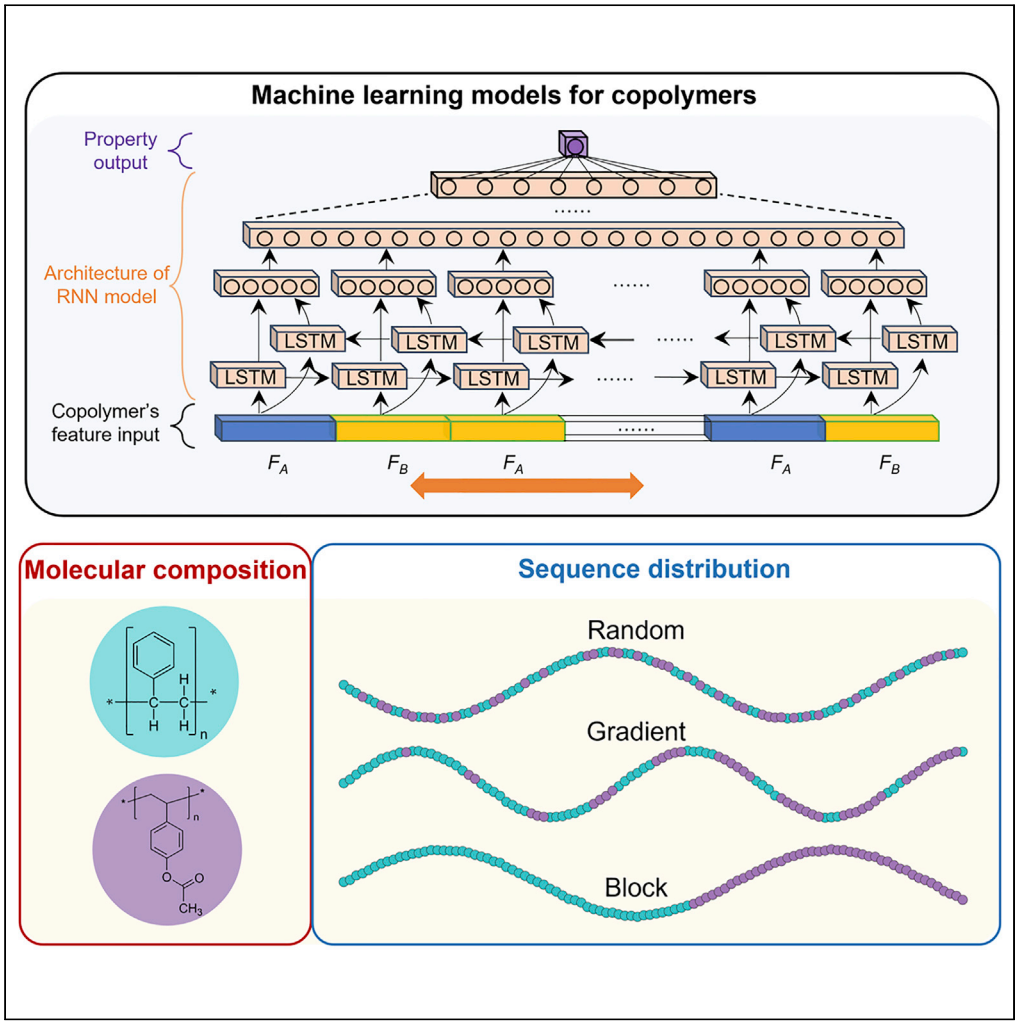
iScience

CellPress
OPEN ACCESS

Article

Machine learning strategies for the structure-property relationship of copolymers



Lei Tao, John Byrnes, Vikas Varshney, Ying Li

ying.3.li@uconn.edu

Highlights

Establish structure-property relationships of copolymer with machine learning (ML)

Incorporate both chemical composition and sequential distribution of copolymers

Analyze various copolymer types with different models in a unified approach

Differentiate the effects of random, block, and gradient patterns of copolymers

Tao et al., iScience 25, 104585
July 15, 2022
https://doi.org/10.1016/j.isci.2022.104585

iScience

CellPress
OPEN ACCESS

Article

Machine learning strategies for the structure-property relationship of copolymers

Lei Tao,[1] John Byrnes,[2] Vikas Varshney,[3] and Ying Li[1,4,5,*]

SUMMARY

Establishing the structure-property relationship is extremely valuable for the molecular design of copolymers. However, machine learning (ML) models can incorporate both chemical composition and sequence distribution of monomers, and have the generalization ability to process various copolymer types (e.g., alternating, random, block, and gradient copolymers) with a unified approach are missing. To address this challenge, we formulate four different ML models for investigation, including a feedforward neural network (FFNN) model, a convolutional neural network (CNN) model, a recurrent neural network (RNN) model, and a combined FFNN/RNN (Fusion) model. We use various copolymer types to systematically validate the performance and generalizability of different models. We find that the RNN architecture that processes the monomer sequence information both forward and backward is a more suitable ML model for copolymers with better generalizability. As a supplement to polymer informatics, our proposed approach provides an efficient way for the evaluation of copolymers.

INTRODUCTION

Polymers are one of the most important material classes that exhibit tremendous modularity in a variety of properties, including thermo-physical properties and thermal stability, chemical resistance, elastic and failure mechanical strength, electronic & optoelectronic properties, and so forth. As their wide range of properties is derived from their diverse molecular structures, understanding polymer's structure-property relationships, namely polymer informatics, is essential for evaluating polymer performance. (Audus and De Pablo, 2017, Chen et al., 2021b; Kim et al., 2018; Doan Tran et al., 2020) In homopolymers consisting of identical monomers, the physical properties are mainly governed by molecular compositions. (Tao et al., 2021b; Ma et al., 2019) For instance, machine learning (ML) methods using chemical inputs of molecular compositions have been successfully applied to predict many homopolymer properties accurately, including glass transition temperature (Tao et al., 2021a, 2021b; Chen et al., 2021a; Kim et al., 2018; Kuenneth et al., 2021a; Ramprasad and Kim, 2019), thermal conductivity (Wu et al., 2019), dielectric constants (Chen et al., 2020), organic photovoltaic properties (Sun et al., 2019; Gómez-Bombarelli et al., 2016; Wheatle et al., 2020), and different transport properties. (Barnett et al., 2020; Liu et al., 2020; Gao et al., 2021a; Yuan et al., 2021). The input for these ML models can be SMILES, fingerprints, and physicochemical descriptors, and so forth that are derived from the geometry and composition of a molecule—namely the molecular composition (monomer chemistry or chemical constituents) of polymers.

Yet, for copolymers that are made of more than one type of monomer, the sequential distribution of monomers along the polymer's backbone also affects these properties significantly (Perry and Sing, 2020; Porel and Alabi, 2014; Meier and Barner-Kowollik, 2019). For example, the copolymer of poly(ethylene terephthalate)/poly(ethylene sebacate) is one of the first studied copolyesters and can be either random or block copolymer. Compared to the random copolymer, its block copolymer has a higher melting point and remarkable elastic properties (Hale Charch and Shivere, 1959). For nylon-6,6/nylon-6 copolymers, its block copolymer leads to a substantially higher tensile strength than its random counterpart. (Kenney, 1968) A similar improvement of mechanical properties has been observed in the case of the 2-ethyl-2-oxazoline/2-nonyl-2-oxazoline copolymer; it displays a higher stiffness with a block arrangement than that with a random arrangement (Fijten et al., 2007) Moreover, the monomer sequence distribution is found to strongly affect the copolymer's other properties such as interfacial activity, (Lefebvre et al., 2005) solid-state properties, (Palermo and Mcneil, 2012) dielectric properties, (Mok et al., 2010) and so forth. However,

[1]Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA

[2]SRI International, San Diego, CA 92131, USA

[3]Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio 45433, USA

[4]Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, CT 06269, USA

[5]Lead contact

*Correspondence: ying.3.li@uconn.edu

https://doi.org/10.1016/j.isci.2022.104585

despite these extensive studies, it is still challenging to evaluate the property and performance of copolymers with different molecular compositions and monomer sequences in an accelerated manner.

Glass transition temperature (T_g) is among the most studied properties in different classes of copolymers. Toward that, several theoretical and empirical equations have been proposed, including the Fox equation (Fox, 1956), the Gordon-Taylor equation (Gordon and Taylor, 1952), and the Gibbs-DiMarzio equation. (Dimarzio and Gibbs, 1959) In the context of copolymers, these equations only consider their molecular compositions but neglect the effect of monomers' sequence distribution (Daimon et al., 1975). Although Barton and Johnston later proposed modified equations that include the monomer's arrangement in the analysis (Barton, 1970; Johnston, 1976), their dyad model has an intrinsic limitation: when the concentration of the AB dyad is low, as is the case for block copolymers, these modified equations are no more applicable (Suzuki and Miyamoto, 1989). To address the limitation of semi-empirical equations, computational methods are utilized, including molecular dynamics (MD) and density functional theory (DFT) simulations (Binder, 1995; Labanowski and Andzelm, 2012). They have demonstrated their advantages in dealing with complex copolymer sequences related to different properties. For example, the composition or sequence dependence of glass transition temperatures (Bejagam et al., 2021), thermal conductivity (Zhou et al., 2021), and interfacial energy (Meenakshisundaram et al., 2017) have been accurately simulated, in which the modeling of random, block, or alternating copolymers directly compare the performances of different copolymer types. Even though computational modeling is a powerful tool to reveal structure-property relationships of copolymers, it has high computational complexity and cost and must be carried out case by case cautiously.

Recently, with the rapid advancements in polymer informatics, the data-driven analysis offers an alternative & efficient solution to build the structure-property relationships for copolymers (Nguyen et al., 2021; Werner et al., 2020; Zhou et al., 2021). Ramprasad and co-workers (Kuenneth et al., 2021b) collected thermal properties of both homopolymers and copolymers to develop a copolymer informatics tool that makes predictions for three thermal properties, including glass transition temperature, melting temperature, and thermal degradation temperature. They assumed all copolymers to be random copolymers; thus, only monomer composition information is included in ML models without considering their sequences. Based on random copolymers and homopolymers, Hanaoka (2020), Leibfarth et al., (Reis et al., 2021), Kosuri et al. (2022), Shi et al. (2021), Tamasi et al. (2022), and Pilania et al. (2019) also to extend their ML models from single-component polymers (homopolymers) to multi-component polymers (copolymers). The importance of the composition information has been emphasized a lot in these ML studies of copolymers. Yet, the sequential distribution of different monomers has not been incorporated into most ML models until recently when Webb and co-workers (Patel et al., 2022) proposed two featurization paradigms that explicitly represent the monomer sequence: a sequence graph that uses edges to indicate the monomer arrangement and a sequence tensor that tracks monomer ordering in copolymers. Their proposed ML models are mainly based on coarse-grained modeling data of copolymers and focus on the arrangement of constitutional units of backbone beads and pendant beads. Although they represent monomers with beads of different types, highlighting the arrangement of bead topologies in copolymer sequence, the specific monomer chemistry is lacking owing to the coarse-grained nature of the beads. To sum up, ML models such as FFNN, CNN, and RNN have been utilized widely for copolymers. They can be applied for alternating, random, or block copolymers. They can also be applied based on coarse-grained modeling results of copolymers to consider the arrangement of bead topologies in copolymer sequence. Investigations on ML models that can incorporate both chemical composition and sequence distribution of monomers, and have generalization ability to process various copolymer types (e.g., alternating, random, block, and gradient copolymers) with a unified approach are missing to the best of authors' knowledge. To fully understand and identify appropriate ML strategies that can use a unified approach for various copolymer types, a systematic investigation of different ML models for copolymer informatics is timely.

