A graph representation of molecular ensembles for polymer property prediction

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2022, **13**, 10486

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 20th May 2022
Accepted 15th August 2022

DOI: 10.1039/d2sc02839e

rsc.li/chemical-science

A graph representation of molecular ensembles for polymer property prediction†

Matteo Aldeghi [a] and Connor W. Coley *[ab]

Synthetic polymers are versatile and widely used materials. Similar to small organic molecules, a large chemical space of such materials is hypothetically accessible. Computational property prediction and virtual screening can accelerate polymer design by prioritizing candidates expected to have favorable properties. However, in contrast to organic molecules, polymers are often not well-defined single structures but an ensemble of similar molecules, which poses unique challenges to traditional chemical representations and machine learning approaches. Here, we introduce a graph representation of molecular ensembles and an associated graph neural network architecture that is tailored to polymer property prediction. We demonstrate that this approach captures critical features of polymeric materials, like chain architecture, monomer stoichiometry, and degree of polymerization, and achieves superior accuracy to off-the-shelf cheminformatics methodologies. While doing so, we built a dataset of simulated electron affinity and ionization potential values for >40k polymers with varying monomer composition, stoichiometry, and chain architecture, which may be used in the development of other tailored machine learning approaches. The dataset and machine learning models presented in this work pave the path toward new classes of algorithms for polymer informatics and, more broadly, introduce a framework for the modeling of molecular ensembles.

Introduction

Synthetic polymers are key components of numerous commodities and play an essential role in our daily lives, with applications ranging from clothing, to electronics and construction, and are used in industries as diverse as automotive, energy, and healthcare.[1–10] This versatility is due to the wide range of properties achievable by tuning a polymer's chemical composition and architecture. The identification of novel copolymers for the delivery of therapeutics cargos,[11–20] or for energy harvesting and storage,[6,21–25] are examples of active areas of research that rely on the availability of a broad range of polymer chemistries.

Machine learning (ML) is now playing a significant role in supporting the discovery and synthesis of new functional organic molecules with specialized applications,[20,26,27] thanks to its ability to capture subtle chemical patterns when enough data is available. The field of polymer informatics has also attracted increasing attention, with a number of studies demonstrating the use of ML for the prediction of thermal,[28–35] thermodynamic,[28,36–38] electronic,[39–44] optical,[41,45,46] and mechanical[41,47] properties of polymers and copolymers. However, while many specialized machine learning approaches have been developed for molecules and sequence-defined polymers like proteins and peptides, polymers characterized by molecular ensembles still rely on off-the-shelf cheminformatics approaches designed for single molecules. This work focuses specifically on the latter class of materials, which cover a considerable fraction of synthetic and natural polymers.

A major challenge in the development of bespoke ML models for polymer property prediction is the lack of a general polymer representation.[48–52] In fact, almost all ML models currently used for polymer property predictions do not capture the ensemble nature of the polymeric material, even when predicting properties of the ensemble rather than sequence-defined oligomers. The vast majority of past studies have relied on molecular representations of repeating units alone, even though such approaches cannot distinguish between alternating, random, block, or graft copolymers. Recent work has tried to obviate this issue by creating cyclic oligomers from which structural fingerprints can be derived.[53] However, this approach would still struggle to distinguish different chain architectures or capture the ensemble of possible monomer sequences.

The challenge of identifying a general polymer representation stems from the fact that, contrary to small organic molecules, many polymers are stochastic objects whose properties

[a]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: ccoley@mit.edu*

[b]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

† Electronic supplementary information (ESI) available: Extended methods, supporting figures and tables. Dataset of computed electron affinity and ionization potential values for 42 966 copolymers (CSV). See https://doi.org/10.1039/d2sc02839e

emerge from the ensemble of molecules that they comprise. A representation that captures this ensemble nature is thus needed to develop tailored and broadly applicable ML models for polymer property prediction. Recently, text-based representations that try to capture this unique aspect of polymer chemistry have been developed. BigSMILES is a text-based representation that builds upon the simplified molecular-input line-entry system (SMILES) representation and is designed specifically to describe the stochastic nature of polymer molecules.[54] Yet, language models based on text-based representations are data inefficient, such that they generally require extensive pretraining, data augmentation, or extremely large dataset sizes to be successful in cheminformatics, making formats like BigSMILES better suited for information storage and retrieval than as a direct input to learning algorithms.[55,56] Representations that more directly capture chemical structure, like fingerprints and graphs, are thus preferred for learning tasks as they typically outperform language models in property prediction tasks when provided with the same amount of data.

In this work, we report the development and validation of a graph-based representation of polymer structure and a weighted directed message passing neural network (wD-MPNN) architecture that learns specialized representations of molecular ensembles for polymer property prediction. To achieve this, we rely on a parametric description of the underlying distribution of molecules that captures its expectation (*i.e.*, the average graph structure of the repeating unit). We test our model on a new dataset of simulated electronic properties of alternating, random, and block copolymers, and achieve superior performance over graph-based representations that only capture monomeric units as well as robust fingerprint-based ML models. We furthermore evaluate the wD-MPNN on an experimental dataset[53] in which the supervised task involves predicting the possible phases of diblock copolymers. In both tasks, we demonstrate that the explicit inclusion of information about aspects of the molecular ensemble like monomer stoichiometries, chain architectures, and average sizes into the network architecture results in improved predictive performance.

Methods

In this section, we first review existing cheminformatics approaches used for polymer structure–property regression, which constitute the baseline representations and models we will benchmark our approach against. We then introduce our proposed graph-based representation of polymers, and a graph neural network (GNN) architecture that uses this representation as input. Finally, we describe the new dataset of computed copolymer properties we used to evaluate both traditional and proposed ML approaches.

Prior work on polymer representations as model baselines

Among the representations often used in polymer informatics are molecular fingerprints, which encode the presence or absence of chemical substructures in a binary vector. This representation is directly applicable to a polymer's repeating unit, although it cannot distinguish between isomeric sequences or different monomer stoichiometries. Stoichiometry can be considered by taking the sum of monomer fingerprints, weighted by the respective ratios.[32,57] Alternatively, count fingerprints, which use vectors of integer values and capture the frequency of different chemical patterns, can be applied to oligomeric molecules constructed in a way to reflect the monomers' stoichiometry. By constructing a short polymer chain, the resulting count fingerprints also capture aspects of the polymer's chain architecture. Note this is only partially true when using binary fingerprints. For instance, a random AB copolymer might have specific patterns that identify A–A, A–B, and B–B connections. A block copolymer will have the same patterns, but A–A and B–B will be more frequent than A–B ones. This frequency difference can be captured by count fingerprints, not by binary ones.

A natural representation for small organic molecules, including the repeating units of synthetic polymers, are molecular graphs in which atoms are represented by the graph vertices and bonds by its edges. GNNs[58] take such representation as input to predict molecular properties,[59–65] and have been applied to polymer property prediction by considering the structure of individual monomers.[30,42] However, standard GNN architectures cannot handle the inherent stochasticity of polymer structures, as they generally model a specific molecule rather than ensembles. While modeling individual monomeric units may be sufficient for homopolymers, in particular linear ones obtained by chain-growth, predicting properties of copolymers require the ability to distinguish between the constitutional isomers resulting in different chain architectures and sequence ensembles.

Mohapatra *et al.*[66] have presented a coarse-grained graph representation for macromolecules, which can capture complex macromolecular topologies. Patel *et al.*[57] have also explored a number of polymer representations, including a similar coarse-grained sequence graph representation. These graph representations can distinguish significantly different macromolecular topologies, but coarse-graining masks information on how monomeric units are bonded to each other. Atomic-level modeling of polymer structure is needed to capture the structure of the connection, which differentiates between structural (*e.g.*, *cis versus trans* bonds, *ortho versus meta* substitutions) and sequence (*e.g.*, head-to-tail *versus* head-to-head or tail-to-tail) isomers. Structural isomers can have vastly different properties. For instance, *trans*-1,4-polyisoprene (gutta-percha) has a regular structure that allows crystallization and results in a rigid material, while *cis*-1,4-polyisoprene (natural rubber) is amorphous and elastic. Sequence isomerism can be important instead for polymers synthesized *via* a step-growth or cationic mechanism, in which the fraction of head-to-tail arrangement can vary based on reactivity and lead to significant differences in polymer properties.

