity Using a Molecular-Dynamics-Based Genetic Algorithm. *J. Chem. Theory Comput.* **2021**, *17*, 3772.
(106) Shi, J.; Quevillon, M. J.; Valença, P. H. A.; Whitmer, J. K. Predicting Adhesive Free Energies of Polymer--Surface Interactions with Machine Learning. *ArXiv (Soft Condensed Matter)*, October 6, **2021**, 211003041, ver. 1.
(107) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tiné, M. R. Predicting Physical-Chemical Properties of Compounds from Molecular Structures by Recursive Neural Networks. *J. Chem. Inf. Model.* **2006**, *46* (5), 2030−2042.
(108) Micheli, A.; Sperduti, A.; Starita, A.; Bianucci, A. M. Analysis of the Internal Representations Developed by Neural Networks for Structures Applied to Quantitative Structure-Activity Relationship Studies of Benzodiazepines. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (1), 202−218.
(109) *RDKit: Open-Source Cheminformatics Software*. https://www.rdkit.org/ (accessed 2021-09-07).
(110) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5* (9), 1523−1531.
(111) Batra, R.; Dai, H.; Huan, T. D.; Chen, L.; Kim, C.; Gutekunst, W. R.; Song, L.; Ramprasad, R. Polymers for Extreme Conditions

Designed Using Syntax-Directed Variational Autoencoders. *Chem. Mater.* **2020**, *32* (24), 10489−10500.
(112) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6* (1), 20952.
(113) Asgari, E.; Mofrad, M. R. K. ProtVec: A Continuous Distributed Representation of Biological Sequences. *PLoS One* **2015**, *10* (11), No. e0141287.
(114) Xiang, Z.; Fan, M.; Tovar, G. V.; Trehern, W.; Yoon, B.-J.; Qian, X.; Arroyave, R.; Qian, X. Physics-Constrained Automatic Feature Engineering for Predictive Modeling in Materials Science. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35* (12), 10414−10421.
(115) Kalidindi, S. R. Feature Engineering of Material Structure for AI-Based Materials Knowledge Systems. *J. Appl. Phys.* **2020**, *128* (4), 041103.
(116) Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ArXiv (Computational Physics)*, July 23, **2021**, 200304919, ver. 5.
(117) Patel, R.; Borca, C.; Webb, M. Featurization Strategies for Polymer Sequence or Composition Design by Machine Learning. *ChemRxiv***2021**. DOI: 10.33774/chemrxiv-2021-m74c8.
(118) Mohapatra, S.; An, J.; Gómez-Bombarelli, R. GLAMOUR: Graph Learning over Macromolecule Representations. *ArXiv (Machine Learning)*, August 23, **2021**, 210302565, ver. 3.
(119) Kojima, T.; Washio, T.; Hara, S.; Koishi, M. Synthesis of Computer Simulation and Machine Learning for Achieving the Best Material Properties of Filled Rubber. *Sci. Rep.* **2020**, *10* (1), 18127.
(120) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-Learning Predictions of Polymer Properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128* (17), 171104.
(121) Bhowmik, R.; Sihn, S.; Pachter, R.; Vernon, J. P. Prediction of the Specific Heat of Polymers from Experimental Data and Machine Learning Methods. *Polymer* **2021**, *220*, 123558.
(122) Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors from Simple Functions to Density Functionals. *Int. J. Quantum Chem.* **2015**, *115* (16), 1115−1128.
(123) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *Npj Comput. Mater.* **2019**, *5* (1), 1−36.
(124) Balachandran, P. V.; Xue, D.; Theiler, J.; Hogden, J.; Lookman, T. Adaptive Strategies for Materials Design Using Uncertainties. *Sci. Rep.* **2016**, *6* (1), 19660.
(125) Vasudevan, R.; Pilania, G.; Balachandran, P. V. Machine Learning for Materials Design and Discovery. *J. Appl. Phys.* **2021**, *129* (7), 070401.
(126) Jørgensen, P. B.; Schmidt, M. N.; Winther, O. Deep Generative Models for Molecular Science. *Mol. Inf.* **2018**, *37* (1−2), 1700133.
