Data-assisted polymer retrosynthesis planning

Cite as: Appl. Phys. Rev. **8**, 031405 (2021); https://doi.org/10.1063/5.0052962
Submitted: 02 April 2021 . Accepted: 14 June 2021 . Published Online: 15 July 2021

Lihua Chen, Joseph Kern, Jordan P. Lightstone, and Rampi Ramprasad

COLLECTIONS

Paper published as part of the special topic on Autonomous (AI-driven) Materials Science

F This paper was selected as Featured

SCI This paper was selected as Scilight

View Online Export Citation CrossMark

Data-assisted polymer retrosynthesis planning

Cite as: Appl. Phys. Rev. **8**, 031405 (2021); doi: 10.1063/5.0052962
Submitted: 2 April 2021 · Accepted: 14 June 2021 ·
Published Online: 15 July 2021

Lihua Chen, Joseph Kern, Jordan P. Lightstone, and Rampi Ramprasad[a)]

AFFILIATIONS

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, USA

Note: This paper is part of the special collection on Autonomous (AI-driven) Materials Science.
[a)]**Author to whom correspondence should be addressed:** rampi.ramprasad@mse.gatech.edu

ABSTRACT

Polymer informatics is being utilized to accelerate polymer discovery. However, the practical realization of the designed polymer is still slow due to synthesis challenges, e.g., difficulties with the identification of potential polymerization mechanisms and optimal reactants/solvents/processing conditions. In the past, synthesis pathways adopted for a target polymer have been heavily dependent on chemical intuition and past experience. To expedite this process, we have developed a data-driven approach to assist in polymer retrosynthesis planning. In this work, a dataset of polymerization reactions was manually accumulated from various resources to extract hundreds of synthetic templates and used as the training set. Further, a similarity metric was adopted to select synthetic templates and similar existing reactants for the new target polymer. Finally, prediction accuracy was measured by comparison with ground truth and/or bench chemists' estimation. The proposed data-driven polymer synthesis recommendation model has been deployed at https://www.polymergenome.org.

I. INTRODUCTION

Polymer informatics approaches are beginning to significantly impact polymer discovery and design.[1–9] Machine learning models for instant property prediction[10,11] are becoming widely available, and advanced design algorithms for generating polymers that meet target property requirements are being actively developed.[12–16] Nevertheless, the next great challenge is polymer synthesis, i.e., "how to make the machine-designed polymers in the lab?" Traditionally, chemical intuition and past experience of chemists steer the design of optimal synthesis strategies for target polymers, such as the determination of optimal reactants, reagents, and processing conditions.[17] If such information and insights pertaining to polymer synthesis strategies can be captured and encoded in a digital framework, and extrapolated for new polymers, the practical realization of new useful polymers may be significantly accelerated.

In the past several decades, computer-assisted retrosynthesis planning has matured in the field of molecular and drug discovery.[18–27] Template-based[22,28] or template-free[23,24,27] machine learning approaches and robotic platforms[26] have been developed for the autonomous synthesis of organic compounds (see more details in Refs. 29 and 30). However, these approaches are still in a state of infancy for polymers, because of unique challenges associated with building data-driven approaches for polymer retrosynthesis planning. Unlike for molecules, only a few polymerization databases are available, such as PolyInfo[8] and NIST Synthetic Polymer MALDI Recipes database.[31] Nevertheless, extensive efforts are required to preprocess the source data for data-driven approaches. Furthermore, reaction templates are non-existent. Additionally, polymers are macromolecules, formed by linking small monomer molecules together. This synthesis processes involve various polymerization mechanisms, multiple steps, and multiple possible linkages between monomers, which is different from the one-step synthesis of small molecules.

In this work, we propose a novel data-driven approach to automate polymer retrosynthesis planning, motivated by the common first question "how have similar polymers been synthesized before?" In this approach, as illustrated in Fig. 1, first, a large dataset of polymerization reactions was accumulated to extract synthesis templates that interpret the chemical reactions between reactant monomers. Second, the dataset and templates were applied to build a polymer retrosynthesis framework to automatically predict the synthesis pathways for new target polymers. In this framework, we employed a similarity metric to select the synthesis templates for target polymers and identify similarly existing reactants. Finally, this approach was demonstrated by testing on previously synthesized polymers, and the prediction accuracy was estimated by comparison with the ground truth (if available with our dataset) or evaluation by bench chemists. In general, the model performance exceeded a precision of 80%. Additionally, we have deployed the developed model at the online Polymer Genome platform (www.polymergenome.org), providing a user-friendly, efficient, and accurate way to predict synthesis paths for new target polymers.

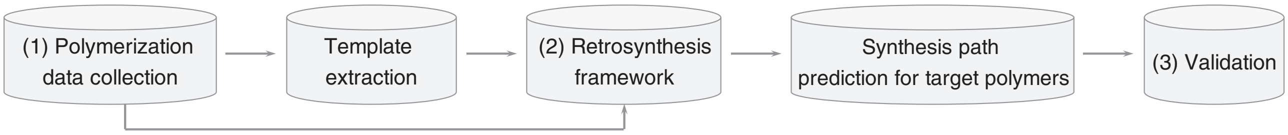

(1) Polymerization data collection
Template extraction
(2) Retrosynthesis framework
Synthesis path prediction for target polymers
(3) Validation

FIG. 1. Polymer retrosynthesis planning via data-driven approach: (1) collecting polymerization data to extract synthesis templates that interpret the chemical reactions between reactant monomers; (2) development of a polymer retrosynthesis framework using the polymerization dataset and templates to automatically predict synthesis pathways for target polymers; (3) validation of the developed approach.

Moving forward, this approach will be further refined by expanding the dataset of polymerization reactions and incorporating other related synthesis information, such as temperature, catalysts, reagents, etc.

II. APPROACH

A. Dataset

Our dataset of polymerization reactions contains 11,448 previously reported polymerization paths, for 9748 homopolymers starting from 8921 reactant molecules (also called monomers). This dataset was manually collected from various resources, including online repositories[8] and published journal articles.[32–34] Three polymerization classes were considered, including condensation (7096 polymers), addition (2267 polymers), and ring-opening (551 polymers). Since many known polymerization paths (1–10) for each polymer are included, the total number of polymerization paths are higher than that of polymers. Furthermore, the polymers and reactant molecules are made up of 12 elements (i.e., C, H, B, O, N, S, P, Si, F, Cl, Br, and I) and a variety of polymer classes. In the present work, the role of other factors, such as solvents, catalysts, and experimental conditions, are neglected.

Additionally, 3582 previously synthesized polymers with unknown polymerization information were considered to validate the developed approach. Based on principal component analysis in Fig. S1 of the supplementary material, we note that this test set has similar chemical compositions with respect to our training dataset. It is worth pointing out that both the training and test datasets are composed of homopolymers and exclude copolymers, polymer blends, ladder, cross-linked, and metal-containing polymers.

B. Template extraction procedure

1. Representations

The synthesis pathways of polymers were encoded into a machine-readable format via reaction SMILES (Simplified Molecular-Input Line-Entry System),[35,36] in the generic form of *product* ≫ *reactants* as shown in Fig. 2. *product* is the final polymer and represented using a modified SMILES notation with [∗] denoting the connection point of repeat units.[11] For example, polyethylene terephthalate (PET) is represented as [∗]CCOC(=O)c1ccc(C(=O)O[∗])cc1. Depending on the polymerization mechanisms, *product* may contain one or two monomer repeat units. *reactants* are molecular monomers, which are described with regular SMILES and separated with "." in the reaction SMILES when the number of reactants is more than one.

In the synthesis templates, SMILES arbitrary target specification (SMARTS) patterns were utilized to represent the reacting atoms in the reaction SMILES, expressed as "[expr:*n*]." "expr" is any legal atomic expression as described below and *n* is the mapping index to track the reacting atoms in product and reactants. Generally, the simpler "expr" was used to represent a more general pattern. For instance, [C:2] represents any aliphatic carbon and 2 is the arbitrary index of C. However, to retain more chemical knowledge and/or accelerate the simulation time, two special rules were applied. First, adjacent atoms (less than 10) of reacting atoms are included in SMARTS to be distinct from other non-reacting parts in the product. Second, since polymers are formed by the reactions between functional groups of reactants, the whole functional group unit for most cases is captured using SMARTS, such as OH ([OH:1]) and C(=O)OH ([OH:3][C:2]=[O:5]) in Fig. 2(a).

2. Template extraction

The three common polymerization mechanisms are condensation, addition, and ring-opening, leading to a unique set of template extraction rules. As illustrated in Fig. 2, in condensation polymerization, used to create polymers such as PET, functional groups involved in molecular monomers react to form chemical linkages accompanied by the formation of byproducts like water. While for addition polymers [e.g., polyethylene (PE)], monomers are joined together without the formation of byproducts. Typically, condensation and addition polymerization require two symmetric/asymmetric type reactants to link together, resulting in two symmetric/asymmetric reacting parts (such as A and B in Fig. 2), thus leading a two-step synthesis process. Consequently, two same/different templates are obtained, which differ from the one-step synthesis of molecules. For instance, PET in Fig. 2(a) has the symmetric A and B reacting parts and the same templates, while PE in Fig. 2(b) needs two different templates to form the symmetric A and B parts (CC single bonds) by linking C=C monomers. More examples are shown in Fig. S2 of the supplementary material. In the case of ring-opening polymerization, the polymer chain is grown by the breakage of a bond in cyclic monomers, for instance, polyethylene oxide (PEO) formed from cyclic ether. As shown in Fig. 2(c), the reacting part A is the ring breakage position. Thus, only one type of reactant is required, and this polymerization may be viewed as a one-step synthesis process.

According to the polymerization mechanisms above and domain knowledge, the templates for 11,488 synthesis paths were extracted using the following rules:

(a) Preprocessing was performed to ensure uniform formats for products and reactants, such as canonicalization to generate clean reaction SMILES.
(b) Starting from reaction SMILES, the reacting atoms in the product and reactants are identified and represented using SMARTS patterns.