To address the above issue, we focus on the applicability and generalization ability of ML models that incorporate the information on both molecular composition and sequence distribution of copolymers. And multiple copolymer types are considered including random, block, alternating, and gradient copolymers. The gist of this study is compared with others in Table 1. ML model's generalization ability is the center of interest in this study, namely whether a proposed ML model can be applied to different copolymer sequence distributions and molecular compositions. Our study doesn't rely on the simulation model of generic beads to consider the chain sequence of polymers. Although more chain-level features such as chain length and branch can also be modulated with coarse-grained (CG) simulations, such CG simulation

Table 1. Comparison of copolymer dataset processed with ML models in literature

Reference	Data origin[a]	Copolymer types[b]	Composition info utilized in ML[c]	Sequence info utilized in ML[d]
(Werner et al., 2020)	CG simulation	Random	CG beads (2 beads)	CG simulated chain length $\leq$ 16
(Kuenneth et al., 2021b)	Experiment	Random	Chemistries (1,569 molecules)	No sequence info
(Hanaoka, 2020)	Experiment	Random	Chemistries (12 molecules)	No sequence info
	Experiment	Random	Chemistries (55 molecules)	No sequence info
(Reis et al., 2021)	Experiment	Random	Chemistries (6 molecules)	No sequence info
(Pilania et al., 2019)	Experiment	Random	Chemistries (16 molecules)	No sequence info
(Shi et al., 2021)	CG simulation	Random	CG beads (2 beads)	CG simulated chain length = 20
(Webb et al., 2020)	CG simulation	Alternate/Random	CG beads (4 beads)	CG simulated chain length = 400
(Patel et al., 2022)	CG simulation	Random	Chemistries (20 molecules)	CG simulated chain length = 20–600
(Patel et al., 2022)	CG simulation	Alternate/Random	CG beads (4 beads)	CG simulated chain length = 400
(Patel et al., 2022)	Experiment	Random	Chemistries (6 molecules)	No sequence info
(Patel et al., 2022)	CG simulation	Random	CG beads (2 beads)	CG simulated chain length = 20
This work	DFT simulation	Alternate	Chemistries (586 molecules)	Sequence pattern of monomers
This work	Experiment	Random	Chemistries (6 molecules)	Sequence pattern of monomers
This work	Experiment	Random/Block	Chemistries (16 molecules)	Sequence pattern of monomers
This work	Experiment	Random/Block + Gradient	Chemistries (1,433 molecules)	Sequence pattern of monomers

[a]Coarse-grained (CG) simulations are based on generic beads of monomers. DFT simulations are based on molecules' chemistries.
[b]Types of copolymers indicate the sequence pattern.
[c]"CG beads" indicate no chemistry information is used to characterize monomers. "Chemistries" indicate the chemistries of monomers are utilized by ML models.
[d]"No sequence info" means there is no sequence information utilized by ML models. CG modeling of different arrangements of beads outputs different chain lengths and sequences. Chain length is not defined for the experimental dataset of copolymers whose sequence pattern is known.

requires extra modeling based on generic beads, and without a direct experimental benchmark for most experimentally reported polymers. A strategy to directly use the experimentally reported polymer representations (monomers/repeat units) and the sequence patterns (characterized by alternating, block, random, or gradient) leads to ML models with the most applicability to various copolymer types in a straightforward and unified manner. To this end, we formulate four ML models based on neural networks: a feedforward neural network (FFNN) model, a convolutional neural network (CNN) model, a recurrent neural network (RNN) model, and a combined FFNN/RNN (Fusion) model. We compare their performance on four distinct datasets related to different physical properties of copolymers and further examine these ML models using recent experimental results of glass transition temperature from gradient copolymers (Kim et al., 2006; Alshehri et al., 2022). Our results reveal the applicability of ML models on various copolymer types and identify the most generalizable model for copolymers' property predictions. Specifically, we find that the CNN and RNN models can be well generalized to different copolymer types. The RNN architecture that processes the monomer sequence information both forward and backward is a more suitable ML model for copolymers. As ML models have better generalization ability, computational efficiency, and architecture flexibility over theoretical equations or molecular simulations, we foresee that the developed ML models will facilitate the evaluation and development of sequence-defined copolymers for many applications, such as thermoplastic elastomers (Guo et al., 2015; Nanjan and Porel, 2019; Meier and Barner-Kowollik, 2019), polyelectrolytes (Wheatle et al., 2020; Jablonka et al., 2021; Sing, 2020), nanofabrication and synthesis, (Tu et al., 2020; Statt et al., 2021), drug delivery (Werner et al., 2020; Deng et al., 2021), and so forth.

RESULTS

Types of copolymers to be investigated

Copolymers have at least two types of monomers and can have different monomer sequence distributions, such as random, alternating, block, and gradient copolymers. With the advancement in synthesis techniques (Badi and Lutz, 2009; Lutz et al., 2013, 2016; Lutz et al., 2013), sequence-defined polymers — polymers where each monomer unit is at a defined position of the chain, similar to proteins and

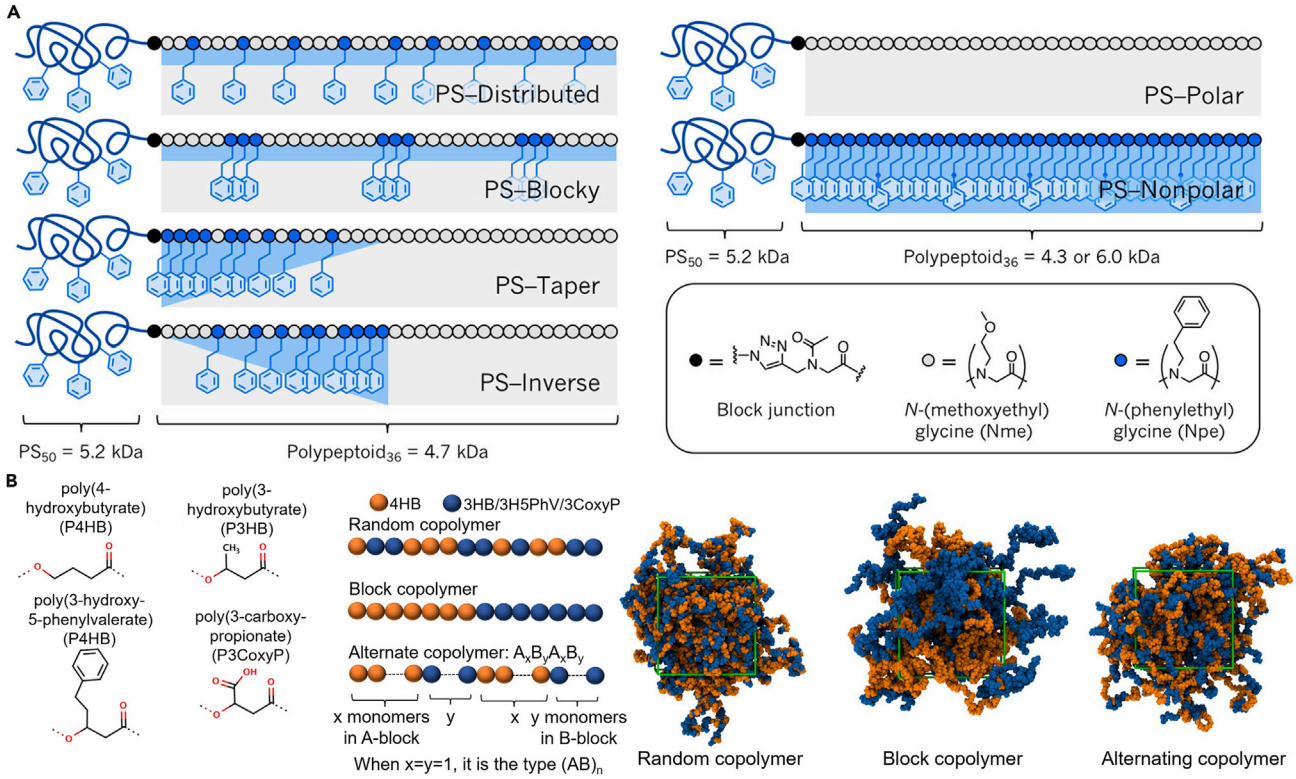


Figure 1. Monomer sequence distributions of different copolymer types

(A) The copolymers composed of two monomers (polar and nonpolar monomers) follow block sequence distributions. The figure is reprinted with permission from ref (Patterson et al., 2019). Copyright 2019 American Chemical Society.

(B) The copolymers composed of two monomers follow random, block, and alternating sequence distributions. The figure is reprinted with permission from ref (Bejagam et al., 2021). Copyright 2021 American Chemical Society.

oligonucleotides — are emerging (Lehto and Wagner, 2014; Fred Dice, 1990; Lupas et al., 1991; Mewes et al., 2002; Kuhlman and Baker, 2000). Compared to classical random and block copolymers, these sequence-defined polymers provide enormous opportunities for materials design, with tailored structural and mechanical properties (Nanjan and Porel, 2019; Leibfarth et al., 2015; Meier and Barner-Kowollik, 2019). Their polymer chain structures have more complex monomer arrangements than homopolymers. If illustrated with two monomers, "A" and "B," Figure 1 features the polymer chains of different copolymers, adapted from References (Patterson et al., 2019; Bejagam et al., 2021). Alternating copolymers have regular alternating units, and the simplest type may be regarded as homopolymers with a repeat unit composed of the two monomers "$(AB)_n$." Random copolymers, on the contrary, have totally unregular sequences. Their two monomers, A and B, are located randomly along the polymer chain. In gradient copolymers, the monomer composition changes gradually from one monomer to the other, and each monomer is predominantly located at one segment of the chain. Unlike gradient copolymers, block copolymers have a chain of different blocks, and each block is composed of the same monomer type. As a result, there is an abrupt change in monomer from one block to another. With these considerations in mind, the main targets of the ML model design for copolymers are to simultaneously incorporate: (1) the monomer's chemical composition into the model; & (2) monomers' sequence information into the model.

Established ML models that are applicable to various copolymer types

To incorporate the information on both molecular composition and sequence distribution of copolymers, four ML models are established whose architectures are suitable for copolymers' feature vectors. Figure 2 shows the architectures of our four ML models: an FFNN model, a CNN model, an RNN model, and an FFNN/RNN (Fusion) model. All models require proper feature engineering of copolymers so that various copolymer types can be processed in a unified manner (see STAR Methods for the feature engineering and architecture details). Although the FFNN model architecture is more applicable to random copolymers, the

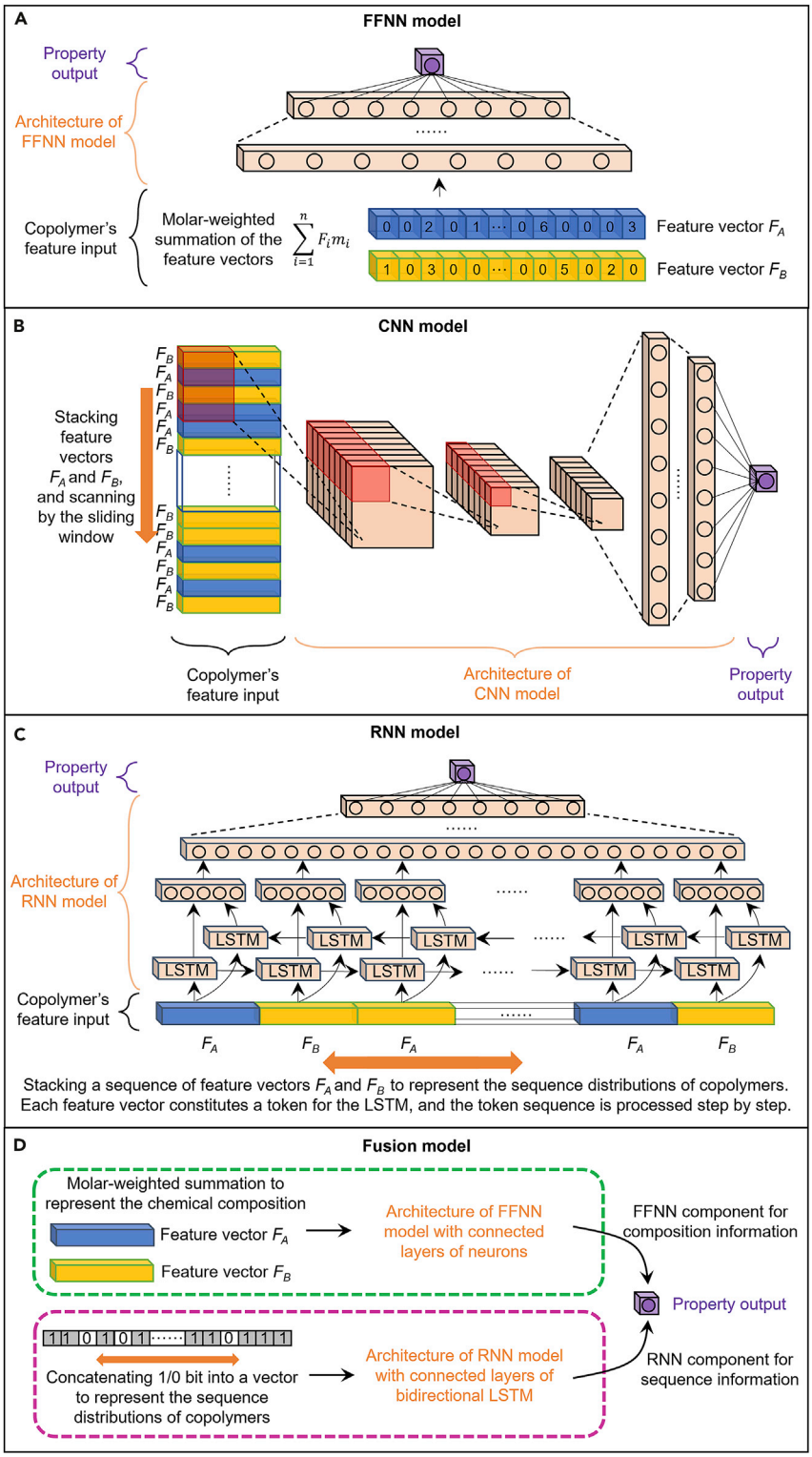


Figure 2. Architectures of four examined machine learning models for copolymers

(A) Feedforward neural network (FFNN) model.