(127) Merz, K. M.; De Fabritiis, G.; Wei, G.-W. Generative Models for Molecular Design. *J. Chem. Inf. Model.* **2020**, *60* (12), 5635−5636.
(128) Amabilino, S.; Pogány, P.; Pickett, S. D.; Green, D. V. S. Guidelines for Recurrent Neural Network Transfer Learning-Based Molecular Generation of Focused Libraries. *J. Chem. Inf. Model.* **2020**, *60* (12), 5699−5713.
(129) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60* (12), 5682−5698.
(130) Boitreaud, J.; Mallet, V.; Oliver, C.; Waldispühl, J. OptiMol: Optimization of Binding Affinities in Chemical Space for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60* (12), 5658−5666.
(131) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **2020**, *32* (12), 4954−4965.
(132) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, 2012.
(133) Murphy, K. P. *Probabilistic Machine Learning: An Introduction*; The MIT Press: Cambridge, MA, 2022.
(134) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825−2830.
(135) *scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation*. https://scikit-learn.org/stable/index.html (accessed 2021-09-15).
(136) *Keras: the Python deep learning API*. https://keras.io/ (accessed 2020-10-06).
(137) *PyTorch*. https://www.pytorch.org (accessed 2021-09-18).
(138) Wolpert, D. H.; Macready, W. G. No Free Lunch Theorems for Optimization. *IEEE Trans. Evol. Comput.* **1997**, *1* (1), 67−82.
(139) Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016**, *7* (1), 11241.
(140) Loeffler, T. D.; Banik, S.; Patra, T. K.; Sternberg, M.; Sankaranarayanan, S. K. R. S. Reinforcement Learning in Discrete Action Space Applied to Inverse Defect Design. *J. Phys. Commun.* **2021**, *5* (3), 031001.
(141) M. Dieb, T.; Ju, S.; Yoshizoe, K.; Hou, Z.; Shiomi, J.; Tsuda, K. MDTS: Automatic Complex Materials Design Using Monte Carlo Tree Search. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 498−503.
(142) Chakraborti, N. Genetic Algorithms in Materials Design and Processing. *Int. Mater. Rev.* **2004**, *49* (3−4), 246−260.
(143) Arora, V.; Bakhshi, A. K. Molecular Designing of Novel Ternary Copolymers of Donor-Acceptor Polymers Using Genetic Algorithm. *Chem. Phys.* **2010**, *373* (3), 307−312.
(144) Mitra, K. Genetic Algorithms in Polymeric Material Production, Design, Processing and Other Applications: A Review. *Int. Mater. Rev.* **2008**, *53* (5), 275−297.
(145) Jaeger, H. M.; de Pablo, J. J. Perspective: Evolutionary Design of Granular Media and Block Copolymer Patterns. *APL Mater.* **2016**, *4* (5), 053209.
(146) Kasat, R. B.; Ray, A. K.; Gupta, S. K. Applications of Genetic Algorithm in Polymer Science and Engineering. *Mater. Manuf. Processes* **2003**, *18* (3), 523−532.
(147) Khaira, G. S.; Qin, J.; Garner, G. P.; Xiong, S.; Wan, L.; Ruiz, R.; Jaeger, H. M.; Nealey, P. F.; de Pablo, J. J. Evolutionary Optimization of Directed Self-Assembly of Triblock Copolymers on Chemically Patterned Substrates. *ACS Macro Lett.* **2014**, *3* (8), 747−752.
(148) Kanters, R. P. F.; Donald, K. J. Cluster: Searching for Unique Low Energy Minima of Structures Using a Novel Implementation of a Genetic Algorithm. *J. Chem. Theory Comput.* **2014**, *10* (12), 5729−5737.
(149) Deaven, D. M.; Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **1995**, *75* (2), 288−291.
(150) Fornleitner, J.; Lo Verso, F.; Kahl, G.; Likos, C. N. Genetic Algorithms Predict Formation of Exotic Ordered Configurations for Two-Component Dipolar Monolayers. *Soft Matter* **2008**, *4* (3), 480−484.
