Creation of Polymer Datasets with Targeted Backbones for Screening of Gas Permeability and Selectivity

Surya Prakash Tiwari*[1,2], Wei Shi[1], Samir Budhathoki[1,2], James Baker[1,2],
David P. Hopkinson[1], Janice A. Steckel[1]

[1] National Energy Technology Laboratory,
626 Cochran Mill Road, Pittsburgh, PA 15236, USA

[2]NETL Support Contractor,
626 Cochran Mill Road, Pittsburgh, PA 15236, USA

Abstract

A simple approach was developed to computationally construct a polymer by composing simplified molecular-input line-entry system (SMILES) strings of a polymer backbone and a molecular fragment. This method was used to create 14 polymer datasets by combining seven polymer backbones and two large molecular datasets (ZINC and QM9). Polymer backbones that were studied include four polydimethylsiloxane (PDMS) based backbones, polyethylene oxide (PEO), poly-allyl glycidyl ether, and polyphosphazene. The generated polymer datasets can be used for various cheminformatics tasks, including high-throughput screening for gas permeability and selectivity. This study used machine learning (ML) models to screen the polymers for CO_2/CH_4 and CO_2/N_2 gas separation using membranes and several polymers of interest were identified.

Keywords: Polymer, backbones, functionalization, molecular fragments, side chains, datasets, machine learning, screening, cheminformatics

.

Introduction

In the field of cheminformatics, large datasets are valuable for various screening and generative tasks. Machine learning (ML) techniques coupled with large collections of data have been successfully applied to materials discovery. With access to large collections of data, machine learning approaches can be used to discern relationships between chemical structures and their properties.[1,2] For these tasks, researchers often represent chemical structures using the simplified molecular-input line-entry system (SMILES) string format.[3] This format allows molecular structures to be specified as text strings, which can be converted into the molecular models.

Although cheminformatics has had a significant impact on molecular chemistry,[2,4–9] its impact on polymer research is rather limited due to the scarcity of accessible polymer datasets in the literature.[10–14] To address this, Luo et al. published the PI1M dataset, consisting of approximately one million polymers for polymer informatics.[15] These polymers, encoded as SMILES strings, were generated using a machine learning (ML) generative model that was trained on a collection of approximately 12,000 polymer structures from the PoLyInfo database[16]. Yang et al.[17] constructed a dataset comprising eight million hypothetical polyimides formed by the polycondensation of known diamines/diisocyanates with dianhydrides from the PubChem library dataset. They also created another dataset containing 1,100 ladder polymers that was generated through the binary combinations of components of existing ladder polymers, supplemented by a recurrent neural network (RNN) model.

Finally, we note that researchers have generated polymer datasets for various ML tasks;[18–20] however, such datasets are often challenging to apply to other classes of polymers or are publicly unavailable. Additionally, recent developments in generative molecular design, such as inverse design methods,[21] have the potential to create molecules and polymers dynamically during ML tasks;[21–27] however, these techniques are more complex to implement than the approach presented in this paper.

This work presents a simple approach to generate large polymer datasets in which the backbone is specified in advance and the side group(s) are varied with different molecular fragments obtained from large chemically diverse datasets of small organic molecules such as ZINC[28,29] and QM9[30]. The molecular fragments are chemically diverse,[2,29,31] leading to the generated polymers being chemically diverse as well.

The generated polymer datasets can be utilized for various cheminformatics purposes, including high-throughput screening of polymers for specific applications. This research focuses on finding suitable polymers for CO_2/CH_4 and CO_2/N_2 gas separation using membranes. Polymers for CO_2/CH_4 separation are widely used in industrial processes such as natural gas purification, biogas improvement, oil production enhancement, and landfill gas cleaning.[32–34] CO_2/N_2 separation polymers are desirable in carbon capture technologies based on polymer membranes, as membrane-based carbon dioxide separation processes offer low energy consumption, ease of operation, and compact design.[35–37]

When evaluating high-performance gas separation polymers, researchers commonly rely on the Robeson upper bounds.[38,39] These bounds demonstrate the trade-off between permeability (Px) and selectivity (Px/Py) for gasses X and Y, and were first identified by Robeson by plotting log10(Px/Py) against log10(Px) for various gas pairs. This approach highlights the challenge of finding polymers that simultaneously exhibit good permeability (Px) and selectivity (Px/Py). The present study employs ML models to identify such high-performance polymers for CO_2/CH_4 and CO_2/N_2 gas pair separations.

