Predicting polymerization reactions via transfer learning using chemical language models

Brenda S. Ferrari 1 , Matteo Manica 2 , Ronaldo Giro 1,* , Teodoro Laino 2,3 , and Mathias B. Steiner 1,*

ABSTRACT

Polymers are candidate materials for a wide range of sustainability applications such as carbon capture and energy storage. However, computational polymer discovery lacks automated analysis of reaction pathways and stability assessment through retro-synthesis. Here, we report the first extension of transformer-based language models to polymerization reactions for both forward and retrosynthesis tasks. To that end, we have curated a polymerization dataset for vinyl polymers covering reactions and retrosynthesis for representative homo-polymers and co-polymers. Overall, we obtain a forward model Top-4 accuracy of 80% and a backward model Top-4 accuracy of 60%. We further analyze the model performance with representative polymerization and retro-synthesis examples and evaluate its prediction quality from a materials science perspective.

Introduction

Polymers have versatile properties and a wide range of applications^{1–3}. The optimization of polymeric materials and the development of new polymers are, however, time-consuming processes. Machine Learning (ML) techniques have been demonstrated to significantly accelerate the discovery process by predicting polymer properties^{4,5} or, more recently, by enabling the automated design and generation of new polymers with predefined target properties^{6–9}. Despite these advances, computational polymer discovery still faces major obstacles. Polymers are macromolecules which are formed by linking up smaller molecular units. Their synthesis typically involves various polymerization steps, with a multitude of possible links between monomer units. The prediction of thermodynamically stable polymer candidates, as well as the determination of a polymer's synthesizability¹⁰, is still affected by critical methodological limitations.

Recently, Caddeo et al.¹¹ reported ML and atomistic approaches for modeling the thermodynamic stability of polymer blends while Chen et al.¹² demonstrated a data-driven approach to automated retrosynthesis of target polymers. Kim et al.¹³ demonstrated the combination of ML-model based generation of new polymer candidates with a synthesizability analysis based on known polymerization reactions and commercially available reactants.

Despite the encouraging progress, significant gaps still exist in both methods and data domains. Currently, ML models do not exist for conducting retro-synthesis analysis on a range of co-polymers, polymer blends, ladder, cross-linked, and metal-containing polymers. Previous research has predominantly focused on homo-polymers, which can be easily represented as strings using the simplified molecular-input line-entry system (SMILES)^{14–16}. The recent development of advanced string representations for

¹IBM Research, Av. República do Chile, 330, CEP 20031-170, Rio de Janeiro, RJ, Brazil.

²IBM Research Europe, Saümerstrasse 4, 8803 Rüschlikon, Switzerland.

³National Center for Competence in Research-Catalysis (NCCR-Catalysis), Switzerland

polymers^{17,18} opens up new opportunities for modeling co-polymers¹⁷ as well as comb, branched, brushed, and star polymers^{9,18–20}.

Another critical issue is that the available polymer reaction datasets do not consider the influence of solvents, catalysts, and experimental conditions. In addition, the data used to train ML models are not always made available publicly, compromising the reproducibility of model predictions. Overall, the lack of open data and open models severely hinders the advancement of computational polymer discovery.

In this work, we report the first extension of a transformer-based language model^{21,22} to polymerization reaction trained on a curated reaction dataset for vinyl polymers. We train the polymerization models for both forward and backward prediction tasks, addressing both homo-polymers and co-polymers consisting of up to two monomers. Our model predicts reactants, as well as reagents, solvents, and catalysts for each step of the retro-synthesis. Finally, we show that our models are able to perform two essential tasks as visualized in Fig.1): (i) given a set of precursors, to predict a polymer product and (ii) given a polymer, to suggest potential disconnections for synthetic strategies. To enable validation and reuse, we have made our models and data available in public repositories.

Results and Discussion

In Fig.2, we visualize the end-to-end workflow for predicting polymerization reactions. The workflow includes dataset preparation and training of reaction and retrosynthesis prediction models, respectively. The training dataset was generated based on the publically available USPTO reaction dataset^{23,24} which contains chemical reactions of organic compounds extracted from US patents issued between 1976 and 2016. For extracting polymerization reactions from the dataset, we have designed a Python tool (see code availability section) that operates based on specific keywords. To ensure the selection of polymerization reactions only, we have employed a manual curation process that involves an individual review step of the reactions chosen by the automated procedure. Overall, we have analyzed 795 data entries for vinyl homo-polymers and co-polymers, respectively, resulting in two distinct datasets containing 3932 and 2965 reactions. These datasets cover all the possible combinations of the 795 reaction examples (details can be found in the Methods section).

In general, polymer properties are determined to a large extent by how the monomer units are interconnected. For the purpose of our study, we have chosen linear chains as topological representations. For accurately predicting polymerization reactions, it is essential to correctly identify and label head and tail positions of the repeat units. To that end, we have adopted two distinct strategies. In the first approach, we have adapted an existing tool for assigning head and tail atoms, referred to as Monomers-to-Polymer (M2P)²⁵. In the second approach, we have developed a Python tool for Head-and-Tail assignment (HTA). We have provided extensive descriptions related to both HTA and M2P workflows in the Methods section. By using the two techniques, we have assigned head and tail atoms to constituent units within our polymer reaction dataset. We have then trained models on the two distinct datasets, labeled HTA and M2P, for comparative analysis of their predictive performance.

The modified M2P method can be applied to oligomers and assigns the positions of head and tail atoms in linkage bonds. The HTA method assigns head and tail atoms within monomers, thus defining the polymeric repeat unit. For facilitating the comparison of the ML models trained with the HTA and M2P datasets, respectively, we have also performed head and tail assignment in oligomers using the HTA routine. Throughout the training phase, the HTA dataset contained both monomers and oligomers, while the M2P dataset contained only oligomers. The inclusion of monomers within the HTA dataset enables the ML model to predict monomeric units of both homopolymers and copolymers. As the M2P dataset contains only oligomers, the respective model is not expected to predict homopolymer reactions correctly.

For reaction and retrosynthesis prediction modeling, we have used the Molecular Transformer architecture introduced by Schwaller et al.^{21,22}. In brief, the model is based on a vanilla transformer architecture²⁶ trained on textual representations of molecules. A Molecular Transformer casts chemical reaction prediction as a language modeling task²⁷. We have encoded chemical reactions as sentences using reaction SMILES representation¹⁴ of reactants, reagents as well as solvents and catalysts, along with the products. We have modeled forward- or retro-reaction predictions as a translation task from one language, i.e., reactants-reagents, to another language, i.e. products. For training purposes, we have formally divided the reaction SMILES into source (reactants and reagents) and target (products) instances. Since HTA and M2P datasets include different target outcomes for the same source instance, we have performed a splitting solely based on the targets. For model training, we have split the datasets on products in 95% for training and 5% for testing to ensure that no polymer (product) appears in both data sets.