(B) Convolutional neural network (CNN) model.

(C) Recurrent neural network (RNN) model.

(D) FFNN and RNN combined (fusion) model. See also Figures S2–S5.

other three model architectures have a better ability to process the sequence information of copolymers such as block and gradient copolymers.

Applications of four ML models on datasets of varying types of copolymers

The applications of these four ML models will be demonstrated one by one on different copolymer datasets. Among the increasing amounts of copolymer data, we have organized four datasets with varying types of copolymers, including alternating, random, and block copolymers. Gradient copolymers are also collected for the further validation of different models subsequently. To examine the generalization ability of ML models, we deliberately include multiple copolymer types and consider different properties. Using various types of datasets and properties lead to a comprehensive and unbiased exploration of ML strategies for copolymers. Table 2 summarizes the information of these four datasets. Dataset one is based on 5000 DFT calculated optoelectronic properties of conjugated polymers by Zwijnenburg et al. (Wilbraham et al., 2019) The target quantities used here are ionization potential (IP) and electron affinity (EA). Dataset two is from an experimental study of the high-contrast ^{19}F magnetic resonance imaging (MRI) agents. The measured ^{19}F nuclear magnetic resonance (NMR) signal-to-noise ratio (SNR) indicates the performance of copolymers as ^{19}F NMR agents (Reis et al., 2021). Dataset three is for specific polyhydroxyalkanoate (PHA)-based polymers and their glass transition temperature T_g (Pilania et al., 2019). Dataset four is collected from a publicly accessible database, PoLyInfo (Otsuka et al., 2011), and consists of more than 6600 copolymers of different classes with experimentally reported T_g values. Together, these four datasets provide diverse copolymer inputs and property targets for our proposed ML models (Figure 2). We should point that that they occupy different areas in chemical spaces as illustrated in Supplemental information Figure S1; such diversities are a prerequisite for evaluating the generalizability of ML models.

We first randomly split each dataset into an 80% training set and 20% testing set during each ML model training. The training set is used to tune the model parameters to obtain a structure-property relationship (see STAR Methods for the model parameters of each ML model). The testing set is used to evaluate the performance of the ML model on previously unseen data. We compute the determination coefficient R^2 to evaluate the predictive performance. It examines how a model predicts an outcome as a percentage and is easier to compare than other indices such as mean absolute error (MAE), mean square error (RMSE), or root-mean-square error (RMSE).

ML models on dataset 1 – Conjugated copolymers with optoelectronic properties

Dataset one is for conjugated binary copolymers whose two composition monomers come from a pool of 586 monomeric units (Wilbraham et al., 2019). All the samples are alternating copolymers in which the regular pattern -A-B-A-B- is assumed. The usual treatment of such alternating copolymer is to connect two monomers into a dimer -AB- as the repeat unit, representing two-monomer copolymers by their homopolymer counterpart. Herein instead of providing the structure of the A/B monomer or using the dimer as the repeat unit, Dataset one is available in trimers as the representation of the copolymer (Figure 3A) from the study of Wilbraham et al. (76). From a perspective of feature engineering, a trimer that connects three monomers -ABA-is the same as a dimer representation because the structural features within and between two monomers are equally preserved. When only trimers or dimers are provided to represent copolymers, we expect the copolymer ML models to be applicable in such special cases as homopolymers. Figure 3A displays some monomers among these 586 monomers, including aromatic dibromides and distannanes, as well as building blocks from the organic photovoltaics. When they are combined into possible copolymer structures in the form of trimers, the connection happens at polymerization positions indicated by the "*" symbol.

As a trimer is used to represent copolymers in Dataset 1, both sequence distribution, as well as the molecular composition, has been embedded in the feature representation. When the copolymer dataset is provided in a homopolymer fashion (A-B-A as a single entity), there are no different monomers given for weighted summation to be applied. Although our FFNN model doesn't have a weighted summation vector to take in, it can directly use the feature vector of the trimer (A-B-A) repeat unit. Similarly, there are no monomers given for CNN and RNN models to consider the monomer arrangement, but the architectures of CNN and RNN do require the stacking of monomers. A working solution is to stack 100 trimers as the feature vector for CNN and RNN, with which their architectures are still appropriate for these special homopolymer-like cases. The performances of these four ML models are compared in Figure 3B. They all demonstrate good performance in handling alternating copolymers. Figure 3C shows the performance of the FFNN model in the

Table 2. Summary of four datasets for copolymers

Datasets	Copolymer type	Number of monomer molecules	Number of data points	Property	Source
1	Alternating	586	5000	Ionization Potential (IP) Electron Affinity (EA)	DFT calculations (Wilbraham et al., 2019)
2	Random	6	271	^{19}F NMR Signal-to-Noise Ratio (SNR)	Experiments (Reis et al., 2021)
3	Block, Random	16	131	Glass Transition Temperature T_g	Experiments (Pilania et al., 2019)
4	Block, Random	1,433	6629	Glass Transition Temperature T_g	Experiments (Otsuka et al., 2011)

See also Figure S1.

literature (Wilbraham et al., 2019) and our four ML models show comparable performance. The neural network model in the literature reports a root-mean-square error (RMSE) of less than 0.12 eV −IP/−EA. To make a direct comparison, we also calculate RMSE for our model and observed it to be around 0.09-0.19 eV. When creating the parity plots colored by point density as in Figure 3C, the similar performance between our models and the literature model is clearly displayed (see Supplemental information Figure S9 for the parity plot colored by point density). Based on the R^2 of the training and testing sets, we found that the RNN is the best model while CNN is the worst among the studied models. As ML models are context-dependent (Patel et al., 2022), a model that performs well on one problem may not work well on a different one. Dataset 1 with regular alternating sequence serves as the most standard test for copolymer ML models, and all these four ML models demonstrate exemplary performance in terms of R^2.

ML models on dataset 2 – Copolymer as ^{19}F MRI agents with signal intensity

Dataset two is developed through numerous experimental-computational cycles for next-generation ^{19}F MRI agents (Reis et al., 2021). Six monomer types are used to synthesize random copolymers while their ^{19}F NMR spectra are examined. Figure 4A illustrates the structure of these six components. The use of partially fluorinated monomers such as trifluoroethyl acrylate (TFEA) with hydrophilic monomers such as poly(ethylene glycol) acrylate (PEGA) provides ^{19}F MRI agents with moderate sensitivity. Figure 4B shows the composition of some samples and their corresponding ^{19}F MRI SNR values. Dataset two features random copolymers with more than two types of monomers. They are more complex than the two-monomer alternating copolymer in Dataset 1, but their molecular composition and monomer's sequence are still manageable by our four ML models.

Dataset two feature vectors for the FFNN model were calculated using the molar-weighted summation of each monomer's feature vector up to six monomer types as required by the specific polymer in the dataset. To include sequence information of the random copolymers in CNN, RNN, and Fusion models, we randomize the stacking of monomers' feature vectors keeping the number of each monomer in the same proportion as their composition in the copolymers. Each random copolymer is represented using one randomized sequence (See Supplemental information Figure S10 for the representativeness of using five randomized sequences to represent a copolymer). The performance of these four ML models on Dataset two of six-monomer random copolymers is compared in Figure 4C. Based on the R^2 of the training and testing set, we observe no significant difference in their predictive performance. RNN and Fusion models are observed to be slightly better than the other two (based on Test R^2).

ML models on dataset 3 – Polyhydroxyalkanoate with glass transition temperature

The previous Dataset one and Dataset two are for alternating and random copolymers, respectively. Their sequence distributions are either regular patterns that can be considered homopolymers or random patterns without specific monomer arrangements. Herein we use Dataset 3, which consists of both random and block polyhydroxyalkanoate copolymers as a new test for our ML models. Polyhydroxyalkanoate is a class of biosynthesized polymers that can be obtained from 150 different types of monomers. Sixteen monomer types are utilized in Dataset three to form random and block copolymers of different compositions and properties. Figure 5A illustrates some of the monomers involved during the synthesis of copolymers in Dataset 3. One of the well-defined features in their structures is: a −C(=O)–* dangling bond is always passivated by a −O–* dangling bond and vice versa.

For FFNN, the molecular composition of block copolymers is considered the same way as random and alternating copolymers by using the molar-weighted summation of each monomer's feature vector to

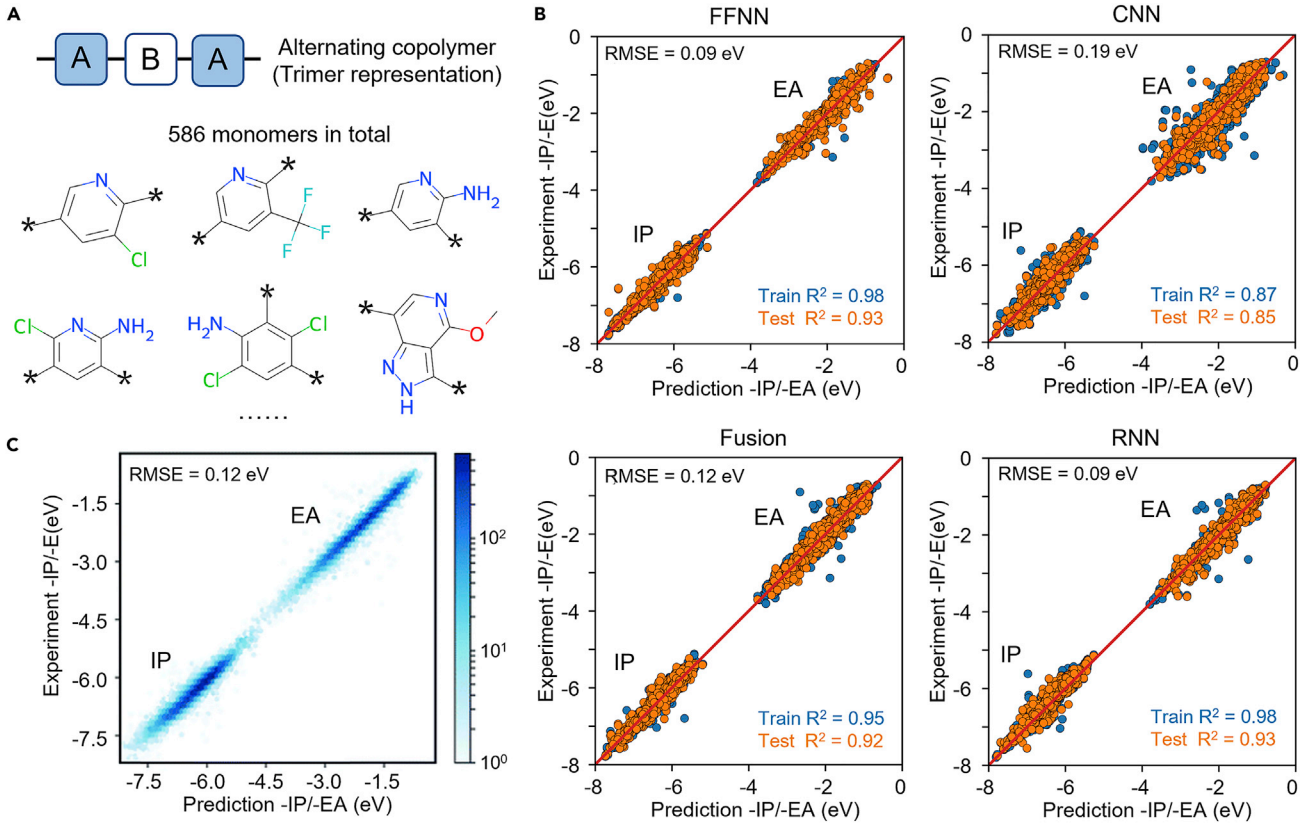


Figure 3. Performance of four ML models on copolymer Dataset 1

(A) The trimer representation of the alternating copolymer and examples of monomers used to build the conjugated copolymer.