In previous studies, various ML models were developed for predicting gas permeabilities and selectivities using experimental data from literature.[13,17,40–42] Rampi et al.[13,40,41] utilized hierarchical fingerprints for model training, while Barnett et al.[42] and Yang et al.[17] employed simpler, bit-based hashed fingerprints. The latter approach was chosen for this study due to its ease of implementation.

Methods

Polymer Dataset Generation

The SRUs[43] (structural repeating units) of the polymers were created by combining the polymer backbones with molecules from the ZINC and QM9 datasets treated as molecular fragments (R) as shown in Figure 1. SMILES notations were used to combine the polymer backbone and R, which can be better explained using an example.

Consider the SMILES of the polymer shown in Figure 1. It can be represented by *O[Si](R)(C)*, where R denotes the molecular fragment to be substituted. A computer program was written to substitute the SMILES of the molecular fragment in place of R, followed by checking whether the resulting composite SMILES represented a valid molecule using RdKit's MolFromSmiles module. If the composite SMILES was invalid, the order of atoms in the fragment group's SMILES was rearranged using RDKit, and the procedure of substitution and validation was repeated up to 20 times until a valid SMILES was obtained. Fragments that did not result in any valid polymer SRUs were discarded.

Combining two SMILES strings often results in a valid SMILES because the SMILES representation does not explicitly include hydrogen atoms. The new SMILES string created can automatically adjust the valency of atoms that are affected by the newly formed bond, making it a valid representation.

It is worth noting that the SMILES of fragments can be rearranged multiple times and added to the backbone, resulting in a greater diversity of chemically unique SRUs. However, in this study, that step was not taken as the datasets created were already quite large.

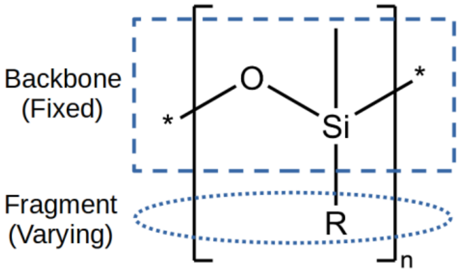

Figure 1. A schematic of a polymer in which the dashed box surrounds a fixed backbone, while the dotted oval indicates varying fragments from the molecular dataset.

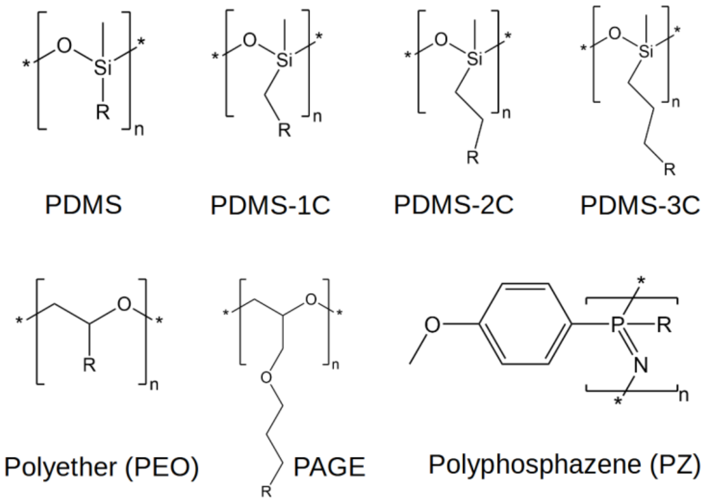

Figure 2. Seven polymer backbones used to create polymer datasets.

