Calculating Pairwise Similarity of Polymer Ensembles via Earth Mover's Distance

Jiale Shi[1], Dylan Walsh[1], Weizhong Zou[1], Nathan J. Rebello[1], Michael E. Deagen[1], Katharina A. Fransen[1], Xian Gao[2], Bradley D. Olsen[1*], Debra J. Audus[3*]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
2. Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, United States
3. Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

*Correspondence: email: bdolsen@mit.edu and debra.audus@nist.gov.

Abstract

Synthetic polymers, in contrast to small molecules and deterministic biomacromolecules, are typically ensembles comprised of polymer chains with varying numbers, lengths, sequences, chemistry, and topologies. While numerous approaches exist for measuring pairwise similarity among small molecules and sequence-defined biomacromolecules, accurately determining the pairwise similarity between two polymer ensembles remains challenging. This work proposes the earth mover's distance (EMD) metric to calculate the pairwise similarity score between two polymer ensembles. EMD offers a greater resolution of chemical differences between polymer ensembles than the averaging method and provides a quantitative numeric value representing the pairwise similarity between polymer ensembles in alignment with chemical intuition. The EMD approach for assessing polymer similarity enhances the development of accurate chemical search algorithms within polymer databases and can improve machine learning techniques for polymer design, optimization, and property prediction.

Keywords

cheminformatics, macromolecules, similarity, earth mover's distance, polymer ensemble, graph edit distance, digital search

Introduction

Polymers, with their wide range of applications and properties, are integral to numerous industries[1] including textiles,[2] water purification,[3, 4] energy,[5] transportation,[6] and health care.[7] As the demand for polymeric materials with bespoke properties continues to grow, understanding the underlying similarities and differences between polymers is essential for the efficient design and optimization of materials.[8-10] The study of polymer similarity not only provides insights into structure-property relationships[11, 12] but also aids in the development of effective search algorithms for polymer databases[13-19] and advances machine learning techniques for property prediction and materials

discovery.[12, 20-37] Despite its importance, quantifying the similarity of polymers remains a challenging task, primarily due to the fact that polymers are ensembles of polymer chains with varying numbers, lengths, sequences, chemistry, and topologies.[38, 39] All of these features can affect polymers' properties and can make similarity studies of polymers more complex than those of well-defined small molecules[40-42] and sequence-defined biomacromolecules.[43, 44]

To compute the similarity between polymer ensembles, researchers typically[45, 46] first embed each molecule in the ensemble, or equivalently convert every polymer chain into a vector, and then average all the embedding vectors to obtain a global embedding vector for the ensemble. Similarity operations (i.e., cosine similarity or jaccard index) are then performed to calculate the similarity between two ensemble embedding vectors, ultimately yielding a similarity between the two polymer ensembles. For instance, Aldeghi et al.[46] utilized this method to derive a global embedding vector for ensembles of polymer chains for block polymers, random polymers, and alternating copolymers. However, this commonly used average method prematurely reduces the dimensionality of the system, eliminating differences among ensembles due to the topological or monomer sequence information.[30, 45-47] This premature loss of key information can result in two distinct ensembles being classified as identical. Furthermore, the design of embedding functions becomes non-trivial when the polymer chains have varying chain lengths or complex nonlinear topologies.

Apart from the average methods, researchers have explored the development of new text-based[48-50] and graph-based stochastic representations[46, 51-53] that respect the unique aspects of polymer chemistry and then utilize these representations for similarity calculations. For instance, BigSMILES[48, 49] is a text-based representation that builds upon the simplified molecular-input line-entry system (SMILES)[54, 55] representation for small molecules and is designed specifically to describe the stochastic nature of polymer molecules. The polymer automaton[51] developed by Lin et al. is a graph-based state machine representation that describes polymers' stochastic features. Aldeghi et al. developed a graph-based representation with "stochastic" edges to describe the average structure of repeat units.[46] These existing text-based and graph-based stochastic representations can be used to calculate the pairwise similarity score, which captures the chemical and topological features contained in a polymer chemical structure diagram.[56] However, these stochastic representations do not specify the weight or probability of each polymer molecule within the ensemble. This probability information can include chain length, composition gradient, stereochemistry, and molecular mass distribution. This additional information is not included in chemical structure representations; rather it is obtained via polymer characterization and linked to chemical structure in data structures such as CRIPT[9] and PolyDAT[57].

This work proposes the earth mover's distance (EMD)[58] to quantitively calculate the similarity of polymer ensembles with greater chemical resolution. Four examples are presented to illustrate the power of EMD in characterizing the similarity between polymer ensembles, including two component copolymer ensembles, first-order Markov linear copolymer ensembles, star-polymer ensembles and polymer ensembles represented by molecular mass distributions (MMDs). The

proposed EMD metric for calculating the pairwise similarity of polymer ensembles offers a higher resolution by avoiding premature dimensionality reduction. It will be shown that EMD yields a more accurate representation of the differences between polymer ensembles and is more consistent with chemical intuition.

Methods

EMD is a well-constructed metric to calculate the similarity of ensembles or distributions. The original application of EMD is an optimization problem where the goal is to minimize the amount of work to move earth from one pile to another. Thus, it can be formulated and solved as a transportation problem. EMD has been successfully applied in multiple fields, including the similarity of inorganic solids,[59, 60] cell-cell similarity inference,[61] and geometric dataset distances.[62] Analogously, the problem here is transforming one polymer ensemble to another polymer ensemble with the minimum amount of work done, which is interpreted as a calculation of dissimilarity. In order to use EMD to calculate the pairwise similarity of polymer ensembles, it is necessary to determine the dissimilarity or distance between each pair of individual polymer chains. There are numerous methods for calculating the dissimilarity or distance between two individual polymer chain, such as sequence alignment algorithms[63, 64] and graph edit distance (GED). Among all these methods, GED stands out as a robust and generalized approach for calculating the pairwise dissimilarity or distance between each pair of individual polymer chains with varying chemistries, lengths, and topologies.

Graph Edit Distance

In this work, each polymer chain in the molecular ensemble is first transformed into a coarse-grained graph representation, where the nodes are molecular fragments, such as repeat units, end groups and linkers, and the edges are the connections between these molecular fragments, as shown in Figure 1a. If the end group is *H, this end group *H is implicit. Canonicalization rules[51] are utilized to ensure the generalization of selecting repeat units as nodes for building the coarse-grained polymer graph representations. GED is then used to calculate the pairwise dissimilarity or distance between each pair of individual polymer chains with one chain selected from each of the two ensembles being compared. GED, first reported by Sanfeliu and Fu[65] in 1983, is a measure of similarity between two graphs g_1 and g_2. The idea behind GED is to find the minimal set of transformations that can transform graph g_1 into graph g_2 by means of edit operations on graph g_1. The set of elementary graph edit operators typically includes insertion, deletion, and substitution of both nodes and edges, as shown in Figure 1. The formula for calculating GED is

$$GED(g_1, g_2) = \min_{(e_1,\ldots,e_k) \in \mathcal{P}(g_1,g_2)} \sum_{i=1}^{k} c(e_i) \qquad (1)$$

where $\mathcal{P}(g_1, g_2)$ denotes the set of edit paths transforming g_1 into graph g_2 and $c(e_i)$ is the cost of each graph edit operation e_i. As shown in Figure 1b, for insertion and deletion of nodes/edges, they add a constant cost to the distance, assumed here to be 1. For node substitution (Figure 1c),

the cost is either the same constant cost as the insertion and deletion costs if the node uses one-hot encoding or equal to the constant cost multiplied by the Tanimoto dissimilarity[43] of the pair of nodes being substituted if the node uses Morgan fingerprint encoding.[43, 46] $GED(g_1, g_2)$ is zero when g_1 and g_2 are identical. GED is symmetric; the minimal cost of transforming graph g_1 into graph g_2 is the same as the minimal cost of transforming graph g_2 into graph g_1.

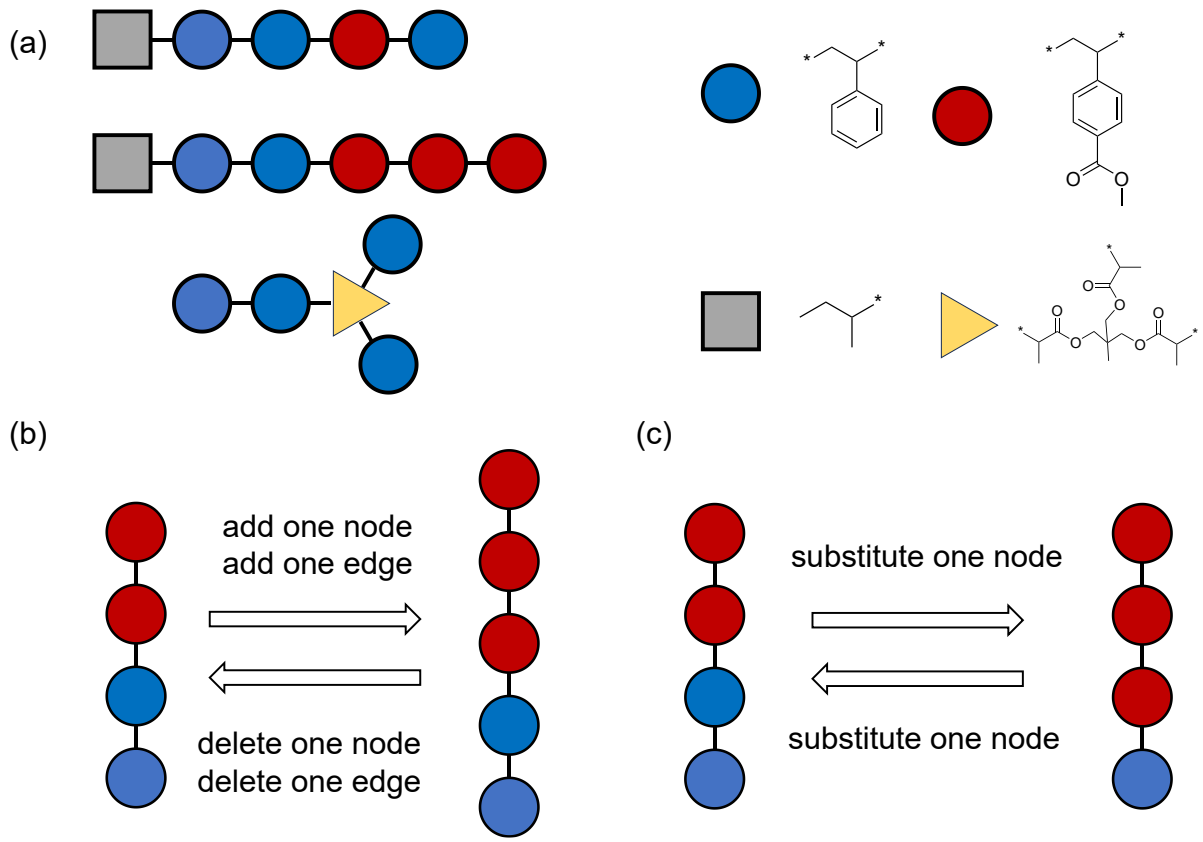


Figure 1: (a) Three examples of the coarse-grained graph representations of polymer chains, where the nodes are molecular fragments, such as repeat units, end groups and linkers, and the edges are the connections between those molecular fragments. Graph edit distance (GED) operations include (b) adding nodes/edges or deleting nodes/edges and (c) substituting nodes.