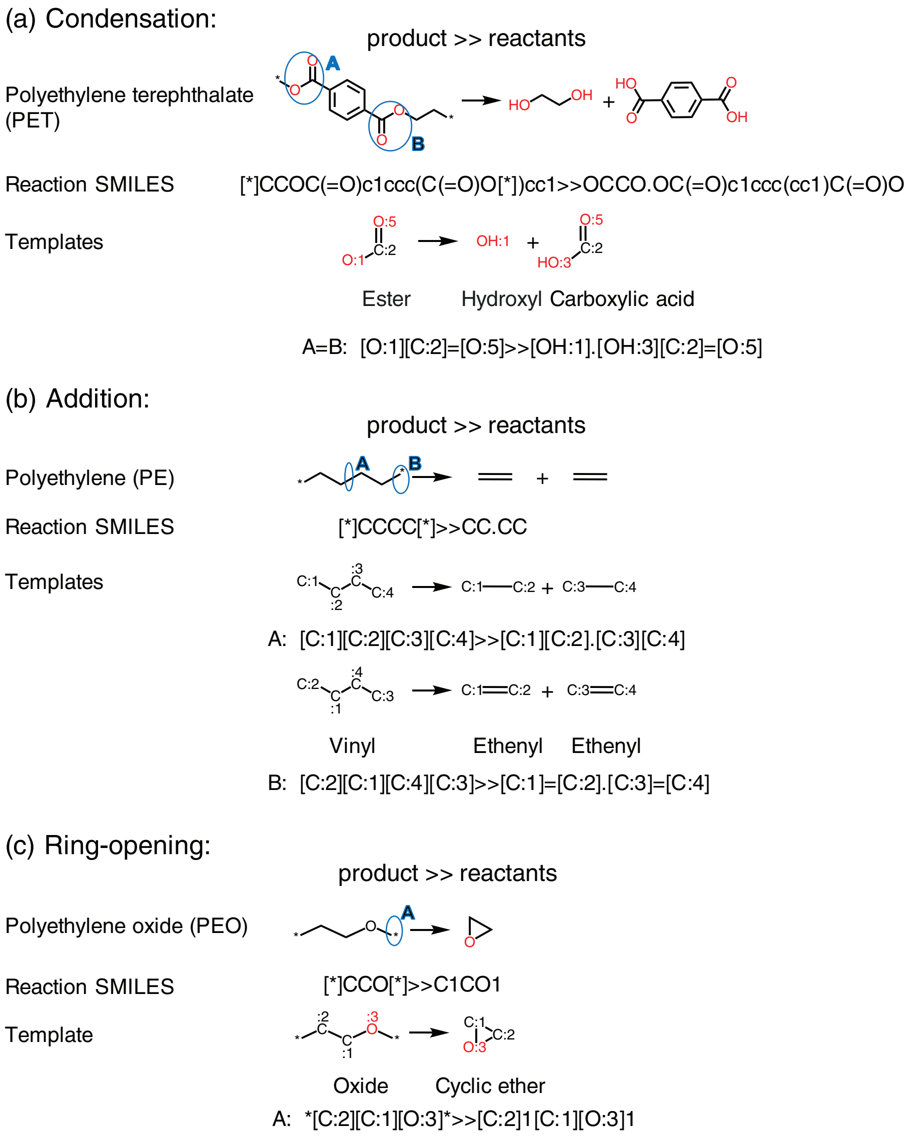


FIG. 2. Template extraction for condensation, addition, and ring-opening polymerization. The synthesis pathways of polymers were represented using reaction SMILES, in the generic form of *product ≫ reactants*. SMARTS patterns were utilized to represent reacting atoms in the synthesis templates. Condensation and addition polymers have two symmetric/asymmetric (A and B) reacting parts, resulting in two same (e.g., PET) or different (e.g., PE) A and B templates. Ring-opening polymerization has one reacting part (A) and one template, such as PEO.

(c) Assigning and matching mapping numbers to corresponding reacting SMARTS in product and reactants was performed next. Taking PET as an example, it is well known that the [O:1] within the ester group ([O:1][C:2]=[O:5]) is from the hydroxy group ([OH:1]) while [C:2]=[O:5] is from carboxylic acid ([OH:3][C:2]=[O:5]). In the templates, the mapping number is unique, but balanced reactions are not required, such as excluding byproducts.

(d) A synthetic template was then formulated next for each polymerization type, as shown in Fig. 2. As mentioned before, two-step synthesis is required for condensation and addition polymerization, resulting in two same or different templates. For instance, the same template [O:1][C:2]=[O:5] ≫ [OH:1].[OH:3][C:2]=[O:5] is obtained for PET in Fig. 2(a), representing that the ester group is formed by hydroxyl and carboxylic acid functional groups, while for PE in Fig. 2(b), we applied two different templates to describe the disconnection process from the product (PE with two repeat units) into two unsaturated ethenyl groups. Template A [C:1][C:2][C:3][C:4] ≫ [C:1][C:2].[C:3][C:4] is applied to break the first single bond (reacting part A) of PE, leading to the formation of [C:2] and [C:3] radicals. It is important to mention that the [C:1] and [C:4] atoms are actually connected by the connection point ∗ of repeat units (reacting part B). Thus, the template B is used to disconnect the [C:1][C:4] single bond to form two ethenyl groups ([C:2][C:1][C:4][C:3] ≫ [C:1]=[C:2].[C:3]=[C:4]). In the ring-opening polymerization templates [Fig. 2(c)], the connection points ∗ are included in the template, because the reacting part A is the breakage of connecting points of repeat units, e.g, ∗[C:2][C:1][O:3]∗ ≫ [C:2]1[C:1][O:3]1 in Fig. 2(c).

(e) To remove duplications, the synthesis templates were screened using the reaction similarity checker implemented in RDKit.

C. Similarity calculation and model prediction ranking

In our approach, a similarity metric was adopted to select synthesis templates for a new target polymer and screen similar existing reactants. S_{polymer} is the similarity between target and previously synthesized polymers, specified by the polymerization type and the synthesis template, while S_{reactant} is the similarity between machine-predicted and known reactants. These two parameters are estimated using the Tanimoto metric, defined as

$$\text{Tanimoto}(x,y) = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i}, \tag{1}$$

where x_i and y_i are the i_{th} fingerprint of polymer (or monomer reactant) x and y, respectively. In this work, polymers were fingerprinted using our hierarchical polymer fingerprinting algorithm, including atomic, block, and chain-level features, the details as described in Ref. 11. Morgan circular fingerprints were applied to fingerprint reactants, as implemented in RDKit.[37]

The retrosynthesis paths for the target polymer are ranked using the synthesizability score (S_{score}), determined by the geometric mean of the similarity scores of polymers and reactants, defined as

$$S_{\text{score}} = \sqrt{S_{\text{polymer}} S^{\text{mean}}_{\text{reactant}}}, \tag{2}$$

$$S^{\text{mean}}_{\text{reactant}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{S^{i}_{\text{reactant}}}}. \tag{3}$$

Here, the harmonic mean ($S^{\text{mean}}_{\text{reactant}}$) is selected to further average the reactant contributions. n refers to the number of reactants: 1 for the ring-opening polymerization and 2 for the condensation and addition polymerization. S_{polymer}, S_{reactant} and $S^{\text{mean}}_{\text{reactant}}$ range from 0 to 1, where 1 (0) indicates identical (completely different) polymers/reactants. As a result, S_{score} is in a range of 0–1, where 1 and 0 indicate a perfect match and mismatch with an existing reaction, respectively.

To evaluate the model performance, two metrics, i.e., prediction ratio (PR) and accuracy ratio (AR), were computed as follows:

$$\text{PR} = N_{\text{prediction}}/N_{\text{total}}, \tag{4}$$

$$\text{AR} = N_{\text{accurate}}/N_{\text{prediction}}. \tag{5}$$

Here, $N_{\text{prediction}}$ and N_{total} are the number of polymers with predictable retrosynthesis paths and the total number of test polymers, respectively. In interpolation experiments, 1145 polymers with known polymerization paths were randomly selected from the training dataset to serve as the test set and provide ground truth. N_{accurate} is the number of polymers whose ground truth was within the top 10 predictions from the model. For statistical purposes, this experiment was repeated five times with random train-test splits. In the extrapolation experiments, N_{accurate} is the number of polymers whose predicted retrosynthesis path was in agreement with the judgment of polymer chemists.

III. RESULTS AND DISCUSSION

A. Synthesis templates

Following the template extraction rules described in Sec. II B, 129 (94 pairs), 314 (309 pairs), and 139 unique templates for addition, condensation, and ring-opening polymerization were obtained, respectively, and listed in Table I. Here, pairs are derived from two same/different templates required in condensation and addition polymerization [see A and B in Figs. 2(a) and 2(b)]. Figure 3 captures the normalized occurrence frequency distribution of these templates, together with the top five templates for each polymerization type.

TABLE I. Summary of synthesis templates. Data in brackets are the number of pairs templates, because two same/different templates are required in condensation and addition polymerization.

Polymerization types	Synthesis steps	Templates
Addition	Two-step	129 (94 pairs)
Condensation	Two-step	314 (309 pairs)
Ring-opening	One-step	139

In the case of condensation, the occurrence frequency of the top five templates reaches around 51%, representing the most common paths to form polyamide, polyester, and polyimide polymers. As shown in Fig. 3, the occurrence of the amide functional group ([N:1][C:3]=[O:6]) is up to 22%, formed by the linkage of the amine function with either acyl chloride or carboxylic acid functional groups. In the former case, [N:1] represents the amine group attached to any elements, while the latter only refers to NH_2. We also note that the ester functional group (the occurrence rate of 29%) is generated by the reaction between the hydroxyl group and acyl chloride (or carboxylic acid). The fifth condensation template is the formation of the imide functional group (occurrence of 6.7%) by the reaction between the acid anhydride and amine groups.

In the case of addition polymerization, the top four templates describe the formation of vinyl, urethane, urea, and diene polymers. The topone vinyl functional group (56% occurrence) is formed by connections between ethenyl groups that contain various compounds in the form of C=CR (R = H, CH_3, Cl, etc.) It is important to point out that a pair of templates was found for the vinyl group (see Fig. 2), but only the second template is shown in Fig. 3 to represent the reaction. The second and third most frequent templates are the formation of urethane and urea, produced by the reaction between isocyanate and hydroxyl (or amine groups), respectively. The fourth addition template is the formation of dienes with two acetylene groups. Unlike the first four templates, the special CCN functional group is infrequent (occurrence ratio of 1.7%), generated by linking the ethenyl and amine groups.

Additionally, there are 139 synthesis templates extracted from 586 ring-opening polymerization paths. Figure 3 shows the top five examples, namely, imines, oxides, dienes, and phosphazenes functional groups. It is noted that ring-opening polymerization is one common way to form the imine functional group (occurrence of 11%) by breaking the cyclic imino ether. The oxide functional group is the product of the breakage of 3- or 4-element cyclic ethers. We also noticed that the breaking of cyclic olefin monomers can result in the diene functional group. The last representative example is phosphazene polymers, which are composed of (P = N) and produced by cyclic phosphazene family monomers (e.g., hexachlorophosphazene).