Polymer Backbones Used

In the study, seven polymer backbone structures were employed, and they are depicted in Figure 2. Four of these backbones were based on the structure of polydimethylsiloxane (PDMS) and included the PDMS backbone, PDMS-1C, PDMS-2C, and PDMS-3C. The PDMS backbone directly connected the R group to the Si atom, while PDMS-1C, PDMS-2C, and PDMS-3C added one, two, and three alkyl groups, respectively, between the R group and the Si atom. In experiments, connecting a side chain or fragment to PDMS-2C and PDMS-3C backbones is considered synthetically more feasible than with PDMS and PDMS-1C backbones. Polymers with the latter two backbones are relatively challenging to synthesize because the reactions used to couple C atoms to Si atoms usually require the presence of a catalyst and involve reacting with a C=C bond.[44,45] Additionally, the PDMS backbone may have unstable bond formation if the side chain has a heteroatom (O, N, S, P, etc.) directly linking the Si and C atoms.

The remaining three backbones utilized in this study were polyethylene oxide (PEO), poly-allyl glycidyl ether (PAGE), and polyphosphazene (PZ). The PAGE backbone is a variant of PEO and is considered easier to synthesize. All the backbones examined in this study have demonstrated good performance in producing polymer membranes.[46–52]

Fragments from Molecular Datasets

The molecular fragments for this study were obtained from two large, chemically diverse datasets of small organic molecules: the shortened ZINC dataset and the shortened QM9 dataset.[29] The ZINC dataset consists of approximately 1.93 million small non-ionic organic molecules, while the QM9 dataset contains approximately 132,000 such molecules (see Table 1). Both datasets were sourced from Github repositories,[53,54] and are also included in the Github repository with the link provided in the SI.

Internal Diversity of Polymer Datasets

The molecular fragments used in the study needed to be diverse to generate a polymer dataset that covers a broad chemical space. One metric to quantify the chemical diversity of a dataset is internal diversity (IntDiv$_p$).[29,31] For a dataset of molecules/polymers A, internal diversity is given by,[29]

$$IntDiv_p(A) = 1 - \sqrt[p]{\frac{1}{|A|^2} \sum_{m1,m2 \in A} T(m1, m2)^p}$$

where, |A| is the size of the dataset, $T(m1, m2)$ is the Tanimoto-similarity of molecules in A with respect to each other, and p can be 1 or 2.[55] The values of internal diversity lie between 0 and 1, with values closer to 1 indicating greater diversity in a dataset. The internal diversities, IntDiv$_1$ and IntDiv$_2$, which correspond to Tanimoto distance and Tanimoto variance, respectively, were calculated for each of the ZINC and QM9 molecular datasets, as well as for the polymer datasets generated from them and for the PI1M dataset.

Development of ML Models for Permeabilities and Selectivities

A database of experimentally measured gas permeabilities in polymers was compiled by gathering data from the *Polymer Gas Separation Membrane Database,*[56,57] verifying the data from original reports, and supplementing it with additional data for various polymers and gases, resulting in a collection of approximately 1,500 polymers with various gas permeability data. Repeat units of the polymers were encoded as SMILES strings and added to the database. This database was utilized in the creation of ML models for predicting CH_4, CO_2, and N_2 permeabilities, as well as CO_2/CH_4 and CO_2/N_2 selectivities in polymers, as discussed in the following sections.

ML Models for Gas Permeability Prediction

First, copolymers and ladder polymers were removed from the polymer database explained in the previous section, resulting in 687, 610, and 725 homopolymers for CO_2, CH_4, and N_2 gas permeabilities, respectively, available for ML. Then, using an in-house code, each SRU SMILES of the cleaned polymer database was repeated to generate an n-mer, also encoded as a SMILES string. The accuracy of the ML model improved with increasing the number of repeat units, but plateaued at n = 3 as shown in Figure SI-2 in the SI. In literature, oligomers with more than five repeat units are known to perform well in property prediction tasks as they partially represent the repeating nature of polymers.[58] Thus, in this study, n was set to 10.

The resulting n-mer encoded as SMILES strings were transformed into RDKit fingerprints[59] using the RDKit library. The target gas permeabilities, measured in Barrers, were converted to a logarithm 10 scale for fitting in the ML model. The ML model used was a Gaussian process regressor (GPR), which was trained using the protocol of Barnett et al.[42] The dataset was randomly shuffled and divided into training and test sets during the fitting process, with 20% of the data reserved for testing (Table 2).