To assess the performance of the Molecular Transformer trained on the two training datasets, we have used the Top-N accuracy metric for both forward and backward prediction models following the method reported in²². We have calculated the model accuracy by considering the number of exact matches between the predicted canonical SMILES and the ground truth in the datasets. The Top-N accuracy considers that the ground truth canonical SMILES was found within the first N suggestions of the model. For example, if the ground truth target was found as the first suggestion in 70 out of 100 examples, it means Top-1 is 70%. While round-trip is the generally preferred method for verifying the performance in the context of single-step retro-synthetic models²², the datasets analyzed in our work link precursors to multiple products. In this case, the round-trip accuracy could be misleading, as multiple forward predictions are still valid for a precursor set and multiple products map to the same precursors. To avoid this, we have used Top-N accuracy for evaluating the performance of both forward and backward models.

In Fig.3, we show the prediction model performance obtained for the two datasets. The M2P dataset shows better performance overall in both forward and backward models, see Fig.3a-b. In backward predictions, we observe the general trend that the higher the number of training steps, the higher the model accuracy. For forward predictions, this trend only manifests in certain intervals of the Top-N range. The accuracy increases monotonously in both forward and backward modes, albeit with different slopes. We observe a sharp accuracy increase in forward model for M2P around Top-3 and HTA around Top-4, respectively. This could be explained by the number of possible reaction outcomes. While M2P provides n reaction outcomes as oligomers built from combination of reagent monomers, HTA also provides the repeat units as product of polymerization. This means that HTA provides n+1 or n+2 results, depending on the number of reagent monomers involved in the reaction. On average, M2P returns 4 possible reaction outcomes while HTA returns 5 or 6.

The observation that the M2P dataset yields superior model performance could be due to the simpler learning process of polymerization rules within this dataset. The M2P algorithm polymerizes monomers in all possible functional groups and chooses a representative structure randomly. Due to the random character of the M2P algorithm, different realizations result in different choices of representative structures, affecting the ML training performance. In comparison, the HTA algorithm identifies reactive sites through the analysis of nucleophile and electrophile atoms, applying the Mulliken's scheme^{28–30} for identifying the most probable structure relating to chemical rules. In other words, M2P structures are a combination of all possible bond connections between monomers, while HTA structures are combinations of all possible connections between reacting sites.

To clarify this point, let us consider how the repeat units in the HTA dataset are linked up to form oligomers. A bond between two vinyl monomers with only secondary carbon atoms may be formed as visualized in the example shown in Fig.4a. We note that the polymeric repeat unit generated by HTA was considered for inclusion into the dataset, however, it was disregarded in the distribution analysis. This is

also the case for oligomers with tertiary carbons.

In case 1, the bond is formed between the carbon atoms at the end of the monomers in the chain. As a result, both head and tail are localized at external atoms of the reaction site. We refer to this connection type as tail-tail. In case 2, head and tail are localized at internal and external carbon positions, respectively. We refer to this connection type as head-tail. Finally, in case 3, the bond occurs between secondary carbon atoms of the double bond. Once polymerized, both head and tail atoms are located at internal carbon atom sites. We refer to this connection type as head-head. By analyzing the case distribution in the dataset for model training, see Fig.4b, we find that the HTA dataset contains 1/3 of each case for oligomers with 3 different combinations while the ratio is 1/2/1 for oligomers with 4 different combinations. The latter can be explained by the twofold possibility in case 2 of bond formation due to the presence of two monomers. Note, that the M2P dataset does not have a fixed case ratio. This is because M2P performs the polymerization for all possible functional groups of the molecular structure, see Fig.4c.

Those differences on the distribution are observed on examples in Fig.4d. For the butadiene isoprene polymer with its four potential polymerizations, the vinyl bond case ratio 1/2/3 representing cases 1, 2 and 3, respectively, see Fig.4a, is 1/2/1 for HTA and 0/2/2 for M2P. Similarly, in the case of allyl methacrylate, we obtain the case ratio 1/2/1 for HTA and 0/2/2 for M2P. In case of M2P, the polymerization is performed by considering all the functional groups of the monomer. The results observed in Fig.3a-b could indicate that the model has learned this pattern efficiently. The larger spread of accuracy values observed in the retro-synthesis model could be due to the specifics of the oligomers.

While we obtain overall better modeling results with M2P, both datasets reveal interesting insights. Despite showing a Top-1 accuracy below 10%, the forward model exhibits Top-4 and Top-6 accuracy around 80%, which suggests a direct relation with the way the two datasets have been compiled. Indeed, by construction, the same set of reactants are associated with multiple polymers. The backward model has a Top-1 accuracy of about 60% for M2P and 40% for HTA. The lower accuracy observed in HTA could be explained by the ease that the model may have learned the polymerization pattern represented in M2P data, as explained previously. We will expand this analysis in the following paragraphs by investigating the usefulness of the model outputs from a materials science perspective.

For our domain applicability analysis, see Methods section for details, we have selected representative polymers from the literature ^{31–38}. A comparison of these reactions reveal product similarities ranging from 0 to 30% for HTA and M2P datasets while reactants similarities range from 0 to 12%, see Supplemental Table S1. Co-polymers show increased similarity values in M2P, about 3-6% higher, attesting to their representation in the training data. Homo-polymers exhibit increased similarity of about 4% in HTA as the dataset includes monomer representations.

Overall, both models correctly predicted 6 out of 8 reactions in Top-4 and could suggest at least one correct monomer in all the examples studied. The HTA based model correctly predicted 3 out of 8 reactions in Top-1 and 4 out of 8 reactions in Top-4, while the M2P based model correctly predicted 1 out of 8 reactions in Top-1 and 2 out of 8 reactions in Top-4. Note, that the HTA based model predominantly matches homo-polymers while M2P matches mainly co-polymers. The pattern is plausible as HTA contains the monomers of all polymers while M2P does only contain oligomers.

For the polymerization example of styrene, see figure 5a), the HTA based model achieves a full SMILES match at Top-1 as well as the representation of a possible oligomer structure, with 2 connected repeat units, at Top-3. In case of the M2P based model, we do not obtain a match for the actual product. The oligomer representation is shown for Top-3 and Top-4. For the polymerization of the co-polymer p(SBMA-nBA), see Figure 5b, the model predicts an exact product match for Top-1, along with the all other bond formation possibilities on Top-2 to Top-4. This means that the model is able to correctly predict the connections in the polymerization reactions. While the HTA model failed to predict the actual result,

the model was able to identify the correct head and tail positions of one of the repeat units (Top-1). In addition, the model suggested fragments of the monomer seen as Top-2 and Top-4.