(151) Chua, A. L.-S.; Benedek, N. A.; Chen, L.; Finnis, M. W.; Sutton, A. P. A Genetic Algorithm for Predicting the Structures of Interfaces in Multicomponent Systems. *Nat. Mater.* **2010**, *9* (5), 418−422.
(152) Coley, D. A. *An Introduction to Genetic Algorithms for Scientists and Engineers*; World Scientific, 1999.
(153) Goldberg, D. E.; Deb, K. *Foundations of Genetic Algorithms (FOGA 1)*; Morgan Kaufmann: San Mateo, CA, 1991.
(154) Kaya, Y.; Uyar, M.; Tekdn, R. A Novel Crossover Operator for Genetic Algorithms: Ring Crossover. *ArXiv (Neural and Evolutionary Computing)*, May 2, **2011**, 11050355, ver. 1.
(155) Tipton, W. W.; Hennig, R. G. A Grand Canonical Genetic Algorithm for the Prediction of Multi-Component Phase Diagrams and Testing of Empirical Potentials. *J. Phys.: Condens. Matter* **2013**, *25* (49), 495401.
(156) Revard, B. C.; Tipton, W. W.; Hennig, R. G. Structure and Stability Prediction of Compounds with Evolutionary Algorithms. In

Prediction and Calculation of Crystal Structures: Methods and Applications; Atahan-Evrenk, S., Aspuru-Guzik, A., Eds.; Topics in Current Chemistry; Springer International Publishing: Cham, 2014; pp 181−222.
(157) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529* (7587), 484−489.
(158) Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S. A Survey of Monte Carlo Tree Search Methods. *IEEE Trans. Comput. Intell. AI Games* **2012**, *4* (1), 1−43.
(159) Kajita, S.; Kinjo, T.; Nishi, T. Autonomous Molecular Design by Monte-Carlo Tree Search and Rapid Evaluations Using Molecular Dynamics Simulations. *Commun. Phys.* **2020**, *3* (1), 1−11.
(160) Dieb, T. M.; Ju, S.; Shiomi, J.; Tsuda, K. Monte Carlo Tree Search for Materials Design and Discovery. *MRS Commun.* **2019**, *9* (2), 532−536.
(161) Kiyohara, S.; Mizoguchi, T. Searching the Stable Segregation Configuration at the Grain Boundary by a Monte Carlo Tree Search. *J. Chem. Phys.* **2018**, *148* (1), 241741.
(162) Kocsis, L.; Szepesvári, C. Bandit Based Monte-Carlo Planning. In *Machine Learning: ECML 2006*; Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Eds.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2006; pp 282−293.
(163) Świechowski, M.; Godlewski, K.; Sawicki, B.; Mańdziuk, J. Monte Carlo Tree Search: A Review of Recent Modifications and Applications. *ArXiv (Artificial Intelligence)*, March 9, **2021**, 210304931, ver. 2.
(164) Patra, T. K.; Loeffler, T. D.; Sankaranarayanan, S. K. R. S. Accelerating Copolymer Inverse Design Using Monte Carlo Tree Search. *Nanoscale* **2020**, *12* (46), 23653−23662.
(165) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.
(166) Griffiths, R.-R.; Hernández-Lobato, J. M. Constrained Bayesian Optimization for Automatic Chemical Design Using Variational Autoencoders. *Chem. Sci.* **2020**, *11* (2), 577−586.
(167) Deshwal, A.; Simon, C.; Doppa, J. R. Bayesian Optimization of Nanoporous Materials. *Mol. Syst. Des. Eng.* **2021**, *6*, 1066.
(168) Zhang, Y.; Apley, D. W.; Chen, W. Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. *Sci. Rep.* **2020**, *10* (1), 4924.
(169) Seko, A.; Togo, A.; Hayashi, H.; Tsuda, K.; Chaput, L.; Tanaka, I. Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization. *Phys. Rev. Lett.* **2015**, *115* (20), 205901.
(170) Ueno, T.; Rhone, T. D.; Hou, Z.; Mizoguchi, T.; Tsuda, K. COMBO: An Efficient Bayesian Optimization Library for Materials Science. *Mater. Discovery* **2016**, *4*, 18−21.