To map $GED(g_1, g_2)$ onto a distance $d(g_1, g_2)$ with the range of $[0, 1)$, first the GED is normalized to be $\frac{GED(g_1,g_2)}{(N_1+N_2)/2}$ and then an exponential decay function is used:[56, 66]

$$d(g_1, g_2) = 1 - \exp\left(-\frac{\alpha \cdot GED(g_1, g_2)}{(N_1 + N_2)/2}\right) \quad (2)$$

where N_i denotes the number of nodes of g_i and α is a tunable parameter with the default value being 1. $d(g_1, g_2)$ is 0 when g_1 and g_2 are identical. $d(g_1, g_2)$ is also symmetric, so $d(g_1, g_2) = d(g_2, g_1)$. The reason for converting the absolute $GED(g_1, g_2)$ to a normalized GED stems from chemical intuition. In the case of explicitly comparing molecular mass distributions, $GED(g_1, g_2)$ is proportional to $|N_1 - N_2|$, so converting to a normalized GED is equivalent to looking at the

percent difference instead of the absolute difference. According to Van Krevelen's book,[67] many properties of polymers, for example, glass-transition temperature[68] and tensile strength,[69] can be described by $X = X_\infty - \frac{A}{M_\mathrm{n}}$, where X is the property considered, X_∞ is the property value at infinite molecular mass, A is a constant, and M_n is the number average molecular mass. This equation suggests that the difference in property values will be larger for the same GED between shorter chains than larger chains. The normalized GED most accurately captures this intuitive trend. The exponential decay on the normalized GED[66] is then used because it constrains the similarity to be between 0 and 1, consistent with prior work about molecular similarity calculations.[40, 70, 71] This provides a sense of scale and allows for two similarities to be compared more easily while maintaining the expected trends from chemical intuition. Although Equation 2 is proposed here as an advantageous metric for $d_{i,j}$, the choice of $d_{i,j}$ can be modified based on users' needs and specific requirements for their scientific problems, using for example absolute graph edit distance[65] or sequence alignment algorithms,[63, 64] without modifying the subsequent EMD calculation.

Earth Mover's Distance

As shown in Figure 2, one polymer ensemble is defined as $P = \{(p_1, w_{p1}), (p_2, w_{p2}), \ldots, (p_i, w_{p_i}), \ldots (p_m, w_{p_m})\}$ having m types of polymer chains, where p_i represents a type of polymer chain and w_{p_i} is its corresponding weight, which can be the mole fraction of this polymer chain in the polymer ensemble. Similarly, the second polymer ensemble $Q = \{(q_1, w_{q1}), (q_2, w_{q2}), \ldots, (q_j, w_{q_j}), \ldots (q_n, w_{q_n})\}$ has n types of polymer chains. The sums of the weights for P and Q are both normalized and equal to one $\left(\text{i.e.}, \sum_{i=1}^{m} w_{p_i} = \sum_{j=1}^{n} w_{q_j} = 1\right)$ and individual weights must be positive.

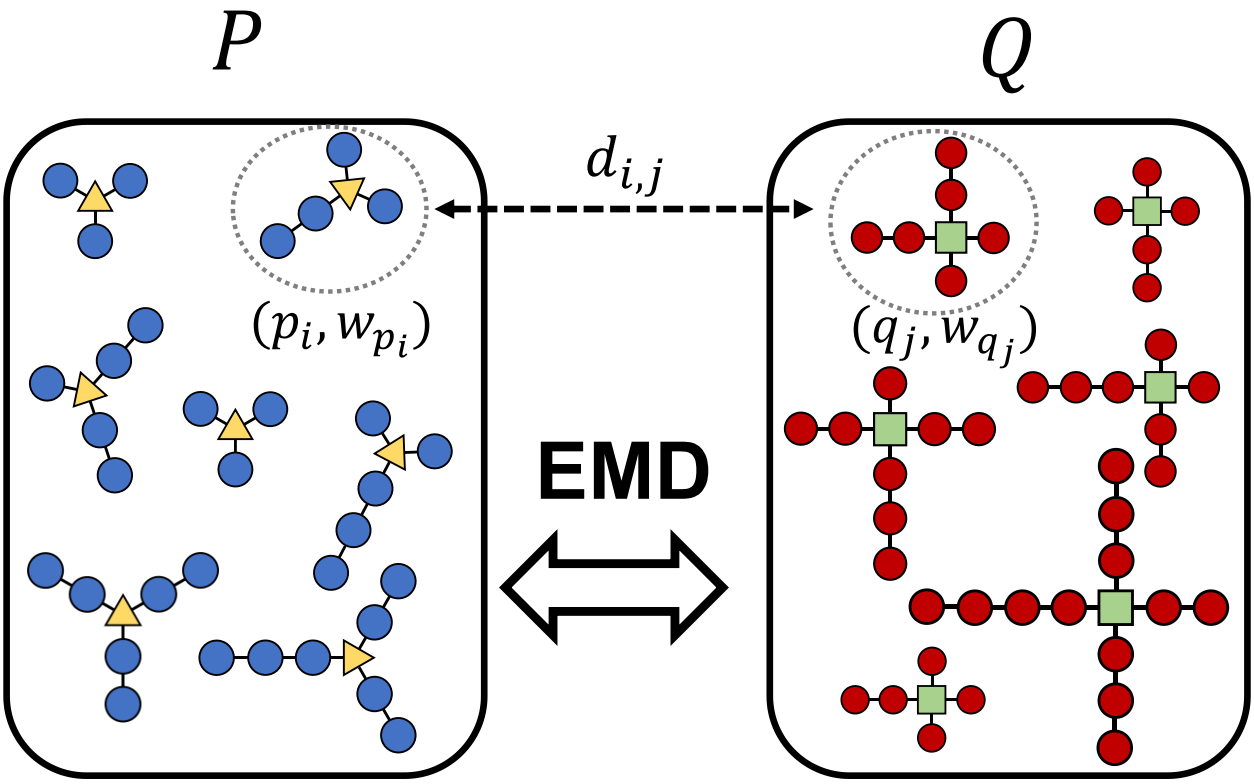


Figure 2: Schematic of earth mover's distance (EMD) for calculating the similarity score between two polymer ensembles, where $P = \{(p_1, w_{p1}), (p_2, w_{p2}), \ldots, (p_i, w_{p_i}), \ldots (p_m, w_{p_m})\}$ has m types of polymer chains and $Q = \{(q_1, w_{q1}), (q_2, w_{q2}), \ldots, (q_j, w_{q_j}), \ldots (q_n, w_{q_n})\}$ has n types of polymer chains. The pairwise dissimilarity or distance $d_{i,j}$ between every individual polymer chains p_i and q_j is calculated through graph edit distance (GED). EMD utilizes w_{p_i}, w_{q_j}, and $d_{i,j}$ to calculate the pairwise ensemble similarity between P and Q.

The pairwise dissimilarity or distance $d_{i,j}$ between every pair of individual polymer chains p_i and q_j is calculated through Equation 2.[40] After w_{p_i}, w_{q_j}, and $d_{i,j}$ are obtained for all the entities in the ensembles, the earth mover's distance (EMD) is determined using Equation 3a along with the constraints as specified in Equations 3b-e.

$$EMD(P,Q) = \frac{\min_F \sum_{i=1}^{m}\sum_{j=1}^{n}(d_{i,j} \cdot f_{i,j})}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}} = \min_F \sum_{i=1}^{m}\sum_{j=1}^{n}(d_{i,j} \cdot f_{i,j}) \quad (3a)$$

$$\text{Subject to } f_{i,j} \geq 0, \forall\, 1 \leq i \leq m, 1 \leq j \leq n \quad (3b)$$

$$\sum_{j=1}^{n} f_{i,j} = w_{p_i}, \forall\, 1 \leq i \leq m \quad (3c)$$

$$\sum_{i=1}^{m} f_{i,j} = w_{q_j}, \forall\, 1 \leq j \leq n \quad (3d)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = \sum_{i=1}^{m} w_{p_i} = \sum_{j=1}^{n} w_{q_j} = 1 \qquad (3e)$$

$f_{i,j}$ is the flow or amount of weight at p_i which is transported to q_j, and $F = [f_{i,j}]$ denotes all the flows between P and Q. Here, $d_{i,j} \cdot f_{i,j}$ is the cost for each individual flow. These equations are coded into Pyomo,[72, 73] an open-source optimization modeling language, and solved with Computational Infrastructure for Operations Research (COIN-OR) Branch-and-Cut (cbc) solver,[74] an open-source mixed integer linear programming solver. Since the sum of the weights is normalized, the minimum overall cost equals the minimum overall distance. All $d_{i,j}$ are bounded between 0 and 1, so EMD is also bounded between 0 and 1, representing the minimum overall distance to convert one polymer ensemble P to another polymer ensemble Q, or equivalently the dissimilarity score. Finally, the pairwise similarity score $S(P, Q)$, for the ensemble pair P and Q, may then be defined as

$$S(P, Q) = 1 - EMD(P, Q) \qquad (4)$$

The value of $S(P, Q)$ is also between 0 and 1. The larger the $S(P, Q)$, the more similar between P and Q. The self ensemble similarity score is 1.