In this work, we adopted both fingerprint- and graph-based representations as baselines approaches. These include Chemprop,[67] an established GNN, and random forest (RF) and fully-connected neural network (NN) models trained on

fingerprint representations. More specifically, Chemprop uses a directed message passing neural network (D-MPNN), a special type of GNN architecture described in more detail later in the Methods. The input for this model was a disconnected graph of the separate monomeric units. The RF and NN models used Extended-Connectivity Fingerprints (ECFP)[68] of length 2048 and radius 2, computed with *RDKit*,[69] as input representation. We tested both binary and count fingerprints, constructed from the monomeric units alone, as well as from an ensemble of oligomeric sequences sampled uniformly at random. In the latter case, we sampled up to 32 octameric sequences while satisfying the stoichiometry and chain architecture of the polymer exactly (*e.g.*, using 6 A monomers and 2 B monomers for a stoichiometric ratio of 3 : 1), computed fingerprints for all resulting oligomers, and averaged them to obtain the input representation. We are not aware of prior work using this sequence sampling approach, but we found it to be the most competitive fingerprint-based baseline. Full details of the baseline approaches tested are in the ESI Extended methods.†

We note that, recently, Patel *et al.*[57] have also explored augmenting fingerprints with sequence-level or topological polymer descriptors. These additional descriptors capture characteristics of the chain architecture, like blockiness, or the distribution of charged or hydrophobic components. This was found to be an effective strategy to incorporate chain-level information into the fingerprint representation and improve results. However, hand-crafting high-level quantitative descriptors that discriminate between chain architectures may not be necessary when these differences can be captured by the underlying graph structure of the copolymer. For example, the count fingerprints with sequence sampling representation described above already carries statistical information on blockiness, as it is encoded by the frequency of A–A/B–B and A–B connections. While user-defined descriptors can expose higher-level properties more directly to the model, lower-level representations of chemical structure provide more flexibility for a ML model to learn directly from raw data.

Graph-based representation of molecular ensembles

Our goal was to expand the architecture of current ML models to capture (i) the recurrent nature of polymers' repeating units, (ii) the different topologies and isomerisms of polymer chains, and (iii) their varying monomer composition and stoichiometry. We thus decided to expand molecular graph representations by incorporating "stochastic" edges to describe the average structure of the repeating unit. In effect, these stochastic edges are bonds weighted by their probability of occurring in the polymer chain (Fig. 1).

In our polymer graph representation, each edge is associated with a weight, $w \in (0, 1]$, according to the probability (or frequency) of the bond being present in each repeating unit. By linking separate monomers with edges where $w \leq 1$, we can capture the recurrent nature of polymer chains as well as the ensemble of possible topologies. Fig. 1a shows such examples for, *e.g.*, alternating, random, and block copolymers with different sequence isomerisms. For homopolymers and simple alternating copolymers where all edges have a weight of one, this representation naturally reduces to a standard graph representation in which the two ends of the repeating unit have been linked. The periodic representation for crystalline materials proposed by Xie and Grossman[70] is also a special case of the ensemble graph representation proposed here.

Directed edges are necessary to handle a more general set of polymer and oligomer topologies than undirected edges can alone (Fig. 1b). Although termini might not exert a strong influence over an overall property of the polymeric material, they provide an apt example to highlight the circumstances that require directed edges. Graph networks learn a hidden representation for each atom in the system based on their neighbors and associated edges. Atoms that connect repeating units mostly have atoms from other repeating units as neighbors, and only infrequently will be connected to the termini. However, atoms that are part of the termini and that connect to the repeating unit always have the repeating unit atoms as neighbors. This asymmetry is needed for a graph network to correctly consider the typical neighborhood of each atom. Some examples of polymer architectures that also require this edge asymmetry are shown in Fig. 1b. In graft copolymers, for instance, where the main chain is not fully saturated, the atoms connecting the main and side chains do not always have each other as neighbors, and they may be so with different relative frequencies.

Graph neural network architecture

The network architecture developed is an extension of the D-MPNN known as Chemprop.[67] MPNNs are a class of GNNs that perform convolutions on graphs while maintaining node-order invariance, and have found broad application for molecular property prediction.[61,71–73] The input of these models is the molecular graph, in which atoms are represented by the graph vertices and bonds by the edges. All nodes and edges are associated with feature vectors that describe the atoms and bonds they represent. These feature vectors are updated iteratively *via* localized convolutions ("message passing" in the more general framework of MPNNs) that involve neighboring atoms and bonds, and result in learnt embeddings for all nodes and edges. A representation for the whole molecule is then obtained by aggregating (*e.g.*, summing) all atom embeddings. This numerical representation of fixed size is then used by a feed-forward neural network to predict the property of interest, and the whole architecture is trained end to end. Because convolutions and all other operations that manipulate the features of the input graph depend on learnable parameters that are updated by gradient descent, the network is encouraged to learn hidden node, edge, and molecular representations that are highly informative for the predictive task at hand.

In D-MPNNs, the messages used to iteratively update feature vectors are associated with directed edges (bonds), rather than with nodes (atoms) as in regular MPNN architectures.[67,74,75] In addition to having shown state-of-the-art performance on molecular property prediction tasks,[67] directed edges are needed for general graph representations of polymers, as