One interesting exception is shown in Supplemental Fig.S1b. In the polymerization of p(xMA), a co-polymer, both models suggested incorrect structures at Top-1. However, the HTA based model generates the correct repeat units for all four predictions, Top-2 being the exact match. The M2P based model merely predicts all possible links between carbon atoms for generating the polymeric bond, and one of the monomers is an exact match. For p(St-BuA), see Supplemental Material, Fig.S1b, the HTA based model predicts the correct repeat units in Top-1 and Top-2. As expected, however, it fails to generate the oligomer. Nevertheless, the M2P based model predicts the correct monomers and the exact match is shown in Top-4.

In the example of Polyvinyl chloride polymerization, see Supplemental Material, Fig.S2a, we observe an interesting model behavior. While neither HTA nor M2P data underwent special processing for monomers/oligomers with protection groups, the model learned to predict output without the protection group. The HTA based model suggested the correct structure for polyvinyl chloride at Top-1, without the protection group. The M2P based model, however, failed to generated an output that resembled the ground-truth structure. In the polymerization of p(DOM-DVB), see Supplemental Fig.S2b, we observe that both models struggles to predict polymers in which monomers have the double bond in the middle of the chain. Nevertheless, both models correctly suggested one of the monomers and its bonds combinations.

Both models correctly predict oligomers formed by monomeric units with halogens, such as chlorine. Since all training data is tagged with a token (Rn) representing the location for the continuation of the chain, all model predictions suggest the formation of monomers with that token in its structure. This is shown in Supplemental Figure S3a for the polymerization of p(tC-tBuM) copolymer. The HTA based model accurately predicts one of the monomers and its combinations while the M2P based model fails this task. Even in the presence of a large number of reactants, catalysts, and solvents, the model is able to correctly predict the polymers, as shown in case of Poly(n-butyl methacrylate), see Supplemental Fig.S3b. As expected for homo-polymers, the HTA model predicts the exact match in Top-1 along with some monomer combinations in Top-2 and Top-3 while the M2P based model predicts the combinations of the monomer in Top-1 to Top-3.

For the curated examples, the HTA based model predicts a higher number of exact matches for the polymer structures in Top-1 (3 out of 8) and Top-4 (4 out of 8), respectively. In cases of incorrect predictions, the model delivered at least one of the monomers correctly. The model trained with M2P data had limitations regarding homo-polymers, as expected. Nevertheless, the M2P model correctly predicts complex co-polymers and a very close match for p(tC-tBuM) copolymer, a pattern not represented in the training dataset. Both models appear to have complementary performance, predicting exact matches for 6 out of 8 reactions and suggesting at least one correct monomer for all the examples studied. For increasing the likelihood of a suitable prediction outcome, we, therefore, recommend the joint utilization of both HTA and M2P based models for domain specific applications

Conclusion

In summary, we have reported the curation of a vinyl polymerization reaction dataset and the training of a Molecular Transformer algorithm for predicting polymerization (forward) and retro-synthesis (backward) reactions. For dataset curation, we have introduced two novel algorithms for assigning head and tail positions, named HTA and M2P. We have applied both algorithms to process 795 data entries for vinyl homo-polymers and co-polymers and produced two separate datasets with 3932 and 2965 reactions, respectively, representing all possible combinations of the 795 reaction examples. Upon training, the

Molecular Transformer exhibits a forward-model (Top-4 and Top-6) accuracy around 80% for both datasets. The retro-model exhibits a Top-1 accuracy of about 60% for the M2P dataset and 40% for the HTA dataset.

We have showcased the capabilities of the models through a case study involving eight reactions. These reactions were selected based on examples provided in the literature. Both models have predicted 6 out of 8 reactions as exact match at Top-4, and suggested at least one correct monomer for all the examples studied. The models work in a complementary manner, as the model trained with the HTA dataset produces better results for homo-polymers while the model trained with the M2P dataset predicts better matches for co-polymers.

Based on our analysis of the strengths and limitations of the Molecular Transformer approach, we expect that extending the model training to include other polymer classes will broaden model applicability and further increase the robustness of prediction outcomes. The lack of available data on polymerization reactions and tools for head and tail assignment were major challenges we have encountered in this work. Therefore, we have made our curated datasets and tools publicly available for reuse and validation.

Methods

Polymerization dataset

The polymerization reactions and polymer names were extracted from a publicly available dataset²³ derived from the patent mining work of Lowe²⁴. This dataset is composed by approximately 1.8M chemical reactions, extracted from 1976 to September 2016 USPTO granted patents. A Python script was developed to automate the data extraction. Only chemical reactions and molecule names that presented the keyword "polymerization" on the experimental procedure text were chosen. After the automated step, a manual validation was performed to remove data entries in which the "polymerization" keyword was related to any information not compatible with the reaction type. In this step the number of data points were reduced from 8.668 to 3.286 possible polymerization reactions. In the Lowe²⁴ dataset, the head and tail atom positions to define the polymer repeat units of polymerization reactions products are missing. How these monomers are linked play an important role in polymer properties³⁹. Since there was no established methodology to perform the assignment of the head and tail in polymer structures represented by SMILES notation, Python tools with two different approaches were developed to perform this task. In the first approach we used an in house developed Python tool, called HTA (Head-and-Tail Assignment), to assign the head and tail atoms (more details see Methods section). In the second approach a modified version of Monomers-to-Polymer (M2P)²⁵ tool was developed to assign the head and tail atoms. These two approaches resulted in two datasets, composed by 795 data entries, related to vinyl homo-polymers and co-polymers with 2 monomer and were properly clean from duplicates and erroneous reactions. Besides the head and tail assignment, another two datasets were generated by describing all the possible product outcomes which are represented by one or two products and the different bond formation between the monomers. The bond formations were performed by the combination of monomers using rdkit.Chem.rdChemReactions method. For that, all the monomers combination were considered according to M2P and HTA algorithms. On the HTA algorithm the monomers were also considered as possible outcome of the reaction. In this sense, regarding the number of results m2p=n and hta=n+1/n+2. This increased the number of reactions from 795 to 3932 and 2965 reactions, for HTA and M2P respectively. In summary, four datasets were generated and two datasets were used to train our model: the all monomers combination datasets for HTA and M2P.