(B) The parity plot of the four ML predicted -IP/-EA versus the DFT values.

The bottom left group is for IP and the top right group is EA. The RMSE of our models is calculated for a direct comparison with the reference.

(C) The parity plot of ML predicted -IP/-EA versus the DFT values, reproduced from reference (Wilbraham et al., 2019) with permission from the Royal Society of Chemistry. See also Figure S9.

obtain the block copolymers' feature vector. For CNN, RNN, and Fusion models, we consider the block copolymer by stacking monomer A and monomer B into two blocks. The number of each monomer is in the same proportion as its molar ratio in the copolymer, but their sequence distribution is in a block-by-block fashion. The parity plots in Figure 5B show that our four ML models can handle the case of block copolymers similarly. Their predictive performance is comparable to the random forest (RF) model in the literature (Figure 5C) (Pilania et al., 2019) We calculate the RMSE and Pearson correlation for our models so that our results can be compared directly with the Reference (Pilania et al., 2019). One important aspect of Dataset three is that it contains a limited number (8) of block copolymer samples among 131 data points. The dominant random copolymers in Dataset three exercise control over ML models' training and performance, which is why the FFNN model that cannot take into account the monomer's sequence of copolymers still performs well in terms of the R^2 of the training and testing sets. In addition, we believe that the molecular composition of copolymers acts as a primary factor that determines T_g is well captured by our FFNN model. If the compositions of two monomers are fixed or somewhat similar, then the sequence distribution of two monomers starts to have noticeable effect on T_g. Examples of such cases are demonstrated in the later discussion in Figure 8.

ML models on dataset 4 – Copolymers in PoLyInfo with glass transition temperature

To better examine these four ML models on a larger dataset containing block copolymers, we collect 6629 copolymers composed of two components from the PoLyInfo dataset (Otsuka et al., 2011). 1,433 monomers are identified from the 6629 block and random polymers in Dataset 4, and the chemical structures of some monomers are illustrated in Figure 6A. Dataset four is composed of random and block copolymers. Although block copolymers are only a small fraction of the total, the decent amount of 331

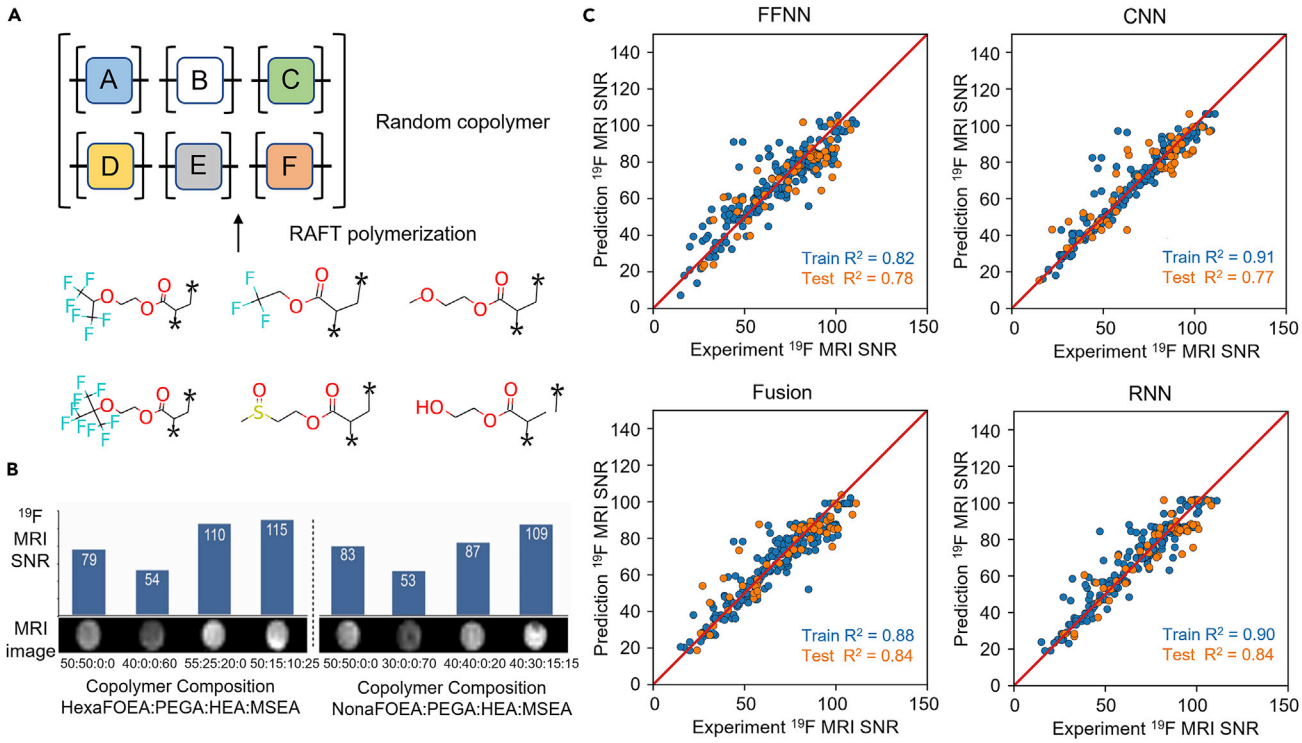


Figure 4. Performance of four ML models on copolymer Dataset 2

(A) The six monomer types polymerize into random copolymers via reversible addition-fragmentation chain transfer (RAFT).

(B) Compositions of eight copolymer samples and their SNR values.

The figure is reprinted with permission from ref (Reis et al., 2021). Copyright 2021 American Chemical Society.

(C) The parity plot of the four ML predicted SNR versus the experimental values. See also Figure S10.

samples is sufficient to activate the effect of the monomer's sequence on the model performance. More diverse chemical space is covered by Dataset 4 (see Supplemental information Figure S1 for its chemical space), and thus, more generalizable ML models can be trained from it. Again, the molecular composition of both random and block copolymers can be considered using the molar-weighted summation of each monomer's feature vector, and the monomer's sequence of block copolymers can be regarded by stacking monomers in blocks. Figure 6B shows that all four ML models can be generalized to this diverse dataset and that too with excellent predictive performance. Based on the high R^2 of the training and testing sets, it appears that these four ML models are comparative when dealing with random and block copolymers. To further investigate the model performances on random copolymers and block copolymers separately, we also calculate their respective train and test R^2 for comparison. It is found that Fusion and RNN model demonstrate slightly better performance on block copolymers, owing to their better abilities for processing sequence information (see Supplemental information Table S1 and Figure S11 for the separate model performance on random and block copolymers). To gauge their generalizability toward copolymers where copolymers' sequence distribution is changed from one type to another along the chain (gradient copolymers), we use them for a new test case of gradient copolymers, as discussed in the following sub-section.

Further validation – Gradient copolymers with glass transition temperature

With advancements in polymerization techniques such as reversible addition–fragmentation transfer polymerization (RAFT) (Matyjaszewski, 2003; Moad, 2015), atom-transfer radical polymerization (ATRP) (Matyjaszewski et al., 2000; Matyjaszewski, 2012), ring-opening metathesis polymerization (ROMP) (Dettmer et al., 2004), and nitroxide-mediated controlled radical polymerization (NM-CRP) (Gray et al., 2004), a variety of gradient copolymers have been synthesized successfully whose properties are observed to be between those of random and block copolymers (Lefebvre et al., 2005; Gray et al., 2002). As experimental measurements of gradient copolymer's properties are not as many as random

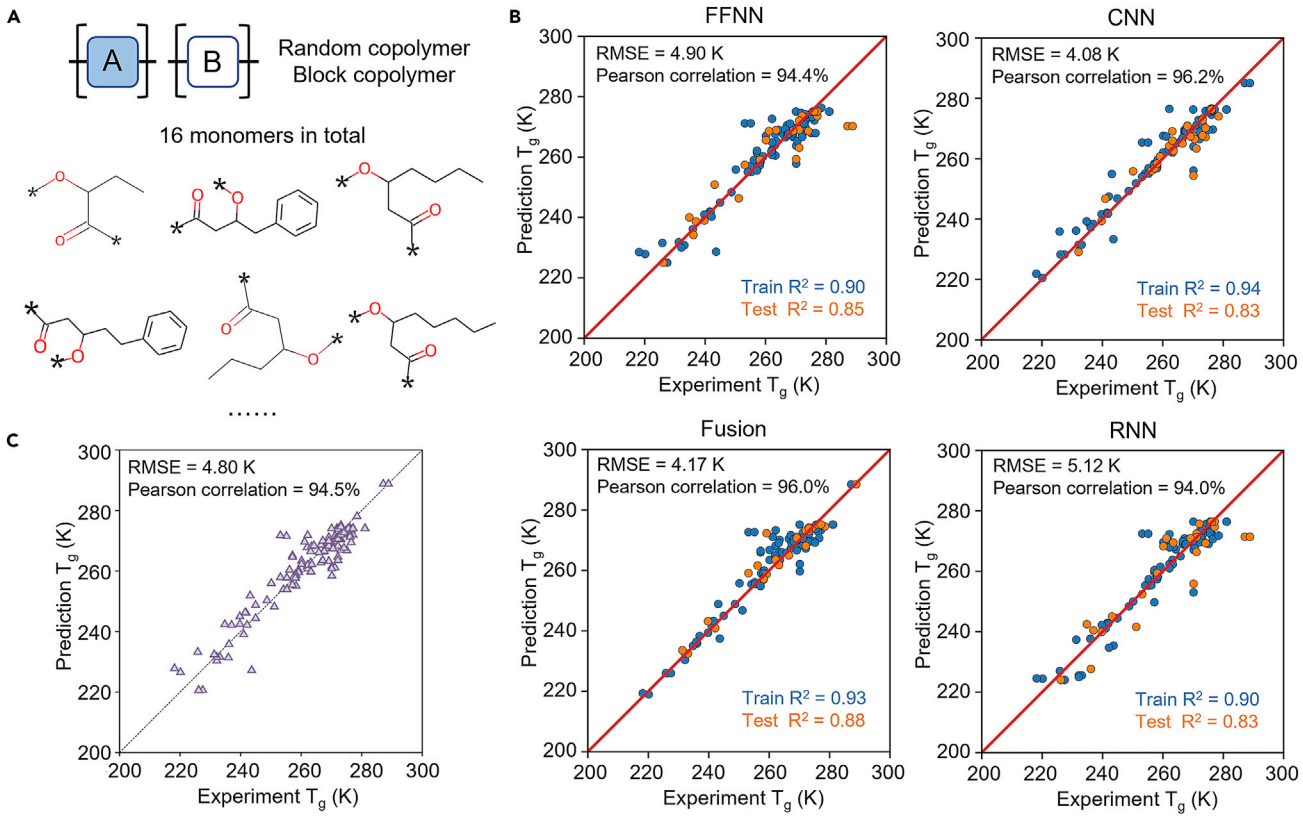


Figure 5. Performance of four ML models on copolymer Dataset 3

(A) Examples of monomers used to form polyhydroxyalkanoate.

(B) The parity plots of four ML predicted T_g versus the experimental values. RMSE and Pearson correlation are calculated for a direct comparison with the reference.

(C) The parity plot of ML predicted T_g versus the experimental values adapted from Reference (Pilania et al., 2019) with permission from the American Chemical Society (2019).

and block copolymers, copolymer ML studies haven't included gradient copolymers in the model training. Currently, the largest and most diverse dataset of copolymers is Dataset 4 with regard to the glass transition temperature T_g. We have obtained four ML models, which are expected to establish the structure-property relationship of T_g, especially for random and block copolymers. As the sequence distribution of gradient copolymers is an intermediate stage between random and block copolymers, we expect that ML models with a good generalization ability should be able to evaluate and predict gradient copolymer properties, although gradient copolymers are not used during the model training. We have found two experimental studies that report the T_g of copolymers when their sequence distributions are random, block, and gradient (Alshehri et al., 2022; Kim et al., 2006). In these studies, the T_g of gradient copolymer is found between the T_g values of its random and block counterparts. The following validation is to examine whether an ML model can predict the same pattern when gradient copolymers are encountered in the test dataset.