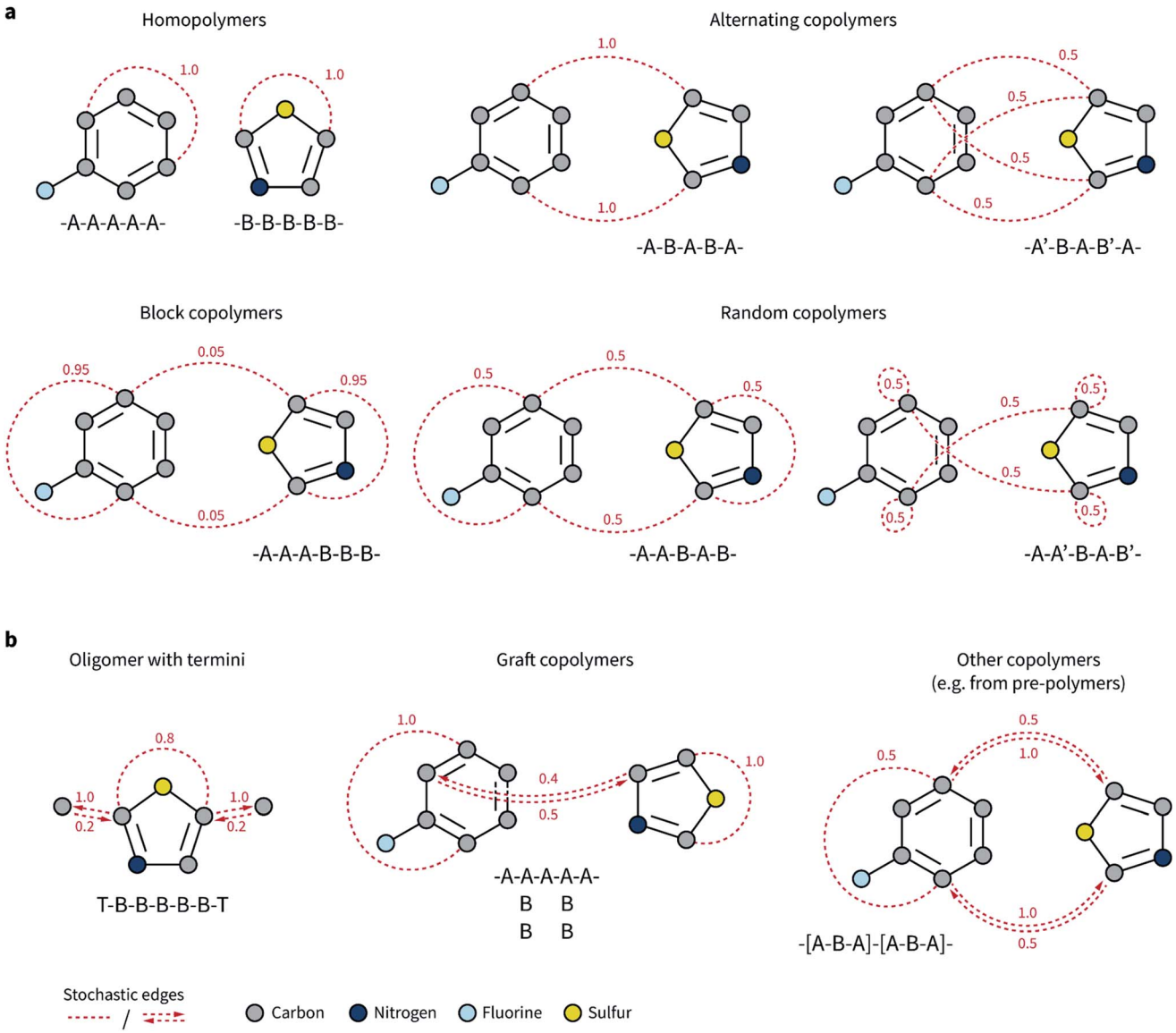

Fig. 1 Graph representation of selected polymer topologies. Stochastic edges are represented by red, dashed lines or arrows and have an associated weight between zero and one. Bead color corresponds to element type as described in the legend in the bottom-left corner of the figure. An example of a polymer sequence corresponding to each topology is displayed underneath each scheme as text, where A and B represent two different monomers, and T is a terminus. The prime symbol indicates a tail-to-head orientation for the monomer (*e.g.*, A–B indicates that the tail of A connects to the head of B, while A–B′ indicates that the tail of A connects to the tail of B). (a) Homopolymers, as well as most alternating, random, and block copolymers can be described as undirected graphs with some stochastic edges, where the probability of the edge reflects the frequency with which the bond is present in the polymer chain. (b) Directed edges enable the representation of a broader range of polymer topologies as graphs, and to capture the effect of termini when needed. Directed edges are necessary when two atoms have different probabilities of being neighbors of each other. The weight of an incoming edge corresponds to the probability that the source node (atom) is a neighbor of the sink node (atom). For instance, in the graft copolymer example, the atom on monomer A has the atom from monomer B as neighbor 40% of the times (two of the five A monomers are connected to B monomers), while the opposite is true with 50% of the times (two out of four B monomers are connected to the main chain of A monomers).

discussed above. Here, we propose to weigh directed edges according to their probability of occurring in the polymer chain. As such, we refer to this graph neural network as a weighted D-MPNN (wD-MPNN). The input provided to the wD-MPNN is the graph of the repeating unit of the polymer, in which each node and edge are associated with a set of atom and bond features, x_v and e_{uv}, respectively (Fig. 2a; details of these features are in the ESI Extended methods†).

A D-MPNN with messages centered on edges learns a hidden representation h_{vu} for each edge in the graph (Fig. 2b). After message passing, a hidden representation for each atom h_v is obtained by considering all of its incoming edges (Fig. 2c). In the wD-MPNN, we weigh each edge message according to its probability of being present in the repeating unit, w_{kv}, both when updating edge and atom representations (Fig. 2b and c). Similarly, in existing D-MPNNs, an overall molecular

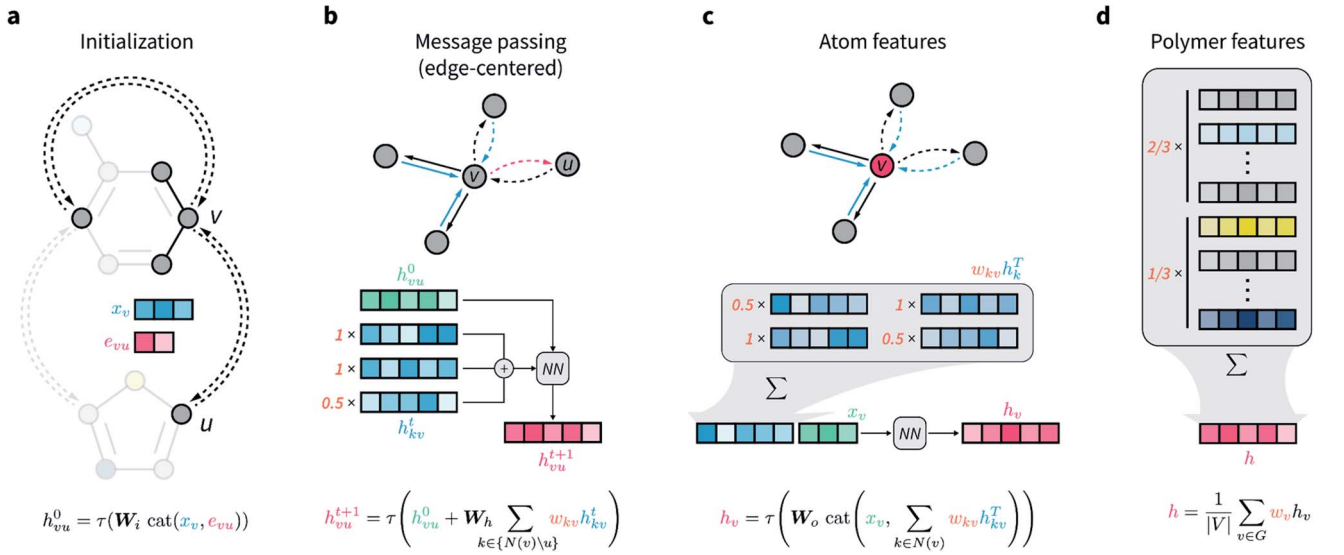

Fig. 2 Architecture of the directed message passing neural network for polymer property prediction. (a) Node and edge features are initialized based on corresponding atomic and bond properties, concatenated and passed through a single neural network layer. (b) Message passing is performed for T steps, in which edge-centered messages of v-outgoing edges are updated based on v-incoming edges. Each message is weighted according to user-specified bond probabilities that reflect the topology of the polymer repeating unit. (c) Updated atom features are obtained by a weighted sum over the features of all v-incoming edges, followed by concatenation with the initial atom features, and transformation *via* a single neural network layer. (d) Overall polymer features are obtained by aggregating all final atomic features *via* a weighted sum or average, where the weights reflect the relative abundance of different substructures (*e.g.*, monomers) in the polymer.

representation h is obtained by averaging or summing over all atom representations h_v. In the wD-MPNN, we weigh each h_v according to the relative abundance (*i.e.*, stoichiometry) of the monomer they belong to (Fig. 2d) to obtain an overall polymer representation h. The aim of incorporating weighted "stochastic" edges and stoichiometry information into the wD-MPNN is to capture a polymer's chain architecture and sequence isomerism by describing its average repeating unit.

The result of the wD-MPNN's processing of the input graph is h, a learned numerical representation of the molecular ensemble that defines a polymer and its properties. This is used as the input of a feed-forward neural network to predict the polymer properties of interest, with the whole architecture being trained end-to-end. Additional details of the wD-MPNN architecture are in the ESI Extended methods,† and an implementation is available on GitHub (see Data availability).

Copolymer dataset

Given the limited amount of publicly-available polymer data with broad coverage of monomer chemistries, chain architectures, and stoichiometries, we built such a dataset *via* computation. We considered the chemical space defined by Bai *et al.*[24] comprising conjugated polymers as photocatalysts for hydrogen production (Fig. 3a). The full set of possible monomer combinations provides $9 \times 682 = 6138$ possible co-polymer compositions. In addition to monomer composition, we considered three chain architectures (alternating, random, and block), and three stoichiometric ratios of monomers (1 : 1, 1 : 3, 3 : 1). For random and block copolymers, all ratios were considered, while for perfectly alternating copolymers only the 1 : 1 stoichiometry was considered. In total, this setup constitutes a space of 42 966 possible copolymers (Fig. 3a).

We took electron affinity (EA) and ionization potential (IP) as the properties to be predicted, and generated ground truth labels by computing these properties with density functional tight-binding methods.[76] Specifically, we followed the protocol proposed by Wilbraham and colleagues,[43] which involves the computation of EA and IP on oligomers (octamers) with xTB,[77] followed by a linear correction based on a calibration against density functional theory (DFT) calculations that used B3LYP density functional[78–81] and the DZP basis-set.[82] For each copolymer, we generated up to 32 sequences and 8 conformers per sequence. In fact, not only random, but also alternating and block copolymers may have multiple possible sequences given the asymmetry of the B monomers, which can result in sequence isomerism. The IP and EA values were Boltzmann averaged across the 8 conformers at 298.15 K, and then averaged across all sequences associated with a specific copolymer (further details in the ESI Extended methods†). Ultimately, this process led to a dataset of 42 966 copolymers with different chain architectures and stoichiometric ratios, each labeled with IP and EA values calculated as averages over the ensemble of sequences and conformations. All ML models were evaluated on the same cross-validation splits of this dataset, which included train, validation, and test sets. Both random and monomer splits were evaluated, as discussed in the Results.

