Quantifying Pairwise Similarity for Complex Polymers

Jiale Shi[1], Nathan J. Rebello[1], Dylan Walsh[1], Weizhong Zou[1], Michael E. Deagen[1], Bruno Salomao Leao[1,2], Debra J. Audus[3‡], Bradley D. Olsen[1‡]

1. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
2. School of Chemical Engineering, University of Campinas, Campinas, Sao Paulo 13083-852, Brazil
3. Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, United States

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

Abstract

Defining the similarity between chemical entities is an essential task in polymer informatics, enabling ranking, clustering, and classification. Despite its importance, pairwise chemical similarity for polymers remains an open problem. Here, a similarity function for polymers with well-defined backbones is designed based on polymers' stochastic graph representations generated from canonical BigSMILES, a structurally-based line notation for describing macromolecules. The stochastic graph representations are separated into three parts: repeat units, end groups, and polymer topology. The earth mover's distance is utilized to calculate the similarity of the repeat units and end groups, while the graph edit distance is used to calculate the similarity of the topology. These three values can be linearly or nonlinearly combined to yield an overall pairwise chemical similarity score for polymers that is largely consistent with the chemical intuition of expert users and is adjustable based on the relative importance of different chemical features for a given similarity problem. This method gives a reliable solution to quantitatively calculate the pairwise chemical similarity score for polymers and represents a vital step toward building search engines and quantitative design tools for polymer data.

Introduction

Polymers are ubiquitous with applications spanning clothing,[1] food,[2] energy,[3] transportation,[4] and health care.[5] This breadth of applications is achieved due to polymers' versatility, low-cost manufacturability, low density and chemical resistance. The massive design space available to polymer chemists leaves an abundance of potentially useful polymers yet to be identified and realized. As new polymers are discovered and current chemistries are manipulated, polymeric data is generated, enabling large polymer databases including PolyInfo,[6] PI1M (A Polymer Informatics Database of about 1 Million Polymers),[7] PolymerGenome,[8] MaterialsMine,[9] Open Macromolecular Genome,[10] and CRIPT (Community Resource for Innovation in Polymer Technology).[11] These databases have the potential to facilitate polymer design.[12–16] However, to accelerate polymer design, these databases must be coupled with additional functionalities.[17,18] For example, ranked search enhances data discoverability, and the ability to find similar polymers which have been previously synthesized, can further enable new polymer chemistries.

Additionally, classification and clustering algorithms are needed to validate, categorize and analyze new input polymer data points.[19] Such tasks are difficult or impossible without a robust similarity scoring method that calculates the magnitude of a chemical change between polymers and quantifies pairwise chemical similarity for polymers.[20]

In the field of cheminformatics, similarity scoring methods are well-established for small molecules. Either the graph structure[21–23] is retained, or it is converted into a vector, known as a fingerprint. [24] Then, either vector or graph similarity metrics, such as Tanimoto[25] and Cosine,[26] may be applied to calculate pairwise molecular similarity.[25] These similarity scoring methods have been used for a variety of tasks such as calculating similarity of entries in a drug molecule library, [27] designing new drug molecules, [28] ranking search results,[29] and calculating the magnitude of a chemical change from one small molecule to another.[25] Specialized machine learning methods also exist for similarity calculations of sequence-defined biomacromolecules such as proteins, peptides, and polysaccharides.[29,30] Both small molecules and sequence-defined biomacromolecules have well-defined deterministic structures that are easily represented by graphs with atoms (or molecular fragments) as nodes and bonds as edges.[29,31–38] In contrast, the vast majority of synthetic polymers are characterized by stochastic graphs that represent molecular ensembles or distributions.[39,40] Previous studies have used monomers and compositions as representations and utilized methods similar to those developed for small molecules to measure pairwise polymer similarity, but those methods can only be applied to polymers with simple topologies, such as homopolymers and copolymers.[41–45] These methods do not take into consideration the variety of topologies and stochastic configurations available to polymers; therefore, it is not possible for these methods to obtain an accurate and meaningful similarity score for polymers with complex topologies and stochastic properties, such as star polymers, graft polymers and segmented polymers.

The first key challenge in developing a broadly applicable polymer similarity metric is developing a representation for the polymer stochastic graph. Aldeghi et al.[40] proposed a graph representation for polymers using stochastic edges. However, the weight of the stochastic edges may not always be available, and when the weight is known, it is an average value that limits expressiveness. Guo et al.[46] proposed PolyGrammar, which is designed for polymer representation and generative modeling; however, the current generation of the PolyGrammar only imitates chain growth polymerization.[46] Recently, Lin et al.[39] demonstrated that polymers have a direct analogy to formal languages, and using this, they were able to develop directed graphs and automata-like deterministic graphs representing polymers. Rather than the graph representing the chemical structure, the graph represents a generating function that, when the graph is traversed, produces all possible molecules in the molecular ensemble.

Here, a method for pairwise similarity scoring of polymers based on an adaptation of Lin et al.'s graph representation [39] is proposed that is broadly applicable to stochastic ensembles across a wide

variety of polymer topologies. First, canonical polymer graph representations are generated with repeat units and end groups as nodes. Then, these graph representations are separated into three parts: repeat units, end groups, and topology. The earth mover's distance (EMD) [47–49] is utilized to calculate the similarity of the repeat units, as well as the end groups. Subsequently, graph edit distance (GED) [30,50,51] is used to calculate the similarity of the topology. Combining similarity scores for the repeat units, end groups, and topology yields an overall pairwise chemical similarity for polymers that is largely consistent with the chemical intuition of expert users and is tunable based on the importance individual users place on specific substructural elements.

Methods

Stochastic Graph Representation

The first step in generating a similarity score is to generate stochastic polymer graphs. The polymer molecular structure (see Figure 1a) is converted to a canonical BigSMILES[52,53] representation, a structurally-based line notation for describing macromolecules (see Figure 1b) following the priority rules of canonicalization procedures from Lin et al.[39] This canonicalization step is essential as it ensures that every polymer has exactly one representation. Without this step, it is possible to generate a similarity score smaller than one for the same polymer, as multiple non-canonicalized BigSMILES can map to the same polymer. Next, the algorithm from Lin et al.[39] parses the canonical BigSMILES and uses connectivity information to build directed graphs, shown in Figure 1c. Each node is labeled with either "Start," "End," a bonding descriptor, a repeat unit SMILES, or an end group SMILES. For the repeat units SMILES and end group SMILES, the symbol, *, is used as a connection point to clearly illustrate which atoms are connected in the polymers and which part of the repeat units belong to pendant groups. For example, one of Polymer A's stochastic objects from its canonical BigSMILES, CC(C[>1])O[<1], is first transferred to CC(C*)O*. Next, CC(C*)O* is transferred to *CC(C)O*, which is more intuitive. The transfer process to a more intuitive SMILES string does not affect the similarity calculation.

For Polymer A, a random copolymer, the directed graph[39] reads from the left end group OCCO* to the stochastic bonding descriptor, which can connect either to poly(propylene glycol) (PPO) with repeat unit *CC(C)O* or poly(ethylene glycol) (PEG) with repeat unit *CCO*. Since H* as an end group is implicit in the canonical BigSMILES, the directed graph does not have a separate node for the right end group H*.[39] As for Polymer B, a diblock copolymer, the directed graph reads from the left stochastic bonding descriptor, which can connect the repeat unit *C(CC)C* or its mirror *CC(CC)*, and then the graph reads the right stochastic bonding descriptor, which can connect the repeat unit *C(C)C* or its mirror *CC(C)*. If the repeat units are symmetric, such as *CC*, two connection paths still exist even though these two connection paths are identical. Therefore, to preserve the topological feature and ensure the robustness of the similarity function, two possible connection paths are retained for symmetric repeat units. For Polymer C, an alternating copolymer, the directed graph reads from the left stochastic bonding descriptor, which connects the repeat unit *C(=O)C(C)CCCC(=O)* or its mirror *C(=O)CCCC(C)C(=O)*, and

then the graph reads the right stochastic bonding descriptor, which connects the second repeat unit *OC(C)CO or its mirror *OCC(C)O*, and finally the graph returns to the left stochastic bonding descriptor. Again, both paths are kept for symmetric repeat units.

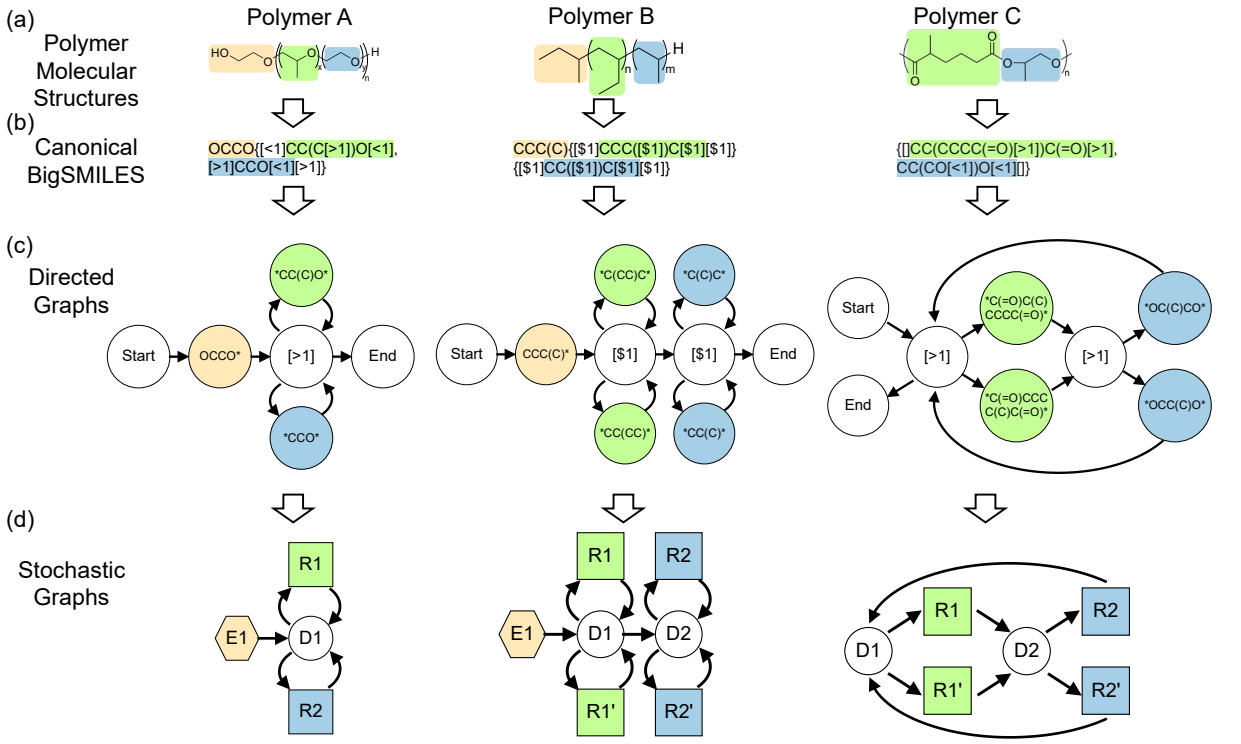


Figure 1: Converting polymer molecular structure*s* into the corresponding stochastic polymer graph representations. (a) A random copolymer (Polymer A), a diblock copolymer (Polymer B), and an alternating copolymer (Polymer C). (b) Canonical BigSMILES representations produced using the canonicalization procedures from Lin et al.[39] (c) The algorithm from Lin et al.[39] parses the canonical BigSMILES and uses the connectivity information to build directed graphs. Each node is labeled with either "Start," "End," a stochastic bonding descriptor, repeat unit SMILES, or end group SMILES. The directed graphs are converted into stochastic graphs in panel (d), where the nodes of "Start" and "End" are removed, stochastic bonding descriptor nodes are represented by circles with indexes (D1, D2, ...), repeat unit SMILES nodes are represented by squares with indexes (R1, R2,...), and end group SMILES nodes are represented by hexagons with indexes (E1, E2,...). The colors of repeat unit SMILES nodes and end group SMILES nodes match the corresponding repeat units and end groups in the canonical BigSMILES representations and directed graph representations.