ML Models for Selectivity Prediction

ML predictions for gas selectivity can be obtained in two ways. The most common method relies on separate models for each of the two gasses to predict their permeabilities; the predicted selectivity is estimated from the ratio of the predicted permeabilities.[17,40–42] An alternative approach is to train models directly on experimental selectivity data, which we found to be more accurate.

In our study, we compared these methods. In method 1, following the methodology explained in the previous section, we trained ML models to predict the permeabilities of CH_4, CO_2, and N_2, and then used the ratios of the predicted permeabilities to calculate the CO_2/CH_4 and CO_2/N_2 selectivities. In method 2, the CO_2/CH_4 and CO_2/N_2 selectivities were calculated from the ratios of the experimental CO_2 and CH_4, and CO_2 and N_2 permeabilities, respectively. GPR models were trained on these selectivity data, using RDKit-fingerprints derived from 10-mer SMILES as input features. A total of 608 polymers were used for CO_2/CH_4 selectivity and 664 polymers for CO_2/N_2 selectivity in the ML training and testing process, with 20% of the data reserved for testing (Table 2). The datasets were randomly shuffled and split into training and test sets during the model fitting process. The results of selectivity predictions using methods 1 and 2 are presented in the results section and in Figure 5.

Use of ML Models for Polymer Screening

Trained ML models were applied to predict the permeabilities of CO_2, CH_4, and N_2, as well as the selectivities of CO_2/CH_4 and CO_2/N_2 across multiple polymer datasets employed in this study. To achieve this, the SRU SMILES of the polymers were first converted into RDKit-fingerprints using the same method employed in training the ML models. These fingerprints were then utilized as inputs to the ML models to predict the gas permeabilities and selectivities of the polymers. Predicted gas permeabilities and selectivities were used to screen for polymers that are predicted to have gas permeation properties above the Robeson upper bound.[38,39]

Results and Discussion

Generated Polymer Datasets

A total of 14 datasets of polymers were created by combining seven different polymer backbones and two small molecule datasets, ZINC and QM9. The generated datasets along with ZINC, QM9 and PI1M are listed in Table 1, and complete datasets can be found in the SI.

Internal Diversity of Polymer Datasets

The internal diversities, IntDiv$_1$ and IntDiv$_2$, of all the datasets considered in this study are displayed in Table 1. Across all datasets, IntDiv$_1$ was found to be slightly greater than IntDiv$_2$. The ZINC dataset had an internal diversity of 0.85, which was lower than the internal diversity of the QM9 dataset, which was calculated to be 0.91. However, this trend was reversed in the corresponding polymer datasets. For instance, ZINC-PDMS-1C had a higher internal diversity of 0.75 compared to QM9-PDMS-1C with an internal diversity of approximately 0.72. On average, the internal diversities of generated polymers from the backbones were close to 0.75, which was lower than the internal diversity of the PI1M dataset, measured to be 0.85. This difference is understandable, as a generated polymer dataset has a uniform backbone in all polymers.

Table 1. Description of datasets along with their internal diversities employed in this study. Number of polymers found above the Robeson bound for various polymer datasets for the CO$_2$/CH$_4$ screening are also included.

Dataset	**Description**	**Data size**	**Reference**	**Internal diversity**		**Number of polymers screened above the Robeson Bound for CO$_2$/CH$_4$**	
				IntDiv$_1$	**IntDiv$_2$**	**All**	**Non-ring**
ZINC	Molecular dataset consisting of small organic molecules.	1,936,962	Polykovskiy et al.[53]	0.856	0.851	-	-
QM9	Molecular dataset consisting of small organic molecules.	131,979	Ramakrishnan et al.[30]	0.918	0.904	-	-
PDMS-ZINC	These polymer datasets were created by combining polymer backbones and molecular datasets, and named accordingly. For example, PDMS-3C-ZINC dataset was created by combining the PDMS-3C polymer backbone and the molecular fragments from the ZINC molecular dataset.	~ 1.93 M each	This work	0.785	0.782	746	0
PDMS-1C-ZINC				0.763	0.76	707	0
PDMS-2C-ZINC				0.745	0.742	953	0
PDMS-3C-ZINC				0.732	0.729	415	0
PEO-ZINC				0.785	0.782	0	0
PAGE-ZINC				0.632	0.695	0	0
PP-ZINC				0.71	0.707	0	0
PDMS-QM9		~ 1.32 K each		0.773	0.768	123	1