Results and Discussions

Example 1: Two Component Polymer Ensemble

EMD provides greater resolution of chemical differences between polymer ensembles than simple sums or averages of the embedding for each polymer chain. The reason is that simply averaging or summing[44] prematurely reduces the dimensionality of the system, eliminating differences among ensembles. In this example, a comparison of two ensembles, each composed of an equal mixture of two equal-length polymer chains, is employed to demonstrate the features of the EMD method for computing pairwise similarity scores, as shown in Figure 3a,b. The two ensembles are denoted $P = \{(p_1, w_{p_1} = 0.5), (p_2, w_{p_2} = 0.5)\}$ where p_1 and p_2 are alternating polymers, and $Q = \{(q_1, w_{q_1} = 0.5), (q_2, w_{q_2} = 0.5)\}$ where q_1 and q_2 are blocky polymers. To compare the two ensembles, each polymer chain is first represented as a vector, also known as an embedding. Specifically, a one-hot encoding method is used, where the blue repeat unit R_0 is represented by $\begin{bmatrix}1\\0\end{bmatrix}$, the red repeat unit R_1 is represented by $\begin{bmatrix}0\\1\end{bmatrix}$. This embedding is easily extended to an arbitrary number of monomers by increasing the dimensionality of the vector. Then the embedding vectors for these polymer chains are $v_{p_1} = \begin{bmatrix}1&0&1&0&1&0&1&0&1\\0&1&0&1&0&1&0&1&0\end{bmatrix}$, $v_{p_2} = \begin{bmatrix}0&1&0&1&0&1&0&1&0\\1&0&1&0&1&0&1&0&1\end{bmatrix}$; $v_{q_1} = \begin{bmatrix}1&1&1&0&0&0&1&1&1\\0&0&0&1&1&1&0&0&0\end{bmatrix}$, $v_{q_2} = \begin{bmatrix}0&0&0&1&1&1&0&0&0\\1&1&1&0&0&0&1&1&1\end{bmatrix}$.

To illustrate the benefits of EMD, the commonly used average method is computed as a benchmark. In this case, a single embedding vector for an ensemble is generated by taking a weighted average of the embedding vectors[35] resulting in a single embedding for the entire ensemble rather than explicitly using the embedding of each constituent. Using this method, the embedding vector V_P for polymer ensemble P is $V_P = \left(v_{p_1} \cdot w_{p_1}\right) + \left(v_{p_2} \cdot w_{p_2}\right) = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$, and the embedding vector V_Q for polymer ensemble Q is $V_Q = \left(v_{q_1} \cdot w_{q_1}\right) + \left(v_{q_2} \cdot w_{q_2}\right) = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}$. Since $V_P \equiv V_Q$, the similarity score between P and Q is one. P and Q are treated identically, regardless of which similarity metric is used in the average method. However, as observed in Figure 3a,b, the two polymer ensembles P and Q are noticeably distinct in terms of their sequences. Thus, the average method fails to capture the dissimilarity between polymer ensembles P and Q.

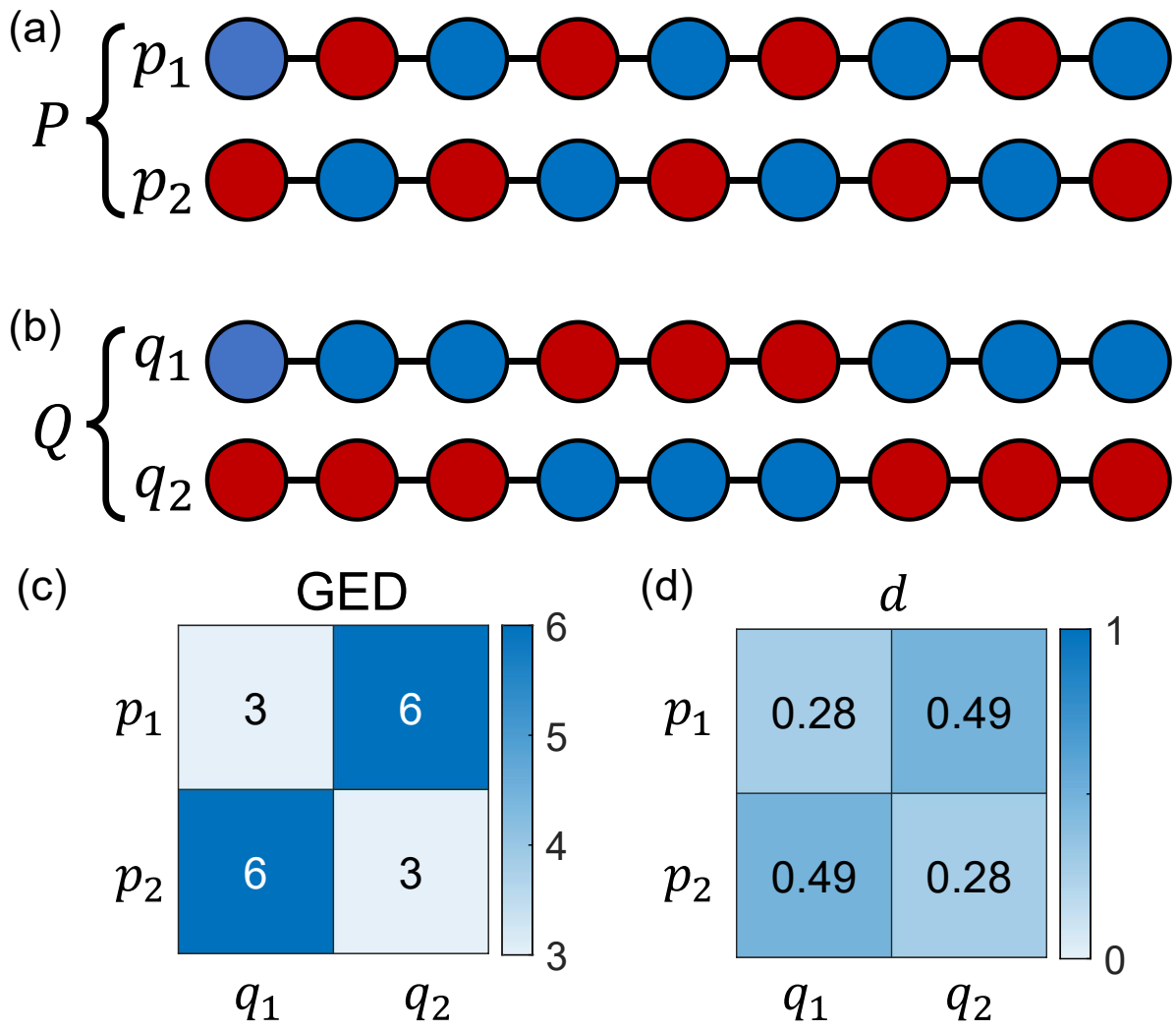


Figure 3: (a) Polymer ensemble $P = \{(p_1, w_{p_1} = 0.5), (p_2, w_{p_2} = 0.5)\}$ and (b) Polymer ensemble $Q = \{(q_1, w_{q_1} = 0.5), (q_2, w_{q_2} = 0.5)\}$. Blue circles represent the repeat unit M_0, and red circles represent the repeat unit M_1. (c) Graph edit distance (GED) matrix and (d) distance (d) matrix between the polymer ensembles P and Q.

Next, the EMD method is used to calculate a quantitative ensemble similarity score. GED matrix and distance matrix ($[d_{i,j}]$) between the polymer ensembles P and Q are calculated, with results shown in Figure 3c,d. Additional details can be found in the Methods section. Using the necessary information about w_{p_i}, w_{q_j}, and $d_{i,j}$, the optimization problem is solved, yielding $EMD(P, Q) = 0.28$ and $S(P, Q) = 0.72$. The EMD method captures the difference and provides a quantitative pairwise similarity score that accurately reflects the similarity between the two ensembles.

Example 2: First-order Markov Copolymer Ensemble

EMD is applied to a more complex system, first-order Markov copolymers,[38, 75] where the primary structure of the copolymer can be treated as a first-order Markov process. The two repeat units are the same as in Example 1, where the blue repeat unit is R_0 and the red repeat unit is R_1. Fixed-length linear polymers are generated using t_{ij}, the transition or conditional probability that a repeat unit of type i is followed by a repeat unit of type j in a linear sequence, with $i, j = R_0, R_1$. As shown in Figure 4a, the transition probability t_{10}, for example, is the probability of forming a $\sim R_1R_0$ from $\sim R_1$ in a copolymer chain where "$\sim$" represents a piece of polymer chain. The t_{ij}s can be used to construct a transition matrix T, which is given by

$$T = \begin{bmatrix} t_{00} & t_{10} \\ t_{01} & t_{11} \end{bmatrix} \tag{5}$$

Due to the rules of probability, the sum of the transition probabilities for the addition to $\sim R_0$ and $\sim R_1$, are each separately equal to 1. Therefore,

$$t_{00} + t_{01} = 1 \tag{6a}$$

$$t_{10} + t_{11} = 1 \tag{6b}$$

The first-order Markov process can thus be specified by two independent parameters: (i) the average fraction of R_1 in a copolymer chain, f_{R_1}

$$f_{R_1} = f_{R_1} \cdot t_{11} + \left(1 - f_{R_1}\right) \cdot t_{01} \tag{7}$$

and (ii) the nontrivial eigenvalue of the transition matrix T, λ, which defines the correlations in the linear repeat unit sequence.