Finally, the directed graphs in Figure 1c are converted into polymer stochastic graphs in Figure 1d where the nodes of "Start" and "End" are removed, stochastic bonding descriptor nodes are represented by circles, repeat unit SMILES nodes are represented by squares, and end group SMILES nodes are represented by hexagons. The colors of repeat unit SMILES nodes and end

group SMILES nodes match the corresponding repeat units and end groups in the canonical BigSMILES representations.

Overview of Similarity Method

Based on this stochastic polymer graphs representation, a method to calculate the pairwise overall chemical similarity between two polymers is proposed, as illustrated in Figure 2. The polymer graph is decomposed into three components: repeat units, end groups and topology. Linkers between stochastic objects are also included into this category of end group. Topology here represents both the local connectivity (the way the monomer units themselves are connected) and the global topology of the graph. Individual similarity metrics are calculated for each component, which are then combined to yield an overall similarity score. The earth mover's distance (EMD) is used to calculate the similarity scores of the repeat units S_{RU} and the end groups S_{EG}. The topological similarity S_{TOP} is then calculated from the stochastic graph representations with all chemical detail removed using graph edit distance (GED). Finally, the overall similarity score S_{OA} between two polymers is generated by combining these three scores via either geometric or arithmetic mean. The details of calculating EMD, GED, and overall similarity score are illustrated in detail in the following sections.

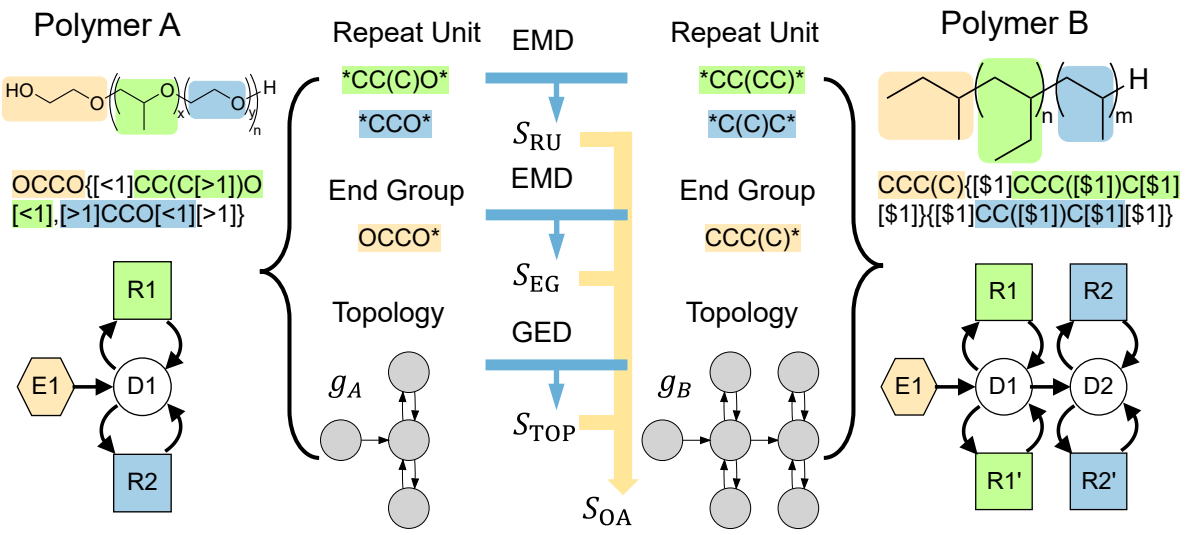


Figure 2: Schematic of the method for calculating the pairwise chemical similarity between two polymers. Using the stochastic graph representation, the polymers are separated into three key features: repeat units, end groups, and topology. Linkers between stochastic objects are also included with end groups, and topology here represents both the local connectivity (the way the monomer units themselves are connected) and the global topology of the graph. The similarity scores for the repeat units S_{RU} and the end groups S_{EG} are calculated via earth mover's distance (EMD) whereas the similarity score for topology S_{TOP} is calculated via graph edit distance (GED).

The overall pairwise similarity score S_{OA} between two polymers is generated by combining these three scores via either geometric or arithmetic mean.

Earth Mover's Distance for S_{RU} and S_{EG}

The workflow of the repeat unit similarity S_{RU} is shown in Figure 3 using Polymer A and Polymer B as an example. The procedure for the end groups is identical to the procedure for repeat units. The first step is to identify the repeat units. Polymer A has two repeat units ($R1_A$ and $R2_A$), and Polymer B has two repeat units ($R1_B$ and $R2_B$). Since the frequencies of the repeat units can vary, the repeat units of each polymer can be conceptualized as a molecular fragment ensemble. Therefore, the problem of calculating S_{RU} is fundamentally a problem of calculating the similarity S between different ensembles or distributions of small molecules, each of which may be computed using existing methods for calculating the pairwise similarity of small molecules.[25] Specifically, molecular fragment ensemble $P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \ldots, (p_i, w_{p_i}), \ldots, (p_m, w_{p_m})\}$ has m molecular fragments, where p_i is a molecular fragment such as a repeat unit or end group and $w_{p_i} > 0$ is the weight, related to the average probability (or frequency, z) of the molecular fragment being present in the polymer. Similarly, the second ensemble $Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \ldots, (q_j, w_{q_j}), \ldots, (q_n, w_{q_n})\}$ has n molecular fragments. The sums of the weights for P and Q are both normalized and equal to one $\sum_{i=1}^{m} w_{p_i} = \sum_{j=1}^{n} w_{q_j} = 1$.

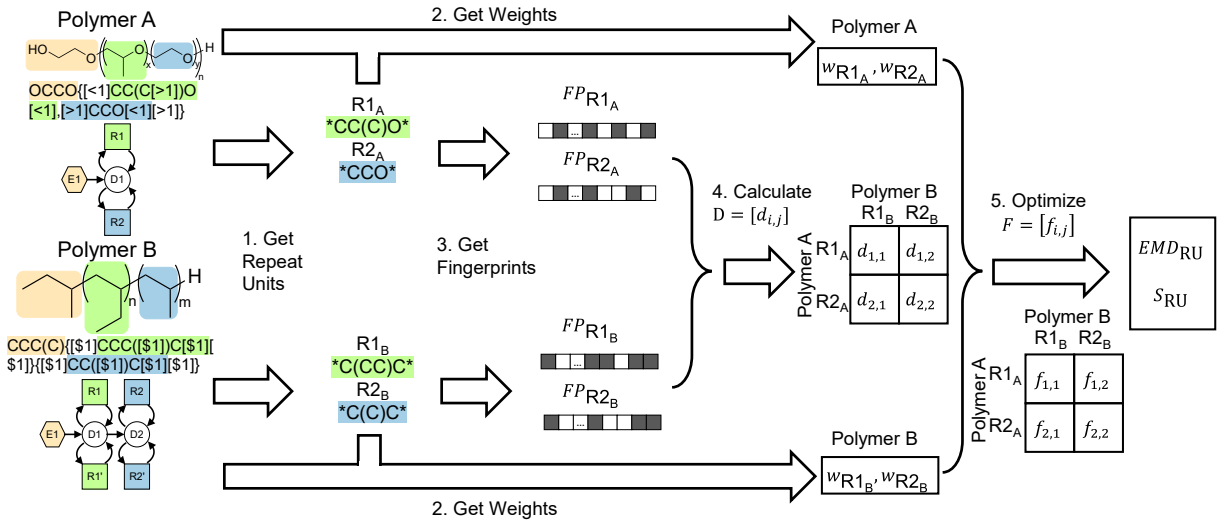


Figure 3: Workflow of earth mover's distance (EMD) calculation for ensemble similarity using the repeat unit sets of Polymers A and B as an example. The first step is to get the repeat units. Polymer A has two repeats ($R1_A$ and $R2_A$) and Polymer B has two repeat units ($R1_B$ and $R2_B$), where the subscripts are used for distinction. The second step is to get the corresponding weight for each repeat unit. The third step is to get the corresponding fingerprints (FPs). The fourth step is to

calculate the set of pairwise distances $D = [d_{i,j}]$ based on the similarity metric. Once the weights and the set of pairwise distance $D = [d_{i,j}]$ are obtained, the fifth step is to optimize the transport flows $F = [f_{i,j}]$ to calculate the distance EMD_{RU} and the similarity score S_{RU}. The procedure for the end groups is identical.

The second step is to obtain the weight of each molecular fragment. Unlike small molecules, whose chemical structure uniquely determines the molar mass, polymers may have varying degrees of polymerization or monomer composition for a given chemical structure. When the composition or degree of polymerization is known, this may be used to determine the weights. For repeat units within the stochastic objects inside the first level of curly brackets, or equivalently at the same level as the backbone when the backbone is present, based on the canonical BigSMILES, the weight w of a repeat unit is directly proportional to the average number of the repeat unit per polymer, z:

$$\frac{w_2}{w_1} = \frac{z_2}{z_1} \qquad (1)$$

If z_i is not specified, then the sum of molecular fragments connected to each stochastic bonding descriptor shares the same relative weight, and each molecular fragment connected with the same stochastic bonding descriptor shares the same relative weight. For example, as shown in Figure 4a, in a random copolymer, R1-*r*-R2, $w_{\mathrm{R1}} = w_{\mathrm{R2}}$. In Figure 4b, a diblock copolymer, R1-*b*-R2, has $w_{\mathrm{R1}} = w_{\mathrm{R2}}$. Figure 4c illustrates a diblock with one block being a random copolymer (R1-*r*-R2)-*b*-R3 such that $w_{\mathrm{R1}} = w_{\mathrm{R2}} = 0.5w_{\mathrm{R3}}$. An alternating copolymer, R1-*alt*-R2 with $w_{\mathrm{R1}} = w_{\mathrm{R2}}$ is shown in Figure 4d.

For repeat units that are the nested stochastic objects based on the canonical BigSMILES, such as the repeat units in the side chain of graft polymers or the repeat units in the macromonomer of segmented polymers (see Figure 4e,f), the lengths, or equivalently the degrees of polymerization, of the nested stochastic objects affect the polymer properties.[54,55] Thus, the length of nested stochastic objects should be included into the weightings. However, if the relative weights are proportional to the frequencies, the influence of the backbone repeat unit may be nearly zero when the nested stochastic objects are long. Therefore, for nested stochastic objects, relative weights between one repeat unit in the backbone and one in the nested stochastic object are given by a logarithmic equation.

$$\frac{w_2}{w_1} = 1 + \ln\frac{z_2}{z_1} \qquad (2)$$

where z_1, and w_1 are the frequency and weight of the backbone repeat unit connected to a nesting stochastic object; z_2, and w_2 are the frequency and weight of the repeat unit in the nesting stochastic object. This modification ensures that the weight monotonically increases with the

monomer frequency while reducing the weight of grafts and segments at high degrees of polymerization. For example, if $z_2 : z_1 = 200 : 1$, then $w_2 : w_1 = 6.3 : 1$ The range of the length of the graft chain or segment part is typically one to hundreds, yielding a range of $\frac{w_2}{w_1}$ of about 1 to 8. If the lengths of grafts or macromonomers are not specified, $\frac{w_2}{w_1} = 4$ is a reasonable choice for a default value.

The weight assignment for the end groups follows the same principles. For example, the same weight is used for both end groups in linear polymers (see Figure 4c). For graft polymers, the weight assignment between end groups at the end of the side chains and the end groups on the backbone also follows Equation 2 because there is one end group per graft side chain. If the degrees of polymerization are not specified, then the weight of the end groups at the end of the side chains is four times the weight of the end groups at the end of the backbone chain, as shown in Figure 4e. With these rules, the similarity algorithm can compare polymers based on chemical structure alone, without any degree of polymerization or composition information. However, if known, this information can be used to improve similarity scoring.