The first validation data come from the experimental study of Alshehri et al. (2022) They prepared copolymers using two monomer types, n-butyl acrylate (nBA) and isobornyl acrylate (IBA). Two homopolymer samples, two random copolymer samples, and five gradient copolymer samples are synthesized. Except for one gradient copolymer sample that is synthesized under a special condition, the other copolymer samples are illustrated in Figure 7A. The name of each sample is accompanied by the composition ratio of the two monomers nBA and IBA. When the composition ratio is 100:0 or 0:100, it means the copolymer is only made of one monomer type, namely homopolymers. With both the molecular composition and monomer's sequence of these samples, we use the aforementioned methods to obtain their feature vectors for ML models. The molar-weighted summation method is used to generate the copolymer feature

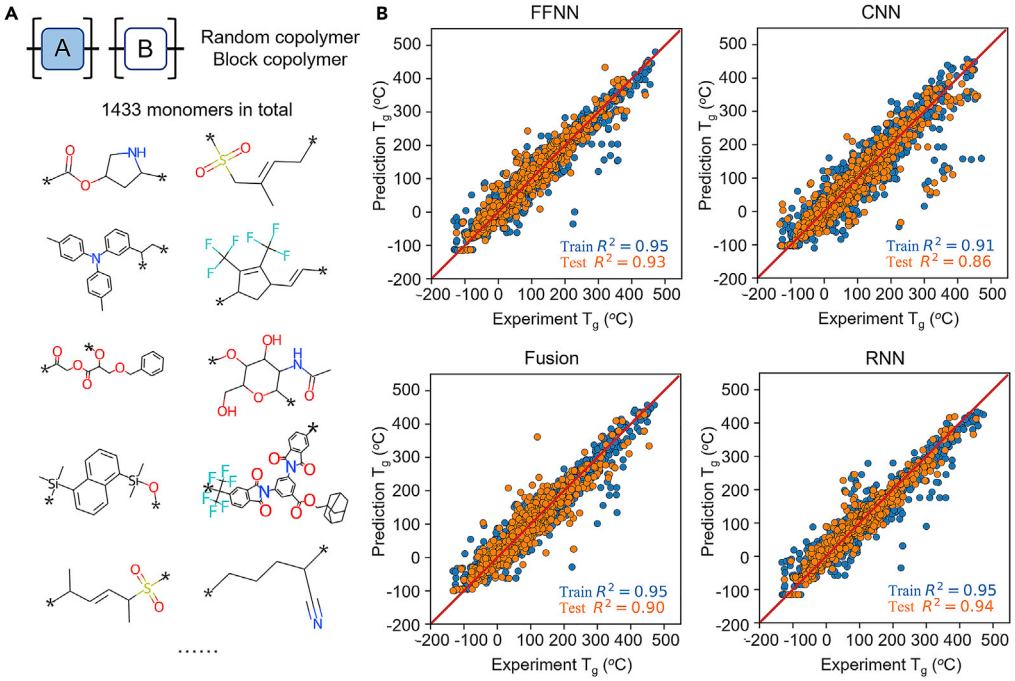


Figure 6. Performance of four ML models on copolymer Dataset 4

(A) Examples of monomers used to form copolymers in Dataset 4.

(B) The parity plots of four ML predicted T_g versus the experimental values. See also Figures S6–S8 and S11.

vector for the FFNN model, and the stacking of two monomers according to the sequence distributions in Figure 7A is used for the CNN, RNN, and Fusion models. The composition ratios of these eight samples are quite different, and experimental results show a trend of T_g along with the change in the composition, as given in Figures 7B and 7C (see Supplemental information Figure S12 for the parity plot comparison of these samples). Although these four ML models haven't seen gradient copolymers in the training dataset, the ML predictions still match well with the experimental trend. We attribute this to the fact that when the composition ratio differs much, the trend of T_g are mostly determined by the change in the composition ratio (as also noted before in Dataset 3). As these four ML models have been confirmed to learn the molecular composition well on Dataset one to four, it is not surprising that they perform well on the new gradient copolymers whose composition ratios are quite different. One exception occurs in Figure 7C for the Fusion model on gradient polymers. The Fusion model has the most complex architecture among the four ML models, which results in a less generalization ability on these new gradient copolymers. It is worth noticing there is a large uncertainty involved in some model's predictions, indicated by the large error bars. A qualitative analysis of the trend match is more reliable here to evaluate the applicability of these ML models. It requires special caution to use these models for quantitative predictions of gradient copolymers.

The second validation data come from the experimental study of Kim et al. (2006) They synthesized the copolymers of styrene (S) and 4-acetoxystyrene (AS), as well as the copolymers of styrene (S) and 4-hydroxystyrene (HS). Figure 8A illustrates six samples with their compositions and sequence distributions. It's worth noting that unlike the previous nBA/IBA copolymers, whose composition ratios are quite different, the samples' composition ratios are roughly the same in this study as shown in Figure 8A. When the composition ratio is kept constant, we expect the sequence distribution to govern-the properties of the samples. As discussed formerly, the monomer's sequence is represented by stacking two monomers in the same distributions in Figure 8A, and the molecular composition is considered using the molar-weighted summation of each monomer's feature vector. Experimental results show that given a composition ratio when the sequence distribution changes from random to gradient copolymer and then to block copolymers, the T_g of the copolymers follow a downward trend, as shown in Figures 8B and 8C. As the four ML models haven't seen the gradient copolymers during their training, the same composition ratio cannot assist ML models in differentiating these samples. Therefore, the ML models' predictions

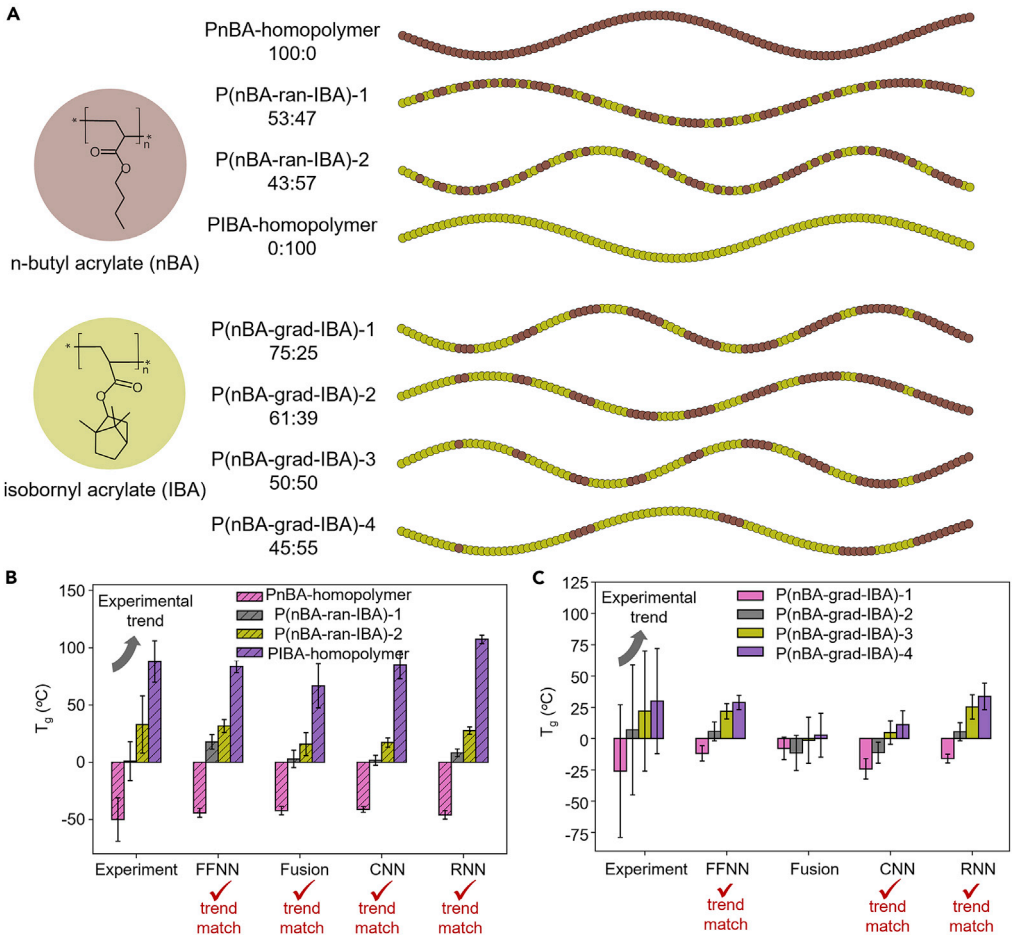


Figure 7. Performance of four ML models on copolymers of n-butyl acrylate (nBA) and isobornyl acrylate (IBA)

(A) The composition and sequence distribution of different nBA-IBA copolymer samples.

(B) The comparison of four ML predicted T_g versus the experimental values for homopolymers and random copolymers.

(C) The comparison of four ML predicted T_g versus the experimental values for gradient copolymers. Data are represented as mean ± SD The error bar of ML models is obtained by calculating the SD of predictions from five independent model training (ensembled average). The error bar of experiments is used to indicate the breadth of glass transition ΔT_g, which does not indicate the SD of experiments for the experimental uncertainty (see Supplemental information Figure S13 for the experimental measurement of ΔT_g). See also Figures S12 and S13.

dominantly rely on their ability to recognize different sequence distributions. Confirming this hypothesis on validation data, we find that CNN and RNN models predict the same downward trend as experimental results (Figures 8B and 8C). Although the FFNN and Fusion model worked well on previous copolymer cases, their limitation is revealed here in their inability to process different sequence distributions unlike that of CNN and RNN models. The CNN architecture uses a sliding window to process the monomer's sequence in one direction. In contrast, the RNN architecture uses bidirectional LSTM to process the sequence information in two directions better. Such strategies are essential for ML models to handle copolymers of different types, simultaneously considering their molecular compositions and monomer's sequences.

DISCUSSION

Copolymers have various types, including alternating, random, block, gradient copolymers, more generally, sequence-defined copolymers. When their monomers follow different sequence distributions, their physical properties are changed accordingly. It is a challenging task to evaluate the properties of copolymer considering both molecular composition and monomer's sequence simultaneously. Inspired by the development of ML models for homopolymers previously, this study examines the applicability of

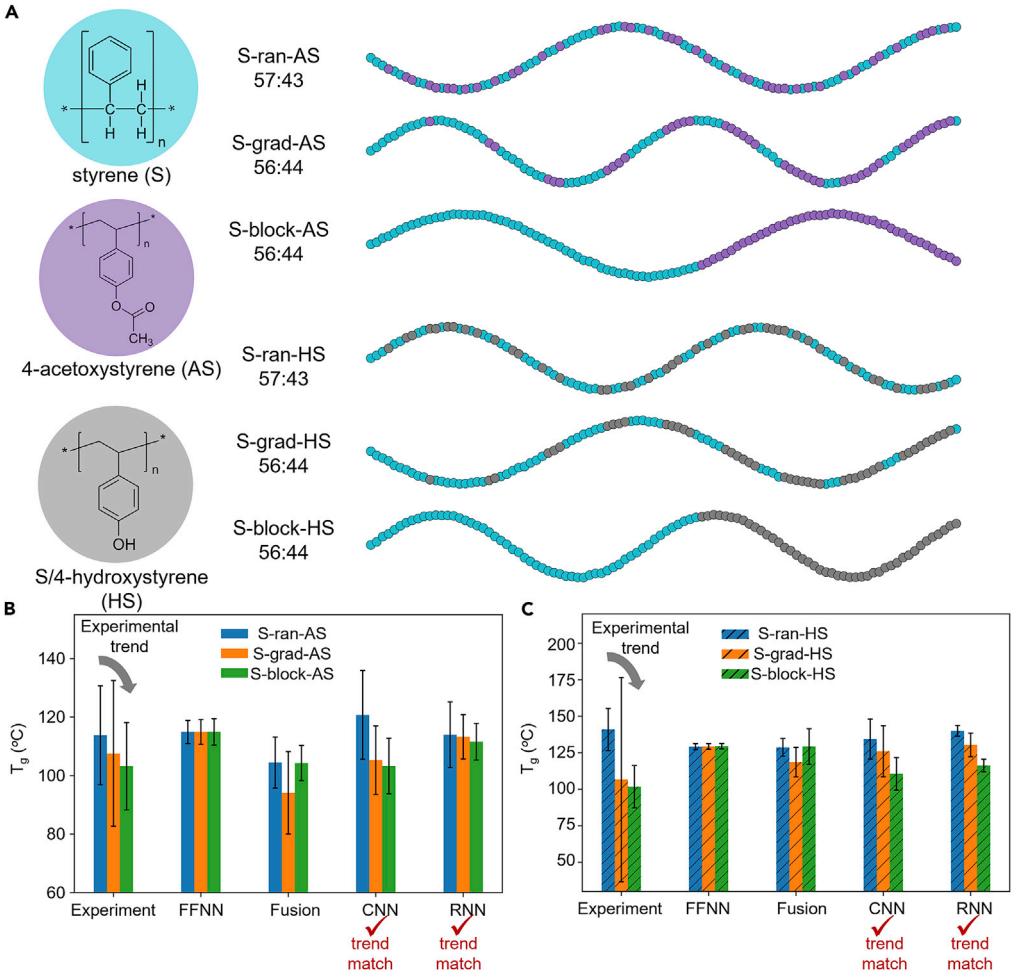


Figure 8. Performance of four ML models on the copolymers of styrene (S) and 4-acetoxystyrene (AS), as well as the copolymers of styrene (S) and 4-hydroxystyrene (HS)

(A) The composition and sequence distribution of different S/AS and S/HS copolymer samples.