(171) González, J. *Gaussian Process Optimization Using GPy*; Sheffield Machine Learning Software, 2021.
(172) Khadilkar, M. R.; Paradiso, S.; Delaney, K. T.; Fredrickson, G. H. Inverse Design of Bulk Morphologies in Multiblock Polymers Using Particle Swarm Optimization. *Macromolecules* **2017**, *50* (17), 6702−6709.
(173) Paradiso, S. P.; Delaney, K. T.; Fredrickson, G. H. Swarm Intelligence Platform for Multiblock Polymer Inverse Formulation Design. *ACS Macro Lett.* **2016**, *5* (8), 972−976.
(174) DeStefano, A. J.; Segalman, R. A.; Davidson, E. C. Where Biology and Traditional Polymers Meet: The Potential of Associating Sequence-Defined Polymers for Materials Science. *JACS Au* **2021**, *1*, 1556.
(175) Perry, S. L.; Sing, C. E. 100th Anniversary of Macromolecular Science Viewpoint: Opportunities in the Physics of Sequence-Defined Polymers. *ACS Macro Lett.* **2020**, *9* (2), 216−225.
(176) Upadhya, R.; Murthy, N. S.; Hoop, C. L.; Kosuri, S.; Nanda, V.; Kohn, J.; Baum, J.; Gormley, A. J. PET-RAFT and SAXS: High Throughput Tools To Study Compactness and Flexibility of Single-Chain Polymer Nanoparticles. *Macromolecules* **2019**, *52* (21), 8295−8304.
(177) Sharma, S.; Kumar, S. K.; Buldyrev, S. V.; Debenedetti, P. G.; Rossky, P. J.; Stanley, H. E. A Coarse-Grained Protein Model in a Water-like Solvent. *Sci. Rep.* **2013**, *3* (1), 1841.
(178) Khokhlov, A. R.; Khalatur, P. G. Conformation-Dependent Sequence Design (Engineering) of AB Copolymers. *Phys. Rev. Lett.* **1999**, *82* (17), 3456−3459.
(179) Statt, A.; Kleeblatt, D. C.; Reinhart, W. F. Unsupervised Learning of Sequence-Specific Aggregation Behavior for a Model Copolymer. *Soft Matter* **2021**, *17* (33), 7697−7707.
(180) Bale, A. A.; Patra, T. K. Sequence Engineering of Copolymers Using Evolutionary Computing. *ArXiv (Materials Science)*, July 14, **2021**, 210706439, ver. 1.
(181) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*; NIPS'11; Curran Associates Inc.: Red Hook, NY, 2011; pp 2546−2554.
(182) Cigana, P.; Favis, B. D.; Jerome, R. Diblock Copolymers as Emulsifying Agents in Polymer Blends: Influence of Molecular Weight, Architecture, and Chemical Composition. *J. Polym. Sci., Part B: Polym. Phys.* **1996**, *34* (9), 1691−1700.
(183) Sundararaj, U.; Macosko, C. W. Drop Breakup and Coalescence in Polymer Blends: The Effects of Concentration and Compatibilization. *Macromolecules* **1995**, *28* (8), 2647−2657.
(184) Macosko, C. W.; Guégan, P.; Khandpur, A. K.; Nakayama, A.; Marechal, P.; Inoue, T. Compatibilizers for Melt Blending: Premade Block Copolymers. *Macromolecules* **1996**, *29* (17), 5590−5598.
(185) Meijer, H. E. H.; Lemstra, P. J.; Elemans, P. H. M. Structured Polymer Blends. *Makromol. Chem., Macromol. Symp.* **1988**, *16* (1), 113−135.
(186) Ruzette, A.-V.; Leibler, L. Block Copolymers in Tomorrow's Plastics. *Nat. Mater.* **2005**, *4* (1), 19−31.
(187) Eastwood, E. A.; Dadmun, M. D. Multiblock Copolymers in the Compatibilization of Polystyrene and Poly(Methyl Methacrylate) Blends: Role of Polymer Architecture. *Macromolecules* **2002**, *35* (13), 5069−5077.