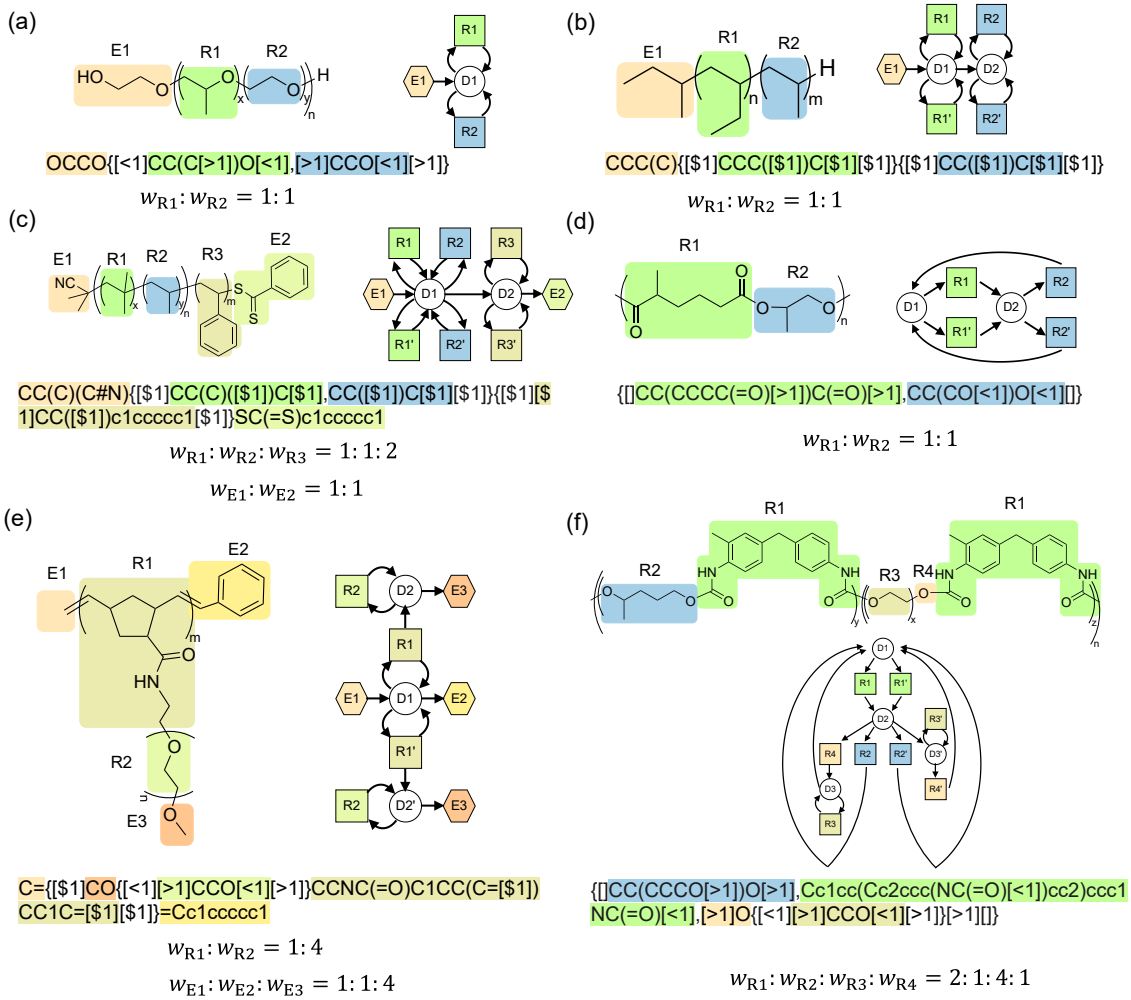


Figure 4: Weight assignment policy for polymers as a function of polymer topology when degrees of polymerization are not specified. Six types of polymer molecular structures, the corresponding canonical BigSMILES and stochastic polymer graph representations are displayed. (a) A random copolymer. (b) A diblock copolymer. (c) A diblock copolymer with one block being a random copolymer (R1-*r*-R2)-*b*-R3. (d) An alternating copolymer. (e) A graft polymer where the monomer on the backbone have a side chain. (f) A segmented polymer where a polymer is nested along the backbone.

Returning to the example in Figure 3 and following the above rules, the two repeat unit ensembles are $P = \{(\mathrm{R1_A}, 0.5), (\mathrm{R2_A}, 0.5)\}$ for Polymer A and $Q = \{(\mathrm{R1_B}, 0.5), (\mathrm{R2_B}, 0.5)\}$ for Polymer B. For Polymer B, $\mathrm{R1_B}$ and $\mathrm{R1_B}$' are different configurations in the polymer chain, but $\mathrm{R1_B}$ and $\mathrm{R1_B}$' are identical when separated from the polymer chain; therefore, the weight of $\mathrm{R1_B}$' is merged to $\mathrm{R1_B}$, and the same for $\mathrm{R2_B}$'.

With the ensembles defined, the earth mover's distance (EMD) is a metric that is well-constructed to calculate the similarity of ensembles or distributions such as these; it has been successfully applied in multiple fields for ensemble similarity calculation, such as the similarity of inorganic

solids,[47] the similarity of biomarker expression levels,[48] and geometric dataset distances.[49] EMD may be conceptualized as the minimal amount of work to transform one distribution into another, and it can be formulated and solved as a transportation problem. Here, the problem is transforming one discrete molecular fragment distribution P to another Q with the minimum amount of work done (EMD), which can be interpreted as a measure of dissimilarity. Therefore, the problem of calculating S_{RU} and S_{EG} is equivalent to calculating the similarity S between different ensembles of molecular fragments, each of which may have pairwise similarity s_{ij} computed using existing methods for calculating the pairwise similarity of small molecules.[25]

Thus, the next step is to determine the pairwise similarity of the individual molecular fragments. First, each molecular fragment is represented by a SMILES (Simplified Molecular-Input Line-Entry System) string[56,57] containing "*" symbols to indicate the interconnections between monomers[7,44]. These SMILES strings are then transformed into fingerprints using Morgan fingerprints (radius = 2, nBits = 2048)[40] as implemented in RDKit[58] (step 3 in Figure 2). Then, the pairwise similarity score between the molecular fragments p_i and q_j, $s_{i,j}$ are calculated using the Tanimoto similarity metric.[25,30] The similarity score $s_{i,j}$ ranges from 0 to 1, where self-similarity is 1. The more similar two molecular fragments p_i and q_j, the larger $s_{i,j}$. Apart from the Morgan fingerprints and Tanimoto similarity metric, different settings for radius, nBits and useChirality of Morgan fingerprints,[30] many other fingerprint embedding functions,[24] molecular graph embedding methods[21–23] and different similarity metrics[25] can be utilized to obtain $s_{i,j}$ without modifying the overarching algorithm for polymer similarity described here.

EMD is inherently a measure of dissimilarity instead of similarity, so first the similarity score $s_{i,j}$ must be converted to a dissimilarity score using[25,30]

$$d_{i,j} = 1 - s_{i,j} \tag{3}$$

as shown in step 4 of Figure 3. After all necessary information is obtained on the w_{p_i}, w_{q_j} and $d_{i,j}$ and for all the entities in the ensembles, the optimized transport flows $F = [f_{i,j}]$ and the EMD are determined using Equation 4a along with the constraints as specified in Equations 4b-e.

$$EMD(P,Q) = \frac{\min_F \sum_{i=1}^{m}\sum_{j=1}^{n}\left(f_{i,j}\cdot d_{i,j}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}} = \min_F \sum_{i=1}^{m}\sum_{j=1}^{n}\left(f_{i,j}\cdot d_{i,j}\right) \tag{4a}$$

$$\text{Subject to } f_{i,j} \geq 0, \text{ for any } 1 \leq i \leq m, 1 \leq j \leq n \tag{4b}$$

$$\sum_{j=1}^{n} f_{i,j} = w_{p_i}, \text{ for any } 1 \leq i \leq m \tag{4c}$$

$$\sum_{i=1}^{m} f_{i,j} = w_{q_j}, \text{ for any } 1 \leq j \leq n \tag{4d}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = \sum_{i=1}^{m} w_{p_i} = \sum_{j=1}^{n} w_{q_j} = 1 \tag{4e}$$

$f_{i,j}$ represents the amount of weight at p_i which is transported to q_j. The sum of all the individual flows originating from p_i is equal to the weight w_{p_i}, and equivalently, the sum of all the individual flows originating from q_j is equal to the weight w_{q_j}, shown in Equations 4cd. Here, $f_{i,j} \cdot d_{i,j}$ is the cost for each individual flow. Thus, EMD represents the minimum overall cost to convert one ensemble P to another ensemble Q.

These equations are coded into Pyomo,[59–61] a python-based, open-source optimization modeling language with a diverse set of optimization capabilities, and solved with COIN-OR Branch-and-Cut (cbc) solver, [62,63] an open-source mixed integer linear programming solver written in C++. $EMD(P,Q)$ is bounded between 0 and 1, representing the dissimilarity score between P and Q.

Finally, the similarity score $S(P,Q)$, for the ensemble pair P and Q, may then be defined as

$$S(P,Q) = 1 - EMD(P,Q) \tag{5}$$

Equation 5 is consistent with Equation 3 relating similarity and dissimilarity for small molecules. The value of $S(P,Q)$ is also between 0 and 1. The smaller $EMD(P,Q)$, the larger $S(P,Q)$, and the higher the similarity between P and Q. Both the similarity between two repeat unit ensembles $S_{\mathrm{RU}}(P,Q)$, and the similarity between two end group ensembles $S_{\mathrm{EG}}(P,Q)$ are calculated through the above EMD method. For the pair Polymer A and Polymer B, $S_{\mathrm{RU}} = 0.28$ and $S_{\mathrm{EG}} = 0.10$ (step 5 in Figure 2). Since not all polymers' molecular representations include the end groups (e.g., rings) and implicit end groups, some complementary rules for the end group similarity scores are set for those situations: (1) if two polymers both do not have end groups, $S_{\mathrm{EG}} = 1$; (2) if one polymer has end groups and the other polymer does not have end groups, then $S_{\mathrm{EG}} = 1$.

EMD provides greater resolution of chemical differences between polymers than simple sums or averages of the Morgan fingerprints or other fingerprints for each repeat unit. The reason is that simply averaging or summing[44] prematurely reduces the dimensionality of the system, eliminating differences among ensembles. Examples are included in the Supporting Information.

Graph Edit Distance for S_{TOP}

The next step is to compute the similarity score for the topology S_{TOP}. Since the chemical details have already been accounted for, the topology can be treated as a homogenous version of the stochastic graph representation where all edges and nodes are treated identically (see grey topology

graphs g_A and g_B in Figure 2). To calculate the similarity between two different polymer topologies, Graph Edit Distance (GED) [30,50,64] is utilized. GED, first reported by Sanfeliu and Fu in 1983,[65] 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[64] typically includes node and edge insertion, deletion, and substitution, although substitution is not considered here since the topology graphs are homogeneous.

$$GED(g_1, g_2) = \min_{(e_1,\ldots,e_k) \in \mathcal{P}(g_1,g_2)} \sum_{i=1}^{k} c(e_i) \quad (6)$$

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. For simplicity, $c(e_i) = 1$, the cost of each graph edit cost is set to be one. $GED(g_1, g_2)$ is zero when g_1 and g_2 are identical. GED is symmetric; the cost of transforming graph g_1 into graph g_2 is the same as the cost of transforming graph g_2 into graph g_1. GED is widely used for similarity measurements in small molecules[66–68] and sequence-defined biomacromolecules.[30,69,70] Using Figure 2 as an example, to transform g_A into g_B, g_A adds three nodes and five edges; therefore the $GED(g_A, g_B) = 8$.

To map GED onto a topological similarity score S_{TOP} with the range of $(0, 1]$, an exponential decay function on the normalized graph edit distance, $\frac{GED(g_1,g_2)}{(|g_1|+|g_2|)/2}$ is used:[51]

$$S_{TOP}(g_1, g_2) = \exp\left(-\frac{\alpha \cdot GED(g_1, g_2)}{(N_1 + N_2)/2}\right) \quad (7)$$

where N_i denotes the number of nodes of g_i; α is a tunable parameter with the default value to be 1. $S_{TOP}(g_1, g_2)$ is 1 when g_1 and g_2 are identical. $S_{TOP}(g_1, g_2)$ is also symmetric, so $S_{TOP}(g_1, g_2) = S_{TOP}(g_1, g_2)$. As shown in Figure 2, for Polymer A and Polymer B, $N_A = 4$ and $N_B = 7$; therefore, $S_{TOP}(g_A, g_B) = 0.23$.