(B) The comparison of four ML predicted T_g versus the experimental values for the S/AS random, gradient, and block copolymers.

(C) The comparison of 4 ML predicted T_g versus the experimental values for the S/AS random, gradient, and block copolymers. Data are represented as mean ± SD The error bar of ML models is obtained by calculating the SD of predictions from five independent model training (ensembled average).

The error bar of the experiment is used to indicate the breadth of glass transition ΔT_g, which does not indicate the SD of experiments for the experimental uncertainty (see Supplemental information Figure S13 for the experimental measurement of ΔT_g). See also Figure S12 and S13.

four ML models on copolymers of different types. Morgan fingerprints indicating substructure's frequency are used as the feature vector of monomers, and four different models are utilized, including FFNN, CNN, RNN, and Fusion models. To adapt the FFNN, CNN, and RNN models for the feature of copolymers, adjustments are made when building their respective architectures: (1) The FFNN model doesn't utilize the sequence distribution of monomers, but uses the molar-weighted summation of each monomer's feature vector to pass the composition ratio of copolymers into the model; (2) The CNN model stacks 100 feature vectors of monomers into a 2D matrix, and uses a sliding window to consider both molecular composition and monomer's sequence; (3) The RNN model stacks 100 feature vectors of monomers into a more extended 1D vector, and uses the bidirectional LSTM to learn both molecular composition and monomer's sequence; (4) The Fusion model uses a more complex architecture to decouples the molecular composition and monomer's sequence into two components, and fuse them into a combined single evaluation.

To test the applicability of these four ML models, their performances on four different datasets are examined including (1) conjugated copolymers with optoelectronic properties; (2) copolymer as ^{19}F MRI agents with signal intensity; (3) polyhydroxyalkanoate with glass transition temperature; and (4) copolymers in PoLyInfo with glass transition temperature. These datasets contain alternating, random, and block copolymers. As these four ML models are able to include molecular composition information, which is the primary factor affecting copolymer's properties, the performance of these ML models on different datasets is comparable. Gradient copolymers from experimental studies are used for further validation to investigate whether developed ML models are sensitive to the change in copolymer's sequence distributions. Among the four ML models, CNN and RNN models are observed to be more generalizable to gradient copolymers because their predictions match well with the experimental trends. It is demonstrated that it is essential for ML models to process the sequence information in copolymers, in addition to their molecular compositions, especially if there are notable changes in the monomer sequence along the chain. The RNN architecture that allows the sequence distribution to be processed both forward and backward is found to be the best-suited model for copolymers with good generalization ability. These ML models focus on the monomer-level fingerprints and monomers' sequence distribution of copolymers. The higher levels of analysis at the microscale or macroscale such as chain topology, crystallization, branch, and so forth are not considered. When the behaviors of copolymers are controlled by microscale or macroscale features, it requires the development of multi-level ML models, such as the recent Multi-Resolution Graph Variational Autoencoders (Gao et al., 2021b). At the monomer level, we expect that our ML models will be further adapted and refined to explore the vast parameter space of sequence-defined copolymers for their molecular engineering and design.

Limitations of the study

This study focuses on the most often used ML architectures that can be applicable to copolymers. Besides the investigated four ML architectures, there are other newly developed models that can process sequential data, like the FFNN with the attention that is considered better than RNN for sequence processing, or the Temporal Convolutional Networks whose architecture is modified to be comparable with RNN for sequence processing. This study doesn't cover all advanced ML architectures although they may also perform well for copolymers. Furthermore, as the performances of ML models are highly problem-dependent, the application of the proposed models on other properties of copolymers needs extra validation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Feature engineering of monomers
  - FFNN model for copolymers
  - CNN model for copolymers
  - RNN model for copolymers
  - Fusion model for copolymers
  - Model parameters and training

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104585.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Air Force Office of Scientific Research through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Dr. Ming-Jen Pan), Air Force Research Laboratory/UES Inc. (FA8650-20-S-5008, PICASSO program), and the National Science Foundation (CMMI-1934829 and CAREER-2046751). Y.L. would also like to thank the support from 3M's Non-Tenured Faculty Award. This research also benefited in part from the computational resources

and staff contributions provided by the Booth Engineering Center for Advanced Technology (BECAT) at the University of Connecticut. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Defense. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (Frontera project and the National Science Foundation award 1818253) for providing HPC resources that have contributed to the research results reported within this article.

AUTHOR CONTRIBUTIONS

Y.L. and J.B. conceived the idea. Y.L., J.B., and V.V. supervised the research. Y.L. and L.T. contributed to the design of the project and data analysis. L.T. collected and analyzed the data, and established ML models. L.T. wrote the first draft of the article, and all authors contributed to revising the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 28, 2022
Revised: May 26, 2022
Accepted: June 7, 2022
Published: July 15, 2022

REFERENCES

Alshehri, I.H., Pahovnik, D., Žagar, E., and Shipp, D.A. (2022). Stepwise gradient copolymers of n-butyl acrylate and isobornyl acrylate by emulsion RAFT copolymerizations. Macromolecules *55*, 391–400. https://doi.org/10.1021/acs.macromol.1c01897.

Audus, D.J., and De Pablo, J.J. (2017). Polymer informatics: opportunities and challenges. ACS Macro Lett. *6*, 1078–1082. https://doi.org/10.1021/acsmacrolett.7b00228.

Badi, N., and Lutz, J.-F. (2009). Sequence control in polymer synthesis. Chem. Soc. Rev. *38*, 3383. https://doi.org/10.1039/b806413j.

Barnett, J.W., Bilchak, C.R., Wang, Y., Benicewicz, B.C., Murdock, L.A., Bereau, T., and Kumar, S.K. (2020). Designing exceptional gas-separation polymer membranes using machine learning. Sci. Adv. *6*, eaaz4301. https://doi.org/10.1126/sciadv.aaz4301.

Barton, J.M. (1970). Relation of glass transition temperature to molecular structure of addition copolymers. J. Polym. Sci. Part C: Polymer Symposia *30*, 573–597. Wiley Online Library.

Bejagam, K.K., Iverson, C.N., Marrone, B.L., and Pilania, G. (2021). Composition and configuration dependence of glass-transition temperature in binary copolymers and blends of polyhydroxyalkanoate biopolymers. Macromolecules *54*, 5618–5628. https://doi.org/10.1021/acs.macromol.1c00135.

Binder, K. (1995). Monte Carlo and Molecular Dynamics Simulations in Polymer Science (Oxford University Press).

Chen, G., Tao, L., and Li, Y. (2021a). Predicting polymers' glass transition temperature by a chemical language processing model. Polymer *13*, 1898. https://doi.org/10.3390/polym13111898.

Chen, L., Kim, C., Batra, R., Lightstone, J.P., Wu, C., Li, Z., Deshmukh, A.A., Wang, Y., Tran, H.D., Vashishta, P., et al. (2020). Frequency-dependent dielectric constant prediction of polymers using machine learning. npj Comput. Mater. *6*, 61. https://doi.org/10.1038/s41524-020-0333-6.

Chen, L., Pilania, G., Batra, R., Huan, T.D., Kim, C., Kuenneth, C., and Ramprasad, R. (2021b). Polymer informatics: current status and critical next steps. Mater. Sci. Eng., R *144*, 100595. https://doi.org/10.1016/j.mser.2020.100595.

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 3642–3649.

Daimon, H., Okitsu, H., and Kumanotani, J. (1975). Glass transition behaviors of random and block copolymers and polymer blends of styrene and cyclododecyl acrylate. I. Glass transition temperatures. Polym. J. *7*, 460–466. https://doi.org/10.1295/polymj.7.460.

Deng, Z., Shi, Q., Tan, J., Hu, J., and Liu, S. (2021). Sequence-defined synthetic polymers for new-generation functional biomaterials. ACS Mater. Lett. *3*, 1339–1356. https://doi.org/10.1021/acsmaterialslett.1c00358.

Dettmer, C.M., Gray, M.K., Torkelson, J.M., and Nguyen, S.T. (2004). Synthesis and functionalization of ROMP-based gradient copolymers of 5-substituted norbornenes. Macromolecules *37*, 5504–5512. https://doi.org/10.1021/ma036002w.

Fred Dice, J. (1990). Peptide sequences that target cytosolic proteins for lysosomal proteolysis. Trends Biochem. Sci. *15*, 305–309. https://doi.org/10.1016/0968-0004(90)90019-8.

Dimarzio, E.A., and Gibbs, J.H. (1959). Glass temperature of copolymers. J. Polym. Sci. *40*, 121–131. https://doi.org/10.1002/pol.1959.1204013609.

Doan Tran, H., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J.P., Gurnani, R., Shetty, P., et al. (2020). Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. *128*, 171104. https://doi.org/10.1063/5.0023759.

Fijten, M.W.M., Kranenburg, J.M., Thijs, H.M.L., Paulus, R.M., Van Lankvelt, B.M., De Hullu, J., Springintveld, M., Thielen, D.J.G., Tweedie, C.A., Hoogenboom, R., et al. (2007). Synthesis and structure− property relationships of Random and block copolymers: a Direct Comparison for Copoly (2-oxazoline) s. Macromolecules *40*, 5879–5886. https://doi.org/10.1021/ma070720r.

Fox, T.G. (1956). Influence of diluent and of copolymer composition on the glass temperature of a poly-mer system. Bull. Am. Phys. Soc. *1*, 123.

Gao, H., Zhong, S., Zhang, W., Igou, T., Berger, E., Reid, E., Zhao, Y., Lambeth, D., Gan, L., Afolabi, M.A., et al. (2021a). Revolutionizing membrane design using machine learning-bayesian optimization. Environ. Sci. Technol. *56*, 2572–2581.

Gao, Z., Wang, X., Blumenfeld Gaines, B., Bi, J., and Song, M. (2021b). A deep molecular generative model based on multi-resolution graph variational Autoencoders. Preprint at chemrxiv. https://doi.org/10.26434/chemrxiv.14692551.v1.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T.D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M.A., Chae, H.S., Einzinger, M., Ha, D.-G., Wu, T., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat. Mater. *15*, 1120–1127. https://doi.org/10.1038/nmat4717.

Gordon, M., and Taylor, J.S. (1952). Ideal copolymers and the second-order transitions of synthetic rubbers. I. Non-crystalline copolymers. J. Appl. Chem. *2*, 493–500. https://doi.org/10.1002/jctb.5010020901.

Gray, M.K., Nguyen, S., Zhou, H., and Torkelson, J.M. (2002). Gradient copolymers produced via nitroxide-mediated controlled radical polymerization. Am. Chem. Soc., Polym. Prepr., Div. Polym. Chem. *43*, 112–113.

Gray, M.K., Zhou, H., Nguyen, S.T., and Torkelson, J.M. (2004). Synthesis and glass transition behavior of high molecular weight styrene/4-acetoxystyene and styrene/4-hydroxystyrene gradient copolymers made via nitroxide-mediated controlled radical polymerization. Macromolecules *37*, 5586–5595. https://doi.org/10.1021/ma0496652.