Data distribution

Both M2P and HTA datasets were sorted by polymer name and repeating unit, the latter alphabetically and by length. All the results for the same polymer were grouped in lists during pre-processing process. The modified M2P tool assign the head and tail atom positions (linkage bounds) in oligomers, while the HTA tool in the monomer, defining the polymeric repeat unit. With the purpose to avoid any bias during the ML training model between the two datasets, we also considered the head and tail assignment with the HTA tool in oligomers. This fact adds another level of complexity: how the repeat units are linked. There are three possible cases: (i) tail-tail; (ii) head-tail and (iii) head-head. For the extraction of the distribution of cases, there were set SMARTS⁴⁰ for each polymerization case and after a dearomatization process, all the SMILES¹⁴ were compared to the SMARTS set, using the RDKit⁴¹ library. SMARTS⁴⁰ is a chemical structure query language for describing molecule patterns. RDKit can import SMARTS queries for use in searching of SMILES patterns. Cases that deviated from the standard SMARTS query pattern (i.e., tertiary carbons that could cause uncertainties on the algorithm) were not considered. After post-processing, both datasets were merged, since only equal polymers were considered on the comparison, and a distribution chart was built with the results.

Applicability domain analysis

The polymers that were used on this case study were manually extracted from the literature^{31–38}. The SMILES representation of polymers were canonicalized using the RDKit⁴¹ package. The fingerprint calculation was performed by defining the fingerprints of the input data and the data used on the Molecular Transformer training using RDKFingerprint⁴¹ followed by the comparison between both datasets. Each input data fingerprint was compared with the fingerprints of the whole training data. The results obtained comprised on the mean of the comparison results and the maximum value on the list. This process was performed separately for reactants/reagents and products.

HTA algorithm

For the head and tail assignment using the HeadTailAssigner (HTA) tool, the reaction SMILES was used as input. However, the algorithm also accepts monomer SMILES as input. Following the pre-processing analysis, the most probable monomer in the reaction string was defined by comparing the products with the reactants. The last step was performed by a fingerprint similarity analysis, using the RDKFingerprint⁴¹ and maxPath=7 and a comparison using Tanimoto Similarity^{41,42}. The vinyl class is the focus of this work, but the algorithm may also identify and assign head and tail of polyamides, polyesters, polyurethanes and polyethers. To define the polymer class, the algorithm searches all the possible functional groups on the molecular structure by substructure match with the SMARTS pattern of each organic function. In a next step, it compares the atomic index of nucleophilicity⁴³ and the functional groups extracted from the monomer. If the monomer smiles has only one functional group, a SMARTS pattern is acquired to classify the polymerization mechanism. If the monomer smiles has two or more functional groups, the priority of polymerization is decided based on the atomic index of nucleophilicity⁴³. The atomic index of nucleophilicity of an atom X involving only the highest occupied molecular orbital (HOMO) n is defined as⁴³:

$$R_X = \frac{\sum_{\alpha}^{X} |C_{\alpha,n}|^2}{(1 - \varepsilon_n)} \tag{1}$$

where $C_{\alpha,n}$ are the molecular orbital expansion coefficients of α th atomic orbital on molecular orbital n (HOMO) and ε_n is the HOMO energy.

The R_X was calculated within STO-3G basis set and with the Mulliken's population analysis^{28–30} scheme. All the quantum states functions were calculated at RHF theory level, using the standard *ab initio* quantum chemistry package GAMESS⁴⁴ version 2021 R2.

In summary, the higher the atomic population value in an atom, higher the atom index of nucleophilicity R_X , which means, the atom has more probability on being the polymerization site⁴³. The condition is set depending on the relation between polymerization class and the functional groups present in the structure. If one atom has a higher R_X but its functional group is not represented in any polymer class, the algorithm is going to keep searching until it finds an atom that is represented in an existing polymer class. After obtaining a match, the functional groups are concatenated up until it is a match with a previously defined class. The mechanism is defined depending on the polymer class described previously. If the class is vinyl and the algorithm detects the presence of an specific catalyst, it may also define if the mechanism is anionic, cationic or radicalar. With all the information obtained previously, the algorithm defines the head and tail by assigning the atom id of the respective nucleophile and electrophile on the functional group responsible for the polymerization.

M2P algorithm

For the head and tail assignment using Monomers to Polymers (M2P)⁴⁵, a modified version of the M2P algorithm was used. According to the authors "The library can generate multiple replicate structures to create polymer chains represented at the atom and bond level. RDKit⁴¹ reaction SMARTS⁴⁰ are used to manipulate the molecular structures and perform in silico reactions. The polymer chemistries available include vinyls, acrylates, esters, amides, imides, and carbonates."⁴⁵. From the source-code, the algorithm was modified to generate head and tail assignments for vinyl polymerization only if the user checks True for the head and tail creation parameter. The vinyl polymerization comprises the initiation, propagation and termination steps with token atoms (Kr, Xe and Rn) used on the reaction SMARTS to define the bond formation site. In the end of the polymerization, these tokens would be deleted, to keep only the polymer product as a result. For the modified version, the token atoms were added on the initiation, propagation and termination step to represent the formation of the head and tail atoms on the polymer. In the end of the polymerization process, these tokens remain on the polymers to represent the head and tail assignment. This treatment was also extended for co-polymers with 3 monomers.

Model training for forward and backward reaction prediction

As model, for both forward and backward reaction prediction, we considered the Molecular Transformer proposed by Schwaller et al.²¹. Encoders follow a standard *transformer* architecture with 6 layers, word vectors and RNN decoders of size 512, the gradient was accumulated 8 times with a maximum vector norm of 0.0, and *adam* was used as an optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.998$). Batch size was set to 4096, and the batch type as well as the gradient normalisation method to *tokens*. The learning rate was set to 2.0 with *noam* as decay method. Dropout and label smoothing (ε) were set to 0.1. Parameter initialisation was disabled and position encoding enabled. All models were trained using a version of OpenNMT⁴⁶ adapted for the Molecular Transformer⁴⁷. Compared to the standard Molecular Transformers we extended the model and tokenizer to handle head and tail representations using noble gasses as additional tokens. We trained models on the datasets generated both with the HTA and the M2P algorithm ad compared the both backward and forward performance.

Data Availability

D. Lowe's dataset 1976_Sep2016_USPTOgrants_cml.7z used to extract the polymerization reaction data is available under the doi:10.6084/m9.figshare.5104873.v1 - at

```
https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_ 1976-Sep2016_/5104873?file=8664364
```

The training dataset file hta_dataset_all_combinations.csv containing polymerization reactions in SMILES format and with the head and tail atoms assigned by the Python tool HTA is available under the doi:10.24435/materialscloud:zw-be - at

```
https://archive.materialscloud.org/record/2023.137
```

The training dataset file m2p_dataset_all_combinations.csv containing polymerization reactions in SMILES format and with the head and tail atoms assigned by the modified version of M2P Python tool is available under the doi:10.24435/materialscloud:zw-be - at

```
https://archive.materialscloud.org/record/2023.137
```

The file trained_models.zip contains the Machine Learning training models (forward and retrosynthesis) and is available under the doi:10.24435/materialscloud:zw-be - at

```
https://archive.materialscloud.org/record/2023.137
```

Code Availability

The code for extracting the polymerization reaction data from Daniel Lowe's dataset is available at: https://github.com/IBM/XLMExtractor-chem-reaction.