$$\lambda = t_{00} + t_{11} - 1 \tag{8}$$

Here copolymer ensembles are studied with the setting $f_{R_0} = f_{R_1} = 0.5$, where the average fractions of R_0 and R_1 are the same in a copolymer chain. Consequently, t_{00}, t_{01}, t_{10} and t_{11} are solely determined by λ.

$$t_{00} = t_{11} = \frac{1 + \lambda}{2} \tag{9a}$$

$$t_{01} = t_{10} = \frac{1 - \lambda}{2} \tag{9b}$$

By modifying the value of λ, different copolymer ensembles P_λ can be generated. In this example, a series of polymer ensembles are generated for $\lambda = -1.0$ to 1.0 in increments of 0.5. All chains have a fixed length $L = 10$. Representative copolymer sequences are illustrated in Figure 4b. At $\lambda = 0$, there are no correlations (memory) along the chain; this is an ideal random copolymer. The case of $\lambda > 0$ corresponds to positive correlations between identical repeat units, meaning that the

last monomer of the polymer chains has a tendency to connect the same type of repeat units (blocky polymers). The case of $\lambda < 0$ corresponds to negative correlations between identical repeat units, meaning chains tend to alternate between R_0 and R_1 repeat units. For the case of $\lambda = -1.0$, the polymer ensemble $P_{\lambda=-1.0}$ has two sequences with equal probability, $R_0R_1R_0R_1R_0R_1R_0R_1R_0R_1$ and $R_1R_0R_1R_0R_1R_0R_1R_0R_1R_0$. Even though these two polymer chains' pairwise GED is zero due to symmetry, they are both kept because they are generated in different Markov processes. For the case of $\lambda = 1.0$, the polymer ensemble $P_{\lambda=1.0}$ has only two sequences with equal probability $R_1R_1R_1R_1R_1R_1R_1R_1R_1R_1$ and $R_0R_0R_0R_0R_0R_0R_0R_0R_0R_0$. Apart from these two special cases, polymer ensembles generated by other λ values are sampled by following the above first-order Markov process for 3×10^7 polymer chains (Discussions of sampling size convergence are included in the Supporting Information).

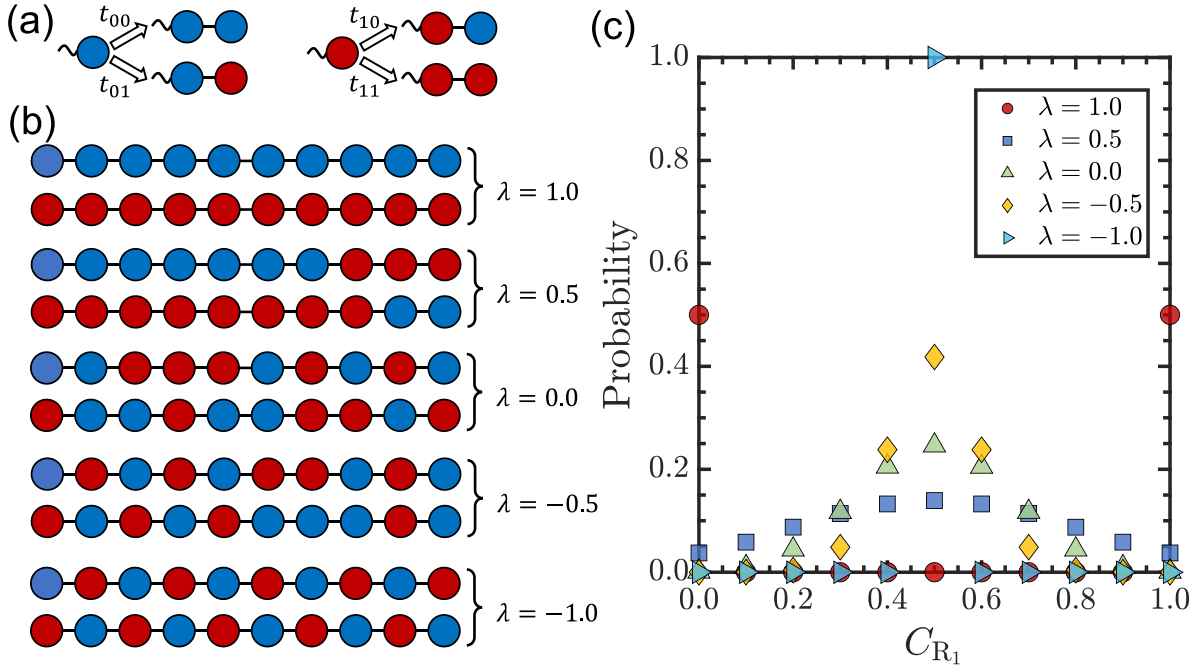


Figure 4: (a) First-order Markov copolymer model. (b) Representative copolymer sequences (Blue circle representing repeat unit R_0 and red circle representing repeat unit R_1) with average mole fraction $f_{R_1} = 0.5$, and different repeat unit sequences (λ value): $\lambda = 1.0$ generates chains either pure R_0 or R_1 ($t_{00} = t_{11} = 1.0$); $\lambda = 0.5$ creates a chain with moderate positive correlations in identical monomers ($t_{00} = t_{11} = 0.75$); $\lambda = 0.0$ is an ideally random chain ($t_{00} = t_{11} = 0.5$); $\lambda = -0.5$ creates a chain with moderate negative correlations in identical monomers ($t_{00} = t_{11} = 0.25$); and $\lambda = -1.0$ is a perfectly alternating chain ($t_{00} = t_{11} = 0$). (c) The distributions of the mole fraction of R_1 in the polymer chain (C_{R_1}) for a series of first-order Markov copolymer ensembles generated from different λ values.

The distributions of the mole fraction of R_1 in the polymer chain (C_{R_1}) for a series of ensembles as a function of the composition variation on a chain basis in the ensemble are shown in Figure 4c. These copolymer ensembles generated by different λ values have distinct chain composition distributions. The ensemble generated at $\lambda = -1.0$ is perfectly alternating copolymer with only two unique sequences where the values of C_{R_1} of these two chains are both 0.5. Therefore, polymer ensemble $P_{\lambda=-1.0}$ has probability 1.0 at $C_{\mathrm{R}_1} = 0.5$. As λ increases, the corresponding probability for $C_{\mathrm{R}_1} = 0.5$ gradually decreases, and the corresponding probability at the two ends ($C_{\mathrm{R}_1} = 0.0$ and $C_{\mathrm{R}_1} = 1.0$) gradually increases. When $\lambda = 1.0$, polymer ensemble $P_{\lambda=1.0}$ has a distribution with the probability of $C_{\mathrm{M}_1} = 0.0$ being 0.5 and the probability of $C_{\mathrm{R}_1} = 1.0$ being 0.5. The distributions are symmetric since the same average fraction $f_{\mathrm{R}_0} = f_{\mathrm{R}_1} = 0.5$.

Similar to Example 1, the one-hot encoding method is utilized to embed the copolymer sequence as a vector, where the blue repeat unit is $\begin{bmatrix}1\\0\end{bmatrix}$ and the red repeat unit is $\begin{bmatrix}0\\1\end{bmatrix}$. For comparison, the average method is employed to compute the mean of all embedding vectors of copolymer chains within each ensemble, the obtained average global vector V for each ensemble is identical: $\begin{bmatrix}0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\\0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\end{bmatrix}$. Consequently, the average method eliminates the ensemble features and fails to accurately characterize the pairwise similarity among these copolymer ensembles, akin to the issue in Example 1.

EMD is then employed to compute the pairwise similarity between a pair of ensembles of first-order Markov copolymers. As illustrated in Figure 5a, using one-hot encoding and setting substitution cost as 1, the value of $S(\lambda_1, \lambda_2)$ between a copolymer ensemble P_{λ_1} and a copolymer ensemble P_{λ_2} from the EMD method is between 0.61 and 1. The lowest similarity score is $S(1.0, -1.0) = S(-1.0, 1.0) = 0.61$, and the highest similarity score is self-similarity $S(\lambda, \lambda) = 1$. The similarity between copolymer ensemble P_{λ_1} and copolymer ensemble P_{λ_2} decreases as the gap $|\lambda_1 - \lambda_2|$ increases, which is consistent with chemical intuition. If repeat units R_0 and R_1represent specific chemical structures, such as "*CC(*)c1ccccc1" (the repeat unit of polystyrene represented in BigSMILES, {[][$]CC(c1ccccc1)[$][]})) and "*CC(*)c1ccc(C(=O)OC)cc1" (the repeat unit of poly(methyl 4-vinylbenzoate), {[][$]CC(c1ccc(C(=O)OC)cc1)[$][]}), then the nodes and edges can be embedded with Morgan fingerprints[46] and the Tanimoto dissimilarity between R_0 and R_1 can be used as the substitution cost.[43] The corresponding pairwise similarity results for this case are shown in Figure 5b. The basic trends match that of one hot encoding, but the range of similarities values is smaller since the two different repeat units are more similar to one another than if one hot encoding is used. For both one-hot encoding and Morgan fingerprint encoding, EMD results follow chemical intuition and provide a quantitative result. Therefore, unlike the average method, EMD method is able to distinguish these first-order Markov copolymer ensembles from one another, thus demonstrating the utility of the method.