Guo, Y., Gao, X., and Luo, Y. (2015). Mechanical properties of gradient copolymers of styrene and n-butyl acrylate. J. Polym. Sci., Part B: Polym. Phys. *53*, 860–868. https://doi.org/10.1002/polb.23709.

Hale Charch, W., and Shivere, J.C. (1959). Part II: elastomeric condensation block copolymers. Textil. Res. J. *29*, 536–540. https://doi.org/10.1177/004051755902900702.

Hanaoka, K. (2020). Deep neural networks for multicomponent molecular systems. ACS Omega *5*, 21042–21053. https://doi.org/10.1021/acsomega.0c02599.

Jablonka, K.M., Jothiappan, G.M., Wang, S., Smit, B., and Yoo, B. (2021). Bias free multiobjective active learning for materials design and discovery. Nat. Commun. *12*, 2312. https://doi.org/10.1038/s41467-021-22437-0.

Jaeger, S., Fulle, S., and Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. J. Chem. Inf. Model. *58*, 27–35. https://doi.org/10.1021/acs.jcim.7b00616.

Johnston, N.W. (1976). Sequence distribution-glass transition effects. J. Macromol. Sci., Rev. Macromol. Chem. *14*, 215–250. https://doi.org/10.1080/15321797608065770.

Kenney, J.F. (1968). Properties of block versus random copolymers. Polym. Eng. Sci. *8*, 216–226. https://doi.org/10.1002/pen.760080307.

Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer genome: a data-powered polymer informatics platform for property predictions. J. Phys. Chem. C *122*, 17575–17585. https://doi.org/10.1021/acs.jpcc.8b02913.

Kim, J., Mok, M.M., Sandoval, R.W., Woo, D.J., and Torkelson, J.M. (2006). Uniquely broad glass transition temperatures of gradient copolymers relative to random and block copolymers containing repulsive comonomers. Macromolecules *39*, 6152–6160. https://doi.org/10.1021/ma061241f.

Kosuri, S., Borca, C.H., Mugnier, H., Tamasi, M., Patel, R.A., Perez, I., Kumar, S., Finkel, Z., Schloss, R., Cai, L., et al. (2022). Machine-Assisted discovery of chondroitinase ABC complexes toward sustained neural regeneration. Adv. Healthcare Mater. 2102101. https://doi.org/10.1002/adhm.202102101.

Kuenneth, C., Rajan, A.C., Tran, H., Chen, L., Kim, C., and Ramprasad, R. (2021a). Polymer informatics with multi-task learning. Patterns *2*, 100238. https://doi.org/10.1016/j.patter.2021.100238.

Kuenneth, C., Schertzer, W., and Ramprasad, R. (2021b). Copolymer informatics with multitask deep neural networks. Macromolecules *54*, 5957–5961. https://doi.org/10.1021/acs.macromol.1c00728.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA *97*, 10383–10388. https://doi.org/10.1073/pnas.97.19.10383.

Labanowski, J.K., and Andzelm, J.W. (2012). Density Functional Methods in Chemistry (Springer Science & Business Media).

Landrum, G. (2013). RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling (Academic Press).

Lefebvre, M.D., Dettmer, C.M., Mcswain, R.L., Xu, C., Davila, J.R., Composto, R.J., Nguyen, S.T., and Shull, K.R. (2005). Effect of sequence distribution on copolymer interfacial activity. Macromolecules *38*, 10494–10502. https://doi.org/10.1021/ma0509762.

Lehto, T., and Wagner, E. (2014). Sequence-defined polymers for the delivery of oligonucleotides. Nanomedicine *9*, 2843–2859. https://doi.org/10.2217/nnm.14.166.

Leibfarth, F.A., Johnson, J.A., and Jamison, T.F. (2015). Scalable synthesis of sequence-defined, unimolecular macromolecules by Flow-IEG. Proc. Natl. Acad. Sci. USA *112*, 10617–10622. https://doi.org/10.1073/pnas.1508599112.

Liu, T., Liu, L., Cui, F., Ding, F., Zhang, Q., and Li, Y. (2020). Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning. J. Mater. Chem. *8*, 21862–21871. https://doi.org/10.1039/d0ta07607d.

Lupas, A., Van Dyke, M., Stock, J., and LupAs, A. (1991). Predicting coiled coils from protein sequences. Science, 1162–1164. https://doi.org/10.1126/science.252.5009.1162.

Lutz, J.-F., Lehn, J.-M., Meijer, E.W., and Matyjaszewski, K. (2016). From precision polymers to complex materials and systems. Nat. Rev. Mater. *1*, 16024. https://doi.org/10.1038/natrevmats.2016.24.

Lutz, J.-F., Ouchi, M., Liu, D.R., and Sawamoto, M. (2013). Sequence-controlled polymers. Science *341*, 1238149. https://doi.org/10.1126/science.1238149.

Ma, R., Liu, Z., Zhang, Q., Liu, Z., and Luo, T. (2019). Evaluating polymer representations via quantifying structure–property relationships. J. Chem. Inf. Model. *59*, 3110–3119. https://doi.org/10.1021/acs.jcim.9b00358.

Matyjaszewski, K. (2003). Controlled/living radical polymerization: state of the art in 2002. In Advances in Controlled/Living Radical Polymerization, pp. 2–9.

Matyjaszewski, K. (2012). Atom transfer radical polymerization (ATRP): current status and future perspectives. Macromolecules *45*, 4015–4039. https://doi.org/10.1021/ma3001719.

Matyjaszewski, K., Ziegler, M.J., Arehart, S.V., Greszta, D., and Pakula, T. (2000). Gradient copolymers by atom transfer radical copolymerization. J. Phys. Org. Chem. *13*, 775–786. https://doi.org/10.1002/1099-1395(200012)13:12<775::aid-poc314>3.0.co;2-d.

Meenakshisundaram, V., Hung, J.-H., Patra, T.K., and Simmons, D.S. (2017). Designing sequence-specific copolymer compatibilizers using a molecular-dynamics-simulation-based genetic algorithm. Macromolecules *50*, 1155–1166. https://doi.org/10.1021/acs.macromol.6b01747.

Meier, M.A.R., and Barner-Kowollik, C. (2019). A new class of materials: sequence-defined macromolecules and their emerging applications. Adv. Mater. *31*, 1806027. https://doi.org/10.1002/adma.201806027.

Mewes, H.-W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. Nucleic Acids Res. *30*, 31–34. https://doi.org/10.1093/nar/30.1.31.

Miccio, L.A., and Schwartz, G.A. (2020a). From chemical structure to quantitative polymer properties prediction through convolutional neural networks. Polymer *193*, 122341. https://doi.org/10.1016/j.polymer.2020.122341.

Miccio, L.A., and Schwartz, G.A. (2020b). Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks. Polymer *203*, 122786. https://doi.org/10.1016/j.polymer.2020.122786.

Moad, G. (2015). RAFT (Reversible addition–fragmentation chain transfer) crosslinking (co) polymerization of multi-olefinic monomers to form polymer networks. Polym. Int. *64*, 15–24. https://doi.org/10.1002/pi.4767.

Mohapatra, S., An, J., and Gómez-Bombarelli, R. (2022). Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. Mach. Learn.: Sci. Technol. *3*, e015028. https://doi.org/10.1088/2632-2153/ac545e.

Mok, M.M., Masser, K.A., Runt, J., and Torkelson, J.M. (2010). Dielectric relaxation spectroscopy of gradient copolymers and block copolymers: comparison of breadths in relaxation time for systems with increasing interphase. Macromolecules *43*, 5740–5748. https://doi.org/10.1021/ma100743s.

Nanjan, P., and Porel, M. (2019). Sequence-defined non-natural polymers: synthesis and applications. Polym. Chem. *10*, 5406–5424. https://doi.org/10.1039/c9py00886a.

Nazarova, A.L., Yang, L., Liu, K., Mishra, A., Kalia, R.K., Nomura, K.-I., Nakano, A., Vashishta, P., and Rajak, P. (2021). Dielectric polymer property prediction using recurrent neural networks with optimizations. J. Chem. Inf. Model. *61*, 2175–2186. https://doi.org/10.1021/acs.jcim.0c01366.

Nguyen, D.T., Tao, L., and Li, Y. (2021). Integration of machine learning and coarse-grained molecular simulations for polymer materials: physical understandings and molecular design. Front. Chem. *9*, 820417. https://doi.org/10.3389/fchem.2021.820417.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., and Yamazaki, M.P.L.I. (2011). Polymer database for polymeric materials design. In 2011 International Conference on Emerging Intelligent Data and Web Technologies (IEEE), pp. 22–29.

Palermo, E.F., and Mcneil, A.J. (2012). Impact of copolymer sequence on solid-state properties for random, gradient and block copolymers containing thiophene and selenophene. Macromolecules *45*, 5948–5955. https://doi.org/10.1021/ma301135n.

Palomba, D., Vazquez, G.E., and Díaz, M.F. (2012). Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures. J. Mol. Graphics Modell. *38*, 137–147. https://doi.org/10.1016/j.jmgm.2012.04.006.

Patel, R.A., Borca, C.H., and Webb, M.A. (2022). Featurization strategies for polymer sequence or composition design by machine learning. Mol. Syst. Des. Eng. *7*, 661–676.

Patterson, A.L., Danielsen, S.P.O., Yu, B., Davidson, E.C., Fredrickson, G.H., and Segalman, R.A. (2019). Sequence effects on block copolymer self-assembly through tuning chain conformation and segregation strength utilizing sequence-defined polypeptoids. Macromolecules *52*, 1277–1286. https://doi.org/10.1021/acs.macromol.8b02298.

Perry, S.L., and Sing, C.E. (2020). 100th anniversary of macromolecular science viewpoint: opportunities in the physics of sequence-defined polymers. ACS Macro Lett. *9*, 216–225. https://doi.org/10.1021/acsmacrolett.0c00002.

Pilania, G., Iverson, C.N., Lookman, T., and Marrone, B.L. (2019). Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. J. Chem. Inf. Model. *59*, 5013–5025. https://doi.org/10.1021/acs.jcim.9b00807.

Porel, M., and Alabi, C.A. (2014). Sequence-defined polymers via orthogonal allyl acrylamide building blocks. J. Am. Chem. Soc. *136*, 13162–13165. https://doi.org/10.1021/ja507262t.

Ramprasad, M., and Kim, C. (2019). Assessing and improving machine learning model predictions of polymer glass transition temperatures. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.02398.

Reis, M., Gusev, F., Taylor, N.G., Chung, S.H., Verber, M.D., Lee, Y.Z., Isayev, O., and Leibfarth, F.A. (2021). Machine-learning-guided discovery of 19F MRI agents enabled by automated copolymer synthesis. J. Am. Chem. Soc. *143*, 17677–17689. https://doi.org/10.1021/jacs.1c08181.

Shi, J., Quevillon, M.J., Valença, P.H.A., and Whitmer, J.K. (2021). Predicting adhesive free energies of polymer–surface interactions with machine learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2110.03041.

Sing, C.E. (2020). Micro-to macro-phase separation transition in sequence-defined coacervates. J. Chem. Phys. *152*, e024902. https://doi.org/10.1063/1.5140756.

Statt, A., Kleeblatt, D.C., and Reinhart, W.F. (2021). Unsupervised learning of sequence-specific aggregation behavior for a model copolymer. Soft Matter *17*, 7697–7707. https://doi.org/10.1039/d1sm01012c.

Sun, W., Zheng, Y., Yang, K., Zhang, Q., Shah, A.A., Wu, Z., Sun, Y., Feng, L., Chen, D., Xiao, Z., et al. (2019). Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Sci. Adv. *5*, eaay4275. https://doi.org/10.1126/sciadv.aay4275.

Suzuki, H., and Miyamoto, T. (1989). A comparative study on barton's and johnston's equations for copolymer glass transition temperature (commemoration issue dedicated to professor hiroshi ibagaki, professor michio kurata, professor ryozo kitamura, on the occasion of their retirments). Bull. Inst. Chem. Res. Kyoto Univ. *66*, 297–311.

Svozil, D., Kvasnicka, V., and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. Chemometr. Intell. Lab. Syst. *39*, 43–62. https://doi.org/10.1016/s0169-7439(97)00061-0.

Tamasi, M., Patel, R., Borca, C., Kosuri, S., Mugnier, H., Upadhya, R., Murthy, N.S., Webb, M., and Gormley, A. (2022). Machine learning on a robotic platform for the design of polymer-protein hybrids. Preprint at chemrxiv. https://doi.org/10.26434/chemrxiv-2022-x2qdz.