The code for assigning the head and tail atoms using quantum chemistry and polymerization mechanisms information is available at:https://github.com/IBM/HeadTailAssign.

The code for assigning the head and tail atoms based on the Monomers to Polymers (M2P) tool is available at: https://github.com/IBM/m2o-head-tail-assign

The code for model training is available at: https://github.com/rxn4chemistry.

References

- 1. Arshad, M., Zubair, M., Rahman, S. S. & Ullah, A. Polymers for advanced applications. In Polymer Science and Nanotechnology, 325–340 (Elsevier, 2020). URL https://doi.org/10.1016/b978-0-12-816806-6.00014-5.
- **2.** Namazi, H. Polymers in our daily life. <u>BioImpacts</u> **7**, 73–74 (2017). URL https://doi.org/10.15171/bi.2017.09.
- **3.** Patel, V. K., Kant, R., Chauhan, P. S. & Bhattacharya, S. Introduction to applications of polymers and polymer composites. In <u>Trends in Applications of Polymers and Polymer Composites</u>, 1–6 (AIP Publishing, 2022). URL https://doi.org/10.1063/9780735424555_001.
- **4.** Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: A data-powered polymer informatics platform for property predictions. The Journal of Physical Chemistry C **122**, 17575–17585 (2018). URL https://doi.org/10.1021/acs.jpcc.8b02913.
- 5. Tran, H. D. et al. Machine-learning predictions of polymer properties with polymer genome.

 Journal of Applied Physics 128, 171104 (2020). URL https://doi.org/10.1063/5.0023759.

- **6.** Kim, C., Batra, R., Chen, L., Tran, H. & Ramprasad, R. Polymer design using genetic algorithm and machine learning. Computational Materials Science **186**, 110067 (2021). URL https://doi.org/10.1016/j.commatsci.2020.110067.
- 7. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. Chemistry of Materials 32, 10489–10500 (2020). URL https://doi.org/10.1021/acs.chemmater.0c033332.
- 8. Giro, R. et al. AI powered, automated discovery of polymer membranes for carbon capture. npj Computational Materials 9 (2023). URL https://doi.org/10.1038/s41524-023-01088-3.
- 9. Park, N. H. et al. Artificial intelligence driven design of catalysts and materials for ring opening polymerization using a domain-specific language. <u>Nature Communications</u> **14**, 3686 (2023). URL https://doi.org/10.1038/s41467-023-39396-3.
- Aziz, A. & Carrasco, J. Towards predictive synthesis of inorganic materials using network science. Frontiers in Chemistry 9 (2021). URL https://doi.org/10.3389/fchem.2021.798838.
- 11. Caddeo, C., Ackermann, J. & Mattoni, A. A theoretical perspective on the thermodynamic stability of polymer blends for solar cells: From experiments to predictive modeling. <u>Solar RRL</u> 6, 2200172 (2022). URL https://doi.org/10.1002/solr.202200172.
- 12. Chen, L., Kern, J., Lightstone, J. P. & Ramprasad, R. Data-assisted polymer retrosynthesis planning. Applied Physics Reviews 8, 031405 (2021). URL https://doi.org/10.1063/5.0052962.
- 13. Kim, S., Schroeder, C. M. & Jackson, N. E. Open macromolecular genome: Generative design of synthetically accessible polymers. <u>ACS Polymers Au</u> (2023). URL https://doi.org/10.1021/acspolymersau.3c00003.
- **14.** Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. <u>Journal of Chemical Information and Modeling</u> **28**, 31–36 (1988). URL https://doi.org/10.1021/ci00057a005.
- 15. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. algorithm for generation of unique SMILES notation. <u>Journal of Chemical Information and Computer Sciences</u> 29, 97–101 (1989). URL https://doi.org/10.1021/ci00062a008.
- **16.** Weininger, D. SMILES. 3. DEPICT. graphical depiction of chemical structures. Journal of Chemical Information and Modeling //doi.org/10.1021/ci00067a005.
- 17. Lin, T.-S. et al. BigSMILES: A structurally-based line notation for describing macromolecules. ACS Central Science 5, 1523-1531 (2019). URL https://doi.org/10.1021/acscentsci.9b00476.
- **18.** Lin, T.-S. et al. PolyDAT: A generic data schema for polymer characterization. <u>Journal of Chemical Information and Modeling</u> **61**, 1150–1163 (2021). URL https://doi.org/ 10.1021/acs.jcim.1c00028.
- 19. Guo, M. et al. Polygrammar: Grammar for digital polymer representation and generation. Advanced Science 9, 2101864 (2022). URL https://doi.org/10.1002/advs. 202101864.

- **20.** Mohapatra, S., An, J. & Gómez-Bombarelli, R. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. Machine Learning: Science and Technology **3**, 015028 (2022). URL https://doi.org/10.1088/2632-2153/ac545e.
- 21. Schwaller, P. et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Central Science 5, 1572–1583 (2019). URL https://doi.org/10.1021/acscentsci.9b00576.
- 22. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. Chemical Science 11, 3316–3325 (2020). URL https://doi.org/10.1039/c9sc05704h.
- 23. Lowe, D. Chemical reactions from US patents (from 1976 to September 2016). https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873. Accessed: 2022-11-9.
- **24.** Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge (2012).
- **25.** Wilson, N., St John, P. & Crowley, M. m2p (monomers to polymers). Tech. Rep., National Renewable Energy Lab.(NREL), Golden, CO (United States) (2020).
- **26.** Vaswani, A. et al. Attention is all you need. Advances in Neural Information Processing Systems **30** (2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- 27. Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M. & Grzybowski, B. A. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses.