Both EA and IP were considerably affected by the varying monomer chemistry, chain architecture, and monomer stoichiometry (Fig. 3b). Overall, however, monomer chemistry and stoichiometry had a larger impact on EA and IP than chain architecture. Note that an overlapping property distribution,

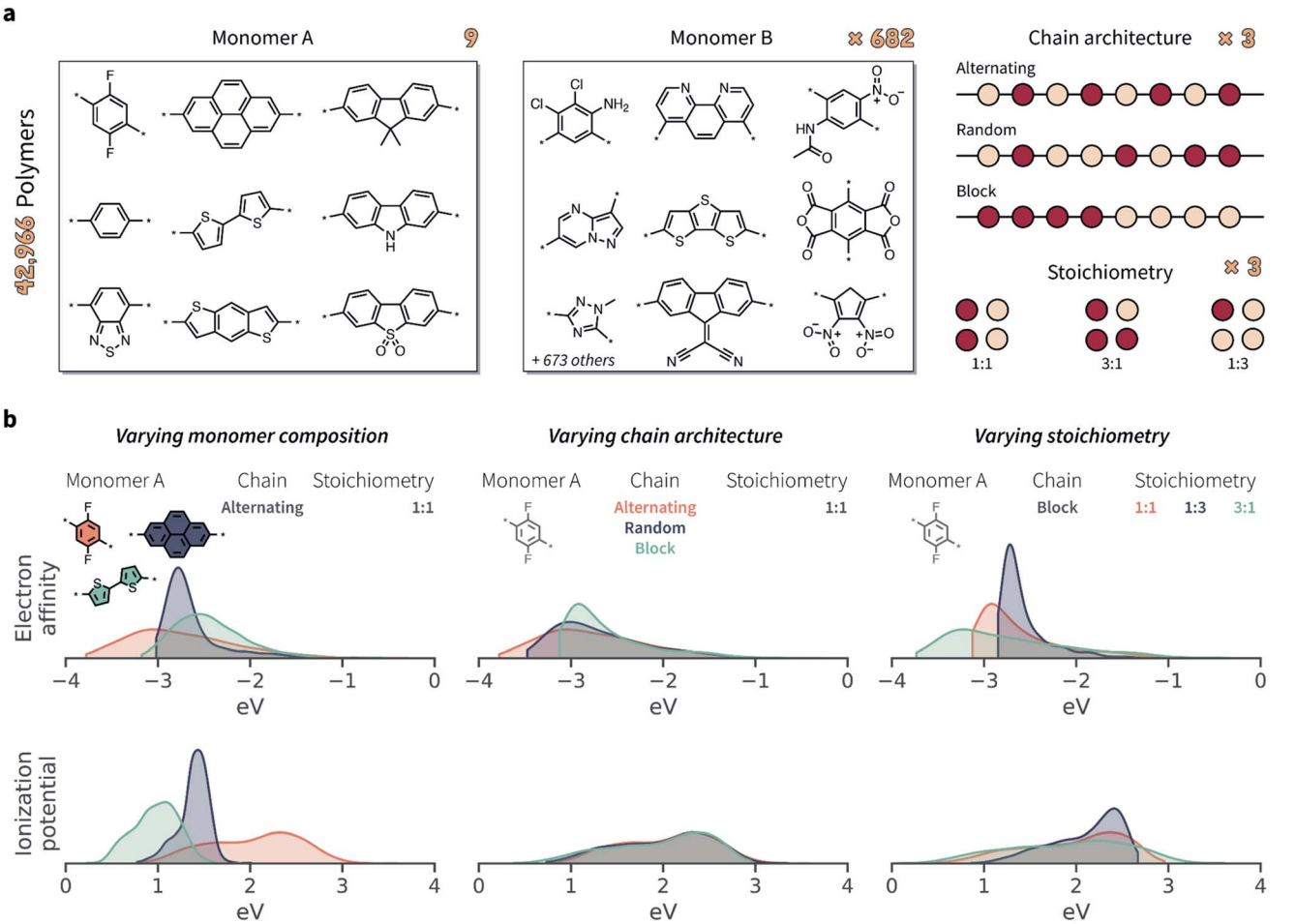

Fig. 3 Polymers and properties of the dataset. (a) Building blocks, chain architectures, and stoichiometries present in the dataset of 42 966 copolymers. The 9 monomers in group A are also present in group B. (b) Probability distributions of electron affinity (EA) and ionization potential (IP) for selected subsets of the copolymer data. The three columns highlight how the property distributions are affected by varying monomer compositions, chain architectures, and stoichiometries. Different monomer chemistry and stoichiometries have the largest impact on EA and IP, while chain architecture had a smaller effect overall. Probability distributions are shown as kernel density estimates, truncated at the respective largest and smallest values.

like that shown in Fig. 3b for the IP of polymers with different chain architectures (given a specific monomer A and stoichiometry), does not also imply no variation across chain architectures. While the overall distributions overlap, and while IP variation might be smaller than for varying monomer compositions and stoichiometries, the IP is still likely to be different between alternating, random, and block polymer sequences.

In addition to the dataset described above, we created two derivative datasets by artificially inflating the importance of (i) chain architecture and (ii) monomer stoichiometry in determining EA and IP. In the first case, given a specific monomer pair and stoichiometry, the standard deviation of EA and IP values was increased by a factor of 5 while maintaining their original mean values. In the second case, the standard deviation of EA and IP values was increased by a factor of 5 for each specific combination of monomer pairs and chain architecture. These artificial datasets were created to highlight how specific attributes of the wD-MPNN architecture capture property changes directly attributable to different chain architectures and stoichiometries.

Results

In the following sections, we present results in which we tested the ability of our wD-MPNN to predict EA and IP across monomer compositions, stoichiometries, and chain architectures, using our newly-built copolymer dataset. We compared the predictive power of this model against that of baselines models that included a D-MPNN[67] and RF and NN models based on fingerprints, and observed significant improvements in predictive ability over all baselines (Table 1). Because the NN results tracked, but were slightly inferior to those obtained with RF, here we will primarily discuss the RF data. We show how, contrary to traditional molecular representations and ML models, the wD-MPNN can discriminate between polymers with the same monomer composition, but different chain architectures and/or monomer stoichiometries. Finally, we demonstrate the use of this model for the prediction of diblock copolymer phases using a recently curated dataset.[53]

Table 1 Performance of all models tested for the prediction of copolymer EA and IP values. Average cross-validated RMSE values are shown. The standard error of the mean is reported in parenthesis and it applies to the least significant digits (*e.g.*, 0.22(1) is equivalent to 0.22 ± 0.01). The best performance across all models for each task is bolded

		Cross-validation split			
		Random split		Monomer split	
Approach		EA (eV)	IP (eV)	EA (eV)	IP (eV)
Monomer repr.	RF, binary FPs	0.19(0)	0.18(0)	0.33(2)	0.36(2)
	RF, count FPs	0.19(0)	0.18(0)	0.31(2)	0.35(3)
	NN, binary FPs	0.22(1)	0.19(0)	0.36(7)	0.30(3)
	NN, count FPs	0.23(0)	0.20(1)	0.26(1)	0.32(3)
	D-MPNN	0.17(0)	0.16(0)	0.20(1)	0.20(2)
Polymer repr.	RF, binary FPs	0.15(0)	0.14(0)	0.31(2)	0.34(2)
	RF, count FPs	0.09(0)	0.08(0)	0.25(3)	0.27(3)
	NN, binary FPs	0.18(0)	0.16(0)	0.28(3)	0.25(2)
	NN, count FPs	0.19(1)	0.14(3)	0.27(3)	0.20(2)
	wD-MPNN	**0.03(0)**	**0.03(0)**	**0.10(1)**	**0.09(2)**

Learned polymer representations provide improved predictive performance

The wD-MPNN architecture achieved higher performance than all other models tested on a random 10-fold cross validation split (Fig. 4a), saturating the performance measures used with an average coefficient of determination (R^2) of 1.00 and a root-mean-square error (RMSE) of 0.03 eV, for both EA and IP predictions. The standard error of the mean was less than 0.005 for both R^2 and RMSE. In our discussion of the results, when uncertainty is not provided, we imply it is less than half of the last significant digit.