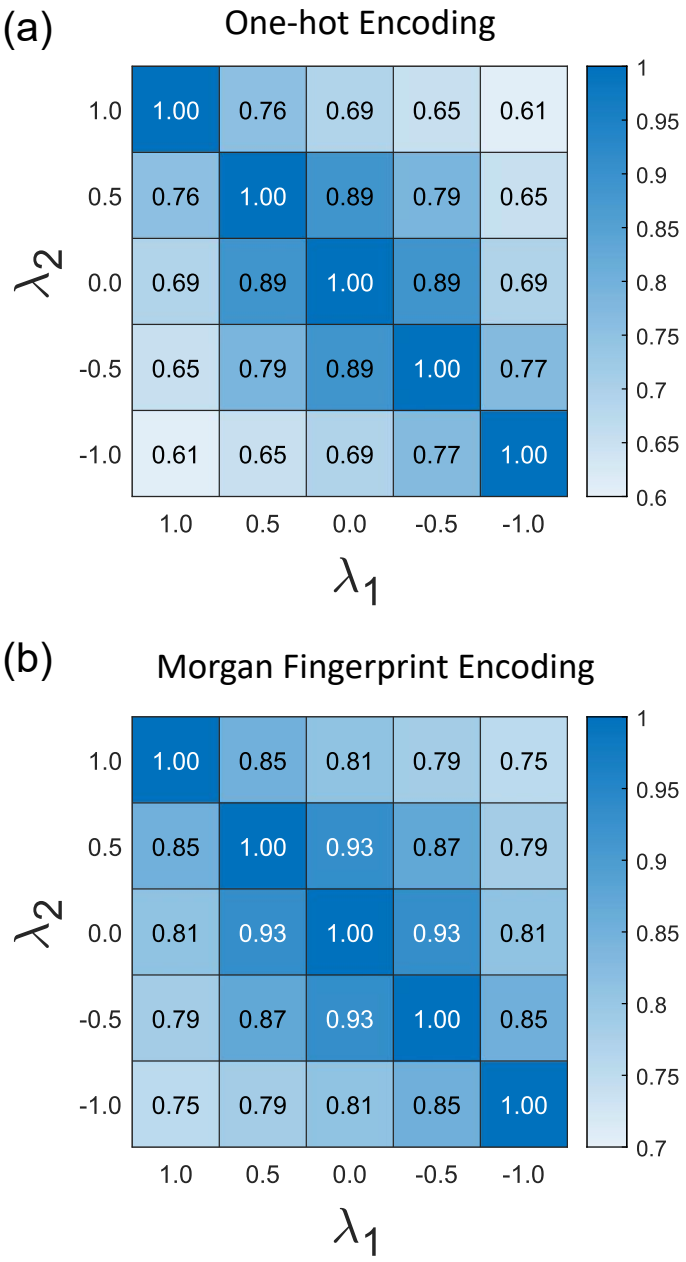


Figure 5: The pairwise similarity score $S(\lambda_1, \lambda_2)$ for first-order Markov copolymer ensembles generated by different λ values under one-hot encoding (a) and Morgan fingerprint encoding (b).

Apart from different chemistry, Example 2 can also be used to describe the pairwise similarity of polymer ensembles with different tacticity, such as atactic, and syndiotactic polypropylene (PP), where R_0 and R_1 represent "*C[C@H](C)*" and "*C[C@@H](C)*". Atactic PP ensemble can be treated as $P_{\lambda=0}$. Syndiotactic PP can be treated as $P_{\lambda=-1.0}$. If using the one-hot encoding, the pairwise similarity result can be found in Figure 5a, $S(\text{atactic PP}, \text{syndiotactic PP}) = 0.69$. Furthermore, if Morgan fingerprint encoding is used to include the detailed chemical structure of

stereochemical centers, then the pairwise similarity result is $S(\text{atactic PP, syndiotactic PP}) = 0.90$.

Example 3: Star Polymer Ensemble

In the two examples discussed above, all the polymer chains are linear and have a constant chain length. However, in the real-world, polymer ensembles are often more complex, featuring varying lengths, topologies and chemistries, such as the eight types of star polymer ensembles in Figure 6a. Take SP-1, a three-arm star polymer ensemble as an example. The probabilities of arm-lengths are assigned to be one, two, and three are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$, respectively. This simple case was chosen as it allows for illustration of the method, while reducing the computational burden of the costly GED method. With these parameters, this three-arm star polymer ensemble SP-1 has ten configurations with the corresponding analytical mole fractions as shown in Figure 6b. The configurations and corresponding mole fractions of the other seven polymer ensembles are given in the Supporting Information. Morgan fingerprints are used for the embedding of nodes and edges while Tanimoto dissimilarity is used as the substitution cost.

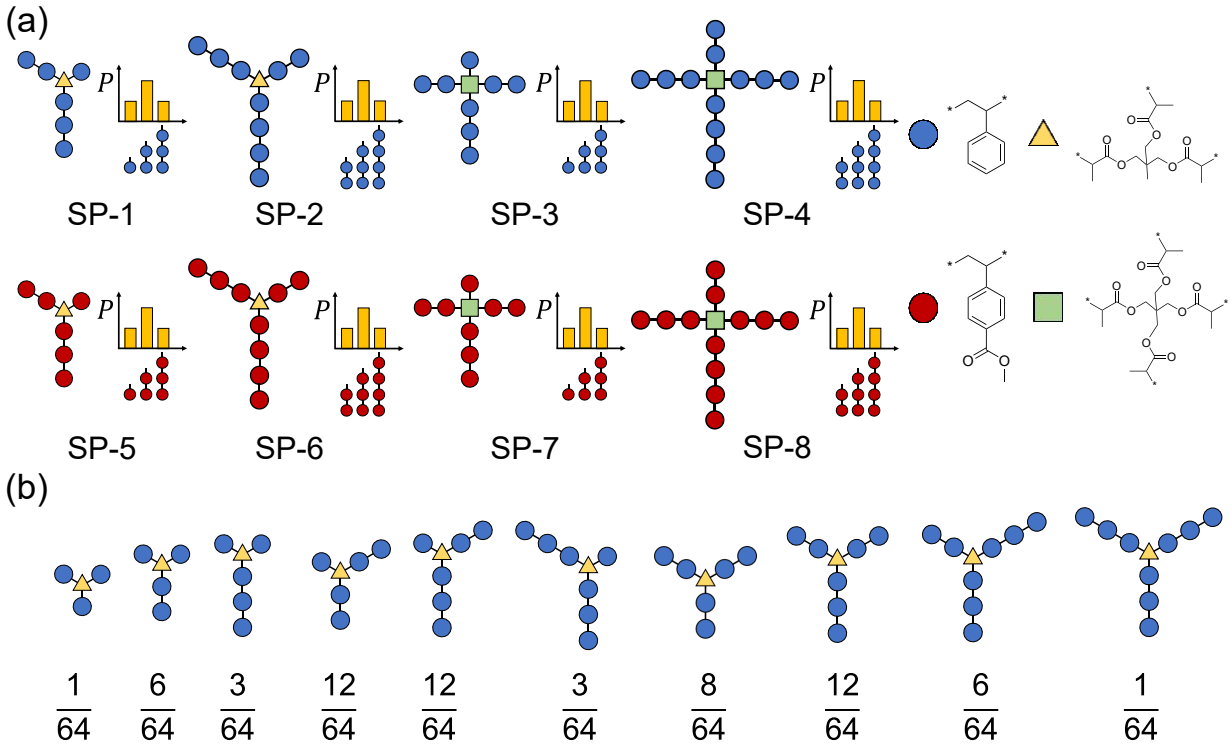
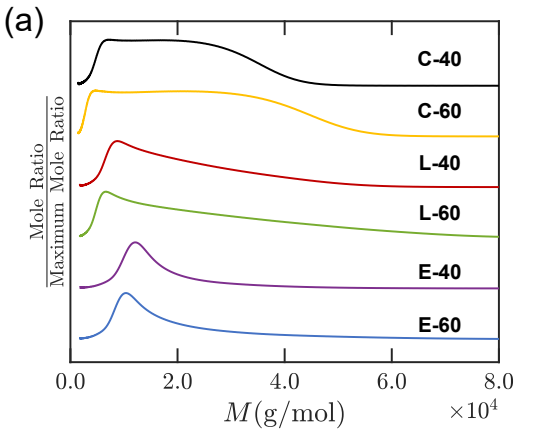


Figure 6: (a) Eight types of star polymer ensembles generated with varying arm lengths, arm numbers, and repeat unit compositions. The blue circle and the red circle represent two types of repeat units. The yellow triangle is a star core group with three connection spots. The green square is a star core group with four connection spots. (b) Ten polymer chains' configurations and the corresponding mole fractions about the three-arm polymer ensemble SP-1 in (a).

The pairwise dissimilarity $d_{i,j}$ between two individual polymer chains is calculated by GED, and then the $d_{i,j}$ and each polymer chain's weight are used to calculate the EMD and similarity score. The similarity results are shown in Figure 7. $S(\text{SP-1}, \text{SP-2})$ reveals the effect of arm length on the similarity score; $S(\text{SP-1}, \text{SP-3})$ illustrates the effect of arm number on similarity score; $S(\text{SP-1}, \text{SP-5})$ demonstrates the impact of repeat units on similarity. $S(\text{SP-1}, \text{SP-2}) > S(\text{SP-1}, \text{SP-4})$ and $S(\text{SP-1}, \text{SP-3}) > S(\text{SP-1}, \text{SP-4})$ because compared with SP-1, SP-4 changes both arm length and arm number, which is consistent with the chemical intuition. The pair SP-1 and SP-8 and the pair SP-4 and SP-5 have the smallest pairwise similarity scores, meaning these pairs are the most different pairs. This is consistent with chemical intuition because the arm length, arm number, and repeat units of SP-1 and SP-8 (or SP-4 and SP-5) are all different.

Figure 7: The pairwise similarity score for eight star polymer ensembles.

Example 4: Polymer Ensembles Represented by Experimental Molecular Mass Distributions
Polymer molecular mass distributions (MMDs)[38, 39, 76-78] exemplify the fact that synthetic polymers are ensembles rather than single, well-defined structures. Six experimental polystyrene MMDs from Kottisch et al.[76] are used to illustrate how EMD can be used to characterize the pairwise similarity between two different MMDs, as displayed in Figure 8a. Kottisch et al. controlled the breadth and shape of polystyrene MMDs by varying initiator (sec-butyllithium) addition rates (constant, linearly ramped, and exponentially ramped) and addition time.[76] For example, C-40 refers to a constant rate of initiator addition with an addition time of 40 min. The parameters of these six MMDs are shown in Figure 8b. Among these MMDs, C-40, L-40 and E-60 have similar number average molar mass (M_{n}) and dispersity ($Đ$) but different shapes illustrated by the asymmetry factor (A_s), skewness (α_3), and kurtosis (α_4).