(188) Anastasiadis, S. H.; Gancarz, I.; Koberstein, J. T. Compatibilizing Effect of Block Copolymers Added to the Polymer/Polymer Interface. *Macromolecules* **1989**, *22* (3), 1449−1453.
(189) Gersappe, D.; Balazs, A. C. Random Copolymers as Effective Compatibilizing Agents. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52* (5), 5061−5064.
(190) Dai, C.-A.; Dair, B. J.; Dai, K. H.; Ober, C. K.; Kramer, E. J.; Hui, C.-Y.; Jelinski, L. W. Reinforcement of Polymer Interfaces with Random Copolymers. *Phys. Rev. Lett.* **1994**, *73* (18), 2472−2475.
(191) Dong, W.; He, M.; Wang, H.; Ren, F.; Zhang, J.; Zhao, X.; Li, Y. PLLA/ABS Blends Compatibilized by Reactive Comb Polymers: Double Tg Depression and Significantly Improved Toughness. *ACS Sustainable Chem. Eng.* **2015**, *3* (10), 2542−2550.
(192) Huan, T. D.; Boggs, S.; Teyssedre, G.; Laurent, C.; Cakmak, M.; Kumar, S.; Ramprasad, R. Advanced Polymeric Dielectrics for High Energy Density Applications. *Prog. Mater. Sci.* **2016**, *83*, 236−269.
(193) Mannodi-Kanakkithodi, A.; Treich, G. M.; Huan, T. D.; Ma, R.; Tefferi, M.; Cao, Y.; Sotzing, G. A.; Ramprasad, R. Rational Co-Design of Polymer Dielectrics for Energy Storage. *Adv. Mater.* **2016**, *28* (30), 6277−6291.
(194) Lorenzini, R. G.; Kline, W. M.; Wang, C. C.; Ramprasad, R.; Sotzing, G. A. The Rational Design of Polyurea & Polyurethane Dielectric Materials. *Polymer* **2013**, *54* (14), 3529−3533.
(195) Sharma, V.; Wang, C.; Lorenzini, R. G.; Ma, R.; Zhu, Q.; Sinkovits, D. W.; Pilania, G.; Oganov, A. R.; Kumar, S.; Sotzing, G. A.; Boggs, S. A.; Ramprasad, R. Rational Design of All Organic Polymer Dielectrics. *Nat. Commun.* **2014**, *5* (1), 1−8.
(196) Ma, R.; Baldwin, A. F.; Wang, C.; Offenbach, I.; Cakmak, M.; Ramprasad, R.; Sotzing, G. A. Rationally Designed Polyimides for High-

Energy Density Capacitor Applications. *ACS Appl. Mater. Interfaces* **2014**, *6* (13), 10445−10451.
(197) Gopakumar, A. M.; Balachandran, P. V.; Xue, D.; Gubernatis, J. E.; Lookman, T. Multi-Objective Optimization for Materials Discovery via Adaptive Design. *Sci. Rep.* **2018**, *8* (1), 3738.
(198) Morris, M. A.; Sung, S. H.; Ketkar, P. M.; Dura, J. A.; Nieuwendaal, R. C.; Epps, T. H. Enhanced Conductivity via Homopolymer-Rich Pathways in Block Polymer-Blended Electrolytes. *Macromolecules* **2019**, *52* (24), 9682−9692.
(199) Cheng, Y.; Yang, J.; Hung, J.-H.; Patra, T. K.; Simmons, D. S. Design Rules for Highly Conductive Polymeric Ionic Liquids from Molecular Dynamics Simulations. *Macromolecules* **2018**, *51* (17), 6630−6644.
(200) Majewski, P. W.; Gopinadhan, M.; Jang, W.-S.; Lutkenhaus, J. L.; Osuji, C. O. Anisotropic Ionic Conductivity in Block Copolymer Membranes by Magnetic Field Alignment. *J. Am. Chem. Soc.* **2010**, *132* (49), 17516−17522.
(201) Sharon, D.; Bennington, P.; Dolejsi, M.; Webb, M. A.; Dong, B. X.; de Pablo, J. J.; Nealey, P. F.; Patel, S. N. Intrinsic Ion Transport Properties of Block Copolymer Electrolytes. *ACS Nano* **2020**, *14* (7), 8902−8914.