B. Retrosynthesis planning pipeline

Next, the aforementioned polymerization dataset and synthesis templates were utilized to develop the polymer retrosynthesis

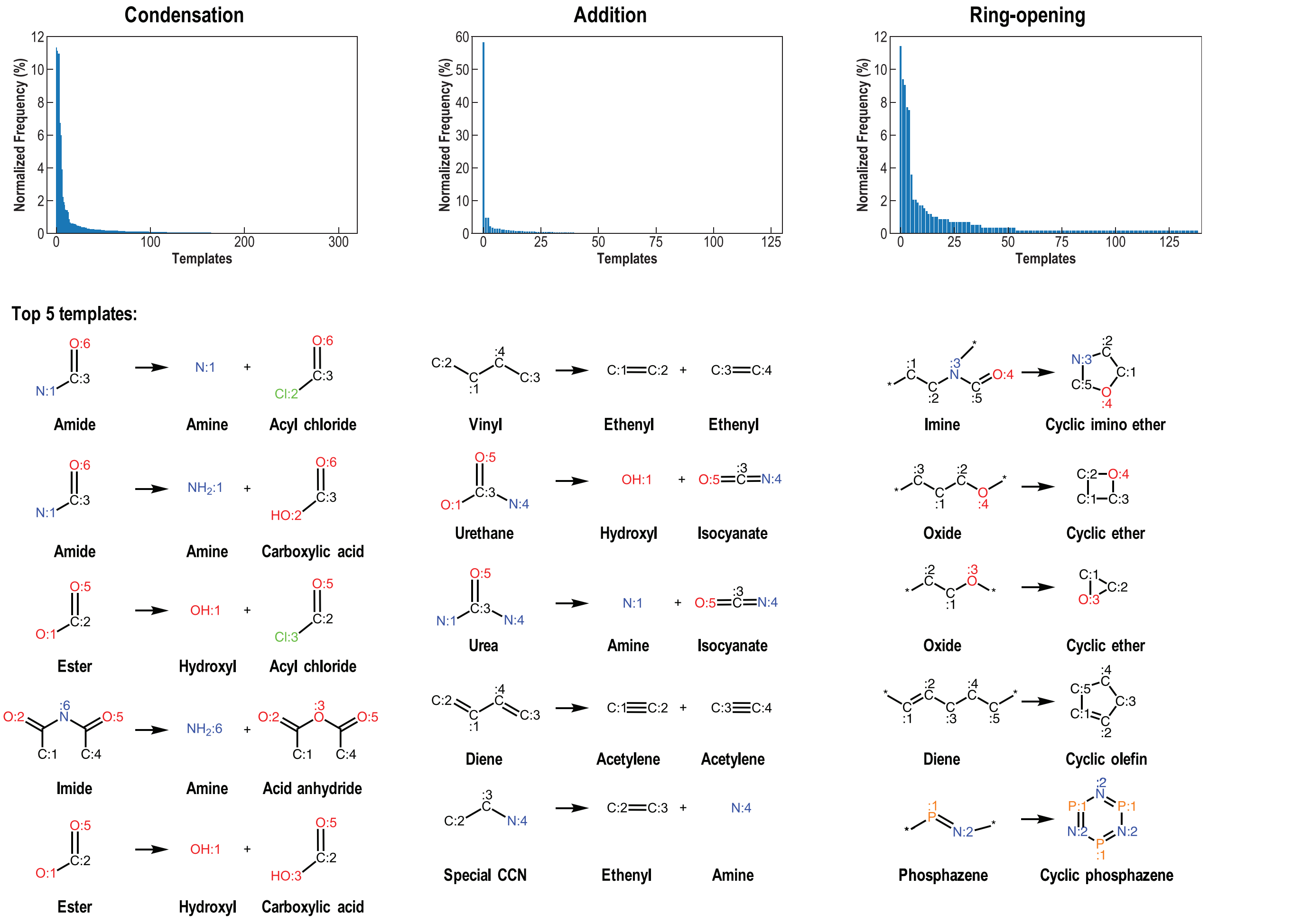


FIG. 3. Normalized occurrence frequency distribution of synthesis templates for condensation, addition and ring-opening polymerization (see Table I), along with the top five templates for each polymerization type. For the formation of the vinyl group, a pair of templates was extracted, but only template B is shown to represent the reaction; see Fig. 2 for the details.

framework. As depicted in Fig. 4, the central tenet of the pipeline is derived from the question of "how similar polymers have been synthesized?" If the polymerization paths of similar polymers have been reported in the literature, it is feasible to apply their synthesis templates to make the target polymer.

Figure 4 illustrates the proposed framework of polymer retrosynthesis planning, along with the outcome of each stage for an example of condensation polymers. Starting from the user-defined target polymer, several similar polymers (say, 10) are selected from the polymerization dataset using the similarity score (S_{polymer}) computed with the Tanimoto metric. This polymerization dataset consists of 11 488 polymerization paths and their corresponding templates. Figure 4 presents the top one similar polymer of the instanced polymer and its polymerization type/template. In the next step [(2) of Fig. 4], the specific polymerization type and templates of each similar polymer are applied to predict reactants of the target polymer (see the example), using "RunReactants" implemented in RDKit. It is followed by another screening procedure [(3) of Fig. 4] to search several (say, five) most similar known reactants from the reactant dataset for machine predicted reactants. These selected reactant candidates are required to contain the reacting parts (e.g., hydroxyl and carboxylic acid functional groups) from the template and ranked using the harmonic mean of the similarity of each reactant [$S_{\text{reactant}}^{\text{mean}}$, computed using Eq. (3)]. In the case of the exampled polymer, there are existing known reactants that are the same as the predicted reactants, resulting in a $S_{\text{reactant}}^{\text{mean}}$ of 1.0. Finally, the retrosynthesis paths of the target polymer are ranked using S_{score}, computed by the geometric mean of S_{polymer} and $S_{\text{reactant}}^{\text{mean}}$ [Eq. (2)], such as the rank 1 S_{score} of 0.97 for the instanced polymer.

There are several points to highlight in terms of the developed pipeline: (1) the reactant dataset includes 8921 reactants and was extracted from 11 488 polymerization paths, which can be significantly increased in size; (2) the preferential number of top similar polymers (N_{top}) positively affects the prediction time and accuracy, but the

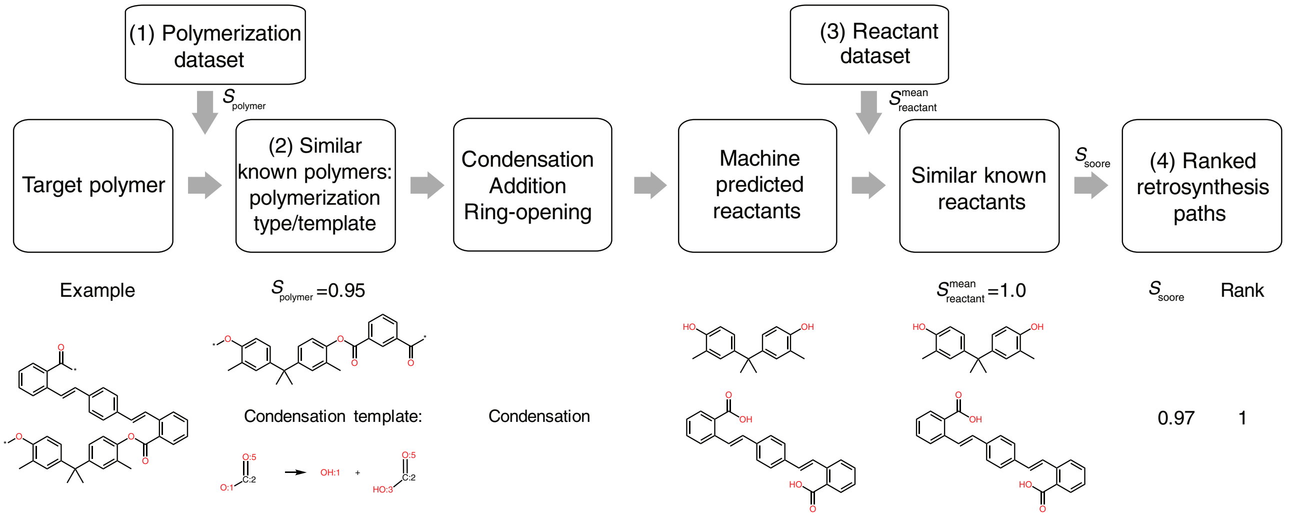


FIG. 4. The framework of polymer retrosynthesis planning. The polymerization dataset includes 11,488 paths and their corresponding synthesis templates. The reactant dataset contains 8921 reactants extracted from the polymerization dataset, which can be significantly increased in size. The synthesizability score (S_{score}) is determined by the geometric mean of $S_{polymer}$ and $S^{mean}_{reactant}$. Here, $S_{polymer}$ is the similarity between the target and the existing polymers. $S^{mean}_{reactant}$ is the harmonic mean of the similarity between the predicted and the existing reactants. The details are described in Sec. II C. Taking a condensation polymer as an example, the outcome of each stage is provided.

prediction accuracy reaches convergence after $N_{top} \geq 10$. The details are discussed in Sec. III C; (3) the accuracy depends on the available synthesis templates and known polymers/reactants, meaning that for some new polymers, the code may predict relatively low S_{score} or may even fail to produce retrosynthesis paths. Nonetheless, this issue can be addressed by increasing the number of polymerization templates, polymers, and reactants in our dataset.

C. Validation

To demonstrate the accuracy and generality of the developed framework shown in Fig. 4, both interpolation and extrapolation experiments have been performed on previously synthesized polymers. The model performance was evaluated using two metrics—prediction (PR) and accuracy (AR) ratios, respectively, computed by Eqs. (4) and (5). Here, the accuracy was estimated by comparison with available ground truth (interpolation experiment) or domain experts' evaluation (extrapolation experiment).

In the interpolation experiment, a group of known polymers with known polymerization information was applied to provide ground truth. We randomly selected 1145 polymers (10%) from the polymerization dataset (11 448 polymerization paths) as the test set. The remaining data served as the training set. Table II summarizes the average prediction ratio (PR), the accuracy ratio (AR), and the simulation time (t), together with the standard deviation from five simulations with the random test set split. Here, AR was determined by comparing the predicted retrosynthesis paths with ground truth.

Given that the selected number of top similar known polymers [N_{top}, step (2) of Fig. 4] determines the applied synthesis templates, we investigated the impact of N_{top} on the prediction accuracy and time. Table II reveals that the average PR and AR increase with larger N_{top} and converges to 98% and 91%, respectively. On the other hand, the average t for each polymer rises in proportion to N_{top}, because the synthesis templates for each top similar polymer are used in step (2) of Fig. 4. The remaining 2% failures are mainly due to the unavailable synthesis templates in the training set. This issue can be tackled by increasing the template set size, while polymers with multiple possible synthesis pathways make contributions to the accuracy discrepancy, in addition to the template limitation. For example, the amide function group can be generated using the amine group and the acyl chloride (or carboxylic acid) group; see Fig. 3.

Figure 5 displays representative examples of predicted retrosynthesis paths for target polymers, together with the expected reactants/template and polymerization type. In the case of the target polymer **a** (condensation), the top two ranked retrosynthesis paths are derived from the same top one similar polymer with a $S_{polymer}$ of 0.95. Using the template of this similar polymer, reactants with acid anhydride and amine functional groups are predicted and retrieved from the known reactant dataset, resulting in the rank 1 S_{score} of 0.97. This proposed synthesis pathway is consistent with the expected path. Moreover, other similar known reactants with acid anhydride functional group ([O:2]=[C:1][N:6][C:4]=[O:5]) are obtained

TABLE II. Performance of the developed retrosynthesis framework in Fig. 4. N_{total} is the total number of test polymers and N_{top} is the selected number of top similar polymers in step (2) of Fig. 4. The prediction (PR) and accuracy (AR) ratios are estimated using Eqs. (4) and (5), respectively. t is the average running time for each polymer. The standard deviation is from five simulations with a random test set split.

Experiment	N_{total}	N_{top}	PR (%)	AR (%)	t (s)
Interpolate	1145	1	93 ± 0.7	60 ± 2	6
		5	97 ± 0.4	86 ± 1	27
		10	98 ± 0.2	89 ± 1	54
		20	98 ± 0.3	91 ± 1	100
Extrapolate	3582	10	92	80[a]	–

[a] Estimated using 100 sampling polymers.