(b)

	M_{n} (kg/mol)	$Đ$	A_S	α_3	α_4
C-40	14.3	1.40	3.58	0.94	3.27
C-60	14.4	1.66			
L-40	14.7	1.43	1.57	1.54	5.59
L-60	15.5	1.71			
E-40	13.9	1.26			
E-60	14.1	1.43	0.34	2.52	11.93

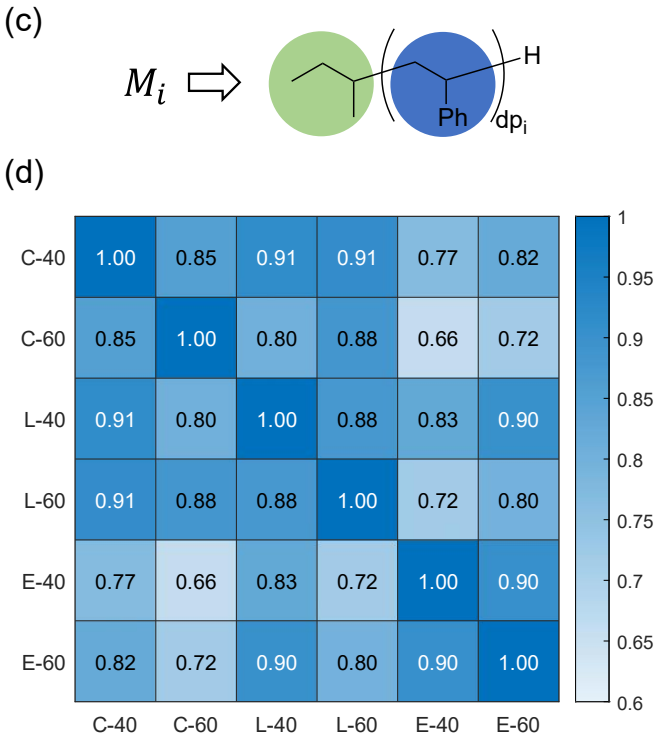


Figure 8: (a) Six polystyrene MMDs from the work of Kottish et al.[76] generated by varying initiator (sec-butyllithium) addition rates (constant (C), linearly ramped (L), and exponentially ramped (E)) and addition time (40 min and 60 min). C-40, for example, refers to a constant rate of initiator addition with an addition time of 40 min. (b) MMDs' parameters include number average molar mass (M_n), dispersity ($Đ$), asymmetry factor (A_s), skewness (α_3), kurtosis (α_4) from Kottish et al.[76] (c) The polymer graph representation is built for every single value of measured molecular mass (M_i) in the MMDs. (d) Pairwise similarity scores for six polystyrene MMDs via EMD.

In this example, the EMD metric is utilized to quantitatively calculate the similarity between two arbitrary distributions. As in the previous examples, each unique polymer molecule is utilized to generate a polymer graph representation. In this case, to make the method generalizable to multi-

parameter distributions (for example molar mass and monomer composition), the polymer graph representation is built for every single value of measured molecular mass (M_i) in the MMDs, as shown in Figure 8c. To do so, the fact that the polymers are linear is used. Since one of the end groups (*H) is implicit, the number of nodes can be computed as

$$N_i = dp_i + 1 \tag{10}$$

where dp_i is the degree of polymerization computed from M_i. The value of 1 corresponds to the other end group, sec-butyl group. Since all of the MMDs in this study are about linear polystyrene, as shown in Figure 8c, adding one repeat unit node means also adding one edge in the graph representation. GED between M_i and M_j can be computed as two times the difference of the degree of polymerization. For nonlinear polymers, this equation may also be a suitable approximation approach if only the MMDs are known.

$$GED\left(M_i, M_j\right) = 2 \times |dp_i - dp_j| \tag{11}$$

Plugging this into Equation 2 yields

$$d_{i,j}\left(M_i, M_j\right) = 1 - \exp\left(-\frac{2 \times |dp_i - dp_j|}{\frac{(dp_i + 1) + \left(dp_j + 1\right)}{2}}\right) \tag{12}$$

For the weight setting of w_i and w_j, the normalized mole fractions are used. EMD takes w_i, w_j and $d_{i,j}$ to calculate the pairwise similarity score values among MMDs. Since the weights are normalized, the EMD should theoretically converge (ignoring experimental error) if the frequency of sampling M_i is increased beyond the sampling of about 1 sec (discussions of MMD sampling frequency convergence are included in the Supporting Information). The pairwise similarity results are shown in Figure 8d. $S(\text{C-40}, \text{C-60})$ is lower than $S(\text{E-40}, \text{E-60})$, which is consistent with chemical intuition that E-40 and E-60 are closer in dispersity. Among polystyrene C-40, L-40, and E-60 MMDs, C-40 and L-40 are the most similar pair, while polystyrene MMDs C-40 and E-60 are the most different. These similarity scores are consistent with the relative rankings of skewness and kurtosis of C-40, L-40, and E-60.

Areas for Future Development

GED is a robust, generalized and powerful tool for calculating the pairwise distance between two individual polymer chains with arbitrary composition, chain length and topology. However, the calculation of exact GED is non-deterministic polynomial-time hard (NP-hard). Even the state-of-the-art algorithms cannot reliably compute the exact GED within reasonable computing time between graphs with more than 16 nodes.[66, 79] If each polymer ensemble has thousands of unique polymer chains, the calculation of EMD between these polymer ensembles requires millions of exact GED calculations. This NP-hard feature of an exact GED calculation renders it especially costly for large graph representations and limits the proposed method to relatively small polymers unless assumptions such as those in Example 4 are used.

The method for selecting repeat units as nodes for building the coarse-grained polymer graph representations needs further improvement. For example, a polyethylene chain with 100 polymerization degrees $(CC)_{100}$ and a hydrogenated 1,4-polybutadiene chain with 50 polymerization degrees, $(CCCC)_{50}$ are treated differently since their monomers are different under the canonicalization priority rules[51] despite the fact that they both represent the same polymer chain with the total length of 200 carbons. Fundamentally, this is an artifact of coarse-graining the polymer chain into monomer units. Comprehensive coarse-graining techniques which satisfy both the repeat unit level and the whole polymer chain level will be developed in the future.

In the study of experimental polymer ensembles, various approximation methods are employed to calculate pairwise similarity scores. These methods aim to construct representative ensembles that reflect the actual states of the polymers, albeit with limited available information (molecular structure representations and MMDs). As experimental characterization techniques for polymers continue to advance in the future, it is anticipated that more comprehensive data will be collected. This additional information will facilitate the creation of more accurate representative ensembles, which in turn will improve the precision of similarity calculations.

Conclusion

Quantifying the pairwise similarity of polymers is a challenging task due to their ensemble nature, consisting of polymer chains with varying quantities, lengths, sequences, chemistry, and topologies. This complexity is greater than that of small molecules with well-defined molecular structures or biomacromolecules with specified sequences. This research leverages the earth mover's distance (EMD) method to quantitatively compute pairwise similarity scores of polymer ensembles. The EMD metric provides enhanced chemical resolution compared to the average method, which may eliminate differences by prematurely reducing system dimensionality. Furthermore, EMD only needs the pair dissimilarity $d_{i,j}$ which can be robustly calculated from graph edit distance, skipping the difficult step of designing comprehensive embedding functions for each polymer chain, especially for those nonlinear polymers with complex topological structures.

Utilizing the EMD metric allows for an accurate and quantitative assessment of chemical similarity between polymer ensembles, and the results have been shown to align with chemical intuition. This method has far-reaching applications in polymer database retrieval systems, including nearest neighbor search queries. It benefits the development of supervised machine learning techniques on polymer properties and provides a robust foundation for future research in polymer design and optimization.[8, 12, 24, 33, 80-85]

Code Availability

Example scripts and information necessary to run and reproduce the examples and the corresponding results contained in this article are posted at the GitHub repository, https://github.com/olsenlabmit/Polymer-Ensemble-Similarity.

Supporting Information

First-order markov copolymer ensembles sampling size convergence for Example 2; weights of the star polymer configurations for Example 3; molecular mass distribution sampling frequency convergence for Example 4.

Acknowledgement

This work was primarily funded by the National Science Foundation Convergence Accelerator award number ITE-2134795. We acknowledge the discussions and suggestions about the earth mover's distance from Yifan Ding from the Department of Computer Science and Engineering, University of Notre Dame. We acknowledge the discussions and suggestions on optimal transport concepts from Xiang Fu from the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. We acknowledge the help of Prof. Brett P. Fors and Jenny Hu from the Department of Chemistry and Chemical Biology, Cornell University, for providing the raw data of polymer molecular mass distributions.

Notes

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

(1) Brandrup, J.; Immergut, E. H.; Grulke, E. A. *Polymer handbook 4th ed*; Wiley New York, 1999.
(2) Provin, A. P.; Regina de Aguiar Dutra, A.; Machado, M. M.; Vieira Cubas, A. L. New materials for clothing: Rethinking possibilities through a sustainability approach - A review. *Journal of Cleaner Production* **2021**, *282*, 124444-124444. DOI: 10.1016/j.jclepro.2020.124444.
(3) Geise, G. M.; Lee, H. S.; Miller, D. J.; Freeman, B. D.; McGrath, J. E.; Paul, D. R. Water Purification by Membranes: The Role of Polymer Science. *Journal of Polymer Science Part B-Polymer Physics* **2010**, *48* (15), 1685-1718. DOI: 10.1002/polb.22037.
(4) Guo, Y. H.; Bae, J.; Fang, Z. W.; Li, P. P.; Zhao, F.; Yu, G. H. Hydrogels and Hydrogel-Derived Materials for Energy and Water Sustainability. *Chem. Rev.* **2020**, *120* (15), 7642-7707. DOI: 10.1021/acs.chemrev.0c00345.
(5) Diao, H.; Yan, F.; Qiu, L.; Lu, J.; Lu, X.; Lin, B.; Li, Q.; Shang, S.; Liu, W.; Liu, J. High Performance Cross-Linked Poly(2-acrylamido-2-methylpropanesulfonic acid)-Based Proton Exchange Membranes for Fuel Cells. *Macromolecules* **2010**, *43* (15), 6398-6405. DOI: 10.1021/ma1010099.