 <u>Angewandte Chemie International Edition</u> 53, 8108–8112 (2014). URL https://doi.org/10.1002/anie.201403708.
- 28. Mulliken, R. S. Electronic population analysis on lcao-mo molecular wave functions. i. The Journal of Chemical Physics 23, 1833-1840 (1955). URL https://doi.org/10.1063/1.1740588.
- **29.** Mulliken, R. S. Electronic population analysis on lcao—mo molecular wave functions. ii. overlap populations, bond orders, and covalent bond energies. The Journal of Chemical Physics **23**, 1841—1846 (1955). URL https://doi.org/10.1063/1.1740589.
- **30.** Mulliken, R. S. Electronic population analysis on lcao-mo molecular wave functions. iv. bonding and antibonding in lcao and valence-bond theories. The Journal of Chemical Physics **23**, 2343–2346 (1955). URL https://doi.org/10.1063/1.1741877.
- **31.** Saleh, N. <u>et al.</u> Surface modifications enhance nanoiron transport and NAPL targeting in saturated porous media. <u>Environmental Engineering Science</u> **24**, 45–57 (2007). URL https://doi.org/10.1089/ees.2007.24.45.
- 32. Francisco-Vieira, L., Benavides, R., Cuara-Diaz, E. & Morales-Acosta, D. Styrene-co-butyl acrylate copolymers with potential application as membranes in PEM fuel cell.

 International Journal of Hydrogen Energy 44, 12492–12499 (2019). URL https://doi.org/10.1016/j.ijhydene.2019.01.181.

- **33.** Concilio, M., Nguyen, N. & Becer, C. R. Oxazoline-methacrylate graft-copolymers with upper critical solution temperature behaviour in yubase oil. <u>Polymer Chemistry</u> (2021). URL https://doi.org/10.1039/d1py00534k.
- **34.** Atta, A. M., Brostow, W., Lobland, H. E. H., Hasan, A.-R. M. & Perez, J. M. Porous polymer oil sorbents based on PET fibers with crosslinked copolymer coatings. <u>RSC Advances</u> **3**, 25849 (2013). URL https://doi.org/10.1039/c3ra44759f.
- **35.** Chen, X.-P. & Qiu, K.-Y. ?living? radical polymerization of styrene with AIBN/FeCl3/PPh3 initiating system via a reverse atom transfer radical polymerization process. Polymer International **49**, 1529–1533 (2000). URL https://doi.org/10.1002/1097-0126 (200011) 49:11<1529:: aid-pi564>3.0.co; 2-b.
- **36.** Ogieglo, W., Wormeester, H., Eichhorn, K.-J., Wessling, M. & Benes, N. E. In situ ellipsometry studies on swelling of thin polymer films: A review. <u>Progress in Polymer Science</u> **42**, 42–78 (2015). URL https://doi.org/10.1016/j.progpolymsci.2014.09.004.
- **37.** Dena, A. S. A., Ali, A. M. & El-Sherbiny, I. M. Surface-imprinted polymers (sips): Advanced materials for bio-recognition. J Natural Sciences Publishing Cor (2020).
- **38.** Ibrahim, K. Towards more controlled poly(n-butyl methacrylate) by atom transfer radical polymerization. European Polymer Journal **39**, 939–944 (2003). URL https://doi.org/10.1016/s0014-3057 (02) 00309-9.
- 39. Zhou, H., Badashah, A., Luo, Z., Liu, F. & Zhao, T. Preparation and property comparison of ortho, meta, and para autocatalytic phthalonitrile compounds with amino group.

 Polymers for Advanced Technologies 22, 1459–1465 (2011). URL https://doi.org/10.1002/pat.2018.
- **40.** SMARTS a language for describing molecular patterns. URL https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
- **41.** RDKit: open-source cheminformatics. https://www.rdkit.org. DOI: 10.5281/zen-odo.591637.
- **42.** Tanimoto, T. T. <u>Elementary mathematical theory of classification and prediction</u> (International Business Machines Corp., 1958).
- **43.** Szczepanik, D. W. & Mrozek, J. Nucleophilicity index based on atomic natural orbitals. <u>Journal of Chemistry</u> **2013**, 1–6 (2013). URL https://doi.org/10.1155/2013/684134.
- **44.** Barca, G. M. J. et al. Recent developments in the general atomic and molecular electronic structure system. The Journal of Chemical Physics 152, 154102 (2020). URL https://doi.org/10.1063/5.0005188.
- **45.** Wilson, N., St John, P. & Crowley, M. Monomers to polymers (m2p) github (2022). URL https://github.com/NREL/m2p.
- **46.** Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. OpenNMT: Open-source toolkit for neural machine translation. In <u>Proceedings of ACL 2017</u>, System Demonstrations, 67–72 (Association for Computational Linguistics, Vancouver, Canada, 2017). URL https://doi.org/10.18653/v1/P17-4012.
- **47.** RXN, I. Onmt adaptation for rxn4chemistry. URL https://github.com/rxn4chemistry/OpenNMT-py.

Acknowledgements

T. L. acknowledges support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

Author Contributions

B. S. F. created and curated the polymerization reaction dataset and co-wrote the manuscript. M. M developed Machine-Learning models and co-wrote the manuscript. R. G. conceived the work and co-wrote the manuscript. T. L. conceived the work and co-wrote the manuscript. M. B. S. conceived the work and co-wrote the manuscript.

Competing financial interests:

The authors declare no competing financial interests.

Additional Information

Supplementary information

Supplementary Information, including Supplementary Table S1 and Supplementary Figures S1-S4, are available as a pdf-file

Correspondence

and requests for materials should be addressed to mathiast@br.ibm.com

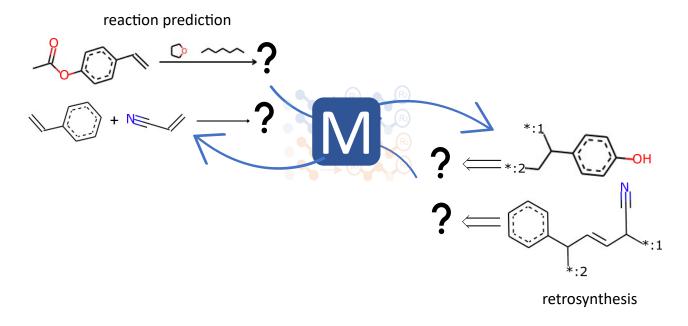


Figure 1. Problem representation. A Molecular Transformer model is being created for answering the following questions: "Given a set of reactants, which polymer could be obtained as product?" and "Given a certain polymer, how could it be synthesized?"

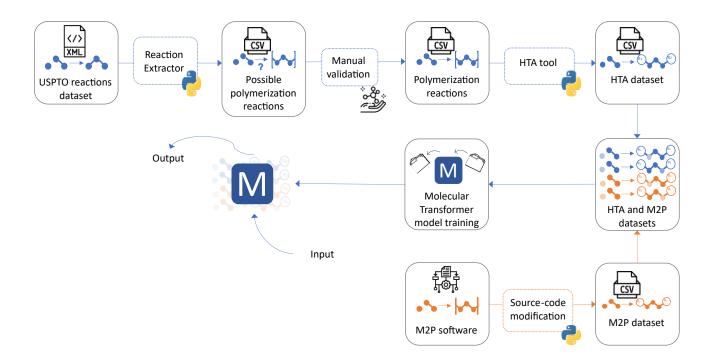


Figure 2. Methodology flowchart. The workflow for predicting polymerization reactions (forward) and retro-synthesis analysis (backward) comprise data preparation and treatment, head and tail assignment with two different methodologies (HTA and M2P), model training and predictions in forward and backward directions.