The baseline D-MPNN model achieved RMSEs roughly six times larger (0.17 and 0.16 eV for EA and IP) than those achieved by the wD-MPNN. The RF models that relied on fingerprints representations of the monomeric units returned a performance only marginally inferior to that of the baseline D-MPNN (Fig. 4c). RF models with both binary and count fingerprints achieved RMSEs of 0.19 eV and 0.18 eV, for EA and IP, respectively. This performance improved substantially when using averaged fingerprints based on sampled oligomer sequences, which better capture chain architecture and monomer stoichiometry. This was especially true for the RF model using count fingerprints, which achieved RMSEs of 0.09 and 0.08 eV, making it the most competitive baseline approach tested. Despite the excellent performance on this dataset, its RMSE was still three times larger than the one achieved with the wD-MPNN model, and its performance overall qualitatively poorer as visible from the parity plots (Fig. 4).

When testing the models on a 9-fold cross-validation where the dataset was split according to the identity of monomer A (Fig. 3), performance decreased, as expected (Fig. S1†).

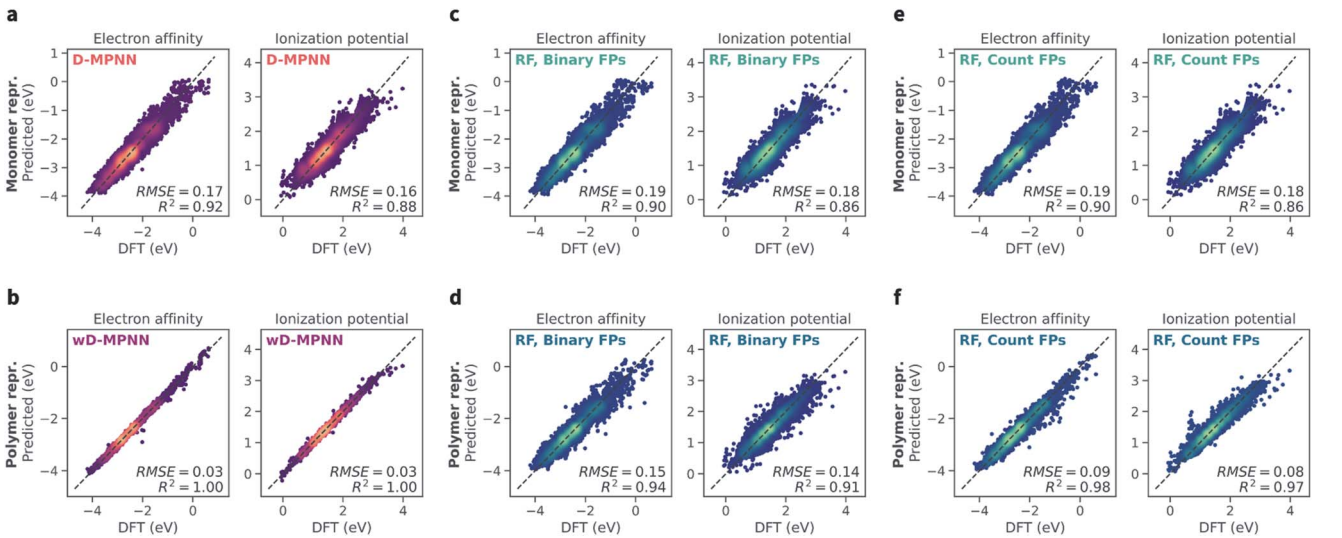

Fig. 4 Performance of the wD-MPNN and baseline models for the prediction of electron affinity (EA) and ionization potential (IP). Each parity plot shows the computed DFT values against the ones predicted by the ML models. The parity line is shown as a black dashed line. The scatter/density plots display the predictions of each model for all folds of the 10-fold cross validation. The color intensity is proportional to the probability density, with brighter colors indicating areas with higher density of points. The average coefficient of determination (R^2) and root-mean-square error (RMSE, in eV) across all folds are shown; the standard error of the mean is not explicitly shown as it is implied as being less than half of the last significant digit used. (a) Performance of the baseline D-MPNN model, which used a graph representation of the monomeric units as input. (b) Performance of the wD-MPNN model, which is augmented with information about chain architecture and monomer stoichiometry. (c) Performance of a RF model that used a binary fingerprint (FP) representation of the monomeric units as input. (d) Performance of a RF model that used a binary fingerprint representation of the polymer as input, which was obtained as the average fingerprint of a set of oligomeric sequences sampled uniformly at random, while satisfying the correct stoichiometry and chain architecture of the specified polymer. (e) Performance of a RF model that used a count fingerprint representation of the monomeric units as input. (f) Performance of a RF model that used a count fingerprint representation of the polymer as input. Equivalent plots for the results obtained with NN models are shown in Fig. S2.† The performance of all models is summarized in Table 1.

However, the wD-MPNN still achieved RMSEs of 0.10 ± 0.01 and 0.09 ± 0.02 eV, indicating strong generalization performance to new monomer identities. In addition, the performance gap with respect to most other methods increased significantly. The baseline D-MPNN achieved RMSE of 0.20 ± 0.01 and 0.20 ± 0.02 eV. Among the RF models, the highest performance was once again achieved by the representation using averaged count fingerprints across sampled oligomeric sequences, but was considerably worse than that of the D-MPNN models, with RMSEs of 0.25 ± 0.03 and 0.27 ± 0.03 eV, for EA and IP, respectively.

Finally, we tested the data efficiency of the D-MPNN model *via* multiple, random dataset splits in which we considered training set sizes that included between 43 and 34 373 polymers (*i.e.*, between 0.1% and 80% of the dataset). While the performance of the most competitive RF model (using count fingerprints and sampled polymer chains) was always above that of the baseline D-MPNN, a cross-over point at ∼1000 data instances was observed for the wD-MPNN architecture, after which its performance overtook that of RF (Fig. S3†).

The better performance of GNNs is likely a consequence of the fact these models can be thought of as generalizations of fingerprints that allow for a more flexible perception of substructures.[59,60] This feature-extraction process that adapts to specific predictive tasks can improve generalization in the large-data limit.[67]

The wD-MPNN captures how polymer properties depend on chain architecture and monomer stoichiometry

The improved performance of the novel wD-MPNN architecture is a direct result of its ability to discriminate between polymers comprised of the same monomeric units, but differing in their relative abundance (*i.e.*, different stoichiometry) and how they connect to one another to form different sequence ensembles (*i.e.*, different chain architecture). To demonstrate how this information is captured by the additional terms (Methods and Fig. 2) provided as inductive biases to the model, we performed ablation studies in which weighted bond information or the terms relative to stoichiometry information were not provided. This resulted in a set of four models: (i) the baseline D-MPNN that is aware of the structure of the separate monomers only, (ii) a D-MPNN that is provided with information of how the monomers may connect to one another to form a chain with specific architecture (alternating, block, or random; information used in the steps shown in Fig. 2b and c), (iii) a D-MPNN that is provided with information about monomer stoichiometry (information used in the step shown in Fig. 2d), and (iv) the full wD-MPNN architecture that is provided with both chain architecture and stoichiometry information.

As discussed above, the baseline D-MPNN model achieved an RMSE of 0.16 eV in the cross-validated prediction of ionization potential (IP). When providing the D-MPNN with information on chain architecture, a small but statistically significant improvement in RMSE was observed, to 0.15 eV. A more substantial improvement was instead observed when the D-MPNN was provided with information on monomer stoichiometries (RMSE = 0.07 eV). This result may have been anticipated given that, overall, monomer stoichiometry was observed to have a larger impact on EA and IP than chain architecture (Fig. 3). Yet, both information on stoichiometry and chain architecture was needed by the default D-MPNN to achieve the highest performance (RMSE = 0.03 eV). Equivalent results were obtained also for EA, both when using RMSE and R^2 as performance measures, and are reported in Table S1.†

While for the specific properties studied here (EA and IP) stoichiometry seemed more important than chain architecture, this is not necessarily the case for other polymer properties. To demonstrate how capturing chain architecture is important in such cases, and to further demonstrate how the wD-MPNN is able to exploit this additional information to achieve superior performance, we created two additional fictitious polymer datasets. These were obtained by artificially inflating the importance of chain architecture and monomer stoichiometry in determining EA and IP (see Methods). While these datasets do not reflect any specific polymeric property, and so we focus on evaluation only in terms of R^2, they provide realistic scenarios in which we can control the relative importance of chain architecture and stoichiometry. When chain architecture was made the primary variable determining the IP values, taking this information into account provided the largest performance boost with respect to the baseline model (R^2 from 0.65 to 0.86; Table 2). Conversely, when stoichiometry was made

Table 2 Effect of capturing chain architecture and monomer stoichiometry information on D-MPNN performance. The average R^2 values obtained from a 10-fold cross validation based on random splits, for the prediction of IP, are shown. Uncertainty is implied as the standard error of the mean was <0.005 in all cases. Under the header "Representation", "monomers" indicates the model was provided with the graph structure of separate monomer units; "chain architecture" indicates the model was provided with information on how the monomer units may connect to one another to form an ensemble of possible sequences, *via* the definition of edge weights, used as shown in Fig. 2b and c; "stoichiometry" indicates the model was provided with information on monomer stoichiometry, which was used to weigh learnt node representations as shown in Fig. 2d. An extended version of this table, with results obtained also for EA and showing RMSE too as performance measure, is available in Table S1

	Representation			
Datasets	Monomers	Monomers + chain architecture	Monomers + stoichiometry	Monomers + chain architecture + stoichiometry
Original dataset	0.88	0.90	0.98	1.00
Inflated chain architecture importance	0.65	0.86	0.71	0.98
Inflated stoichiometry importance	0.26	0.27	0.97	0.99

artificially even more important, models that did not take it into account could not achieve R^2 values above 0.27, while those that did achieved R^2 equal or above 0.97. Importantly, in both cases, in which either chain architecture or stoichiometry provided only minimal information, the full wD-MPNN model was able to focus on the most important of the two and always achieve the highest performance of all models tested (R^2 of 0.98 and 0.99).