PDMS-1C-QM9				0.733	0.73	47	0
PDMS-2C-QM9				0.704	0.701	24	0
PDMS-3C-QM9				0.68	0.677	10	0
PEO-QM9				0.778	0.772	0	0
PAGE-QM9				0.633	0.63	0	0
PP-QM9				0.638	0.634	0	0
PI1M	Polymer dataset generated using a ML generative model.	~ 1 M	Luo et al.[15]	0.855	0.843	13	2

Table 2. Evaluation of ML fitting for training and testing of various gas permeabilities and selectivities.

Fitting entity	Training			Test		
	Data Size	**MAE**	**R^2**	**Data Size**	**MAE**	**R^2**
Log10 CO_2 permeability	550	0.08	0.99	137	0.38	0.81
Log10 CH_4 permeability	488	0.10	0.99	122	0.42	0.82
Log10 N_2 permeability	580	0.09	0.99	145	0.42	0.84
CO_2/CH_4 selectivity	486	0.99	0.99	122	7.58	0.59
CO_2/N_2 selectivity	531	0.96	0.97	133	3.81	0.60

ML Models for Gas Permeabilities

Figure 3 presents ML models utilized for the training and testing of gas permeabilities for CO_2, CH_4, and N_2. Table 2 provides an overview of the model evaluations. The training results showed high levels of accuracy, with coefficients of determination (R^2) of 0.99 and mean absolute errors (MAE) of 0.08, 0.10, and 0.09 for CO_2, CH_4, and N_2, respectively. Similarly, the testing phase demonstrated reasonable performance, with R^2 values of 0.81, 0.82, and 0.84 and corresponding MAEs of 0.38, 0.42, and 0.42 for CO_2, CH_4, and N_2, respectively.

Prediction of Gas Selectivities

Gas selectivities were calculated using two methods: method 1 involved taking the ratio of predicted permeabilities, while method 2 involved fitting a separate ML model to predict selectivities. Figure 4 shows the ML fittings for CO_2/CH_4 and CO_2/N_2 selectivities using method 2, and Table 2 provides an overview of the model evaluations. The training R^2 values for both selectivities were strong (>0.97), indicating a good fit between the models and the training data. However, the test R^2 values for CO_2/CH_4 and CO_2/N_2 selectivities were lower (0.59 and 0.60, respectively) due to the limited data available for fitting selectivity. Despite this, the corresponding MAE values of 7.58 and 3.81, respectively, were acceptable for identifying and screening potential polymers in the study.

The accuracy of both methods for predicting selectivities was evaluated against experimental selectivities for CO_2/CH_4 and CO_2/N_2 on untrained data (Figure 5). Both methods showed comparable accuracy for CO_2/CH_4, with R^2 of approximately 0.6 and MAE of approximately 8. However, for CO_2/N_2, method 2 outperformed method 1 with an R^2 of 0.60 and MAE of 3.81, compared to method 1 with an R^2 of 0.44 and MAE of 4.57. The key difference between the two methods was the presence of large error bars in gas selectivities for both gas pairs in method 1. This was due to error propagation that occurred during the division of predicted permeabilities to calculate the selectivity. Such errors may lead to over or underestimated selectivity values, particularly noticeable in large and diverse polymer datasets. This is demonstrated in Figure 6, using the ZINC-PDMS polymer dataset, where method 1 displayed a higher variance of gas selectivities for both gas pairs compared to method 2. This trend was observed across all polymer datasets analyzed in this study.

Considering the more accurate selectivities obtained from the separately fitted ML model for selectivity (method 2), we present the screening results for CO_2/CH_4 and CO_2/N_2 using this approach.