(202) Chintapalli, M.; Chen, X. C.; Thelen, J. L.; Teran, A. A.; Wang, X.; Garetz, B. A.; Balsara, N. P. Effect of Grain Size on the Ionic Conductivity of a Block Copolymer Electrolyte. *Macromolecules* **2014**, *47* (15), 5424−5431.
(203) Devaux, D.; Harry, K. J.; Parkinson, D. Y.; Yuan, R.; Hallinan, D. T.; MacDowell, A. A.; Balsara, N. P. Failure Mode of Lithium Metal Batteries with a Block Copolymer Electrolyte Analyzed by X-Ray Microtomography. *J. Electrochem. Soc.* **2015**, *162* (7), A1301−A1309.
(204) Bowden, M. J. Polymers for Electronic and Photonic Applications. In *Electronic and Photonic Applications of Polymers*; Advances in Chemistry; American Chemical Society, 1988; Vol. *218*, pp 1−73.
(205) Wei, X.; Luo, T. The Effect of the Block Ratio on the Thermal Conductivity of Amorphous Polyethylene-Polypropylene (PE-PP) Diblock Copolymers. *Phys. Chem. Chem. Phys.* **2018**, *20* (31), 20534−20539.
(206) Han, Z.; Fina, A. Thermal Conductivity of Carbon Nanotubes and Their Polymer Nanocomposites: A Review. *Prog. Polym. Sci.* **2011**, *36* (7), 914−944.
(207) Kim, G.-H.; Lee, D.; Shanker, A.; Shao, L.; Kwon, M. S.; Gidley, D.; Kim, J.; Pipe, K. P. High Thermal Conductivity in Amorphous Polymer Blends by Engineered Interchain Interactions. *Nat. Mater.* **2015**, *14* (3), 295−300.
(208) Wei, X.; Zhang, T.; Luo, T. Chain Conformation-Dependent Thermal Conductivity of Amorphous Polymer Blends: The Impact of Inter- and Intra-Chain Interactions. *Phys. Chem. Chem. Phys.* **2016**, *18* (47), 32146−32154.
(209) Choy, C. L.; Luk, W. H.; Chen, F. C. Thermal Conductivity of Highly Oriented Polyethylene. *Polymer* **1978**, *19* (2), 155−162.
(210) Freeman, B. D. Basis of Permeability/Selectivity Tradeoff Relations in Polymeric Gas Separation Membranes. *Macromolecules* **1999**, *32* (2), 375−380.
(211) Comesaña-Gándara, B.; Chen, J.; Bezzu, C. G.; Carta, M.; Rose, I.; Ferrari, M.-C.; Esposito, E.; Fuoco, A.; Jansen, J. C.; McKeown, N. B. Redefining the Robeson Upper Bounds for CO2/CH4 and CO2/N2 Separations Using a Series of Ultrapermeable Benzotriptycene-Based Polymers of Intrinsic Microporosity. *Energy Environ. Sci.* **2019**, *12* (9), 2733−2740.
(212) Robeson, L. M. The Upper Bound Revisited. *J. Membr. Sci.* **2008**, *320* (1), 390−400.
(213) Robeson, L. M. Correlation of Separation Factor versus Permeability for Polymeric Membranes. *J. Membr. Sci.* **1991**, *62* (2), 165−185.
(214) Park, H. B.; Kamcev, J.; Robeson, L. M.; Elimelech, M.; Freeman, B. D. Maximizing the Right Stuff: The Trade-off between Membrane Permeability and Selectivity. *Science* **2017**, DOI: 10.1126/science.aab0530.
(215) Adhikari, S.; Nikoubashman, A.; Leibler, L.; Rubinstein, M.; Midya, J.; Kumar, S. K. Gas Transport in Interacting Planar Brushes. *ACS Polym. Au* **2021**, *1* (1), 39−46.
(216) Bilchak, C. R.; Huang, Y.; Benicewicz, B. C.; Durning, C. J.; Kumar, S. K. High-Frequency Mechanical Behavior of Pure Polymer-Grafted Nanoparticle Constructs. *ACS Macro Lett.* **2019**, *8* (3), 294−298.