Predicting diblock copolymer phases from the polymers' chemistry

We further evaluated the wD-MPNN architecture on an experimental dataset that has been recently compiled by Arora *et al.*[53] This dataset provides the phase behavior of 50 diblock copolymers corresponding to a set of 32 homopolymers and copolymers. It reports the observed copolymer phases (lamellae, hexagonal-packed cylinders, body-centered cubic spheres, a cubic gyroid, or disordered) for various relative volume fractions and molar masses, for a total of 4780 entries. Each entry may be associated with more than one phase, such that the task can be defined as a binary multi-label classification task with five labels, one for each of the phases that can be observed.

The wD-MPNN model was provided with the monomer graphs for both blocks, how these may connect to each other *via* stochastic edges, and the mole fraction of each block (Fig. 2). Here, we also provided the overall copolymer size by scaling the molecular embeddings h by $1 + \log(N)$, where N is the degree of polymerization. The scaling factor thus has no effect for chain lengths of one, reducing naturally to the default D-MPNN implementation.

Overall, the wD-MPNN achieved a classification performance, as measured by the area under the precision–recall curve (PRC),[83] of 0.68 ± 0.01 in a 10-fold cross-validation based on stratified splits (Fig. 5). Given that some phases are more common than others, resulting in the five labels being imbalanced, the PRC of a random classifier is expected to be 0.23. When the chain architecture, stoichiometry, and degree of polymerization are not provided to the model, performance drops significantly to a PRC of 0.47 ± 0.01. Considering each of these aspects of the polymer structure improves performance (Fig. 5). When information on chain architecture was provided, *via* weighted edges, the D-MPNN achieved a PRC of 0.49 ± 0.01; when information on polymer size was provided, by scaling molecular embeddings with the degree of polymerization, a PRC of 0.52 ± 0.01 was achieved; and when information on monomer stoichiometry was provided, by scaling atom embeddings with mole fractions, a PRC of 0.67 ± 0.01 was achieved.

From the results above it emerges how, for this task, the mole fraction of each block is the most informative feature of the polymer. This may be expected given that mole fractions highly correlate with the volume fractions of the two blocks, which is an important factor determining the copolymer phase. In particular, it has been observed that for this dataset very high classification performance can be achieved based *solely* on knowledge of the volume fractions.[53] A RF model trained only on mole fractions achieved a PRC of 0.69 ± 0.01 (Fig. S4†), and

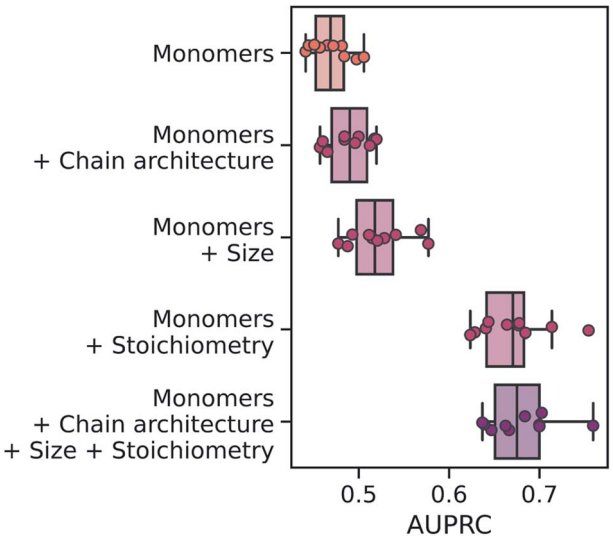

Fig. 5 Performance of the wD-MPNN and ablated architectures on the classification of diblock copolymer phases. Performance is measured as the average area under the precision–recall curve (PRC) across the five labels (phases) to be predicted. Each marker reflects the average PRC value, for one fold of a 10-fold cross validation, across the five classification labels. The boxes show the first, second, and third quartiles of the data, and the whiskers extend to up to 1.5 times the interquartile range.

0.71 ± 0.01 when using solely volume fractions, both of which are better than the wD-MPNN. The highest performance on this task was achieved by a RF model that used count fingerprints with sequence sampling, as well as stoichiometry and size information (PRC of 0.74 ± 0.01; Fig. S4†). It is important to note how this is a unique scenario, as for most properties of interest there will not be a simple univariate relationship between the property being predicted and an easily computed or measured variable that also does not depend on the chemistry of the copolymer (Extended discussion, Note S1†). Nevertheless, the relative performance of a structure-based representation in Fig. 5 demonstrates the advantages of the wD-MPNN over a monomer-only D-MPNN (Extended discussion, Note S2†).

Discussion

The lack of suitable representations for molecular ensembles is a key obstacle to the development of supervised learning algorithms for polymer informatics. While we have taken a first step toward tailored representations and models for polymer property prediction, these are by no means complete. In particular, the proposed approach only captures the expectation of a molecular ensemble, and not its dispersity.[84,85] The way in which the chain architecture of polymers is described *via* weighted edges is representative of an average repeating unit. As such, this representation would not distinguish between gradient[86,87] and block copolymers with the same average block length. Similarly, the use of the average chain length alone as a scaling factor for the molecular embeddings results in

a model that cannot distinguish between polymers with equal average size but different polydispersity. Graph network architectures that better capture the heterogeneity of molecular ensembles could be explored in the future by, for instance, expanding the parametric approach used in this work to higher-order moments beyond the mean of the distribution of connections. Despite these limitations, the approach developed represents a first step toward better ML representations of materials that are composed of large molecular ensembles.

A hypothetical alternative strategy would be to train a model to predict the properties of sequence-defined structures only, and to then examine the ensemble of values corresponding to an ensemble of structures; this may be viable for computed properties where each constituent oligomer has a calculable property, but it does not naturally extend to experimental datasets where only one aggregate property is measured for the ensemble of structures (Extended discussion, Note S3†).

In the copolymer phase prediction task, we incorporated information on polymer size in the wD-MPNN architecture by scaling the learnt polymer embeddings. However, there are alternative approaches that could be explored with suitable datasets. Another way to incorporate size information explicitly into the model would be to append the degree of polymerization, or the molar mass, to the embedding vector h after message passing. This also provides a general means to have the wD-MPNN consider information about process parameters. However, when information about the termini is available, the weights w_{kv} associated with the stochastic edges of the termini, together with the weights w_v reflecting the stoichiometric ratio between different building blocks, in principle would already capture average chain length implicitly. As more copolymer datasets become available, one could explore multiple ways to integrate size information into the wD-MPNN architecture and study the performance and generality of different approaches.

To further advance these ML models and the field of polymer informatics, data availability is fundamental. Here, we have built and provided a computational dataset of over 40 000 polymers that may be used to further develop tailored ML models for polymer property prediction. Yet, more such datasets are needed to increase the diversity of polymer prediction tasks, each of which might uniquely be affected by different aspects of the ensemble of molecules defining the material. While expensive, properties computed *via* electronic structure[42,43] or molecular dynamics calculations[88] provide a means to obtain comprehensive and relatively noise-free datasets to establish the first generation of ML models designed specifically for polymers. Despite not yet being readily available, properties obtained *via* atomistic molecular dynamics simulations may be especially complementary to the dataset provided here, as they may more strongly depend on intramolecular interactions, conformational ensembles, and chain length. In the meantime, it will be important to create open-access databases of experimentally-measured properties available to the community, and in machine readable format, reflecting similar and established initiatives in other chemistry fields.[89–97] Indeed, we have observed that thousands of training points may be required to fully take advantage of the expressivity of these more flexible graph architectures (Fig. S3†). Efforts like PolyInfo[98] and Polymer Genome[28] attempt to tackle this challenge, but the data in these databases is not truly open access. Open initiatives that aim at building findable as well as accessible databases, like the Community Resource for Innovation in Polymer Technology (CRIPT),[99] will likely play an increasingly important role in enabling tailored ML models for polymer informatics.