Although the calculation of an exact GED is non-deterministic polynomial-time hard (NP-hard), the size of the topological graph is relatively compact unlike the graph representations used for sequence-defined biomacromolecules, which can be very complex[30] for large molar masses. Additionally, the chemical details are dropped from the stochastic topological graph, and the exact GED is calculated on a homogenous version of the stochastic graph representation where all edges and nodes are treated identically, which dramatically reduces computational complexity and cost. Therefore, computing the exact GED for the stochastic topological graph for polymers represented in this compact fashion is computationally tractable.

Overall Pairwise Chemical Similarity Score

From the above EMD and GED calculations, three similarity scores are obtained: S_{RU} for repeat units, S_{EG} for end groups, and S_{TOP} for topology. To calculate the overall similarity score S_{OA}, a weighted geometric average is proposed:

$$S_{\mathrm{OA}} = S_{\mathrm{RU}}^{W_{\mathrm{RU}}} \cdot S_{\mathrm{TOP}}^{W_{\mathrm{TOP}}} \cdot S_{\mathrm{EG}}^{W_{\mathrm{EG}}} \tag{8}$$

where $W_{\mathrm{RU}} + W_{\mathrm{TOP}} + W_{\mathrm{EG}} = 1$ are the weights for the repeat units, topology and end groups, respectively. These weights can be tuned to suit the user's target application. For simplicity, reasonable defaults of $W_{\mathrm{RU}} = 0.475$, $W_{\mathrm{TOP}} = 0.475$, and $W_{\mathrm{EG}} = 0.05$ are chosen. The choice of $W_{\mathrm{EG}} = 0.05$ was motivated by the low frequency of the end group relative to the repeat unit. This choice results in $\frac{W_{\mathrm{RU}}}{W_{\mathrm{EG}}} = 9.5$, about one order of magnitude. The choice of equal setting for W_{RU} and W_{TOP} is grounded in the idea that both repeat units and topology are essential for capturing polymer similarity based on chemical intuition. The repeat units reflect what the types of monomers comprise the polymers, and the repeat units can influence physical properties, such as glass transition temperature and density. The topology reflects how the monomers are connected in the polymer chains and what synthesis routines are used for polymerization. The topology also significantly impacts physical properties, such as viscosity and phase behavior. For additional freedom for user-specific cases, a weighted arithmetic mean can also be used to calculate the overall similarity score:

$$S_{\mathrm{OA}} = S_{\mathrm{RU}} \cdot W_{\mathrm{RU}} + S_{\mathrm{TOP}} \cdot W_{\mathrm{TOP}} + S_{\mathrm{EG}} \cdot W_{\mathrm{EG}} \tag{9}$$

where $W_{\mathrm{RU}} + W_{\mathrm{TOP}} + W_{\mathrm{EG}} = 1$ are the weights in the arithmetic function.

Applying these equations along with different weight choices for Polymer A and Polymer B from Figure 2 yields $S_{\mathrm{OA}}(\text{Polymer A, Polymer B}) = 0.243$ with the weighted geometric mean and $S_{\mathrm{OA}}(\text{Polymer A, Polymer B}) = 0.248$ with the weighted arithmetic mean. For this case, the results are similar for different mean functions because S_{RU} and S_{TOP} which occupy the major weights are similar. If S_{RU} and S_{TOP} are more distinct, then the choice of mean function evidently affects the S_{OA}. In the Results and Discussion section, the geometric mean is used as the default as it weighs very small similarities more heavily. Weighted arithmetic mean values are provided in the Supporting Information for completeness. In all the following cases, the weight settings $W_{\mathrm{RU}} = 0.475$, $W_{\mathrm{TOP}} = 0.475$, and $W_{\mathrm{EG}} = 0.05$ are used.

Results and Discussion

Case 1: Varying Repeat Units

Case 1 illustrates the computation of the pairwise similarity score of polymers with the same topological graph representation and end groups but different repeat units, shown in Figure 5a-d. These polymer examples are collected and modified from Shim et al.[71] All four polymers are

diblock copolymers, and they have an identical stochastic topological graph representation with nodes' and edges' details shown in Figure 5e, where the colors of nodes match with the repeat units and end groups, and the directions of the edges match with the connection paths. Therefore, all the pairwise $S_{TOP} = 1.0$ and $S_{EG} = 1.0$, and S_{RU} determines the overall pairwise similarity S_{OA}. The results are shown in Figure 5f,g. Taking C1-1 as a reference, the similarity order is C1-3 > C1-2 > C1-4; this is consistent with chemical intuition since adding more functional groups increases the dissimilarity between monomers and adding functional groups to the simpler monomers results in larger dissimilarity.

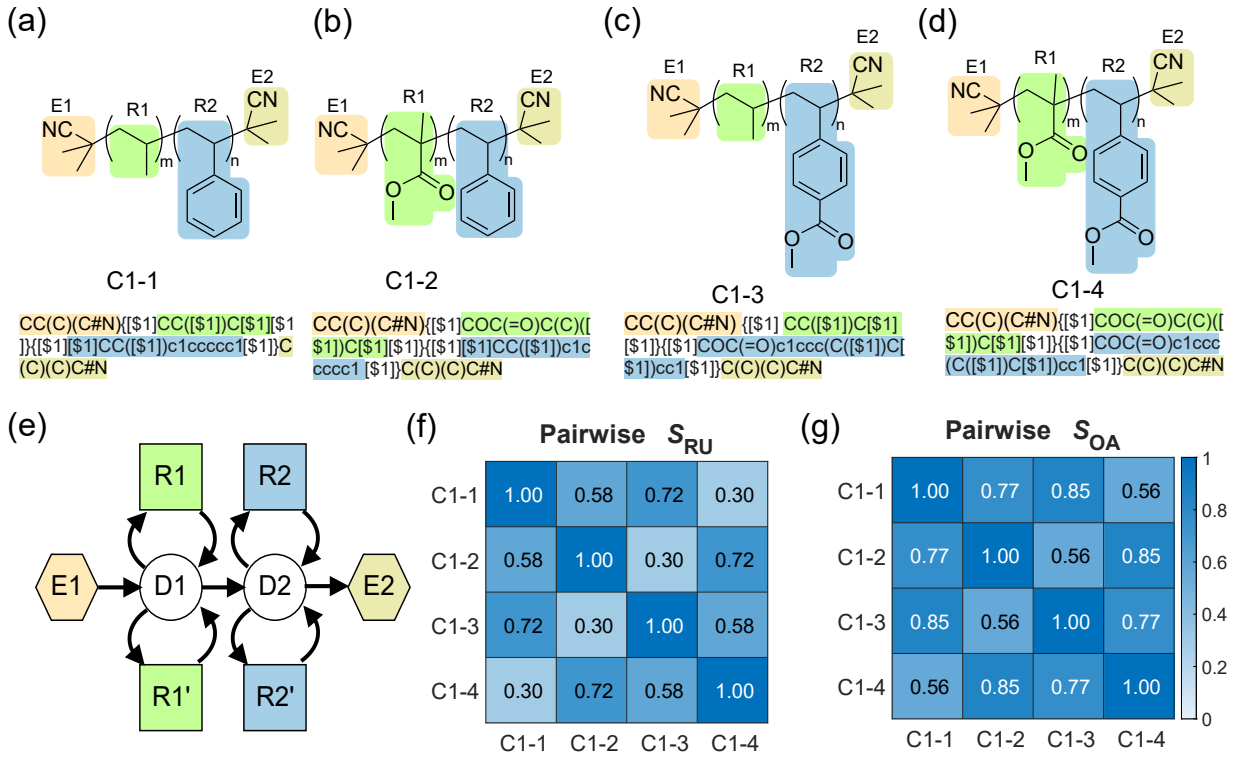


Figure 5: Four diblock polymers and the corresponding canonical BigSMILES, C1-1 (a), C1-2 (b), C1-3 (c), C1-4 (d) which have the same topological graph representation (as shown in (e)) and end groups but different repeat units. The polymerization degrees, m and n are not specified, so that all the repeat units share equal weight. (f) Pairwise repeat unit ensemble S_{RU} and (g) overall similarity S_{OA} for four diblock polymers in Case 1.

Case 2: Varying Topologies

Apart from the repeat units, polymers' topology can also largely affect the polymer's properties in many aspects. Case 2 compares the pairwise similarity score of polymers that have the same repeat units but different topological graph representations, as shown in Figure 6a-d. These examples of reversible addition-fragmentation chain transfer (RAFT) [72,73] polystyrenes (C2-1: one-arm, C2-2:

two-arm, C2-3: three-arm, C2-4: four-arm) are collected and modified from Altintas et al. [72] and Zayas et al.[73]

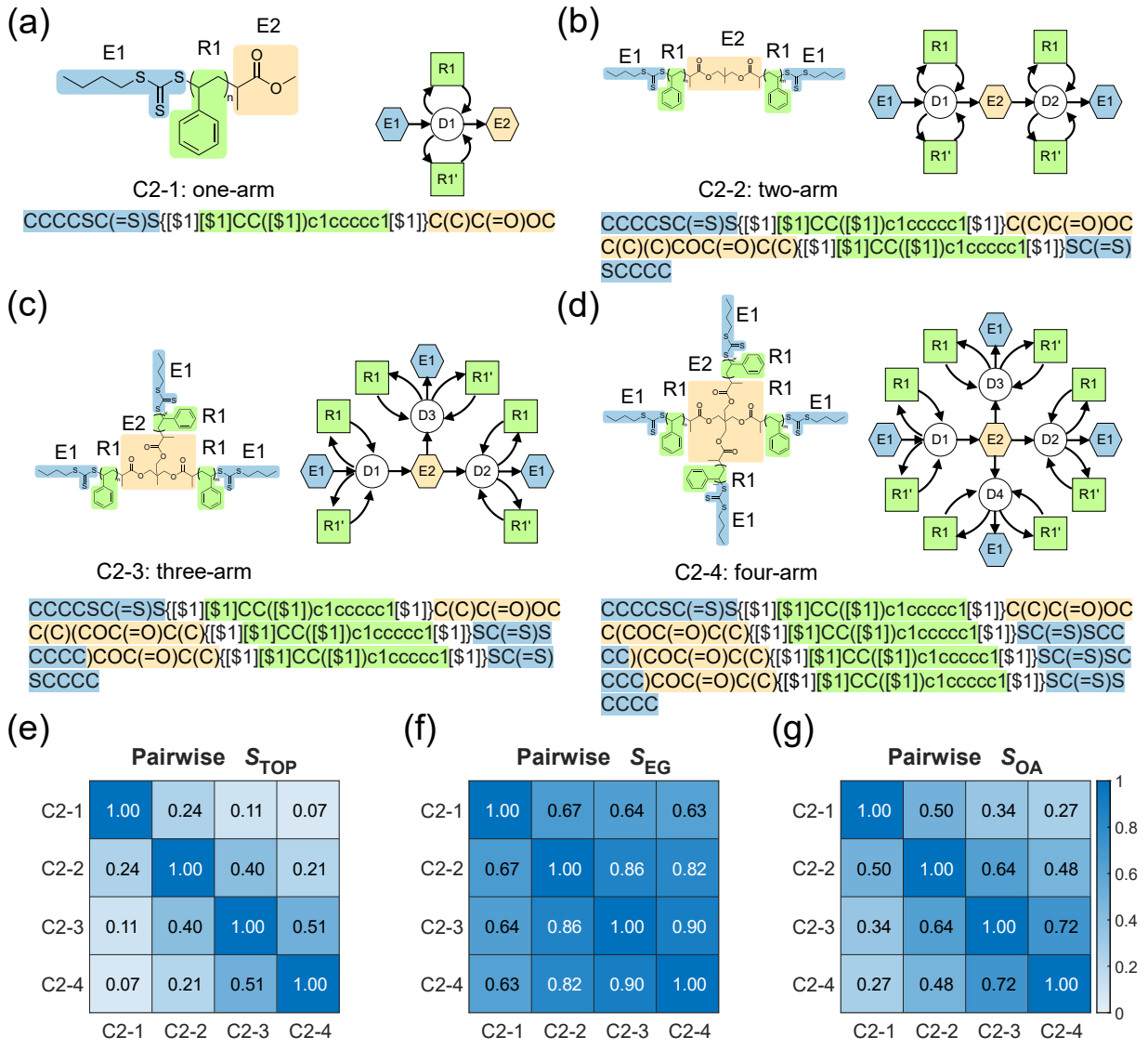


Figure 6: Four RAFT polystyrenes, the corresponding canonical BigSMILES and stochastic graph representations. (a) A one-arm polymer C2-1, (b) a two-arm polymer C2-2, (c) a three-arm star polymer C2-3, and (d) a four-arm star polymer C2-4. (e) Pairwise topological similarity S_{TOP}, (f) end group ensemble similarity S_{EG}, and (g) overall similarity S_{OA} for four RAFT polymers in Case 2.