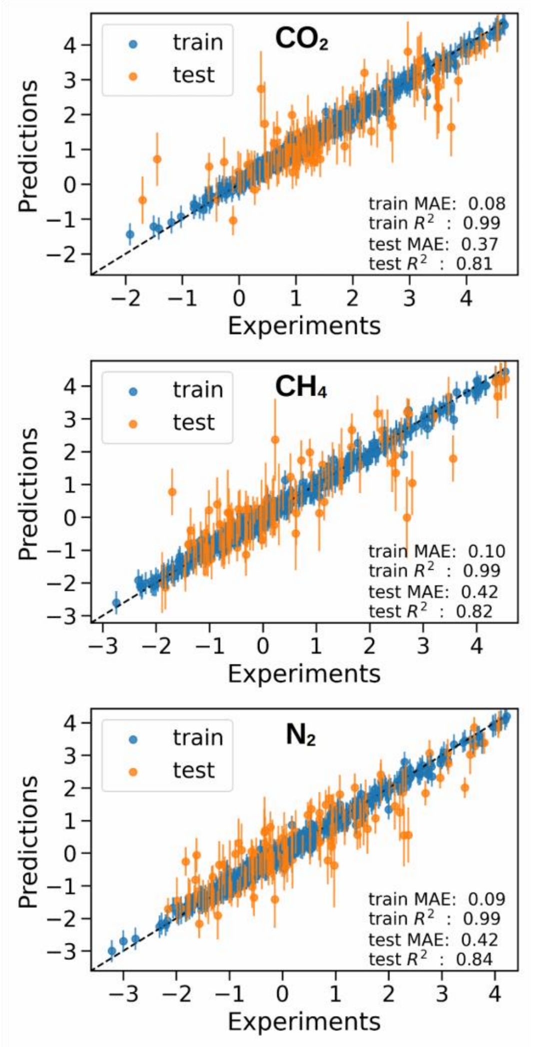

Figure 3. Fittings for log10s of CO_2, CH_4, and N_2 permeabilities. Blue and orange points correspond to train and test datasets, respectively.

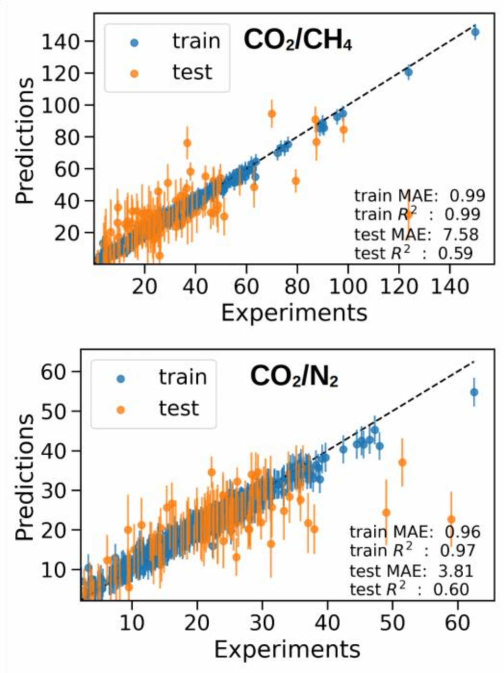

Figure 4. Fittings for CO_2/CH_4 and CO_2/N_2 selectivities. Blue and orange points correspond to train and test datasets, respectively.