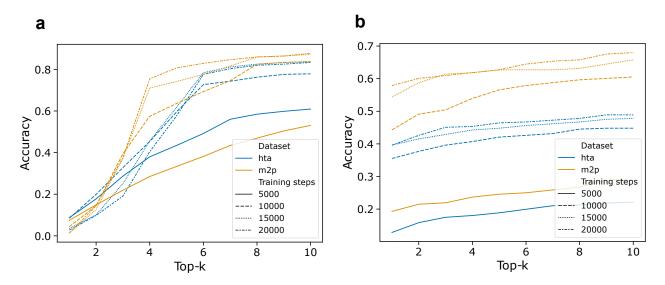


Figure 3. Prediction model performance. a) Polymerization reaction prediction (forward model) accuracy. b) Retro-synthesis prediction (backward model) accuracy.

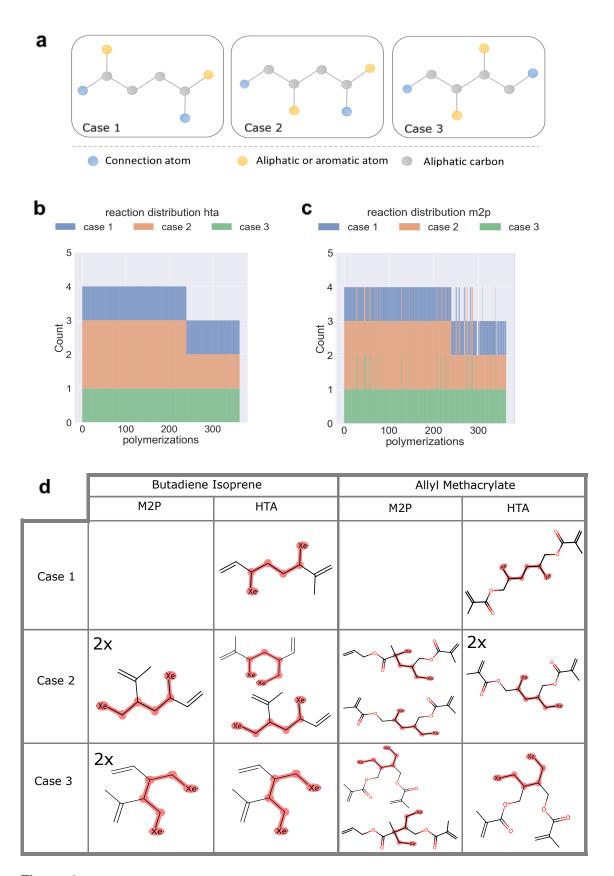


Figure 4. Data representation. "2X" representation means the same structure appears twice. a) SMARTS representation of the vinyl bond formation. b) Comparative distribution of HTA data. c) Comparative distribution of M2P data. d) Examples of Butadiene Isoprene and Allyl Methacrylate.

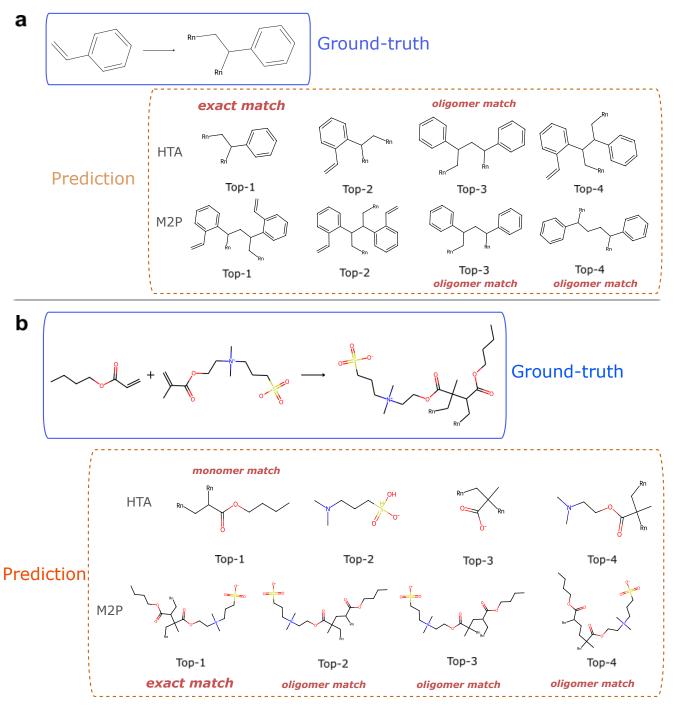


Figure 5. Representative examples. Model predictions using the Molecular Transformer trained on HTA and M2P datasets, respectively. Catalysts, solvents, and stochiometry are not shown. a) Polystyrene. b) p(SBMA-nBA) copolymer. In 2D molecules representations carbon atoms are in black, oxygen and hydroxyl in red, nitrogen in dark blue, and sulfur in yellow. The connection points of polymer repeat units are represented with Rn atoms.

SUPPLEMENTARY INFORMATION

Results and Discussion

We provide supplemental table and figures to support the discussion on the main manuscript. These figures demonstrate the performance of the trained Molecular Transformer mode. The polymers that were considered on these case studies were manually extracted from the literature followed by the applicability domain analysis (for more details see Methods section).

M2P reactants			HTA reactants		
name	total mean	total maximum	name	total mean	total maximum
p(St-BuA)	0.12	0.15	p(St-BuA)	0.12	0.15
Polystyrene	0.11	0.15	Polystyrene	0.11	0.15
p(DOM-DVB)	0.10	0.13	p(DOM-DVB)	0.10	0.13
p(SBMA-nBA)	0.08	0.14	p(SBMA-nBA)	0.08	0.14
p(xMA)	0.08	0.14	p(xMA)	0.08	0.14
Poly(n-butyl			Poly(n-butyl		
methacrylate)	0.07	0.13	methacrylate)	0.07	0.14
Polyvinyl chloride	0.06	0.08	Polyvinyl chloride	0.06	0.07
p(tC-tBuM)	0.05	0.12	p(tC-tBuM)	0.05	0.12

M2P products			HTA products		
name	total mean	total maximum	name	total mean	total maximum
p(DOM-DVB)	0.28	0.28	p(St-BuA)	0.24	0.24
p(xMA)	0.28	0.28	p(xMA)	0.23	0.23
p(St-BuA)	0.28	0.28	p(DOM-DVB)	0.23	0.23
			Poly(n-butyl		
p(SBMA-nBA)	0.23	0.23	methacrylate)	0.21	0.21
Poly(n-butyl					
methacrylate)	0.18	0.18	p(SBMA-nBA)	0.19	0.19
p(tC-tBuM)	0.15	0.15	Polystyrene	0.19	0.19
Polystyrene	0.14	0.14	p(tC-tBuM)	0.13	0.13
Polyvinyl chloride	0.08	0.08	Polyvinyl chloride	0.08	0.08

Table S1. Applicability domain analysis results for HTA and M2P datasets.