The wD-MPNN model described in this work is particularly useful in polymer design campaigns in which exploring a broad range of monomer chemistries and compositions, chain architectures, and polymer sizes is of interest. When this is not the case, however, and one would like to focus on a small set of monomers and a well-defined chain architecture (*e.g.*, only alternating copolymers, or even sequence-defined polymers), the use of such a model is not necessarily advantageous with respect to more traditional ML models. Indeed, if a ML algorithm is not required to distinguish between polymers with, *e.g.*, different chain architectures, average sizes, or monomer stoichiometries, then the structure of the monomers alone or the use of hand-crafted descriptors will be sufficient. Furthermore, the availability of highly informative descriptors or proxy observables may obviate the need for a deep learning model, as we noticed for the task of predicting the phases of diblock copolymers. Finally, the model choice might also be forced by data availability. As discussed in the Results section, for the task of predicting EA and IP we found that with fewer than ~1000 data instances the wD-MPNN did not provide an advantage over a RF model. Only when >1000 examples were provided for training did the wD-MPNN overtake the performance seen for RF (Fig. S3 and Extended discussion, Note S4†).

Conclusion

In this work we have developed a graph representation of molecular ensembles and an associated wD-MPNN architecture with immediate relevance to polymer property prediction. We have shown how this approach captures critical features of polymeric materials, like chain architecture, monomer stoichiometry, and expected size, to achieve superior accuracy with respect to baseline approaches that disregard this information. We have furthermore developed competitive baseline models based on random forest, count fingerprints, and sequence sampling. To evaluate the performance of the different ML models, we generated a dataset with electron affinity and ionization potential values for over 40k polymers with varying monomer composition, stoichiometry, and chain architecture *via* ~15 million single point energy calculations. Both this dataset and the ML models developed constitute a positive step toward next-generation algorithms for polymer informatics and provide an avenue for modeling the properties of molecular ensembles.

Data availability

The wD-MPNN model developed is available on GitHub at **https://github.com/coleygroup/polymer-chemprop** (v1.4.0-

polymer). The dataset with electron affinity and ionization potential values for 42 966 copolymers is provided as part of the ESI† as a CSV file, and on GitHub at **https://github.com/coleygroup/polymer-chemprop-data**. Jupyter notebooks and Python scripts needed to create the dataset and reproduce the results of this manuscript are also available in the same GitHub repository.

Author contributions

M. A. and C. W. C. conceptualized and planned the research. M. A. wrote the code, performed the experiments, and analyzed the data. M. A. and C. W. C. interpreted the results and wrote the manuscript. C. W. C. supervised the work.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

M. A. and C. W. C. thank Guangqi Wu, Bradley D. Olsen, Hursh V. Sureka, and Katharina Fransen for useful discussions on polymer chemistry. M. A. thanks Thijs Stuyver for help and discussions related to the electronic structure calculations, Rebecca Neeser for suggestions on data visualization, and John Bradshaw and Samuel Goldman for discussions on the results of this work. Research reported in this publication was supported by NIGMS of the National Institutes of Health under award number R21GM141616. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1 M. M. Coleman and P. C. Painter, *Fundamentals of Polymer Science: An Introductory Text*, Taylor & Francis, 2nd edn, 1998.
2 N. G. McCrum, C. P. Buckley, C. B. Bucknall and C. B. Bucknall, *Principles of Polymer Engineering*, Oxford University Press, 1997.
3 A.-V. Ruzette and L. Leibler, *Nat. Mater.*, 2005, **4**, 19–31.
4 K. Kataoka, A. Harada and Y. Nagasaki, *Adv. Drug Delivery Rev.*, 2001, **47**, 113–131.
5 M. T. Islam, M. M. Rahman and N.-U.-S. Mazumder, *Front. Text. Mater.*, 2020, 13–59.
6 J. Lopez, D. G. Mackanic, Y. Cui and Z. Bao, *Nat. Rev. Mater.*, 2019, **4**, 312–330.
7 A. Patil, A. Patel and R. Purohit, *Mater. Today: Proc.*, 2017, **4**, 3807–3815.
8 Y. K. Sung and S. W. Kim, *Biomater. Res.*, 2020, **24**, 12.
9 H.-J. Huang, Y.-L. Tsai, S.-H. Lin and S.-H. Hsu, *J. Biomed. Sci.*, 2019, **26**, 73.
10 W. B. Liechty, D. R. Kryscio, B. V. Slaughter and N. A. Peppas, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 149–173.
11 R. Kumar, N. Le, Z. Tan, M. E. Brown, S. Jiang and T. M. Reineke, *ACS Nano*, 2020, **14**(12), 17626–17639.
12 R. Kumar, N. Le, F. Oviedo, M. E. Brown and T. M. Reineke, *JACS Au*, 2022, **2**, 428–442.
13 R. Kumar, C. F. Santa Chalarca, M. R. Bockman, C. Van Bruggen, C. J. Grimme, R. J. Dalal, M. G. Hanson, J. K. Hexum and T. M. Reineke, *Chem. Rev.*, 2021, **121**, 11527–11652.
14 P. Bannigan, M. Aldeghi, Z. Bao, F. Häse, A. Aspuru-Guzik and C. Allen, *Adv. Drug Delivery Rev.*, 2021, **175**, 113806.
15 Z. Tan, Y. Jiang, M. S. Ganewatta, R. Kumar, A. Keith, K. Twaroski, T. Pengo, J. Tolar, T. P. Lodge and T. M. Reineke, *Macromolecules*, 2019, **52**(21), 8197–8206.
16 M. J. Mitchell, M. M. Billingsley, R. M. Haley, M. E. Wechsler, N. A. Peppas and R. Langer, *Nat. Rev. Drug Discovery*, 2021, **20**, 101–124.
17 A. C. Obermeyer, C. E. Mills, X.-H. Dong, R. J. Flores and B. D. Olsen, *Soft Matter*, 2016, **12**, 3570–3581.
18 K. Ulbrich, K. Holá, V. Šubr, A. Bakandritsos, J. Tuček and R. Zbořil, *Chem. Rev.*, 2016, **116**, 5338–5431.
19 P. Bannigan, Z. Bao, R. Hickman, M. Aldeghi, F. Häse, A. Aspuru-Guzik and C. Allen, *ChemRxiv*, 2022, DOI: **10.26434/chemrxiv-2021-mxrxw-v2**.
20 S. Kosuri, C. H. Borca, H. Mugnier, M. Tamasi, R. A. Patel, I. Perez, S. Kumar, Z. Finkel, R. Schloss, L. Cai, M. L. Yarmush, M. A. Webb and A. J. Gormley, *Adv. Healthcare Mater.*, 2022, **11**, e2102101.
21 A. C. Mayer, S. R. Scully, B. E. Hardin, M. W. Rowell and M. D. McGehee, *Mater. Today*, 2007, **10**, 28–33.
22 C. Lee, S. Lee, G.-U. Kim, W. Lee and B. J. Kim, *Chem. Rev.*, 2019, **119**, 8028–8086.
23 S. Muench, A. Wild, C. Friebe, B. Häupler, T. Janoschka and U. S. Schubert, *Chem. Rev.*, 2016, **116**, 9438–9484.
24 Y. Bai, L. Wilbraham, B. J. Slater, M. A. Zwijnenburg, R. S. Sprick and A. I. Cooper, *J. Am. Chem. Soc.*, 2019, **141**, 9063–9071.
25 Y. Wang, A. Vogel, M. Sachs, R. S. Sprick, L. Wilbraham, S. J. A. Moniz, R. Godin, M. A. Zwijnenburg, J. R. Durrant, A. I. Cooper and J. Tang, *Nat. Energy*, 2019, **4**, 746–760.
26 M. J. Tamasi, R. A. Patel, C. H. Borca, S. Kosuri, H. Mugnier, R. Upadhya, N. S. Murthy, M. A. Webb and A. J. Gormley, *Adv. Mater.*, 2022, e2201809.
27 S. Mohapatra, N. Hartrampf, M. Poskus, A. Loas, R. Gómez-Bombarelli and B. L. Pentelute, *ACS Cent. Sci.*, 2020, **6**, 2277–2286.
28 C. Kim, A. Chandrasekaran, T. D. Huan, D. Das and R. Ramprasad, *J. Phys. Chem. C*, 2018, **122**, 17575–17585.
29 A. Jha, A. Chandrasekaran, C. Kim and R. Ramprasad, *Modell. Simul. Mater. Sci. Eng.*, 2019, **27**, 024002.
30 L. Tao, V. Varshney and Y. Li, *J. Chem. Inf. Model.*, 2021, **61**, 5395–5413.
31 J. Park, Y. Shim, F. Lee, A. Rammohan, S. Goyal, M. Shim, C. Jeong and D. S. Kim, *ACS Polym. Au*, 2022, **2**(4), 213–222.
32 C. Kuenneth, W. Schertzer and R. Ramprasad, *Macromolecules*, 2021, **54**, 5957–5961.
33 L. Tao, G. Chen and Y. Li, *Patterns*, 2021, **2**, 100225.
34 S. Wu, Y. Kondo, M.-A. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi,