All four RAFT polymers have the same styrene repeat unit. Therefore, the pairwise $S_{RU} = 1$ for repeat unit ensembles between all polymer pairs. However, as shown in Figure 6a-d, these four polymers have different stochastic graph representations. The results of the pairwise S_{TOP} (see Figure 6e) reflect these topological differences. Taking the one-arm polymer C2-1 as a reference, S_{TOP} decreases from C2-2 to C2-4, showing that graph edit distance intuitively increases as the

difference in the number of arms increases. The absolute graph edit distance, $GED(\text{C2-1, C2-2}) = GED(\text{C2-2, C2-3}) = GED(\text{C2-3, C2-4})$, but S_{TOP} is determined by the normalized graph edit distance. Therefore, the neighbor pairwise similarity score increases with the increasing number of arms: $S_{\text{TOP}}(\text{C2-3, C2-4}) > S_{\text{TOP}}(\text{C2-2, C2-3}) > S_{\text{TOP}}(\text{C2-1, C2-2})$. This feature is also chemically intuitive; when the number of arms is low, adding an arm is a large change in topology, but when the number of arms is high, adding an arm leads to a smaller change in the topology. The end group ensembles are also slightly different because of the chemical structure changes to the core of the star (see Figure 6f). The ranking of the overall similarity score S_{OA} (see Figure 6g) which is mainly determined by S_{TOP}, follows the same trends as S_{TOP}.

Case 3: Varying Both Repeat Units and Topologies

In many real-world applications, one is interested in the similarity between polymers that have both different chemistries and different topologies. Three block copolymers which have both different repeat units and different topological graph representations are shown in Figure 7a-c: C3-1, a diblock polymer; C3-2, a triblock polymer; and C3-3, a tetrablock polymer. The default weight assignment, $W_{\text{RU}} = 0.475$, $W_{\text{TOP}} = 0.475$, and $W_{\text{EG}} = 0.05$ is one suitable option, but this case also illustrates how modifying the values of W_{RU} and W_{TOP} can change the overall pairwise similarity scores S_{OA} and affect the final ranking. For simplification, all three block polymers have the same end groups and $W_{\text{EG}} = 0.05$ is held constant.

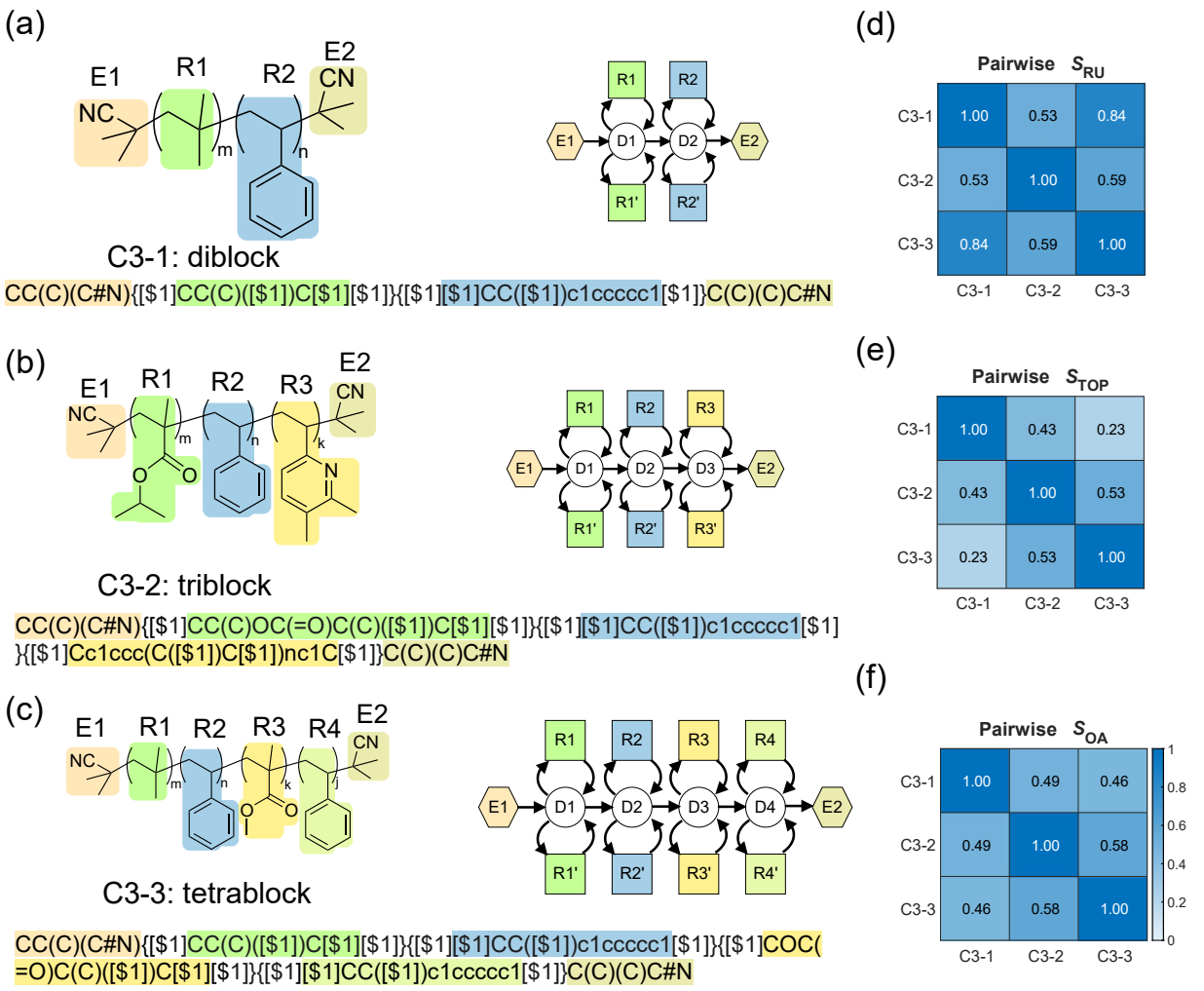


Figure 7: Three block polymers, the corresponding canonical BigSMILES and stochastic graph representations. (a) a diblock polymer C3-1, (b) a triblock polymer EC-2, and (c) a tetrablock polymer C3-3. (d) Pairwise repeat ensemble similarity S_{RU}, (e) topological similarity S_{TOP}, and (f) overall similarity S_{OA} for three block polymers in Case 3.

The results of pairwise repeat unit similarity scores S_{RU} are shown in Figure 7d, and the results of the pairwise topological similarity score S_{TOP} are shown in Figure 7e. C3-1, C3-2, C3-3 have the same end groups; therefore, the pairwise $S_{EG} = 1$ for all pairs. If C3-1 is taken as the reference, the repeat units of C3-3 are closer to C3-1's than C3-2's based on their chemical structures in Figure 7a-c. Therefore, C3-3 tetrablock polymer has a higher S_{RU} than C3-2 triblock polymer. With respect to topology, GED increases with increasing difference in block number. Therefore, C3-3 tetrablock polymer has a lower S_{TOP} than C3-2 triblock polymer. Therefore, the similarity ranking of S_{RU} is opposite to the similarity ranking of S_{TOP} for this case. In this situation,

modifying the values of W_{RU} and W_{TOP} can change the final ranking order of S_{OA} (see Supporting Information), thus demonstrating the flexibility of the polymer similarity method proposed.

Case 4: Graft Copolymers

The polymer similarity method can be applied to complex polymer architectures. Case 4 demonstrates similarity scoring for graft polymers (see Figure 8a-d) collected from Walsh et al.[55] and Su et al. [74] Here, degrees of polymerization are unspecified; therefore, the molecular fragment weights are the defaults (see Figure 4f). Similarity calculations for graft polymers with specified degrees of polymerization are included in the Supporting Information.

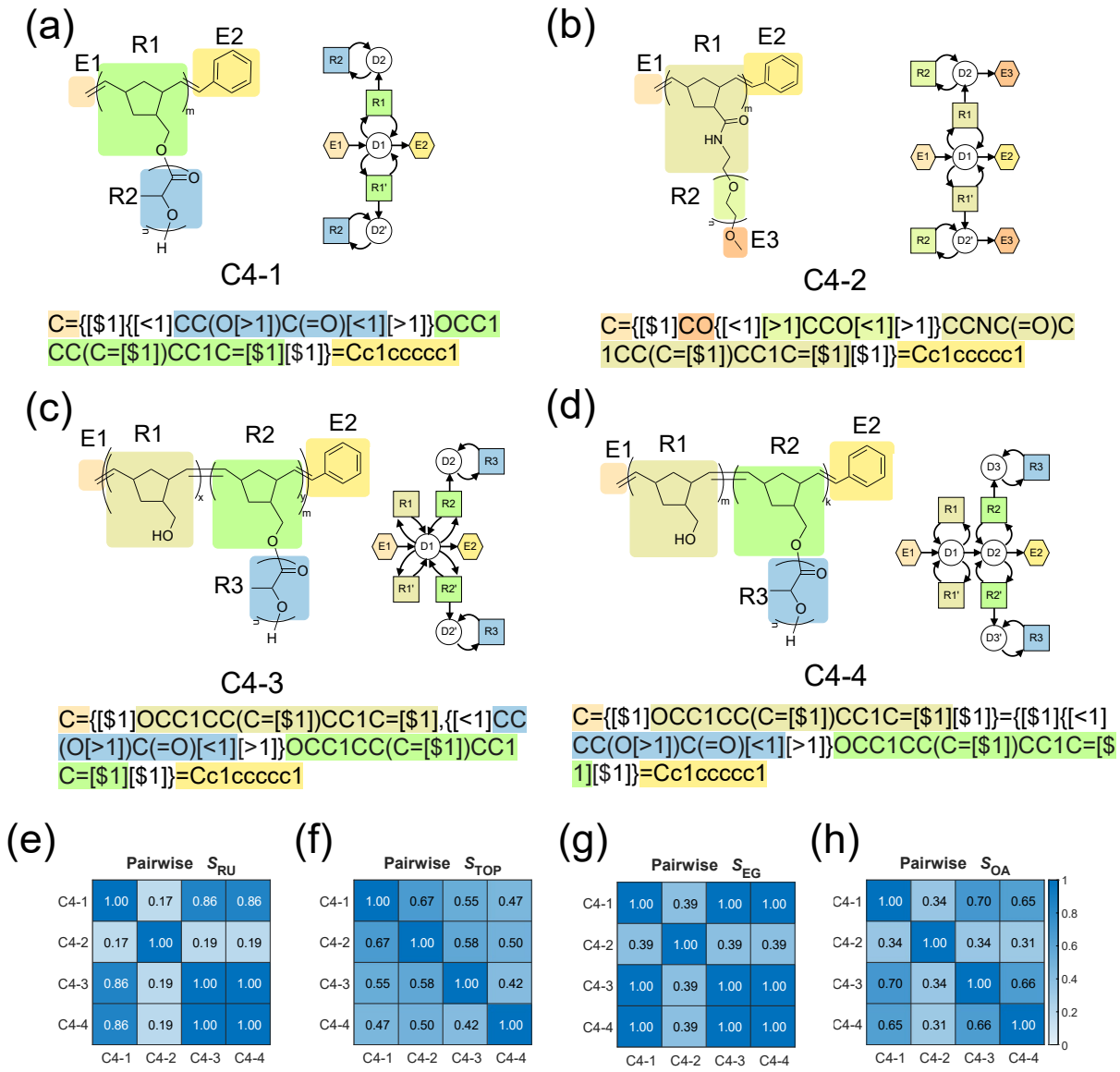


Figure 8: Four graft polymers, the corresponding canonical BigSMILES and stochastic graph representations. (a) C4-1 and (b) C4-2 are homo graft polymers but have different main chain repeat units and side chain repeat units. (c) C4-3 is a random graft copolymer, where one monomer on the main chain has a side chain, and the other monomer on the backbone does not have a side

chain. (d) C4-4 is a diblock graft copolymer, where one monomer on the main chain has a side chain but the other monomer does not. The polymerization degrees are not specified for all five graft polymers, so the molecular fragment weights are the defaults in Figure 4f. (e) Pairwise repeat ensemble similarity S_{RU}, (f) topological similarity S_{TOP}, (g) repeat ensemble similarity S_{EG} and (h) overall similarity S_{OA} for four graft polymers in Case 4.