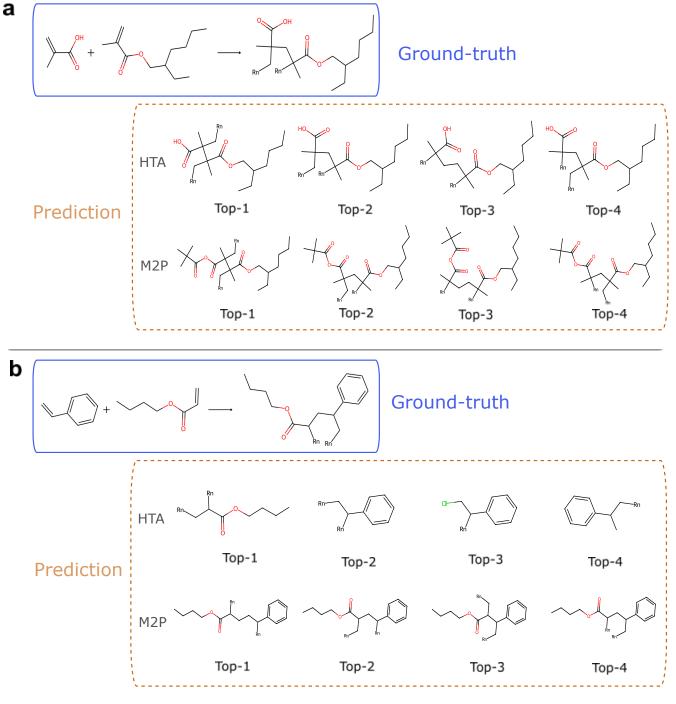


Figure S1. Example of prediction using Molecular Transformer model trained with HTA and M2P data. Catalysts, solvents and stoichiometry not shown. a) p(xMA) copolymer. b) p(St-BuA) copolymer. In 2D molecules representations carbon atoms are in black and oxygen in red. The connection points of polymer repeat units are represented with Rn atoms.

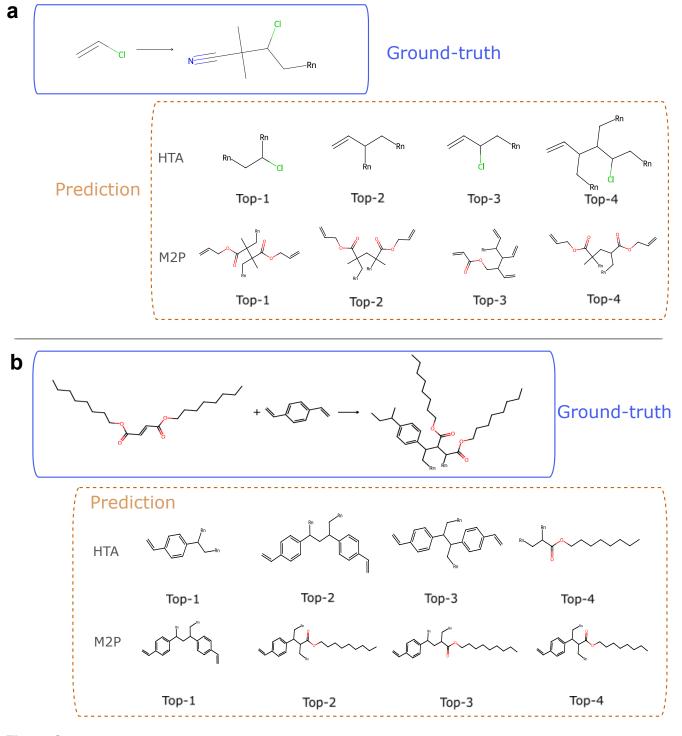


Figure S2. Example of prediction using Molecular Transformer model trained with HTA and M2P data. Catalysts, solvents and stoichiometry not shown. a) Polyvinyl chloride. b) p(DOM-DVB) copolymer. In 2D molecules representations carbon atoms are in black, oxygen in red, nitrogen in dark blue, and chloride in green. The connection points of polymer repeat units are represented with Rn atoms.

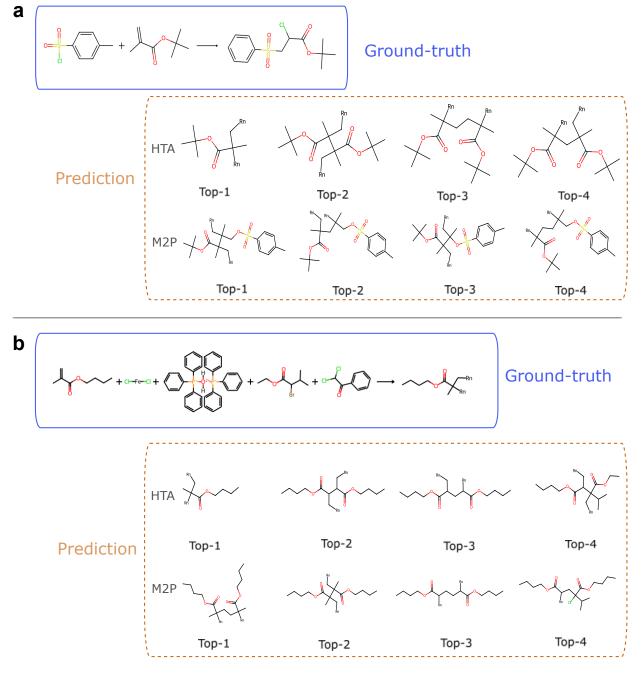


Figure S3. Example of prediction using Molecular Transformer model trained with HTA and M2P data. Catalysts, solvents and stoichiometry not shown. a) p(tC-tBuM) copolymer. b) Poly(n-butyl methacrylate). In 2D molecules representations carbon atoms are in black, oxygen in red, chloride in green, phosphorus in orange, boron in brown and sulfur in yellow. The connection points of polymer repeat units are represented with Rn atoms.