C. Schick, J. Morikawa and R. Yoshida, *npj Comput. Mater.*, 2019, **5**, 1–11.
35 H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, *ACS Cent. Sci.*, 2019, **5**, 1717–1730.
36 S. Venkatram, C. Kim, A. Chandrasekaran and R. Ramprasad, *J. Chem. Inf. Model.*, 2019, **59**, 4188–4194.
37 M. Wang, Q. Xu, H. Tang and J. Jiang, *ACS Appl. Mater. Interfaces*, 2022, **14**, 8427–8436.
38 J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau and S. K. Kumar, *Sci. Adv.*, 2020, **6**, eaaz4301.
39 L. Wilbraham, R. S. Sprick, K. E. Jelfs and M. A. Zwijnenburg, *Chem. Sci.*, 2019, **10**, 4973–4984.
40 A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan and R. Ramprasad, *Comput. Mater. Sci.*, 2020, **172**, 109286.
41 H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty, M. Ramprasad, J. Laws, M. Shelton and R. Ramprasad, *J. Appl. Phys.*, 2020, **128**, 171104.
42 P. C. St. John, C. Phillips, T. W. Kemper, A. N. Wilson, Y. Guan, M. F. Crowley, M. R. Nimlos and R. E. Larsen, *J. Chem. Phys.*, 2019, **150**, 234111.
43 L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs and M. A. Zwijnenburg, *J. Chem. Inf. Model.*, 2018, **58**, 2450–2459.
44 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, *Sci. Rep.*, 2016, **6**, 1–10.
45 L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, P. Vashishta, G. A. Sotzing, Y. Cao and R. Ramprasad, *npj Comput. Mater.*, 2020, **6**, 1–9.
46 M. Reis, F. Gusev, N. G. Taylor, S. H. Chung, M. D. Verber, Y. Z. Lee, O. Isayev and F. A. Leibfarth, *J. Am. Chem. Soc.*, 2021, **143**, 17677–17689.
47 C. Kuenneth, A. C. Rajan, H. Tran, L. Chen, C. Kim and R. Ramprasad, *Patterns Prejudice*, 2021, **2**, 100238.
48 L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth and R. Ramprasad, *Mater. Sci. Eng., R*, 2021, **144**, 100595.
49 S. Wu, H. Yamada, Y. Hayashi, M. Zamengo and R. Yoshida, Potentials and challenges of polymer informatics: exploiting machine learning for polymer design, DOI: **10.48550/arXiv.2010.07683**.
50 D. J. Audus and J. J. de Pablo, *ACS Macro Lett.*, 2017, **6**, 1078–1082.
51 J. S. Peerless, N. J. B. Milliken, T. J. Oweida, M. D. Manning and Y. G. Yingling, *Adv. Theory Simul.*, 2019, **2**, 1800129.
52 T. K. Patra, *ACS Polym. Au*, 2022, **2**, 8–26.
53 A. Arora, T.-S. Lin, N. J. Rebello, S. H. M. Av-Ron, H. Mochigase and B. D. Olsen, *ACS Macro Lett.*, 2021, **10**, 1339–1345.
54 T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen and B. D. Olsen, *ACS Cent. Sci.*, 2019, **5**, 1523–1531.
55 S. Chithrananda, G. Grand and B. Ramsundar, *Machine Learning for Molecules Workshop at NeurIPS*, 2020.
56 B. Fabian, T. Edlich, H. Gaspar, M. H. S. Segler, J. Meyers, M. Fiscato and M. Ahmed, *Machine Learning for Molecules Workshop at NeurIPS*, 2020.
57 R. A. Patel, C. H. Borca and M. A. Webb, *Mol. Syst. Des. Eng.*, 2022, **7**(6), 661–676.
58 J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li and M. Sun, *AI Open*, 2020, **1**, 57–81.
59 D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, in *Advances in Neural Information Processing Systems*, ed. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, Curran Associates, Inc., 2015, vol. 28.
60 S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 595–608.
61 J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **181**, 475–483.
62 B. Sánchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik and A. B. Wiltschko, arXiv:1910.10685, 2019.
63 M. Tsubaki, K. Tomii and J. Sese, *Bioinformatics*, 2018, **35**, 309–318.
64 D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu and T. Hou, *J. Cheminf.*, 2021, **13**, 12.
65 K. McCloskey, E. A. Sigel, S. Kearnes, L. Xue, X. Tian, D. Moccia, D. Gikunju, S. Bazzaz, B. Chan, M. A. Clark, J. W. Cuozzo, M.-A. Guié, J. P. Guilinger, C. Huguet, C. D. Hupp, A. D. Keefe, C. J. Mulhern, Y. Zhang and P. Riley, *J. Med. Chem.*, 2020, **63**, 8857–8866.
66 S. Mohapatra, J. An and R. Gómez-Bombarelli, *Machine Learning: Science and Technology*, 2022, **3**, 015028.
67 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *J. Chem. Inf. Model.*, 2019, **59**, 3370–3388.
68 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
69 G. A. Landrum, *RDKit: Open-Source Cheminformatics*.
70 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
71 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017, pp. 1263–1272, **https://JMLR.org**.
72 E. Heid and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**(9), 2101–2110.
73 D. Flam-Shepherd, T. C. Wu, P. Friederich and A. Aspuru-Guzik, *Machine Learning: Science and Technology*, 2021, **2**, 045009.
74 H. Dai, B. Dai and L. Song, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016, pp. 2702–2711, **https://JMLR.org**.
75 J. Gasteiger, J. Groß and S. Günnemann, in *International Conference on Learning Representations*, 2020.

76 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**(2), e1493.
77 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
78 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **59**, 1200.
79 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
80 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
81 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
82 A. Schäfer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571–2577.
83 T. Saito and M. Rehmsmeier, *PLoS One*, 2015, **10**, e0118432.
84 N. A. Lynd and M. A. Hillmyer, *Macromolecules*, 2005, **38**, 8803–8810.
85 D. T. Gentekos, L. N. Dupuis and B. P. Fors, *J. Am. Chem. Soc.*, 2016, **138**, 1848–1851.
86 E. Grune, M. Appold, A. H. E. Müller, M. Gallei and H. Frey, *ACS Macro Lett.*, 2018, **7**, 807–810.
87 M. M. Alam, K. S. Jack, D. J. T. Hill, A. K. Whittaker and H. Peng, *Eur. Polym. J.*, 2019, **116**, 394–414.
88 M. A. Webb, N. E. Jackson, P. S. Gil and J. J. de Pablo, *Sci. Adv.*, 2020, **6**, eabc6216.
89 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
90 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
91 M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *Nucleic Acids Res.*, 2015, **44**, D1045–D1053.
92 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
93 C. Zardecki, S. Dutta, D. S. Goodsell, M. Voigt and S. K. Burley, *J. Chem. Educ.*, 2016, **93**, 569–575.
94 H. M. Berman and L. M. Gierasch, *J. Biol. Chem.*, 2021, **296**, 100608.
95 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
96 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
97 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
98 S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu and M. Yamazaki, in *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011, pp. 22–29, **https://ieeexplore.ieee.org**.
99 *A Community Resource for Innovation in Polymer Technology*, **https://cript.mit.edu/**.