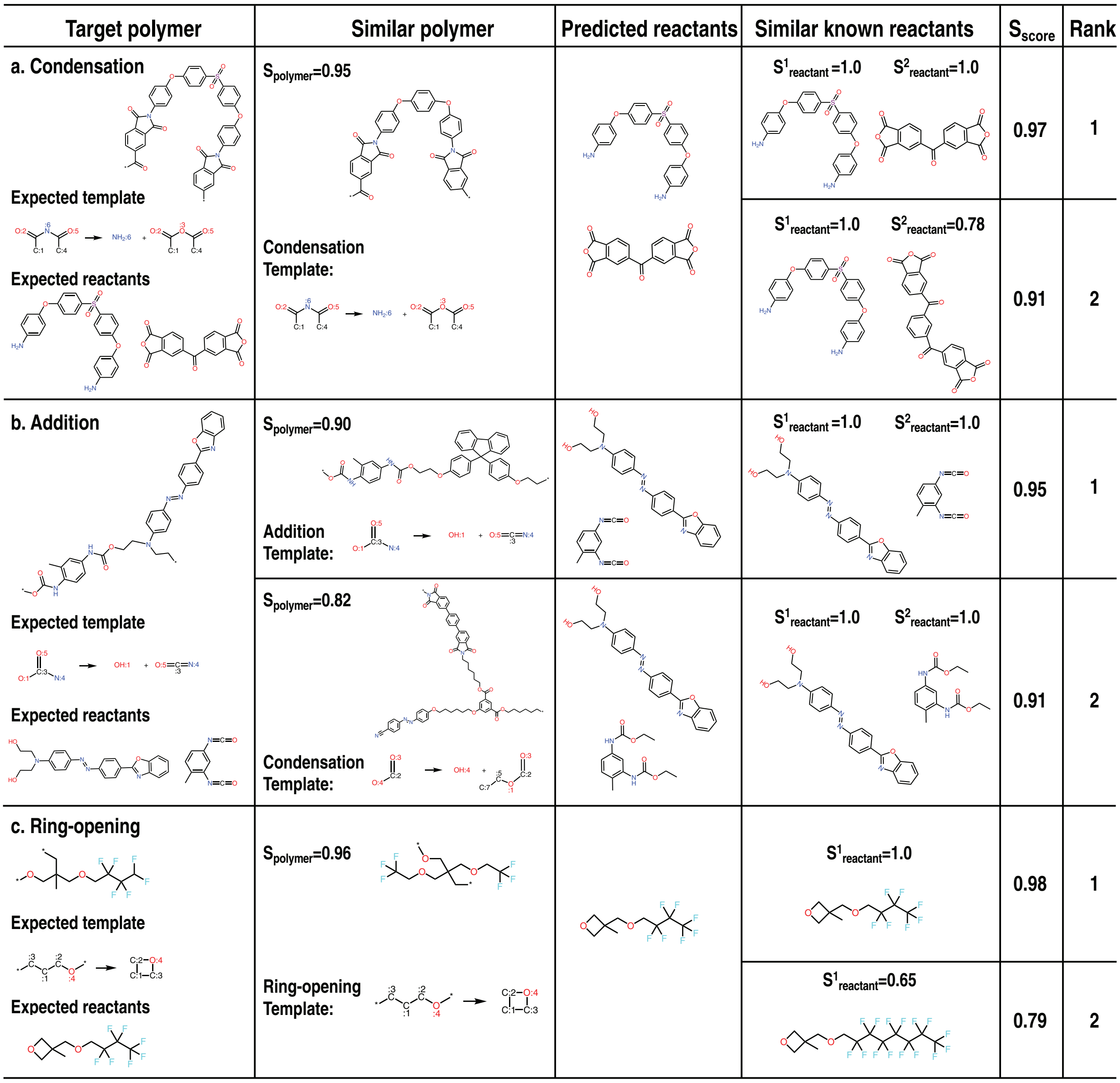


FIG. 5. Predicted synthesis paths of representative condensation, addition, and ring-opening polymers in the interpolation experiment. The expected polymerization information of these polymers, i.e., synthesis templates and reactants, is also provided.

with a slightly lower S^2_{reactant} score (0.78) and thus a rank two S_{score} of 0.91. For the target polymer **b** (addition), the top two similar polymers with different polymerization types and templates were screened from the training set. As can be seen from Fig. 5(b), the top one similar polymer has a S_{polymer} of 0.90 and addition polymerization type/templates, leading to the agreement between the predicted and expected reactants. The second most similar polymer has a S_{polymer} of 0.82 with the condensation polymerization type. In this scenario, the ester functional group was used as the linkage of the polymer, leading to the generation of reactants with hydroxyl and carboxylic functional groups. This is probably an alternative way to synthesize the target polymer **b** with the approval of a synthetic chemist. The third example is a typical ring-opening polymer **c**. Figure 5(c) reveals that an existing polymer of S_{polymer} leads to accurately predicted reactants that agree with ground truth and high rank one $\mathrm{S}_{\text{score}}$ of 0.96. There are other similarly known reactants with a S^1_{reactant} of 0.65, resulting in a rank two $\mathrm{S}_{\text{score}}$ of 0.79.

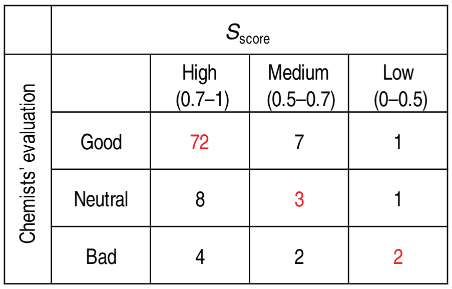
Next, we move on to discuss the extrapolation experiment using 3582 previously synthesized polymers with meager polymerization information (resulting in unknown synthesis templates). These polymers were collected from various literature sources[8,38] and described in Sec. II A. In this experiment, the whole polymerization dataset (11 488 paths) was utilized as the training set. As listed in Table II, the retrosynthesis paths of 92% polymers were predicted. The remaining 8% failed cases are mainly induced by unavailable synthesis templates in the dataset and uncommon reported polymers.

To further measure the prediction accuracy, the rank one retrosynthesis path for 100 randomly selected test polymers has been evaluated by polymer chemists. By considering the reliability of the predicted polymerization mechanism and reactant functional groups, three accuracy levels were defined—good, neutral, and bad. Good (bad) prediction indicates accepted (unrealistic) retrosynthesis paths, and the neutral prediction is intermediate to good and bad. To avoid bias, similar reactants and $S_{\rm score}$ information were not provided to chemists. The evaluation of results for 100 polymers are provided in the Appendix A, containing the target polymer, machine predicted polymerization type, and reactants, together with the evaluation of polymer chemists.

Figure 6 shows the confusion matrix for the 100 sampling polymers. To compare with chemists' evaluation, we further classified $S_{\rm score}$ into three levels: high $(0.7 < S_{\rm score} \leq 1)$, medium $(0.5 \leq S_{\rm score} \leq 0.7)$, and low $(S_{\rm score} < 0.5)$. We note that 96 polymers have high or medium $S_{\rm score}$, being consistent with the fact that they are previously synthesized polymers. Further, the prediction of 80 polymers was estimated as "good" by chemists, leading to an AR of 80% in Table II. Among them, 72, 7, and 1 polymers are predicted to have high, medium, and low $S_{\rm score}$, respectively, by our model. This finding indicates that the $S_{\rm score}$ is a promising way to estimate polymer synthesizability. Additionally, Table II reveals that 12 polymers with high $S_{\rm score}$ have the neutral (8) and bad (4) evaluation from chemists. This discrepancy may be caused by limited model accuracy, intrinsic noise of the polymerization dataset, or limited experience of chemists for the polymerizations considered. Even so, the dominance of the diagonal terms in Fig. 6 indicates acceptable model performance.

IV. CONCLUSION AND OUTLOOK

In summary, a novel and powerful data-driven approach was developed for polymer retrosynthesis planning. This work involves polymerization dataset accumulation, synthesis template extraction and automated polymer retrosynthesis planning paradigm development. It is worth pointing out that the hand-crafted synthesis templates encode the chemical reactions between reactant molecules, rather than the formation of reactant molecules themselves. With this approach, synthesis paths of new target polymers can be predicted and ranked using the synthesizability score, providing synthesis guidance for chemists. The accuracy of the model was measured by comparison of the model predictions with the ground truth and/or polymer chemists' evaluation. Further, the developed framework solely depends on the polymerization dataset devoid of any model parameters. These unique advantages also lead to some concerns. For example, polymers outside the dataset may have high prediction uncertainty or relatively low synthesizability score. These issues can be further addressed either by expanding polymerization and template dataset sizes or by applying advanced algorithms to learn new templates. Given the problem of data sparsity, it is recommended that the community report successful and/or failed polymerization paths of designed polymers in papers/online repository to facilitate the development of data-driven approaches. Additionally, the present work focuses on the reactant monomer prediction and disregards the reagents, experimental conditions, and other factors affecting the polymerization (such as the reactivity of functional groups). These aspects will be considered in future studies.



Chemists' evaluation	$S_{\rm score}$		
	High (0.7–1)	Medium (0.5–0.7)	Low (0–0.5)
Good	72	7	1
Neutral	8	3	1
Bad	4	2	2

FIG. 6. Confusion matrix of computed $S_{\rm score}$ and chemists' evaluation for 100 sampling polymers in the extrapolation experiment. $S_{\rm score}$ was classified into high, medium, and low levels. The chemists' evaluation also includes three accuracy levels: good, neutral, and bad. Good (bad) prediction means accepted (unrealistic) retrosynthesis paths, and neutral prediction is intermediate to good and bad.

Although the work can still be refined, we believe that vital initial steps have been taken to accelerate polymer synthesis and discovery due to two aspects. First, the developed model is implemented in the online Polymer Genome platform (https://www.polymergenome.org), leading to accessible ranked retrosynthesis paths and predicted properties for user-defined polymers. Second, it is now possible to utilize extracted synthesis templates coupled with various polymer design algorithms[12,13] to generate synthesis-friendly polymers with desired properties for specific applications.

SUPPLEMENTARY MATERIAL

See the supplementary material for the principal component analysis of the chemical space of polymerization and test datasets (Fig. S1), and more examples of extracted synthesis templates (Fig. S2). Appendix includes the evaluation of results for 100 sampling polymers, containing the target polymer, machine predicted polymerization type/reactants, and the evaluation of polymer chemists.

ACKNOWLEDGMENTS

This work is supported by the Office of Naval Research through two Multi-University Research Initiative (MURI) Grants: N00014-17-2656 and N00014-20-1-2586. We thank Ajinkya Deshmukh for help with the evaluation of predicted results and Gregory Sotzing for insightful discussions. We also thank Chiho Kim and Mark Danielson for their help in deploying the developed model using Application Programming Interface.