Pairwise similarity scores are shown in Figure 8e-g. C4-1 and C4-2 have different repeat units on both backbones and side chains, but have similar topology graphs. C4-1 and C4-3 have a significant overlapping on the repeat units, but have more different topologies. Therefore, from chemical intuition, $S_{\mathrm{RU}}(\text{C4-1, C4-2}) < S_{\mathrm{RU}}(\text{C4-1, C4-3})$ and $S_{\mathrm{TOP}}(\text{C4-1, C4-2}) > S_{\mathrm{TOP}}(\text{C4-1, C4-3})$. The quantitative scores in Figure 8e,f are consistent with this chemical intuition. Using the default weights, the overall similarity score $S_{\mathrm{OA}}(\text{C4-1, C4-2}) < S_{\mathrm{OA}}(\text{C4-1, C4-3})$ due to the larger difference in S_{RU} for the pair C4-1 and C4-2. C4-3 and C4-4 have the same repeat units and end groups; only C4-3 is a random copolymer whereas C4-4 is a block copolymer. Taking C4-1 as a reference, $S_{\mathrm{RU}}(\text{C4-1, C4-3}) = S_{\mathrm{RU}}(\text{C4-1, C4-4})$, but $S_{\mathrm{TOP}}(\text{C4-1, C4-3}) > S_{\mathrm{TOP}}(\text{C4-1, C4-4})$. Thus, the order of the overall similarity score is $S_{\mathrm{OA}}(\text{C4-1, C4-3}) > S_{\mathrm{OA}}(\text{C4-1, C4-4})$. This is equivalent to the statement that a homopolymer is closer to a random copolymer than a diblock polymer assuming the same repeat units. While the above examples follow intuition, there are many other examples for which a clear, intuitive answer does not exist. The method presented here provides a quantitative similarity score for all cases and when available, is consistent with intuition.

Case 5: Segmented Polymers

Examples of segmented polyurethanes (see Figure 9a-d) are collected from Szczepańczyk et al. [75] with the symmetric isocyanates and chain extenders modified to be asymmetric to clarify the topological graphs shown in Figure 9e, specifically, that R1 and its mirror R1' are chemically distinct. For simplicity, it is assumed that the degrees of polymerization (x, y, z, n) are not specified, but calculations including degrees of polymerization are included in the Supporting Information. The comparison between C5-1 and C5-2 quantifies the impact of changing isocyanate, the comparison between C5-1 and C5-3 quantifies the impact of changing polyol, and the comparison between C5-1 and C5-4 quantifies the impact of changing the chain extenders (see Figure 9a-d). Since the weight of the repeat unit in the macromonomer is larger than the weights of backbone repeat units, changing the repeat units in the macromonomer leads to a larger effect on similarity. Thus, taking C5-1 as reference, C5-3 is the least similar (see Figure 9f,g). Additionally, for C5-2, C5-3 and C5-4, each of them only has one different component compared to C5-1, while each of them has two different components from the other two. For example, C5-2 has different R1 from C5-1, while C5-2 has different R1 and R3 from C5-3, and C5-2 has different R1 and R2 from C5-4. Therefore, for each of C5-2, C5-3 and C5-4, the similarity score with C5-1 is always larger than the pairwise similarity score with the other two. For instance, the overall similarity

S_{OA}(C5-2, C5-1) > S_{OA}(C5-2, C5-3) and S_{OA}(C5-2, C5-1) > S_{OA}(C5-2, C5-4) (see Figure 9g). These results are consistent with the chemical intuition.

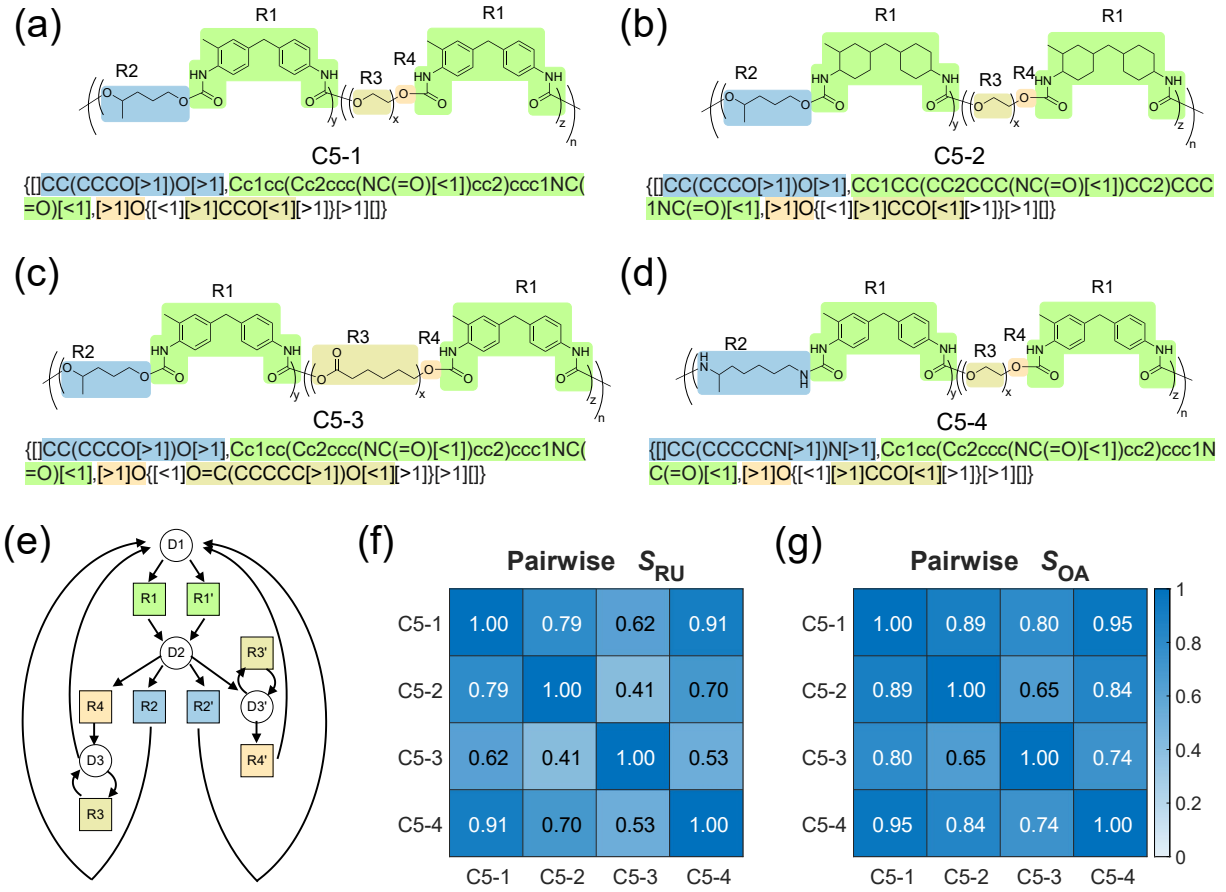


Figure 9: Four segmented polymers and the corresponding canonical BigSMILES. (a) C5-1 and (b) C5-2 have different isocyanates. C5-1 and (c) C5-3 have different polyols. C5-1 and (d) C5-4 have different chain extenders. (e) is the stochastic graph representation of all four segmented polyurethanes. (f) Pairwise repeat ensemble similarity S_{RU}, (g) overall similarity S_{OA} for four segmented polymers in Case 5.

Case 6: Unspecified Chemical Groups

In some cases, molecular fragments have variable groups, commonly called "R-groups", shown in Figure 10. The similarity calculation first identifies the functional groups or chains that R-groups represent and then takes only other remaining molecular fragment structures into consideration; therefore, polymer C6-1 has a similarity of 1 with all other polymers illustrated in Figure 10.

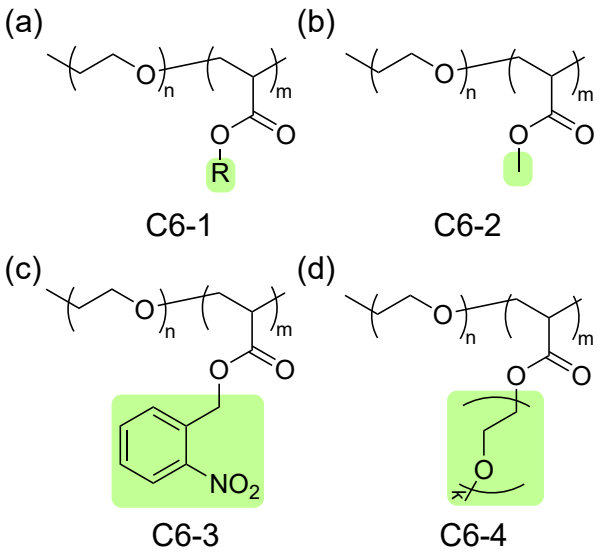


Figure 10: (a) Polymer C6-1, a polymer with an unspecified "R-group". (b) Polymer C6-2, (c) Polymer C6-3, and (d) Polymer C6-4 are all possible polymer candidates that Polymer C6-1 represents.

Fine-Ranking Targets When the Same Overall Similarity Score Occurs

One widespread use for similarity scores is to rank target molecules with respect to their similarity to a query molecule. The similarity methods developed herein are suitable for this purpose, but in many cases, such as the prior examples Figure 10 and Figure S2b, ties are possible. If two targets' similarity scores S_{OA} are the same, tiebreaking rules, such as those listed below, may be implemented in order to produce a single preferred ordinal list of similarities.

1. For polymers with the same overall similarity score S_{OA}, prioritize component similarity scores, S_{RU}, S_{TOP} and S_{EG}, in the order of their weights. For instance, in Figure S2b where S_{OA}(C3-1,C3-2) = S_{OA}(C3-1, C3-3) at the weight setting $W_{\mathrm{RU}} = 0.53$, $W_{\mathrm{TOP}} = 0.42$ and $W_{\mathrm{EG}} = 0.05$, so fine-ranking is carried out by prioritizing S_{RU} yielding C3-3 before C3-2 when taking C3-1 as reference.

2. For pairs that are still tied, rank according to the total number of heavy atoms in the canonicalized BigSMILES strings. Targets with a larger number of heavy atoms occupy the higher priority order. Figure 10, for example, the results of similarity order is C6-3 > C6-4 > C6-2 when taking C6-1 as reference.

3. If the total number of heavy atoms is tied, rankings may be performed by individual atom types in the order of atoms with larger atomic numbers.

4. Finally, alphabetized canonicalized BigSMILES[39] can break any remaining ties.

Areas for Future Development

One key application of our work is ranking; for this application, a computationally efficient algorithm is essential. Thus, several compromises were made to ensure that the methods developed here can be immediately used. Specifically, repeat units, end groups, and topology are separated, and the nodes' chemical details in topology are ignored in GED calculation, resulting in a loss of the chemical connectivity between nodes. For example, this method gives a similarity score of one for ABC and ACB triblock copolymers. For cases where these fine-grained distinctions matter and computational speed can be compromised, this limitation can be solved by including the nodes' chemical details in the GED calculation.[30] Another simplification is that only the average frequencies of the repeat units based on their average polymerization degrees are used in the EMD calculation; thus, the dispersity is ignored. For instance, EMD cannot distinguish a RAFT four-arm star polystyrene with equal arm length and a RAFT four-arm star polystyrene with various arm lengths for each arm[76,77] where the sums of the four arm-length of these two polymers are equal. EMD cannot distinguish a random copolymer and a gradient polymer which have the same repeat units and compositions since BigSMILES representations which are used to generate the stochastic graph representations cannot distinguish them. Additionally, EMD cannot distinguish bottlebrush polymers with hourglass, football, bowtie, and sphere architecture profiles for the graft side chains[55] where the sums of their whole graft side chain length are the same. Again, this simplification ensures the method is computationally efficient.