Tao, L., Chen, G., and Li, Y. (2021a). Machine learning discovery of high-temperature polymers. Patterns *2*, 100225. https://doi.org/10.1016/j.patter.2021.100225.

Tao, L., Varshney, V., and Li, Y. (2021b). Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. J. Chem. Inf. Model. *61*, 5395–5413. https://doi.org/10.1021/acs.jcim.1c01031.

Todeschini, R., and Consonni, V. (2008). Handbook of Molecular Descriptors (John Wiley & Sons).

Tu, K.H., Huang, H., Lee, S., Lee, W., Sun, Z., Alexander-Katz, A., and Ross, C.A. (2020). Machine learning predictions of block copolymer self-assembly. Adv. Mater. *32*, 2005713. https://doi.org/10.1002/adma.202005713.

Webb, M.A., Jackson, N.E., Gil, P.S., and De Pablo, J.J. (2020). Targeted sequence design within the coarse-grained polymer genome. Sci. Adv. *6*, eabc6216. https://doi.org/10.1126/sciadv.abc6216.

Werner, M., Guo, Y., and Baulin, V.A. (2020). Neural network learns physical rules for copolymer translocation through amphiphilic barriers. npj Comput. Mater. *6*, 72. https://doi.org/10.1038/s41524-020-0318-5.

Wheatle, B.K., Fuentes, E.F., Lynd, N.A., and Ganesan, V. (2020). Design of polymer blend electrolytes through a machine learning approach. Macromolecules *53*, 9449–9459. https://doi.org/10.1021/acs.macromol.0c01547.

Wilbraham, L., Sprick, R.S., Jelfs, K.E., and Zwijnenburg, M.A. (2019). Mapping binary copolymer property space with neural networks. Chem. Sci. *10*, 4973–4984. https://doi.org/10.1039/c8sc05710a.

Wu, S., Kondo, Y., Kakimoto, M.-A., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., et al. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput. Mater. *5*, 66. https://doi.org/10.1038/s41524-019-0203-2.

Yuan, Q., Longo, M., Thornton, A.W., Mckeown, N.B., Comesaña-Gándara, B., Jansen, J.C., and Jelfs, K.E. (2021). Imputation of missing gas permeability data for polymer membranes using machine learning. J. Membr. Sci. *627*, 119207. https://doi.org/10.1016/j.memsci.2021.119207.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1409.2329.

Zhou, T., Wu, Z., Chilukoti, H.K., and Müller-Plathe, F. (2021). Sequence-engineering polyethylene–polypropylene copolymers with high thermal conductivity using a molecular-dynamics-based genetic algorithm. J. Chem. Theor. Comput. *17*, 3772–3782. https://doi.org/10.1021/acs.jctc.1c00134.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python version 3.7	Python Software Foundation	https://www.python.org
Tensorflow 2.3.0	Open-Source Software	https://www.tensorflow.org/
RDKit	Open-Source Software	https://www.rdkit.org/
Model codes	Github	https://github.com/figotj/Copolymer

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ying Li (ying.3.li@uconn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data from publications and open website.
- All original code has been deposited at https://github.com/figotj/Copolymer and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Feature engineering of monomers

For each ML model, a feature vector must be defined for each polymer molecule. As homopolymers are composed of one monomer type (or repeat unit), this monomer's feature vector contains all the composition information of the homopolymer. However, for a copolymer that is made of two monomer types, "A" and "B", the two monomers' feature vector F_A and F_B are both required to contain all the composition information of the copolymer. Feature vectors can be obtained from physiochemical descriptors(Todeschini and Consonni, 2008), fingerprints(Tao et al., 2021b), molecular graphs(Mohapatra et al., 2022), unsupervised molecular embeddings(Jaeger et al., 2018), or supervised embeddings(Gómez-Bombarelli et al., 2016). Supervised embeddings, used for example in Graph Convolutional Neural Networks, are learned specifically for the given task and often have the highest performance, but they require much larger datasets than we will use in order to avoid overtraining. Based on the successful application of our improved Morgan fingerprint for homopolymers, (Tao et al., 2021a, 2021b), this study only utilizes the fingerprint-based feature vectors – which is sufficient to take the composition information into account. Compared to the standard Morgan fingerprint that uses 1/0 (on/off or one-hot encoding) bit in the feature vector to indicate the occurrence of a specific substructure, our improved Morgan fingerprint uses integers in the feature vector to also indicate the number of occurrences of each substructure, which is more informative than the standard Morgan fingerprint(Tao et al., 2021a). The substructure is obtained using the Daylight-like fingerprinting algorithm as implemented in RDKit package(Landrum, 2013) with radius 3. Labeling the number of occurrences for substructures doesn't encode the microscale level feature of polymers such as average chain length or molecular weight. As the reported T_g in the experimental dataset is considered to be the saturated value of the glass transition temperature of a certain polymer, the effect of molecular weight is not explicitly represented in the fingerprints.

FFNN model for copolymers

FFNN is composed of neurons connected layer by layer(Svozil et al., 1997). For homopolymer studies, it is one of the most widely used ML models that have established the structure-property relationship satisfactorily. (Palomba et al., 2012, Miccio and Schwartz, 2020b; Tao et al., 2021a; Ma et al., 2019) It accepts the feature vector as an input and use it to predict the target property of a polymer, which makes it suitable for copolymers as well. The feature vector of a copolymer can be calculated as the molar-weighted summation of each monomer's feature vector: $F_{AB} = F_A m_A + F_B m_B$ as shown in Figure 2A, where F is the feature vector and m is the molar ratio. The subscripts A and B represent the monomers A and B, respectively. The molar ratio used in this study is assumed to be the final ratio of different monomer composition in the chain of copolymers. Reactivity ratio is not discussed in this study as the reaction stage of copolymers will complicate the structure-property problem significantly. It should be noted that while the model considers copolymer's composition as detailed above, the information of copolymers' sequence distribution is missing. The FFNN model is applied on copolymers by Kuenneth et al. (2021b), when assuming all copolymers to be random copolymers. Since random copolymers are a combination of two components without a specific monomer sequence, the FFNN model that uses the weighted summed-up feature vector F_{AB} should be well suited for them.

CNN model for copolymers

CNN model contains convolutional layers that are connected by a set of filters (Figure 2B)(Ciregan et al., 2012). It is also very effective for homopolymer ML predictions (Tao et al., 2021b; Miccio and Schwartz, 2020a). For copolymers, the application of CNN is feasible if the feature vector of copolymers is appropriately constructed. Patel et al. (2022) and Webb et al. (2020) have applied CNN to copolymers. To consider the sequence effect of copolymers, we align two monomers' feature vectors F_A and F_B into a 2D matrix, so that the alignment explicitly represents the sequence distribution of copolymers. We stack 100 monomers in total to form the 2D matrix, and determine the number of each monomer in the same proportion as their composition in the copolymer, e.g., stacking 65 F_A and 35 F_B if the molar ration of the copolymer A:B is 65:35. Stacking them in different sequence will correspond to different copolymer types. In this way, the CNN model is suitable for the alternating, random, gradient, and block copolymers. When CNN's filter window slides in two directions (along the length of feature vectors and along the stacking of feature vectors), both molecular composition and monomer's sequence information are passed to the CNN layer. Because the relative positioning of the bits along the length of feature vectors has no sequence meaning, restricting the filter width to be equal to the length of the feature vectors also passes the molecular composition information to the CNN model. Compared to such a setup like 1D CNN, it is more flexible to allow the filter size to be optimized as in the pristine CNN model for image recognition, in which filter windows sliding in two directions.

RNN model for copolymers

RNN contains neurons that accept sequential data like words in a sentence (Zaremba et al., 2014). It is designed to predict the next tokens in the sequence given past tokens for natural language processing. RNN has also been applied successfully for homopolymer ML predictions (Nazarova et al., 2021; Tao et al., 2021b; Chen et al., 2021a). Its intrinsic ability to process sequential data makes it an ideal option for copolymer problems in particular, considering the sequence of different monomers. Our RNN model uses bidirectional long short-term memory (LSTM) architecture to accept copolymers' feature vectors. As shown in Figure 2C, we align different monomers' feature vectors into a sequence of feature vector and pass it to the model, such as a connection of 65 F_A and 35 F_B if the molar ratio of the copolymer A:B is 65:35. Each feature vector constitutes a token for the LSTM. As the token sequence is processed step by step, the RNN model successfully learns both the molecular composition from each monomer and the monomer's sequence from the connection of different monomers. LSTM on copolymers is explored by Patel et al. (2022) and Webb et al. (2020). It is noted that CNN and RNN models represent the monomer's sequence of copolymers by stacking monomers' feature vectors into a specific order. The difference is that the CNN samples part of a feature vector together with the same part of its neighbors, whereas the RNN samples each feature vector as an independent token. The advantage of RNN over CNN is that the bidirectional LSTM architecture allows the sequence information to be processed both forward and backward and that the meaning of the token is processed separately from the sequence information. The bidirectionality is important as one can imagine that the featurization of the polymer chain can start from either end. On the other hand, the filter window in CNN only slides one way along the direction of stacking.

Fusion model for copolymers

Lastly, the fusion model in Figure 2D is a combination of the above FFNN and RNN models. Its FFNN component is used to extract molecular composition using the weighted sum of feature vectors as discussed above. Its RNN component is used to represent the monomer's sequence in copolymers. Since the feature vectors of monomers have been utilized in the FFNN component, the RNN component can only use vectors of 1/0 bit to represent the sequence distributions. We use "1" for monomer "A" and "0" for monomer "B"; and instead of stacking the 100 feature vectors F_A or F_B, we stack 100 bits to represent the sequence distribution, such as a connection of 65 "1" and 35 "0" if the molar ratio of the copolymer A:B is 65:35. The fusion model decouples the molecular composition and monomer's sequence into respective FFNN and RNN components. The architecture of the fusion model is more complex than the other models, but the RNN component in fusion model has an easier input of a 1/0 bit vector to process rather than full feature vectors F_A and F_B.

Model parameters and training

Data split is carried out with the train_test_split function of scikit-learn 1.0.2. Each model is trained on an 80% training set and tested on a 20% testing set with tensorflow 2.3.0. To optimize the architecture of ML models, the Random Search Tuner in Keras 2.4.3 is used to explore the hyperparameter space (including the number of layers, the number of neurons, the kernel size, etc.). 100 combinations of hyperparameters are explored to find the best one. Each combination is executed once to measure its performance. The objective of the search is to find the hyperparameter that has the minimum mean_squared_error on the testing set. The optimized hyperparameters for each ML model is listed in the model parameters table (see Supplemental information Figures S2–S8 for the scheme of each model, and a test of the effect of utilizing data augmentation, padding layer, and 1D filter on the performance of CNN).

Machine learning model parameters

Models	Parameters
FFNN	2 hidden layers; 24 neurons for the first layer with 'ReLU' activation function; 64 neurons for the second layer with 'ReLU' activation function; batch_size = 128; epochs = 100
CNN	3 Conv2D layers; • filters = 8, kernel_size = (10, 10), strides = (1, 1), 'ReLU' activation function for the first layer; • filters = 8, kernel_size = (4, 4), strides = (1, 1), 'ReLU' activation function for the second layer; • filters = 8, kernel_size = (3, 3), strides = (1, 1), 'ReLU' activation function for the third layer; Followed by 1 MaxPooling2D layer, pool_size = (2, 2); 1 dropout layer with rate = 0.3; batch_size = 4; epochs = 200
RNN	2 bidirectional LSTM layers; 20 neurons for each layer; 1 time-distributed layer; 1 reshape layer; batch_size = 4; epochs = 120
Fusion	FFNN component has 2 hidden layers; 8 neurons for each layer with 'ReLU' activation function. RNN component has 2 bidirectional LSTM layers; 20 neurons for each layer; 1 timedistributed layer. A concatenate layer combines two components; then 1 hidden layer of 8 neurons with 'ReLU' activation function; batch_size = 32; epochs = 300

With the optimized hyperparameter, the performance of each model on different datasets can be obtained, represented by train R^2 and test R^2. After the model architecture is finalized for the largest Dataset 4, the stacking ensemble method is used. A model architecture is trained on Dataset 4 five times to generate five separate learners. Averaging the predictions of the 5 learners generates an ensemble prediction for a new copolymer. The standard deviation of the 5 learners' predictions indicates the error range of the prediction.