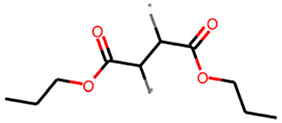
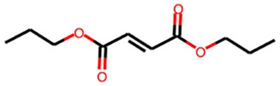
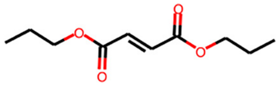
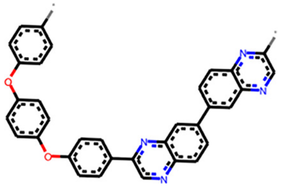
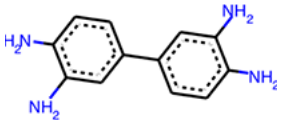
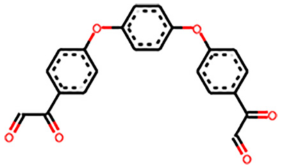
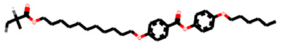
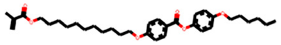
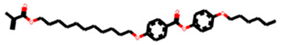
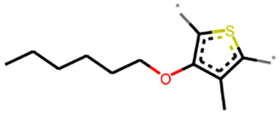
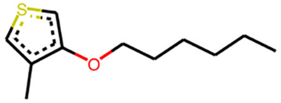
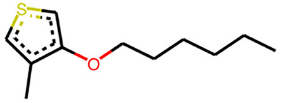
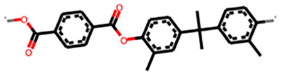
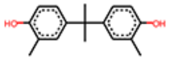
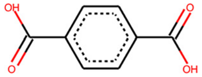
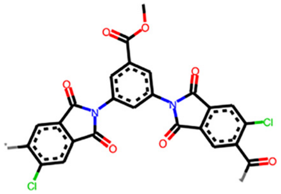
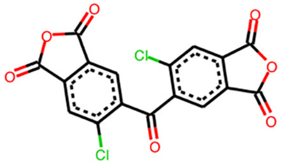
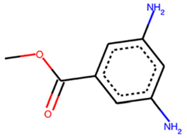
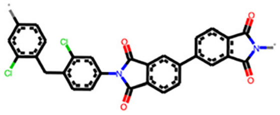
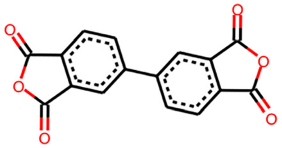
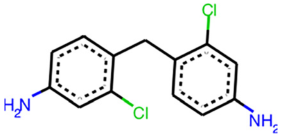
The authors declare no Conflict of Interest.

DATA AVAILABILITY

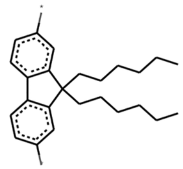
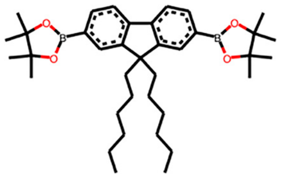
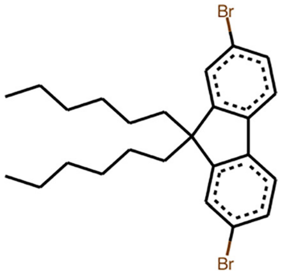
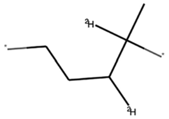
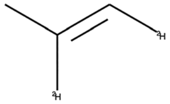
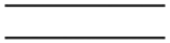
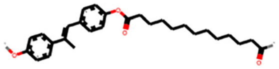
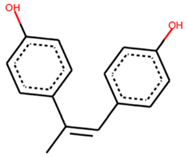
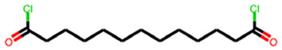
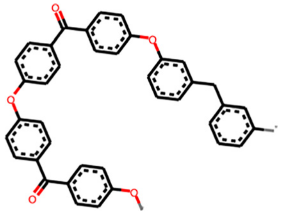
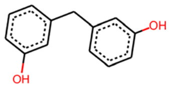
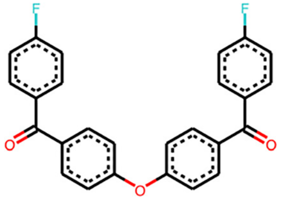
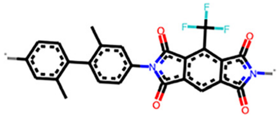
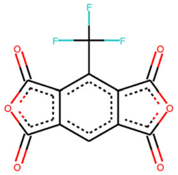
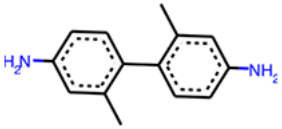
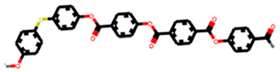
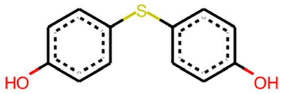
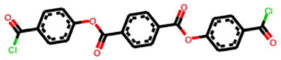
The data used in the present work can be obtained from the PolyInfo database.[8]

APPENDIX:

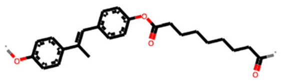
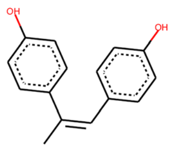
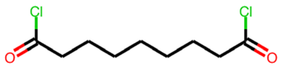
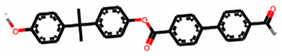
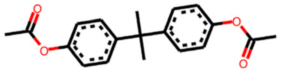
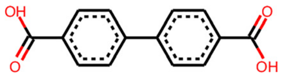
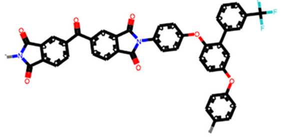
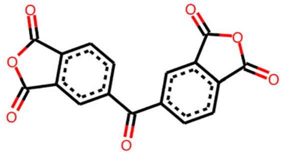
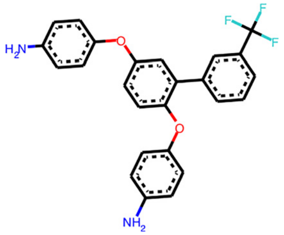
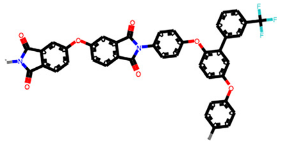
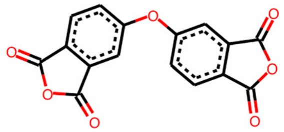
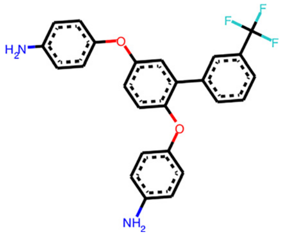
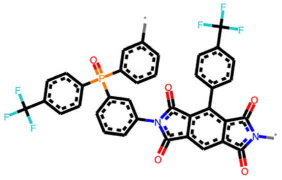
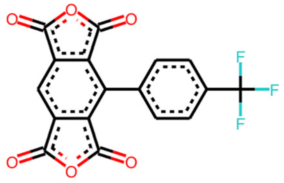
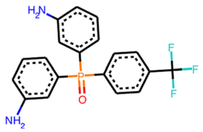
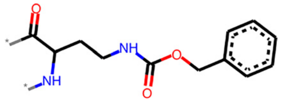
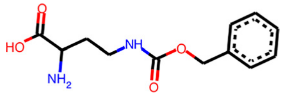
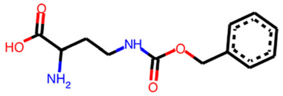
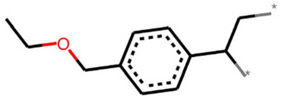
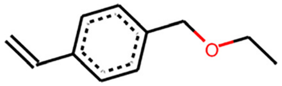
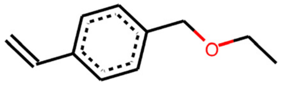
Sampling set1: 100 polymers, evaluated by polymer chemist.

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
0		Addition			Good
1		Condensation			Neutral
2		Addition			Good
3		Condensation			Good
4		Condensation			Good
5		Condensation			Good
6		Condensation			Good

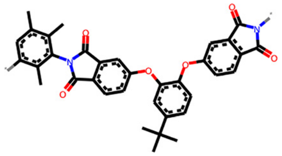
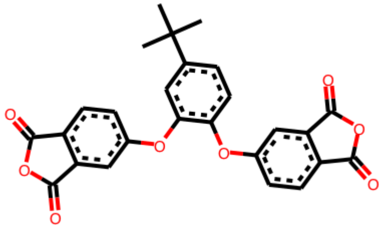
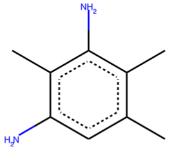
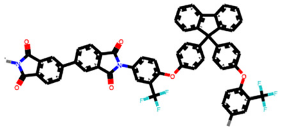
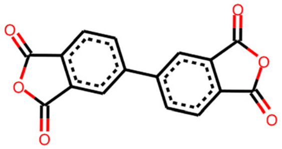
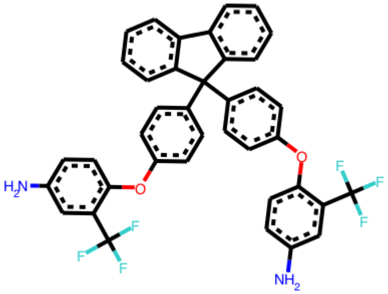
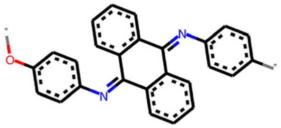
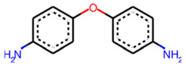
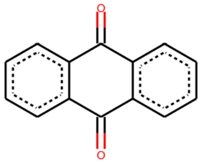
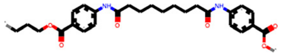
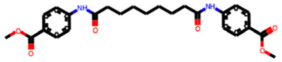
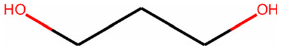
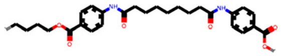
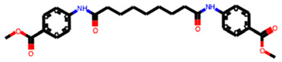
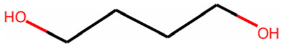
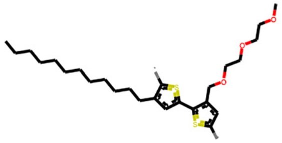
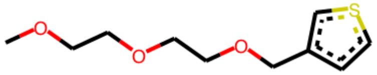
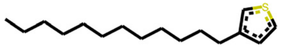
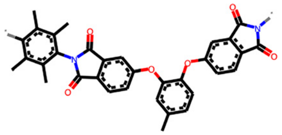
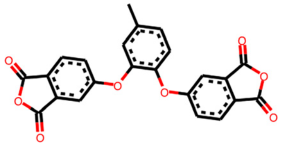
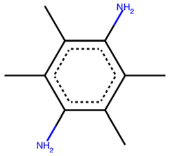
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
7		Condensation			Neutral
8		Addition			Bad
9		Condensation			Good
10		Condensation			Good
11		Condensation			Good
12		Condensation			Good