Another limitation of this work is that it requires a canonical BigSMILES to generate a deterministic stochastic graph. Since the current BigSMILES canonicalization from Lin et al.[39] is limited to linear polymers and thus cannot handle network polymers and branched polymers, the method only applies to polymers with a well-defined backbone. Without canonicalization, multiple graph representations and monomer sets are possible for a single polymer, which could lead to a similarity score smaller than one even when two polymers are identical. Once a canonicalization method for branched and network cases is available, they can be implemented using the same methods described herein.

Finally, tacticity has significant effects on the physical properties of polymers, such as crystallization, melting temperature, solubility, and mechanical properties; however, the treatment of tacticity by fingerprinting algorithms can cause challenges for similarity scoring. The influence of tacticity on pairwise similarity calculation is studied in the Supporting Information, using an example of four polypropylenes with the pure head-to-tail configuration and different tacticities (two stereoisomers of isotactic polypropylene, syndiotactic polypropylene, and atactic polypropylene). The results show that the two stereoisomers of isotactic polypropylene have the highest similarity, and isotactic polypropylene and syndiotactic polypropylene are closer to each other when compared to atactic polypropylene, which is chemically intuitive and constant with the crystallinity and melting temperature. However, the Morgan fingerprint treats the two

stereoisomers of isotactic polypropylene asymmetrically and overly differently, which results in two areas for further improvement. Firstly, the similarity scores between the two stereoisomers of isotactic polypropylene and syndiotactic polypropylene are found to be similar but not identical, contrary to the expected chemical intuition. Secondly, the similarity between the two stereoisomers of isotactic polypropylene is expected to be closer to one. One potential solution is to develop a different embedding method for molecular fragments, which can treat stereoisomers symmetrically and with less differentiation, but that is beyond the scope of this work.

Conclusion

This work quantitively calculates pairwise chemical similarity by first developing the polymers' stochastic graph representation and then utilizing two similarity measurements, Earth Mover's Distance (EMD) and Graph Edit Distance (GED). The EMD metric captures the similarity of repeat units and end groups by computing the similarity score between individual molecular fragments according to their chemical structures building on current methods for small molecular similarity calculations. EMD preserves the molecular fragments' chemical characteristics better than simply averaging or summing the fingerprints. The GED metric captures the topological similarity to illustrate how the two polymers are similar in their topological connections. A series of cases illustrate the flexibility and utility of this method across a wide range of polymer chemistries. While there is no ground truth for polymer similarity, the method produces results that are consistent with chemical intuition across all explored cases.

The similarity metric proposed herein gives a solution to calculate the chemical pairwise similarity score, which enables the sorting of retrieved database entries based on a query polymer, as well as the detection of abnormal data for polymer data validation. Additionally, the quantitative similarity scores can be used to cluster or catalog polymer data and improve polymer discovery. Therefore, this method is an essential contribution to the field of polymer informatics.

Code Availability

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

Acknowledgement

This work was primarily funded by the National Science Foundation Convergence Accelerator award number ITE-2134795. We acknowledge the discussion with Melody Morris, Haley Beech, Katharina Fransen, Natalie Mamrol, Sarah Av-Ron, Alexis Hocken, Ameya Rao, Devosmita Sen, Clara Troyano-Valls, and Alex Zappi, especially on the suggestions of polishing figures.

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) 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. *J Clean Prod* **2021**, *282*, 124444. https://doi.org/10.1016/J.JCLEPRO.2020.124444.

(2) Norton, M. Tackling the Challenge of Packaging Plastic in the Environment. *Chemistry: A European Journal* **2020**, *26* (35), 7737–7739. https://doi.org/10.1002/chem.202001890.

(3) 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. https://doi.org/10.1021/ma1010099.

(4) Yadav, R.; Tirumali, M.; Wang, X.; Naebe, M.; Kandasubramanian, B. Polymer Composite for Antistatic Application in Aerospace. *Defence Technology* **2020**, *16* (1), 107–118. https://doi.org/10.1016/j.dt.2019.04.008.

(5) Stenzel, M. H. Glycopolymers for Drug Delivery: Opportunities and Challenges. *Macromolecules* **2022**, *55* (12), 4867–4890. https://doi.org/10.1021/acs.macromol.2c00557.

(6) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. *Proceedings - 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011* **2011**, 22–29. https://doi.org/10.1109/EIDWT.2011.13.

(7) Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J Chem Inf Model* **2020**, *60* (10), 4684–4690. https://doi.org/10.1021/acs.jcim.0c00726.

(8) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-Learning Predictions of Polymer Properties with Polymer Genome. *J Appl Phys* **2020**, *128* (17). https://doi.org/10.1063/5.0023759.

(9) *MaterialsMine: An open-source, user-friendly materials data resource guided by FAIR principles | Tetherless World Constellation*. https://tw.rpi.edu/project/materialsmine-open-source-user-friendly-materials-data-resource-guided-fair-principles.

(10) Kim, S.; Schroeder, C. M.; Jackson, N. E. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. *ACS Polymers Au* **2023**. https://doi.org/10.1021/ACSPOLYMERSAU.3C00003.

(11) 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.; Olsen, B. D. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. *ACS Cent Sci* **2023**. https://doi.org/10.1021/acscentsci.3c00011.

(12) Sha, W.; Li, Y.; Tang, S.; Tian, J.; Zhao, Y.; Guo, Y.; Zhang, W.; Zhang, X.; Lu, S.; Cao, Y.; Cheng, S. Machine Learning in Polymer Informatics. *InfoMat* **2021**, *3* (4), 353–361. https://doi.org/10.1002/inf2.12167.

(13) Hatakeyama-Sato, K. Recent Advances and Challenges in Experiment-Oriented Polymer Informatics. *Polymer Journal 2022 55:2* **2022**, *55* (2), 117–131. https://doi.org/10.1038/S41428-022-00734-9.

(14) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci Adv* **2022**, *8* (29), 9545. https://doi.org/10.1126/sciadv.abn9545.

(15) Tao, L.; Arbaugh, T.; Byrnes, J.; Varshney, V.; Li, Y. Unified Machine Learning Protocol for Copolymer Structure-Property Predictions. *STAR Protoc* **2022**, *3* (4), 101875. https://doi.org/10.1016/J.XPRO.2022.101875.

(16) 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 Lett* **2020**, *9* (8), 1086–1094. https://doi.org/10.1021/acsmacrolett.0c00264.

(17) Brinson, L. C.; Bartolo, L. M.; Blaiszik, B.; Elbert, D.; Foster, I.; Strachan, A.; Voorhees, P. W. Community Action on FAIR Data Will Fuel a Revolution in Materials Research. *MRS Bulletin 2023* **2023**, 1–5. https://doi.org/10.1557/S43577-023-00498-4.

(18) Hu, B.; Lin, A.; Brinson, L. C. ChemProps: A RESTful API Enabled Database for Composite Polymer Name Standardization. *J Cheminform* **2021**, *13* (1), 22. https://doi.org/10.1186/s13321-021-00502-6.

(19) 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. https://doi.org/10.1007/s10618-020-00733-5.

(20) 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. https://doi.org/10.1016/j.mser.2020.100595.

(21) Wang, X.; Li, Z.; Jiang, M.; Wang, S.; Zhang, S.; Wei, Z. Molecule Property Prediction Based on Spatial Graph Embedding. *J Chem Inf Model* **2019**, *59* (9), 3817–3828. https://doi.org/10.1021/acs.jcim.9b00410.

(22) Wang, H.; Kaddour, J.; Liu, S.; Tang, J.; Kusner, M.; Lasenby, J.; Liu, Q.; Contribution, E. Evaluating Self-Supervised Learning for Molecular Graph Embeddings. **2022**. https://doi.org/10.48550/arxiv.2206.08005.

(23) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. PMLR November 21, 2020, pp 8818–8827. https://proceedings.mlr.press/v119/shi20d.html (accessed 2022-12-16).

(24) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71* (C), 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

(25) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J Cheminform* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

(26) Öztürk, H.; Ozkirimli, E.; Özgür, A. A Comparative Study of SMILES-Based Compound Similarity Functions for Drug-Target Interaction Prediction. *BMC Bioinformatics* **2016**, *17* (1), 128. https://doi.org/10.1186/s12859-016-0977-x.

(27) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org Biomol Chem* **2004**, *2* (22), 3204. https://doi.org/10.1039/b409813g.

(28) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J Med Chem* **2014**, *57* (8), 3186–3204. https://doi.org/10.1021/jm401411z.

(29) 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. *Comput Struct Biotechnol J* **2021**, *19*, 1541–1556. https://doi.org/10.1016/j.csbj.2021.03.004.

(30) Mohapatra, S.; An, J.; Gómez-Bombarelli, R. Chemistry-Informed Macromolecule Graph Representation for Similarity Computation, Unsupervised and Supervised Learning. *Mach Learn Sci Technol* **2022**, *3* (1). https://doi.org/10.1088/2632-2153/ac545e.

(31) Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. Targeted Sequence Design within the Coarse-Grained Polymer Genome. *Sci Adv* **2020**, *6* (43). https://doi.org/10.1126/sciadv.abc6216.

(32) Patel, R. A.; Borca, C. H.; Webb, M. A. Featurization Strategies for Polymer Sequence or Composition Design by Machine Learning. *Mol Syst Des Eng* **2022**, *7* (6), 661–676. https://doi.org/10.1039/D1ME00160D.

(33) Shi, J.; Quevillon, M. J.; Amorim Valença, P. H.; Whitmer, J. K. Predicting Adhesive Free Energies of Polymer–Surface Interactions with Machine Learning. *ACS Appl Mater Interfaces* **2022**, *14* (32), 37161–37169. https://doi.org/10.1021/acsami.2c08891.

(34) 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. https://doi.org/10.1039/D1SM01012C.

(35) Statt, A.; Casademunt, H.; Brangwynne, C. P.; Panagiotopoulos, A. Z. Model for Disordered Proteins with Strongly Sequence-Dependent Liquid Phase Behavior. *J Chem Phys* **2020**, *152* (7), 075101. https://doi.org/10.1063/1.5141095.

(36) Bhattacharya, D.; Kleeblatt, D. C.; Statt, A.; Reinhart, W. F. Predicting Aggregate Morphology of Sequence-Defined Macromolecules with Recurrent Neural Networks. *Soft Matter* **2022**, *18* (27), 5037–5051. https://doi.org/10.1039/d2sm00452f.

(37) Huo, Z.; Arora, S.; Kong, V. A.; Myrga, B. J.; Statt, A.; Laaser, J. E. Effect of Polymer Composition and Morphology on Mechanochemical Activation in Nanostructured Triblock Copolymers. **2022**. https://doi.org/10.26434/CHEMRXIV-2022-65Z24.

(38) 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**, *16*, 4. https://doi.org/10.1021/ACS.JCTC.2C01314.

(39) Lin, T.-S.; Rebello, N. J.; Lee, G.-H.; Morris, M. A.; Olsen, B. D. Canonicalizing BigSMILES for Polymers with Defined Backbones. *ACS Polymers Au* **2022**, *2* (6), 486–500. https://doi.org/10.1021/acspolymersau.2c00009.

(40) Aldeghi, M.; Coley, C. W. A Graph Representation of Molecular Ensembles for Polymer Property Prediction. *Chem Sci* **2022**, *13* (35), 10486–10498. https://doi.org/10.1039/d2sc02839e.

(41) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-Learning Predictions of Polymer Properties with Polymer Genome. *J Appl Phys* **2020**, *128* (17). https://doi.org/10.1063/5.0023759.

(42) Kuenneth, C.; Ramprasad, R. PolyBERT: A Chemical Language Model to Enable Fully Machine-Driven Ultrafast Polymer Informatics. **2022**.

(43) 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. https://doi.org/10.1021/acs.chemmater.1c02061.