(6) Yadav, R.; Tirumali, M.; Wang, X.; Naebe, M.; Kandasubramanian, B. Polymer composite for antistatic application in aerospace. *Defence Technology* **2020**, *16* (1), 107-118. DOI: 10.1016/j.dt.2019.04.008.
(7) Stenzel, M. H. Glycopolymers for Drug Delivery: Opportunities and Challenges. *Macromolecules* **2022**, *55* (12), 4867-4890. DOI: 10.1021/acs.macromol.2c00557.
(8) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Letters* **2017**, *6* (10), 1078-1082. DOI: 10.1021/acsmacrolett.7b00228.
(9) Walsh, D. J.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M. E.; Mysona, J.; Lin, T.-S.; de Pablo, J. J.; Jensen, K. F.; Audus, D. J.; et al. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. *ACS Central Sci.* **2023**. DOI: 10.1021/acscentsci.3c00011.
(10) Deagen, M. E.; Walsh, D. J.; Audus, D. J.; Kroenlein, K.; de Pablo, J. J.; Aou, K.; Chard, K.; Jensen, K. F.; Olsen, B. D. Networks and interfaces as catalysts for polymer materials innovation. *Cell Rep. Phys. Sci.* **2022**, *3* (11), 101126-101126. DOI: 10.1016/j.xcrp.2022.101126.
(11) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889-2919. DOI: 10.1021/cr200066h.
(12) Ma, R. M.; Liu, Z. Y.; Zhang, Q. W.; Liu, Z. Y.; Luo, T. F. Evaluating Polymer Representations via Quantifying Structure-Property Relationships. *J. Chem Inf. Model.* **2019**, *59* (7), 3110-3119. DOI: 10.1021/acs.jcim.9b00358.
(13) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011/9//, 2011; IEEE: pp 22-29. DOI: 10.1109/EIDWT.2011.13.
(14) Ma, R. M.; Luo, T. F. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem Inf. Model.* **2020**, *60* (10), 4684-4690. DOI: 10.1021/acs.jcim.0c00726.
(15) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; et al. Machine-learning predictions of polymer properties with Polymer Genome. *Journal of Applied Physics* **2020**, *128* (17). DOI: 10.1063/5.0023759.
(16) Kim, S.; Schroeder, C. M.; Jackson, N. E. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. *ACS Polym. Au*. DOI: 10.1021/acspolymersau.3c00003.
(17) McGuinness, D.; Brinson, C.; Chen, W.; Daraio, C.; Rudin, C.; Schadler, L.; Cowan, R.; McCusker, J.; Stouffer, S.; Keshan, N. MaterialsMine: An open-source, user-friendly materials data resource guided by FAIR principles. **2022**. (acccessed 9/21/2023).
(18) Zhao, H.; Li, X.; Zhang, Y.; Schadler, L. S.; Chen, W.; Brinson, L. C. Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design. *APL Materials* **2016**, *4* (5). DOI: 10.1063/1.4943679.
(19) Brinson, L. C.; Deagen, M.; Chen, W.; McCusker, J.; McGuinness, D. L.; Schadler, L. S.; Palmeri, M.; Ghumman, U.; Lin, A.; Hu, B. Polymer Nanocomposite Data: Curation, Frameworks, Access, and Potential for Discovery and Design. *ACS Macro Letters* **2020**, *9* (8), 1086-1094. DOI: 10.1021/acsmacrolett.0c00264.
(20) Gurnani, R.; Kamal, D.; Tran, H.; Sahu, H.; Scharm, K.; Ashraf, U.; Ramprasad, R. polyG2G: A Novel Machine Learning Algorithm Applied to the Generative Design of Polymer Dielectrics. *Chemistry of Materials* **2021**, *33* (17), 7008-7016. DOI: 10.1021/acs.chemmater.1c02061.
(21) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14* (1), 4099. DOI: 10.1038/s41467-023-39868-6.

(22) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Science Advances* **2020**, *6* (43). DOI: 10.1126/sciadv.abc6216.
(23) Patel, R. A.; Borca, C. H.; Webb, M. A. Featurization strategies for polymer sequence or composition design by machine learning. *Molecular Systems Design & Engineering* **2022**, *7* (6), 661-676. DOI: 10.1039/D1ME00160D.
(24) Zhang, Y.; Xu, X. J. Machine learning glass transition temperature of polymers. *Heliyon* **2020**, *6* (10), 7. DOI: 10.1016/j.heliyon.2020.e05055.
(25) Lin, C.; Wang, P.-H.; Hsiao, Y.; Chan, Y.-T.; Engler, A. C.; Pitera, J. W.; Sanders, D. P.; Cheng, J.; Tseng, Y. J. Essential Step Toward Mining Big Polymer Data: PolyName2Structure, Mapping Polymer Names to Structures. *ACS Applied Polymer Materials* **2020**, *2* (8), 3107-3113. DOI: 10.1021/acsapm.0c00273.
(26) Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem Inf. Model.* **2021**, *61* (11), 5395-5413. DOI: 10.1021/acs.jcim.1c01031.
(27) Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2021**, *2* (4), 15. DOI: 10.1016/j.patter.2021.100225.
(28) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: Current status and critical next steps. *Materials Science and Engineering: R: Reports* **2021**, *144*, 100595-100595. DOI: 10.1016/j.mser.2020.100595.
(29) Statt, A.; Kleeblatt, D. C.; Reinhart, W. F. Unsupervised learning of sequence-specific aggregation behavior for a model copolymer. *Soft Matter* **2021**, *17* (33), 7697-7707. DOI: 10.1039/D1SM01012C.
(30) Shi, J.; Quevillon, M. J.; Amorim Valença, P. H.; Whitmer, J. K. Predicting Adhesive Free Energies of Polymer–Surface Interactions with Machine Learning. *ACS Applied Materials & Interfaces* **2022**, *14* (32), 37161-37169. DOI: 10.1021/acsami.2c08891.
(31) Gormley, A. J.; Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials* **2021**, *6* (8), 642-644. DOI: 10.1038/s41578-021-00282-3.
(32) Tamasi, M. J.; Patel, R. A.; Borca, C. H.; Kosuri, S.; Mugnier, H.; Upadhya, R.; Murthy, N. S.; Webb, M. A.; Gormley, A. J. Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids. *Advanced Materials* **2022**, *34* (30), 2201809. DOI: 10.1002/adma.202201809.
(33) Patra, T. K. Data-Driven Methods for Accelerating Polymer Design. *ACS Polym. Au* **2022**, *2* (1), 8-26. DOI: 10.1021/acspolymersau.1c00035.
(34) Arora, A.; Lin, T. S.; Rebello, N. J.; Av-Ron, S. H. M.; Mochigase, H.; Olsen, B. D. Random Forest Predictor for Diblock Copolymer Phase Behavior. *ACS Macro Letters* **2021**, *10* (11), 1339-1345. DOI: 10.1021/acsmacrolett.1c00521.
(35) Wu, Z.; Jayaraman, A. Machine Learning-Enhanced Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) for Analyzing Fibrillar Structures in Polymer Solutions. *Macromolecules* **2022**, *55* (24), 11076-11091. DOI: 10.1021/acs.macromol.2c02165.
(36) Heil, C. M.; Patil, A.; Dhinojwala, A.; Jayaraman, A. Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) with Machine Learning Enhancement to Determine Structure of Nanoparticle Mixtures and Solutions. *ACS Central Sci.* **2022**, *8* (7), 996-1007. DOI: 10.1021/acscentsci.2c00382.
(37) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* **2022**, *8* (29), 9545-9545. DOI: 10.1126/SCIADV.ABN9545.

(38) Odian, G. *Principles of Polymerization*; Wiley, 2004. DOI: 10.1002/047147875X.
(39) Hiemenz, P. C.; Lodge, T. P. *Polymer Chemistry*; CRC Press, 2007. DOI: 10.1201/9781420018271.
(40) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **2015**, *7* (1), 20-20. DOI: 10.1186/s13321-015-0069-3.
(41) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (8), 3186-3204. DOI: 10.1021/jm401411z.
(42) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry* **2004**, *2* (22), 3204-3204. DOI: 10.1039/b409813g.
(43) Mohapatra, S.; An, J.; Gomez-Bombarelli, R. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. *Mach. Learn.-Sci. Technol.* **2022**, *3* (1), 11. DOI: 10.1088/2632-2153/ac545e.
(44) Lim, S.; Lu, Y.; Cho, C. Y.; Sung, I.; Kim, J.; Kim, Y.; Park, S.; Kim, S. A review on compound-protein interaction prediction methods: Data, format, representation and model. *Computational and Structural Biotechnology Journal* **2021**, *19*, 1541-1556. DOI: 10.1016/j.csbj.2021.03.004.
(45) Kuenneth, C.; Schertzer, W.; Ramprasad, R. Copolymer Informatics with Multitask Deep Neural Networks. *Macromolecules* **2021**, *54* (13), 5957-5961. DOI: 10.1021/acs.macromol.1c00728.
(46) Aldeghi, M.; Coley, C. W. A graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **2022**, *13* (35), 10486-10498. DOI: 10.1039/d2sc02839e.
(47) Shi, J.; Albreiki, F.; Colón, Y. J.; Srivastava, S.; Whitmer, J. K. Transfer Learning Facilitates the Prediction of Polymer-Surface Adhesion Strength. *J. Chem. Theory Comput.* **2023**, *19* (14), 4631-4640. DOI: 10.1021/acs.jctc.2c01314.
(48) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; et al. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Central Sci.* **2019**, *5* (9), 1523-1531. DOI: 10.1021/acscentsci.9b00476.
(49) Zou, W.; Martell Monterroza, A.; Yao, Y.; Millik, S. C.; Cencer, M. M.; Rebello, N. J.; Beech, H. K.; Morris, M. A.; Lin, T.-S.; Castano, C. S.; et al. Extending BigSMILES to non-covalent bonds in supramolecular polymer assemblies. *Chem. Sci.* **2022**, *13* (41), 12045-12055. DOI: 10.1039/D2SC02257E.
(50) Pablo, L. S.; Dylan, W.; Bradley, O.; Juan, d. Generative BigSMILES: An Extension for Polymer Informatics, Computer Simulations & ML/AI. **2023**. DOI: 10.26434/chemrxiv-2023-xv1kf (acccessed 2023/08/09).
(51) Lin, T.-S.; Rebello, N. J.; Lee, G.-H.; Morris, M. A.; Olsen, B. D. Canonicalizing BigSMILES for Polymers with Defined Backbones. *ACS Polym. Au* **2022**, *2* (6), 486-500. DOI: 10.1021/acspolymersau.2c00009.
(52) Guo, M.; Shou, W.; Makatura, L.; Erps, T.; Foshey, M.; Matusik, W. Polygrammar: Grammar for Digital Polymer Representation and Generation. *Adv. Sci.* **2022**, *9* (23), 2101864-2101864. DOI: 10.1002/advs.202101864.
(53) Park, N. H.; Manica, M.; Born, J.; Hedrick, J. L.; Erdmann, T.; Zubarev, D. Y.; Adell-Mill, N.; Arrechea, P. L. Artificial intelligence driven design of catalysts and materials for ring opening polymerization using a domain-specific language. *Nat. Commun.* **2023**, *14* (1), 3686. DOI: 10.1038/s41467-023-39396-3.