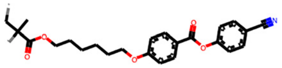
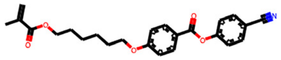
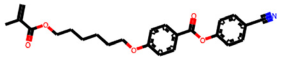
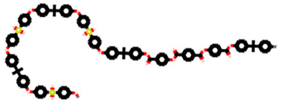
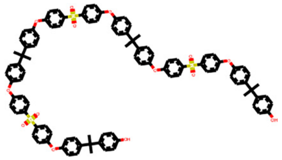
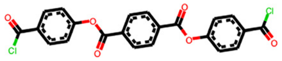
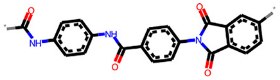
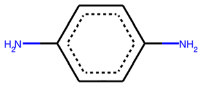
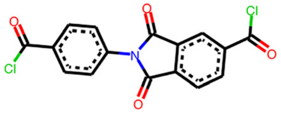
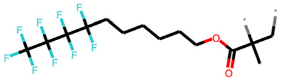
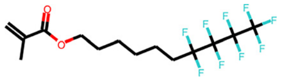
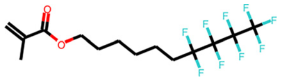
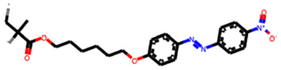
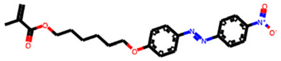
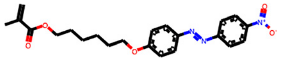
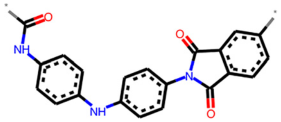
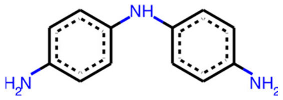
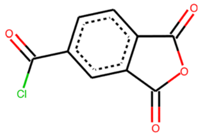
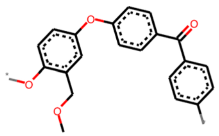
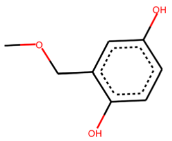
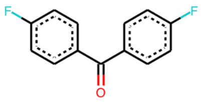
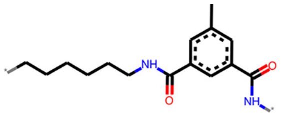
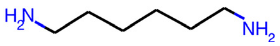
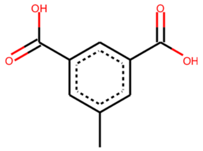
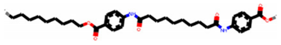
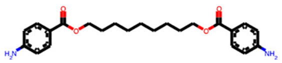
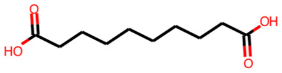
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
13		Condensation			Good
14		Condensation			Good
15		Condensation			Good
16		Condensation			Good
17		Condensation			Good
18		Condensation			Good
19		Addition			Good

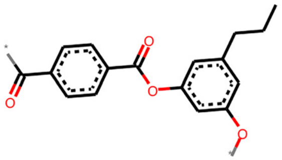
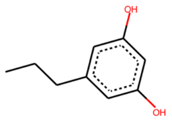
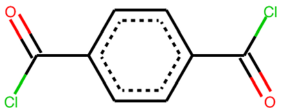
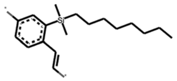
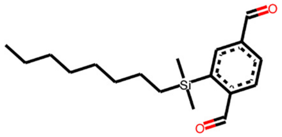
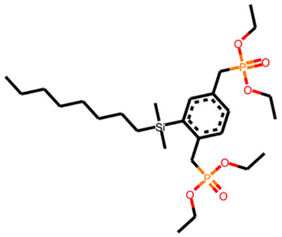
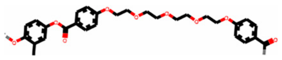
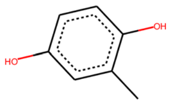
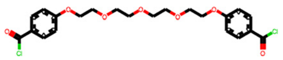
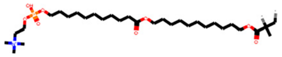
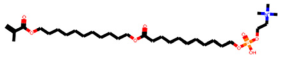
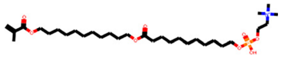
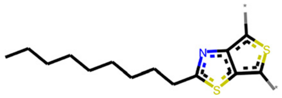
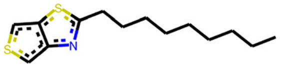
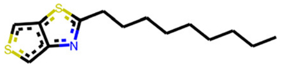
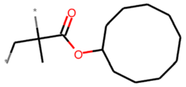
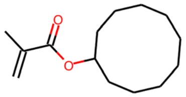
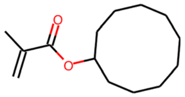
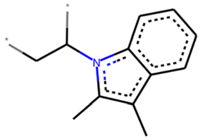
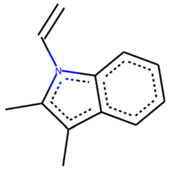
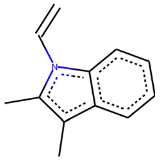
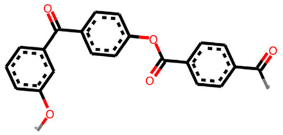
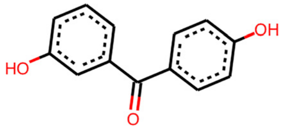
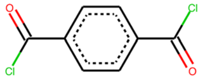
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
20		Condensation			Good
21		Condensation			Good
22		Condensation			Good
23		Condensation			Good
24		Condensation			Good
25		Condensation			Good
26		Condensation			Good

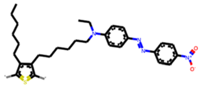
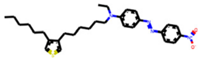
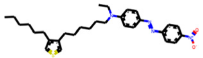
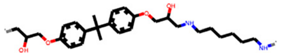
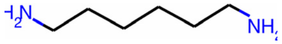
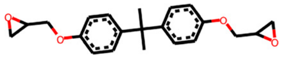
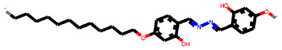
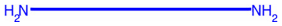
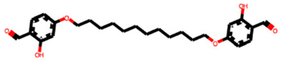
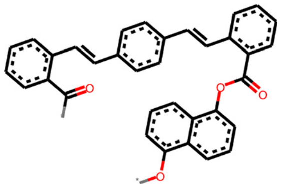
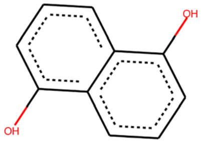
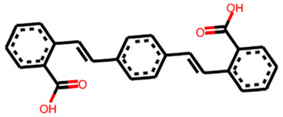
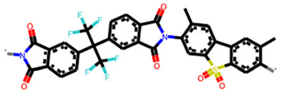
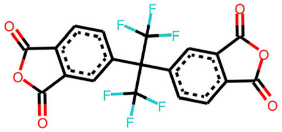
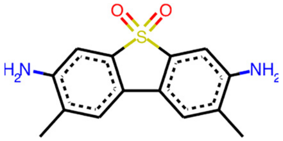
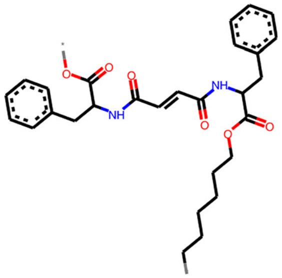
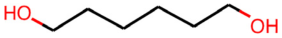
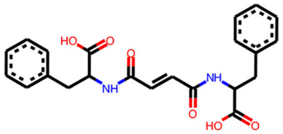
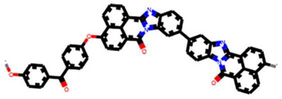
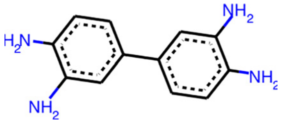
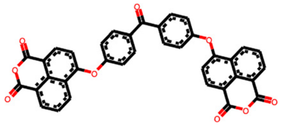
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
27		Addition			Good
28		Condensation			Good
29		Condensation			Good
30		Addition			Good
31		Addition			Good
32		Condensation			Good
33		Condensation			Good
34		Condensation			Good
35		Condensation			Good

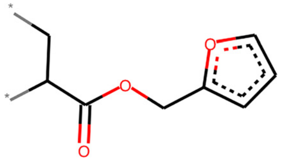
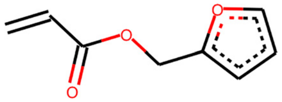
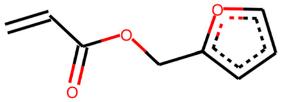
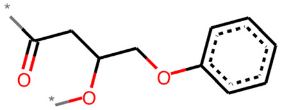
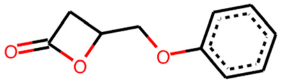
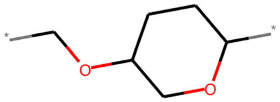
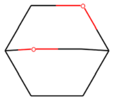
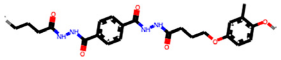
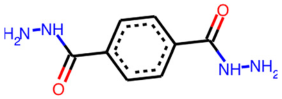
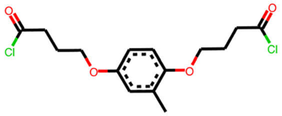
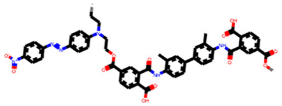
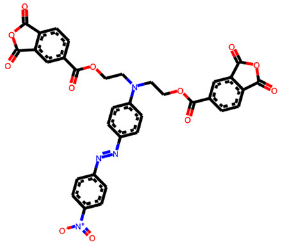
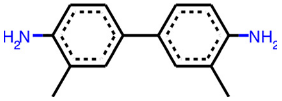
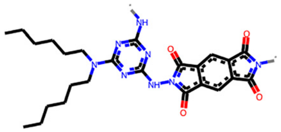
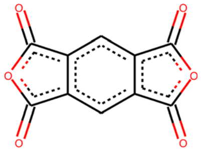
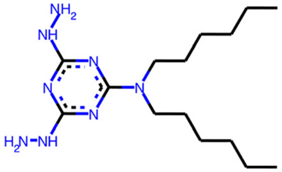
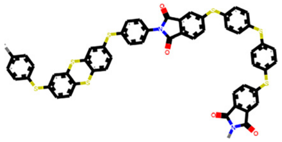
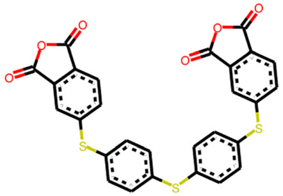
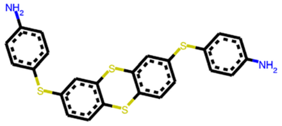
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
36		Condensation			Good
37		Condensation			Neutral
38		Condensation			Good
39		Addition			Good
40		Condensation			Good
41		Addition			Good
42		Addition			Good
43		Condensation			Good

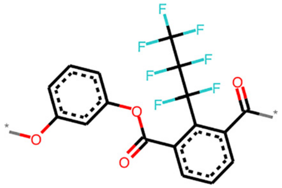
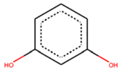
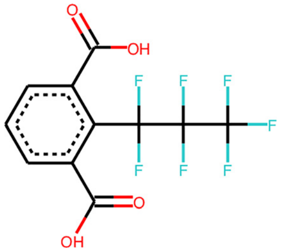
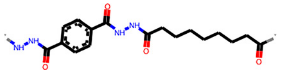
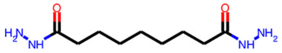
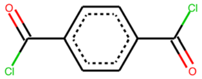
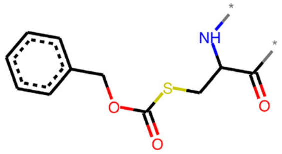
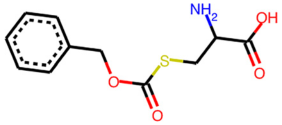
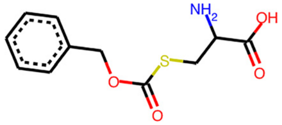
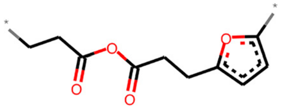
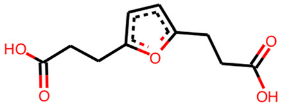
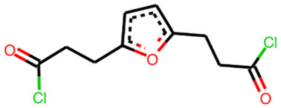
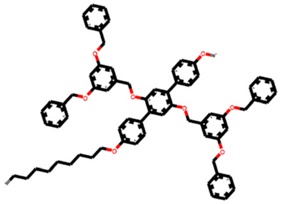
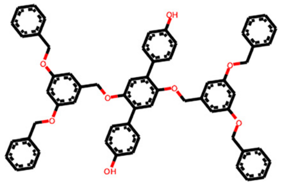
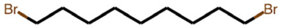
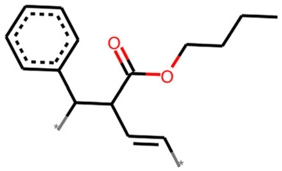
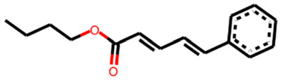
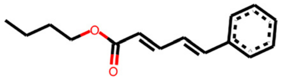
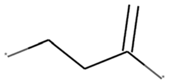
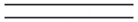
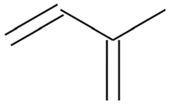
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
44		Condensation			Good
45		Addition			Good
46		Condensation			Good
47		Condensation			Good
48		Condensation			Good
49		Condensation			Good
50		Condensation			Good