(44) Kuenneth, C.; Schertzer, W.; Ramprasad, R. Copolymer Informatics with Multitask Deep Neural Networks. *Macromolecules* **2021**, *54* (13), 5957–5961. https://doi.org/10.1021/acs.macromol.1c00728.

(45) Chen, L.; Kern, J.; Lightstone, J. P.; Ramprasad, R. Data-Assisted Polymer Retrosynthesis Planning. *Appl Phys Rev* **2021**, *8* (3). https://doi.org/10.1063/5.0052962.

(46) Guo, M.; Shou, W.; Makatura, L.; Erps, T.; Foshey, M.; Matusik, W. Polygrammar: Grammar for Digital Polymer Representation and Generation. *Advanced Science* **2022**, *9* (23), 2101864. https://doi.org/10.1002/advs.202101864.

(47) 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. https://doi.org/10.1021/acs.chemmater.0c03381.

(48) Orlova, D. Y.; Zimmerman, N.; Meehan, S.; Meehan, C.; Waters, J.; Ghosn, E. E. B.; Filatenkov, A.; Kolyagin, G. A.; Gernez, Y.; Tsuda, S.; Moore, W.; Moss, R. B.; Herzenberg, L. A.; Walther, G. Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLoS One* **2016**, *11* (3), e0151859. https://doi.org/10.1371/journal.pone.0151859.

(49) Alvarez-Melis, D.; Fusi, N. Geometric Dataset Distances via Optimal Transport. *Adv Neural Inf Process Syst* **2020**, *2020-December*. https://doi.org/10.48550/arxiv.2002.02923.

(50) Sanfeliu, A.; systems, K. F.-I. transactions on; man, undefined; and, undefined; 1983, undefined. A Distance Measure between Attributed Relational Graphs for Pattern Recognition. *ieeexplore.ieee.org*.

(51) Bai, Y.; Ding, H.; Bian, S.; Chen, T.; Sun, Y.; Wang, W. SimGNN: A Neural Network Approach to Fast Graph Similarity Computation. **2018**.

(52) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent Sci* **2019**, *5* (9), 1523–1531. https://doi.org/10.1021/acscentsci.9b00476.

(53) 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.; Kalow, J. A.; Craig, S. L.; Nelson, A.; Moore, J. S.; Olsen, B. D. Extending BigSMILES to Non-Covalent Bonds in Supramolecular Polymer Assemblies. *Chem Sci* **2022**, *13* (41), 12045–12055. https://doi.org/10.1039/D2SC02257E.

(54) Xu, X.; Douglas, J. F.; Xu, W.-S. Influence of Side-Chain Length and Relative Rigidities of Backbone and Side Chains on Glass Formation of Branched Polymers. *Macromolecules* **2021**, *54* (13), 6327–6341. https://doi.org/10.1021/acs.macromol.1c00834.

(55) Walsh, D. J.; Dutta, S.; Sing, C. E.; Guironnet, D. Engineering of Molecular Geometry in Bottlebrush Polymers. *Macromolecules* **2019**, *52* (13), 4847–4857. https://doi.org/10.1021/acs.macromol.9b00845.

(56) 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. https://doi.org/10.1021/ci00057a005.

(57) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J Chem Inf Comput Sci* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.

(58) *RDKit*. https://www.rdkit.org/.

(59) *Pyomo*. http://www.pyomo.org/.

(60) 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 Optimization and Its Applications; Springer International Publishing: Cham, 2017; Vol. 67. https://doi.org/10.1007/978-3-319-58821-6.

(61) 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. https://doi.org/10.1007/S12532-011-0026-8.

(62) *coin-or/Cbc: COIN-OR Branch-and-Cut solver*. https://github.com/coin-or/Cbc.

(63) *coin-or/Cbc: Release releases/2.10.8 | Zenodo*. https://zenodo.org/record/6522795#.Y5d2fH3MIQ8.

(64) Serratosa, F. Redefining the Graph Edit Distance. *SN Comput Sci* **2021**, *2* (6), 438. https://doi.org/10.1007/s42979-021-00792-5.

(65) Sanfeliu, A.; Fu, K.-S. A Distance Measure between Attributed Relational Graphs for Pattern Recognition. *IEEE Trans Syst Man Cybern* **1983**, *SMC-13* (3), 353–362. https://doi.org/10.1109/TSMC.1983.6313167.

(66) Garcia-Hernandez, C.; Fernández, A.; Serratosa, F. Ligand-Based Virtual Screening Using Graph Edit Distance as Molecular Similarity Measure. *J Chem Inf Model* **2019**, *59* (4), 1410–1421. https://doi.org/10.1021/acs.jcim.8b00820.

(67) Gao, X.; Xiao, B.; Tao, D.; Li, X. A Survey of Graph Edit Distance. *Pattern Analysis and Applications* **2010**, *13* (1), 113–129. https://doi.org/10.1007/s10044-008-0141-y.

(68) Garcia-Hernandez, C.; Fernández, A.; Serratosa, F. Learning the Edit Costs of Graph Edit Distance Applied to Ligand-Based Virtual Screening. *Curr Top Med Chem* **2020**, *20* (18), 1582–1592. https://doi.org/10.2174/1568026620666200603122000.

(69) Ibragimov, R.; Malek, M.; Baumbach, J.; Guo, J. Multiple Graph Edit Distance - Simultaneous Topological Alignment of Multiple Protein-Protein Interaction Networks with an Evolutionary Algorithm. *GECCO 2014 - Proceedings of the 2014 Genetic and Evolutionary Computation Conference* **2014**, 277–283. https://doi.org/10.1145/2576768.2598390.

(70) Ibragimov, R.; Malek, M.; Guo, J.; Baumbach, J. GEDEVO: An Evolutionary Graph Edit Distance Algorithm for Biological Network Alignment. *DROPS-IDN/4229* **2013**, *34*, 68–79. https://doi.org/10.4230/OASICS.GCB.2013.68.

(71) Shim, J.; Bates, F. S.; Lodge, T. P. Superlattice by Charged Block Copolymer Self-Assembly. *Nat Commun* **2019**, *10* (1), 1–7. https://doi.org/10.1038/s41467-019-10141-z.

(72) Altintas, O.; Abbasi, M.; Riazi, K.; Goldmann, A. S.; Dingenouts, N.; Wilhelm, M.; Barner-Kowollik, C. Stability of Star-Shaped RAFT Polystyrenes under Mechanical and Thermal Stress. *Polym Chem* **2014**, *5* (17), 5009–5019. https://doi.org/10.1039/C4PY00484A.

(73) Zayas, H. A.; Truong, N. P.; Valade, D.; Jia, Z.; Monteiro, M. J. Narrow Molecular Weight and Particle Size Distributions of Polystyrene 4-Arm Stars Synthesized by RAFT-Mediated Miniemulsions. *Polym Chem* **2013**, *4* (3), 592–599. https://doi.org/10.1039/C2PY20709E.

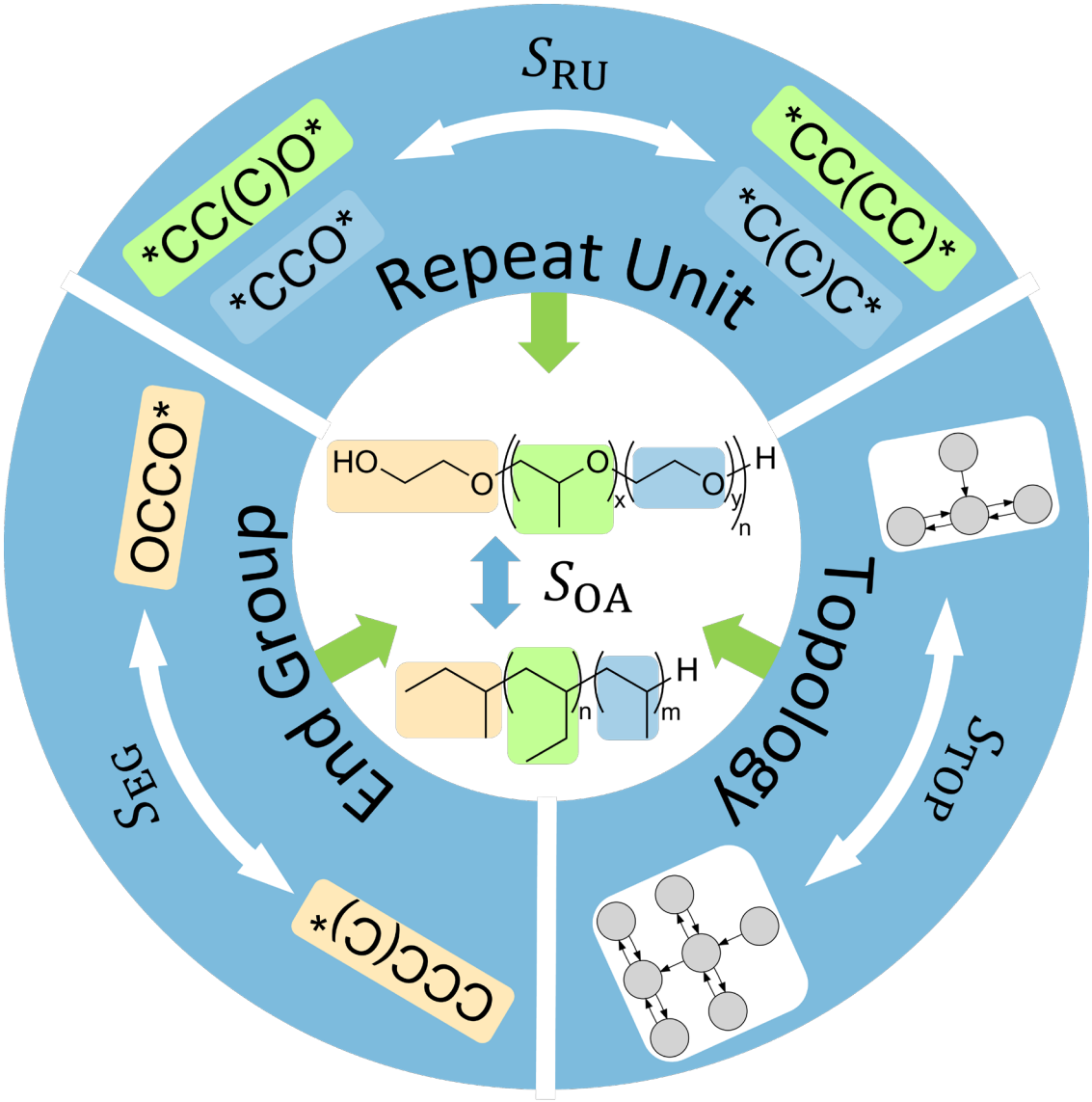
(74) Su, L.; Heo, G. S.; Lin, Y.; Dong, M.; Zhang, S.; Chen, Y.; Sun, G.; Wooley, K. L. Syntheses of Triblock Bottlebrush Polymers through Sequential ROMPs: Expanding the Functionalities of Molecular Brushes. *J Polym Sci A Polym Chem* **2017**, *55* (18), 2966–2970. https://doi.org/10.1002/pola.28647.

(75) Szczepańczyk, P.; Szlachta, M.; Złocista-Szewczyk, N.; Chłopek, J.; Pielichowska, K. Recent Developments in Polyurethane-Based Materials for Bone Tissue Engineering. *Polymers* **2021**, *13* (6), 946. https://doi.org/10.3390/polym13060946.

(76) Wu, W.; Wang, W.; Li, J. Star Polymers: Advances in Biomedical Applications. *Prog Polym Sci* **2015**, *46*, 55–85. https://doi.org/10.1016/j.progpolymsci.2015.02.002.

(77) Higashihara, T.; Hayashi, M.; Hirao, A. Synthesis of Well-Defined Star-Branched Polymers by Stepwise Iterative Methodology Using Living Anionic Polymerization. *Prog Polym Sci* **2011**, *36* (3), 323–375. https://doi.org/10.1016/j.progpolymsci.2010.08.001.

TOC


S_{RU}
CC(C)O
CCO
Repeat Unit
CC(CC)
C(C)C
OCCO*
End Group
S_{EG}
CCC(C)*
S_{OA}
Topology
S_{TOP}
HO
O
O
O
H
x
y
n
n
m
H