(217) Everaers, R.; Karimi-Varzaneh, H. A.; Fleck, F.; Hojdis, N.; Svaneborg, C. Kremer-Grest Models for Commodity Polymer Melts: Linking Theory, Experiment, and Simulation at the Kuhn Scale. *Macromolecules* **2020**, *53* (6), 1901−1916.
(218) DeCost, B. L.; Hattrick-Simpers, J. R.; Trautt, Z.; Kusne, A. G.; Campo, E.; Green, M. L. Scientific AI in Materials Science: A Path to a Sustainable and Scalable Paradigm. *Mach. Learn. Sci. Technol.* **2020**, *1* (3), 033001.
(219) Rodrigues, J. F.; Florea, L.; de Oliveira, M. C. F.; Diamond, D.; Oliveira, O. N. Big Data and Machine Learning for Materials Science. *Discovery Mater.* **2021**, *1* (1), 12.
(220) Gormley, A. J.; Webb, M. A. Machine Learning in Combinatorial Polymer Chemistry. *Nat. Rev. Mater.* **2021**, *6* (8), 642−644.
(221) de Pablo, J. J.; Jones, B.; Kovacs, C. L.; Ozolins, V.; Ramirez, A. P. The Materials Genome Initiative, the Interplay of Experiment, Theory and Computation. *Curr. Opin. Solid State Mater. Sci.* **2014**, *18* (2), 99−117.
(222) *Materials Genome Initiative*. https://www.mgi.gov/ (accessed 2020-09-05).
(223) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; DeLongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Rodriguez, E.; Stevanovic, V.; Suntivich, J.; Thornton, K.; Zhao, J.-C. New Frontiers for the Materials Genome Initiative. *Npj Comput. Mater.* **2019**, *5* (1), 1−23.
(224) DIT *FAIR Data*. https://www.dit.ie/dsrh/data/fairdata/ (accessed 2021-07-22).
(225) Draxl, C.; Scheffler, M. Big Data-Driven Materials Science and Its FAIR Data Infrastructure. In *Handbook of Materials Modeling: Methods: Theory and Modeling*; Andreoni, W., Yip, S., Eds.; Springer International Publishing: Cham, 2020; pp 49−73.
(226) Draxl, C.; Scheffler, M. NOMAD: The FAIR Concept for Big Data-Driven Materials Science. *MRS Bull.* **2018**, *43* (9), 676−682.
(227) Zhang, X.; Zhao, J.; Ye, C.; Lai, T.-Y.; Snyder, C. R.; Karim, A.; Cavicchi, K. A.; Simmons, D. S. Dynamical Correlations for Statistical Copolymers from High-Throughput Broad-Band Dielectric Spectroscopy. *ACS Comb. Sci.* **2019**, *21* (4), 276−299.
(228) 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* (7815), 237−241.
(229) Li, J.; Tu, Y.; Liu, R.; Lu, Y.; Zhu, X. Toward "On-Demand" Materials Synthesis and Scientific Discovery through Intelligent Robots. *Adv. Sci.* **2020**, *7* (7), 1901957.
(230) Zitzler, E.; Thiele, L.; Laumanns, M.; Fonseca, C. M.; da Fonseca, V. G. Performance Assessment of Multiobjective Optimizers: An Analysis and Review. *IEEE Trans. Evol. Comput.* **2003**, *7* (2), 117−132.
(231) Tušar, T.; Filipič, B. Visualization of Pareto Front Approximations in Evolutionary Multiobjective Optimization: A Critical Review and the Prosection Method. *IEEE Trans. Evol. Comput.* **2015**, *19* (2), 225−245.
(232) Agrawal, G.; Bloebaum, C.; Lewis, K.; Chugh, K.; Huang, C.-H.; Parashar, S. Intuitive Visualization of Pareto Frontier for Multiobjective Optimization in N-Dimensional Performance Space. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*; Multidisciplinary Analysis Optimization Conferences; American Institute of Aeronautics and Astronautics, 2004.
(233) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias Free Multiobjective Active Learning for Materials Design and Discovery. *Nat. Commun.* **2021**, *12* (1), 2312.