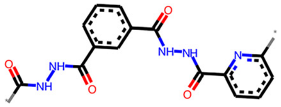
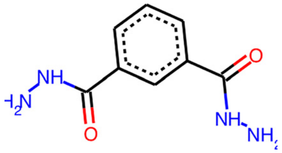
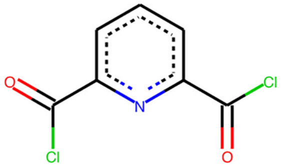
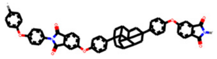
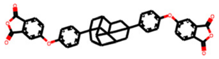
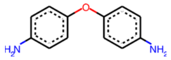
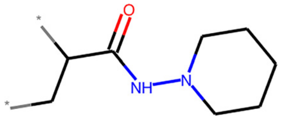
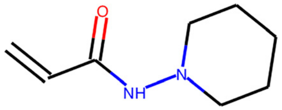
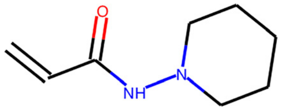
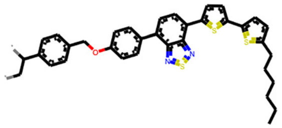
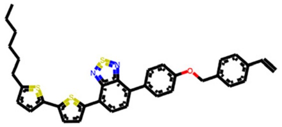
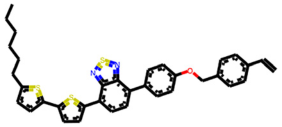
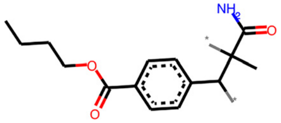
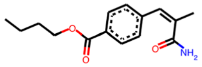
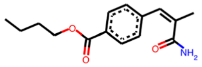
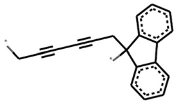
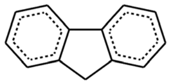
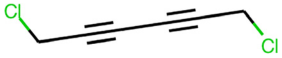
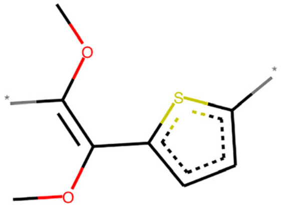
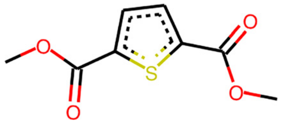
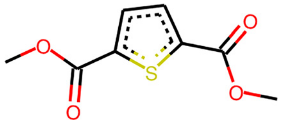
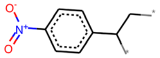
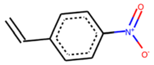
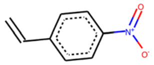
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
51		Addition			Good
52		Ring-opening		-	Good
53		Ring-opening		-	Good
54		Condensation			Good
55		Addition			Good
56		Condensation			Good
57		Condensation			Good

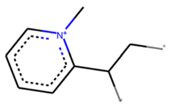
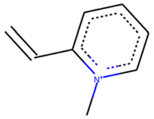
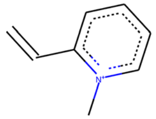
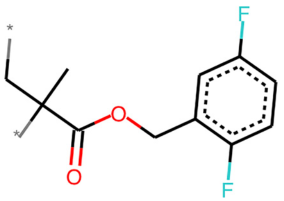
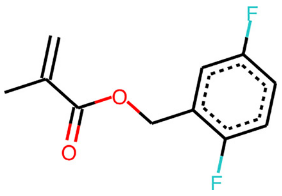
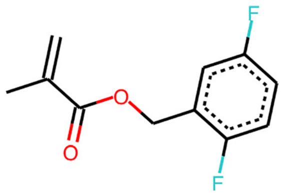
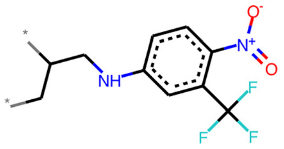
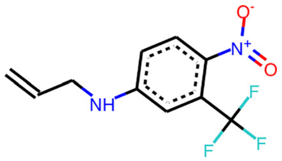
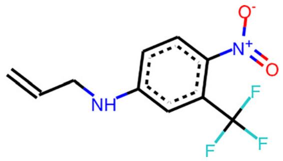
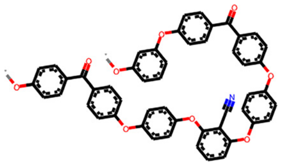
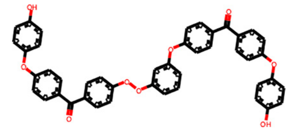
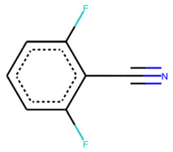
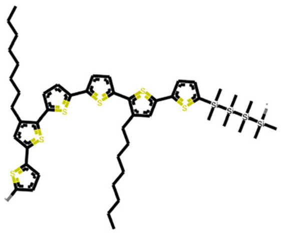
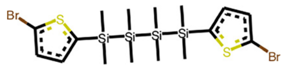
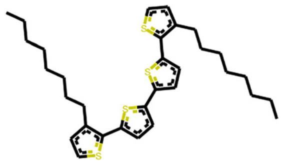
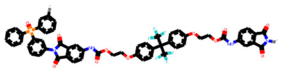
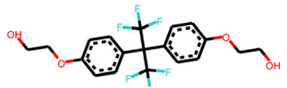
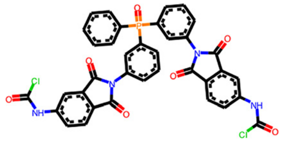
(Continued.)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
58		Condensation			Good
59		Condensation			Good
60		Condensation			Good
61		Condensation			Bad
62		Condensation			Neutral
63		Addition			Neutral
64		Addition			Good

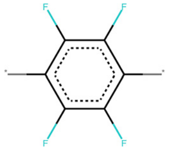
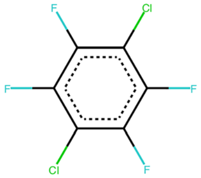
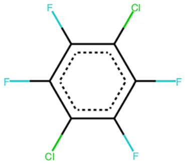
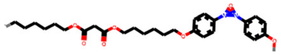
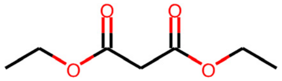
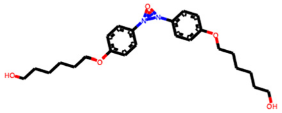
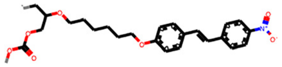
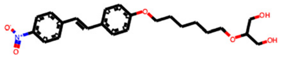
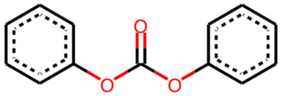
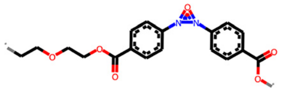
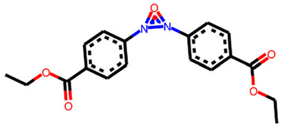
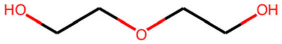
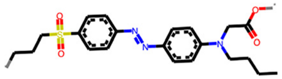
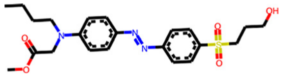
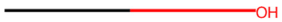
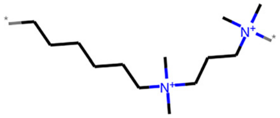
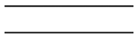
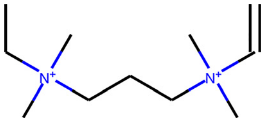
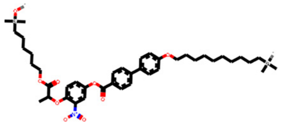
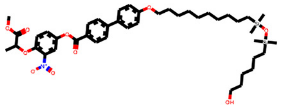
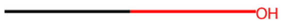
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
65		Condensation			Good
66		Condensation			Good
67		Addition			Good
68		Addition			Neutral
69		Addition			Good
70		Condensation			Bad
71		Addition			Bad
72		Addition			Good

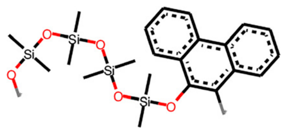
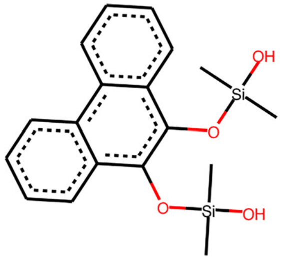
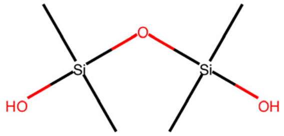
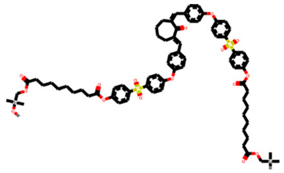
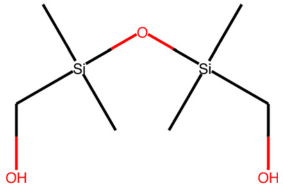
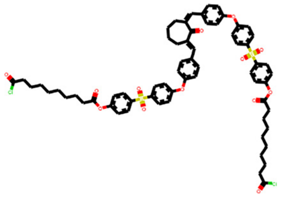
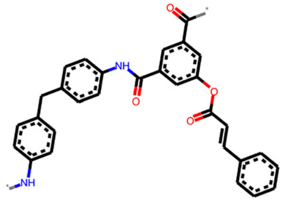
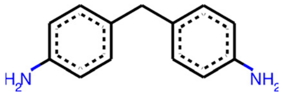
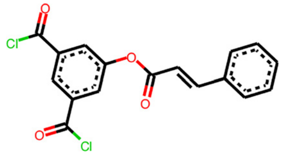
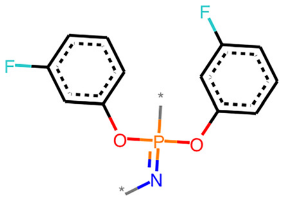
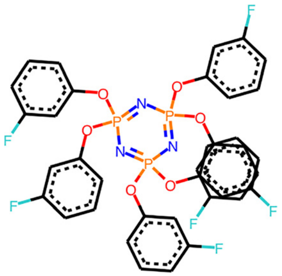
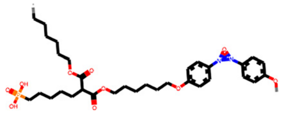
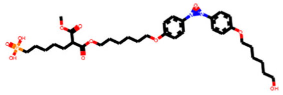
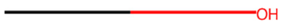
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
73		Addition			Good
74		Addition			Good
75		Addition			Good
76		Condensation			Good
77		Condensation			Good
78		Condensation			Good