(54) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem Inf. Model.* **1988**, *28* (1), 31-36. DOI: 10.1021/ci00057a005.
(55) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989**, *29* (2), 97-101. DOI: 10.1021/ci00062a008.
(56) Shi, J.; Rebello, N. J.; Walsh, D.; Zou, W.; Deagen, M. E.; Leão, B. S.; Audus, D. J.; Olsen, B. D. Quantifying Pairwise Similarity for Complex Polymers. *Macromolecules* **2023**, *56* (18), 7344-7357. DOI: 10.1021/acs.macromol.3c00761.
(57) Lin, T.-S.; Rebello, N. J.; Beech, H. K.; Wang, Z.; El-Zaatari, B.; Lundberg, D. J.; Johnson, J. A.; Kalow, J. A.; Craig, S. L.; Olsen, B. D. PolyDAT: A Generic Data Schema for Polymer Characterization. *J. Chem Inf. Model.* **2021**, *61* (3), 1150-1163. DOI: 10.1021/acs.jcim.1c00028.
(58) Rubner, Y.; Tomasi, C.; Guibas, L. J. Earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* **2000**, *40* (2), 99-121. DOI: 10.1023/A:1026543900054.
(59) Hargreaves, C. J.; Dyer, M. S.; Gaultois, M. W.; Kurlin, V. A.; Rosseinsky, M. J. The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions. *Chemistry of Materials* **2020**, *32* (24), 10610-10620. DOI: 10.1021/acs.chemmater.0c03381.
(60) Durdy, S.; Hargreaves, C. J.; Dennison, M.; Wagg, B.; Moran, M.; Newnham, J. A.; Gaultois, M. W.; Rosseinsky, M. J.; Dyer, M. S. The Liverpool materials discovery server: a suite of computational tools for the collaborative discovery of materials. *Digital Discovery* **2023**, *2* (5), 1601-1611. DOI: 10.1039/D3DD00093A.
(61) Huizing, G.-J.; Peyr, G.; Cantini, L. Optimal transport improves cell-cell similarity inference in single-cell omics data. *Bioinformatics* **2022**, *38* (8), 2169-2177. DOI: 10.1093/bioinformatics/btac084.
(62) Alvarez-Melis, D.; Fusi, N. Geometric Dataset Distances via Optimal Transport. *Advances in Neural Information Processing Systems* **2020**, *33*, 21428--21439.
(63) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* **1981**, *147* (1), 195-197. DOI: 10.1016/0022-2836(81)90087-5.
(64) Eddy, S. R. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **2004**, *22* (8), 1035-1036. DOI: 10.1038/nbt0804-1035.
(65) Sanfeliu, A.; Fu, K.-S. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics* **1983**, *SMC-13* (3), 353-362. DOI: 10.1109/TSMC.1983.6313167.
(66) Bai, Y. S.; Ding, H.; Bian, S.; Chen, T.; Sun, Y. Z.; Wang, W.; Acm. SimGNN: A Neural Network Approach to Fast Graph Similarity Computation. In *12th ACM International Conference on Web Search and Data Mining (WSDM)*, Melbourne, AUSTRALIA, Feb 11-15, 2019; Assoc Computing Machinery: NEW YORK, 2019; pp 384-392. DOI: 10.1145/3289600.3290967.
(67) Van Krevelen, D. W.; Te Nijenhuis, K. Chapter 2 - Typology of Polymers. In *Properties of Polymers (Fourth Edition)*, Van Krevelen, D. W., Te Nijenhuis, K. Eds.; Elsevier, 2009; pp 7-47.
(68) Fox, T. G., Jr.; Flory, P. J. Second‐Order Transition Temperatures and Related Properties of Polystyrene. I. Influence of Molecular Weight. *Journal of Applied Physics* **1950**, *21* (6), 581-591. DOI: 10.1063/1.1699711.
(69) Balani, K.; Verma, V.; Agarwal, A.; Narayan, R. Physical, Thermal, and Mechanical Properties of Polymers. In *Biosurfaces: a materials science and engineering perspective*, 2014; pp 329-344.

(70) Chen, L.; Kern, J.; Lightstone, J. P.; Ramprasad, R. Data-assisted polymer retrosynthesis planning. *Applied Physics Reviews* **2021**, *8* (3). DOI: 10.1063/5.0052962.
(71) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Central Sci.* **2017**, *3* (12), 1237-1245. DOI: 10.1021/acscentsci.7b00355.
(72) Hart, W. E.; Watson, J.-P.; Woodruff, D. L.; Hart, W. E.; Watson, J. P.; Woodruff, D. L. Pyomo: modeling and solving mathematical programs in Python. *Mathematical Programming Computation 2011 3:3* **2011**, *3* (3), 219-260. DOI: 10.1007/S12532-011-0026-8.
(73) Hart, W. E.; Laird, C. D.; Watson, J.-P.; Woodruff, D. L.; Hackebeil, G. A.; Nicholson, B. L.; Siirola, J. D. *Pyomo — Optimization Modeling in Python*; Springer International Publishing, 2017. DOI: 10.1007/978-3-319-58821-6.
(74) coin-or/Cbc: Release releases/2.10.8 | Zenodo.
(75) Rumyantsev, A. M.; Jackson, N. E.; Yu, B.; Ting, J. M.; Chen, W.; Tirrell, M. V.; De Pablo, J. J. Controlling Complex Coacervation via Random Polyelectrolyte Sequences. *ACS Macro Letters* **2019**, *8* (10), 1296-1302. DOI: 10.1021/acsmacrolett.9b00494.
(76) Kottisch, V.; Gentekos, D. T.; Fors, B. P. "Shaping" the Future of Molecular Weight Distributions in Anionic Polymerization. *ACS Macro Letters* **2016**, *5* (7), 796-800. DOI: 10.1021/acsmacrolett.6b00392.
(77) Gentekos, D. T.; Sifri, R. J.; Fors, B. P. Controlling polymer properties through the shape of the molecular-weight distribution. *Nature Reviews Materials* **2019**, *4*, 761-774. DOI: 10.1038/s41578-019-0138-8.
(78) Sifri, R. J.; Padilla-Vélez, O.; Coates, G. W.; Fors, B. P. Controlling the Shape of Molecular Weight Distributions in Coordination Polymerization and Its Impact on Physical Properties. *J. Am. Chem. Soc.* **2020**, *142* (3), 1443-1448. DOI: 10.1021/jacs.9b11462.
(79) Blumenthal, D. B.; Gamper, J. On the exact computation of the graph edit distance. *Pattern Recognition Letters* **2020**, *134*, 46-57. DOI: 10.1016/j.patrec.2018.05.002.
(80) Ma, G.; Ahmed, N. K.; Willke, T. L.; Yu, P. S. Deep graph similarity learning: a survey. *Data Min. Knowl. Discov.* **2021**, *35*, 688-725. DOI: 10.1007/s10618-020-00733-5.
(81) Antoniuk, E. R.; Li, P.; Kailkhura, B.; Hiszpanski, A. M. Representing Polymers as Periodic Graphs with Learned Descriptors for Accurate Polymer Property Predictions. *J. Chem Inf. Model.* **2022**, *62* (22), 5435--5445. DOI: 10.1021/acs.jcim.2c00875.
(82) Ramprasad, M.; Kim, C. Assessing and improving machine learning model predictions of polymer glass transition temperatures. *Journal of Emerging Investigations* **2020**, *3*, 1-5. DOI: 10.59720/19-097.
(83) Peerless, J. S.; Milliken, N. J. B.; Oweida, T. J.; Manning, M. D.; Yingling, Y. G. Soft Matter Informatics: Current Progress and Challenges. *Advanced Theory and Simulations* **2019**, *2* (1), 1800129-1800129. DOI: 10.1002/adts.201800129.
(84) Lin, C.; Wang, P. H.; Hsiao, Y.; Chan, Y. T.; Engler, A. C.; Pitera, J. W.; Sanders, D. P.; Cheng, J.; Tseng, Y. J. Essential Step Toward Mining Big Polymer Data: PolyName2Structure, Mapping Polymer Names to Structures. *ACS Applied Polymer Materials* **2020**, *2* (8), 3107-3113. DOI: 10.1021/acsapm.0c00273.
(85) Wu, S.; Yamada, H.; Hayashi, Y.; Zamengo, M.; Yoshida, R. Potentials and challenges of polymer informatics: exploiting machine learning for polymer design. *arXiv preprint arXiv:2010.07683* **2020**. DOI: 10.48550/arxiv.2010.07683 (acccessed 2023-10-2).

TOC Graph