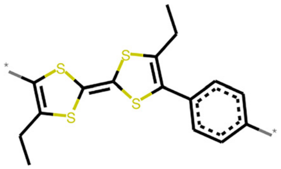
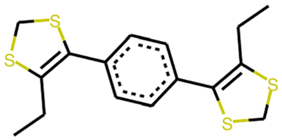
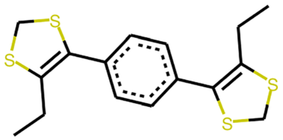
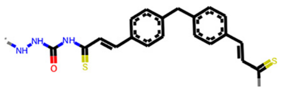
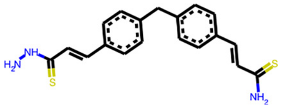
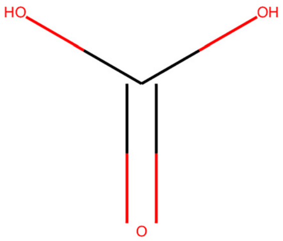
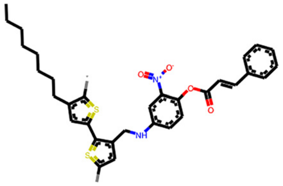
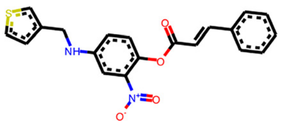
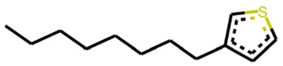
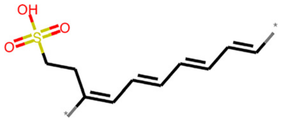
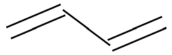
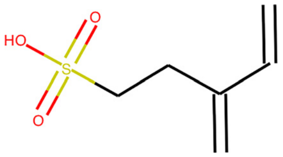
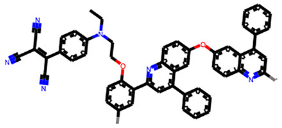
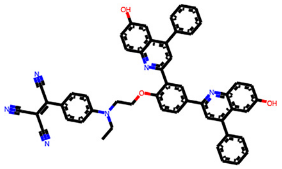
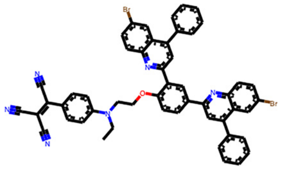
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
79		Condensation			Neutral
80		Condensation			Good
81		Condensation			Neutral
82		Condensation			Good
83		Condensation			Good
84		Addition			Good
85		Condensation			Good

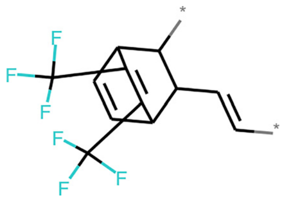
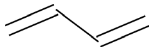
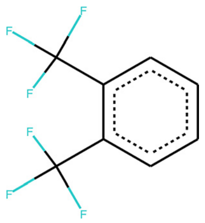
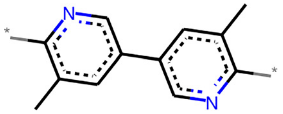
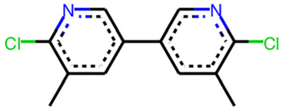
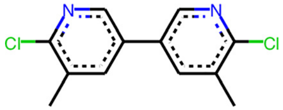
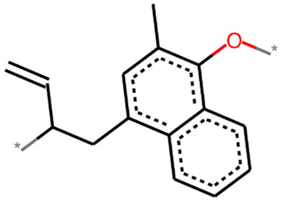
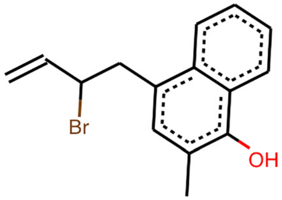
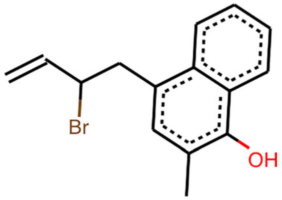
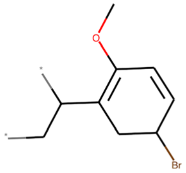
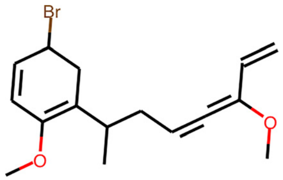
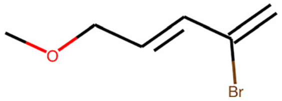
(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
86		Condensation			Good
87		Condensation			Good
88		Condensation			Good
89		Ring-opening		-	Good
90		Condensation			Good

(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
91		Condensation			Bad
92		Condensation			Bad
93		Condensation			Neutral
94		Addition			Neutral
95		Condensation			Neutral

(*Continued.*)

	Target_polymer	Polymerization_type	Predicted_reactant1	Predicted_reactant2	Evaluation
96		Addition			Bad
97		Condensation			Good
98		Condensation			Neutral
99		Addition			Bad

REFERENCES

[1] R. Batra, L. Song, and R. Ramprasad, "Emerging materials intelligence ecosystems propelled by machine learning," Nat. Rev. Mater. **2020**, 1–24, available at https://www.nature.com/articles/s41578-020-00255-y.

[2] L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, "Polymer informatics: Current status and critical next steps," Mat. Sci. Eng. R. **144**, 100595 (2021).

[3] J. S. Peerless, N. J. Milliken, T. J. Oweida, M. D. Manning, and Y. G. Yingling, "Soft matter informatics: Current progress and challenges," Adv. Theory Simul. **2**, 1800129 (2019).

[4] N. Adams and P. Murray-Rust, "Engineering polymer informatics: Towards the computer-aided design of polymers," Macromol. Rapid Commun. **29**, 615–632 (2008).

[5] D. J. Audus and J. J. de Pablo, "Polymer informatics: Opportunities and challenges," ACS Macro Lett. **6**, 1078–1082 (2017).

[6] A. Mannodi-Kanakkithodi, G. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. Sotzing, and R. Ramprasad, "Rational co-design of polymer dielectrics for energy storage," Adv. Mater. **28**, 6277–6291 (2016).

[7] Y. Wu, J. Guo, R. Sun, and J. Min, "Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells," NPJ Comput. Mater. **6**, 1–8 (2020).

[8] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, "Polyinfo: Polymer database for polymeric materials design," in *2011 International Conference on Emerging Intelligent Data and Web Technologies* (IEEE, 2011) pp. 22–29.

[9] See Polymer Property Predictor and Database for information about polymer property data (03/01/2021).

[10] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, "Polymer genome: A data-powered polymer informatics platform for property predictions," J. Phys. Chem. C **122**, 17575–17585 (2018).

[11] H. Doan Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, and P. Shetty, "Machine-learning predictions of polymer properties with polymer genome," J. Appl. Phys. **128**, 171104 (2020).

[12] C. Kim, R. Batra, L. Chen, H. Tran, and R. Ramprasad, "Polymer design using genetic algorithm and machine learning," Comput. Mater. Sci. **186**, 110067 (2020).

[13] R. Batra, H. Dai, H. Tran, L. Chen, C. Kim, G. Will, L. Song, and R. Ramprasad, "Polymers for extreme conditions designed using syntax-directed variational autoencoders," Chem. Mater. **32**, 10489–10500 (2020).

[14] R. Ma and T. Luo, "Pi1m: A benchmark database for polymer informatics," J. Chem. Inf. Modeling **60**, 4684–4690 (2020).

[15] J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, "Designing exceptional gas-separation polymer membranes using machine learning," Sci. Adv. **6**, eaaz4301 (2020).

[16] L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, P. Vashishta *et al.*, "Frequency-dependent dielectric constant prediction of polymers using machine learning," NPJ Comput. Mater. **6**, 1–9 (2020).

[17] D. Braun, H. Cherdron, M. Rehahn, H. Ritter, and B. Voit, *Polymer Synthesis: Theory and Practice: Fundamentals, Methods, Experiments* (Springer Science & Business Media, 2012).

[18] E. J. Corey, A. K. Long, and S. D. Rubenstein, "Computer-assisted analysis in organic synthesis," Sci. **228**, 408–418 (1985).

[19] L. Achenie, V. Venkatasubramanian, and R. Gani, *Computer Aided Molecular Design: Theory and Practice* (Elsevier, 2002).

[20] L. Y. Ng, F. K. Chong, and N. G. Chemmangattuvalappil, "Challenges and opportunities in computer-aided molecular design," Comput. Chem. Eng. **81**, 115–129 (2015).

[21] M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic ai," Nature **555**, 604–610 (2018).

[22] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," ACS Cent. Sci. **3**, 434–443 (2017).

[23] W. Jin, C. Coley, R. Barzilay, and T. Jaakkola, "Predicting organic reaction outcomes with weisfeiler-lehman network," arXiv preprint arXiv:1709.04555 (2017).

[24] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," Chem. Sci. **10**, 370–377 (2019).

[25] H. Dai, C. Li, C. Coley, B. Dai, and L. Song, "Retrosynthesis prediction with conditional graph logic network," in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019) pp. 8872–8882.

[26] C. W. Coley, D. A. Thomas, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao *et al.*, "A robotic platform for flow synthesis of organic compounds informed by ai planning," Science **365**, eaax1566 (2019).

[27] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," ACS Cent. Sci. **5**, 1572–1583 (2019).

[28] M. H. Segler and M. P. Waller, "Neural-symbolic machine learning for retrosynthesis and reaction prediction," Chem. Eur. J. **23**, 5966–5971 (2017).

[29] O. Engkvist, P.-O. Norrby, N. Selmi, Y-h Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard, and L. A. Smyth, "Computational prediction of chemical reactions: Current status and outlook," Drug Discov. Today **23**, 1203–1218 (2018).

[30] C. W. Coley, W. H. Green, and K. F. Jensen, "Machine learning in computer-aided synthesis planning," Acc. Chem. Res. **51**, 1281–1289 (2018).

[31] "NIST Synthetic Polymer MALDI Recipes Database," https://maldi.nist.gov. Search for synthesis recipes of polymers (03/15/2021).

[32] R. Ma, A. F. Baldwin, C. Wang, I. Offenbach, M. Cakmak, R. Ramprasad, and G. A. Sotzing, "Rationally designed polyimides for high-energy density capacitor applications," ACS Appl. Mater. Interfaces **6**, 10445–10451 (2014).

[33] Z. Li, G. M. Treich, M. Tefferi, C. Wu, S. Nasreen, S. K. Scheirey, R. Ramprasad, G. A. Sotzing, and Y. Cao, "High energy density and high efficiency all-organic polymers with enhanced dipolar polarization," J. Mater. Chem. A **7**, 15026–15030 (2019).

[34] C. Wu, A. Deshmukh, Z. Li, L. Chen, A. Alamri, Y. Wang, R. Ramprasad, G. A. Sotzing, and Y. Cao, "Flexible temperature-invariant polymer dielectrics with large bandgap," Adv. Mater. **32**, 2000499 (2020).

[35] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).

[36] T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen, and B. D. Olsen, "BigSMILES: A structurally-based line notation for describing macromolecules," ACS Cent. Sci. **5**, 1523–1531 (2019).

[37] "RDKit, open source toolkit for cheminformatics." See details of RDKit functionality from Python (11/20/2020).

[38] J. Mark, *Polymer Data Handbook* (Oxford University Press, 1999).