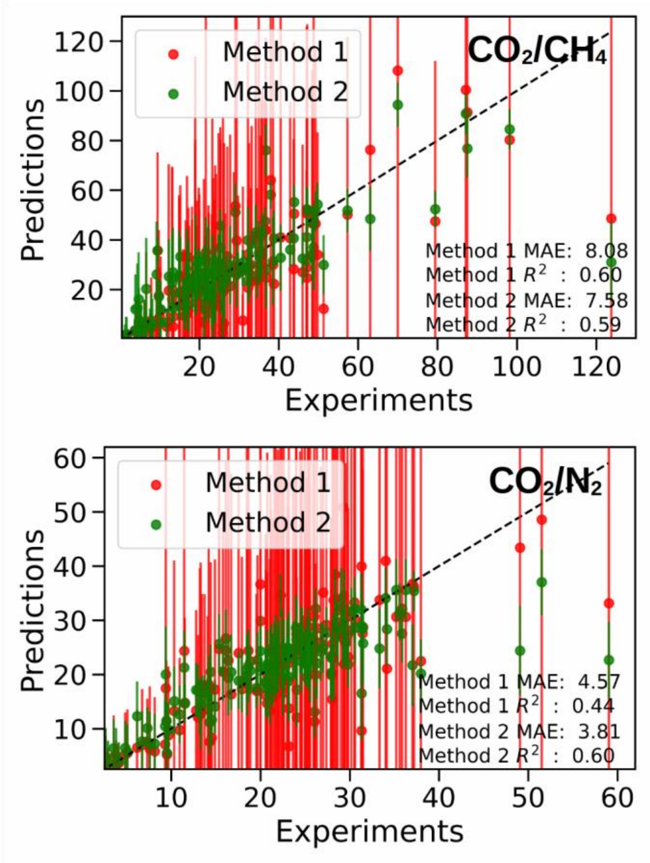

Figure 5. Evaluations of selectivities obtained from taking the ratio of the predicted permeabilities (method 1) versus fitting a separate ML model to predict selectivities (method 2) against the experimental values. The figure shows the results for both methods for the CO_2/CH_4 and CO_2/N_2 gas pairs. Note that the error bars from method 1 were too large, therefore, not shown in full.

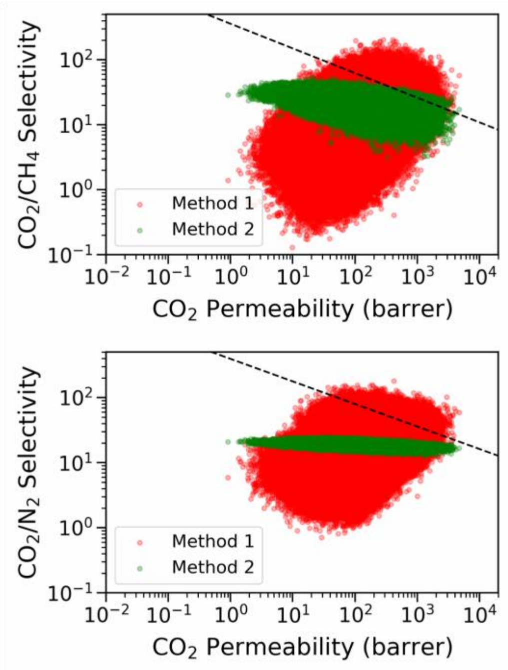

Figure 6. Comparison of screening results for the PDMS-ZINC polymer dataset, using two different methods: Method (1) predicts permeabilities and divides them to obtain selectivities, and method (2) predicts selectivities via separately fitted ML models. The figure shows the results for both methods for the CO_2/CH_4 and CO_2/N_2 gas pairs.

Screening of Polymer Datasets

Using the ML models described (vide supra), predictions of CO_2/CH_4 and CO_2/N_2 permeabilities and selectivities were made for the 14 generated polymer datasets and the PI1M dataset, comprising a total of 15.4 million polymers.

CO_2/CH_4 Screening

The results of CO_2/CH_4 selectivity versus CO_2 permeability predictions are shown in Figure 7. The number of polymers above the Robeson bound for all datasets is displayed in Table 1. The datasets that contain PDMS and its derivative backbones tend to have higher permeabilities and the highest number of polymers above the Robeson bound, with a total of 3025. In contrast, datasets containing PEO, PZ, and PAGE backbones showed no polymers above the Robeson bound. The spread of predictions in the permeability-selectivity space was found to be related to the size of the backbones. Larger backbones had a larger impact on the polymer, leading to a smaller spread in predictions. This trend was observed across the PDMS series backbones, where the spread decreased with an increase in the number of C atoms. The PZ backbone, being the largest, showed the smallest spread in predictions.

Finally, the PI1M dataset demonstrated the broadest spread in the permeability-selectivity space due to its high internal diversity. Despite this, only 13 polymers were found to be above the Robeson bound, which was the key objective of this screening study. This highlights the importance of considering multiple datasets in the screening process, as relying solely on a dataset with high internal diversity may result in missing the edge cases of interest.

Additionally, it's important to note that polymers containing rings are typically glassy at room temperature and may not be ideal for membrane applications.[60] As such, the number of non-ring polymers above the Robeson bound was also recorded in Table 1. Only three non-ring polymers were found to lie above the Robeson bound. A full list of screened polymers is included in the SI.

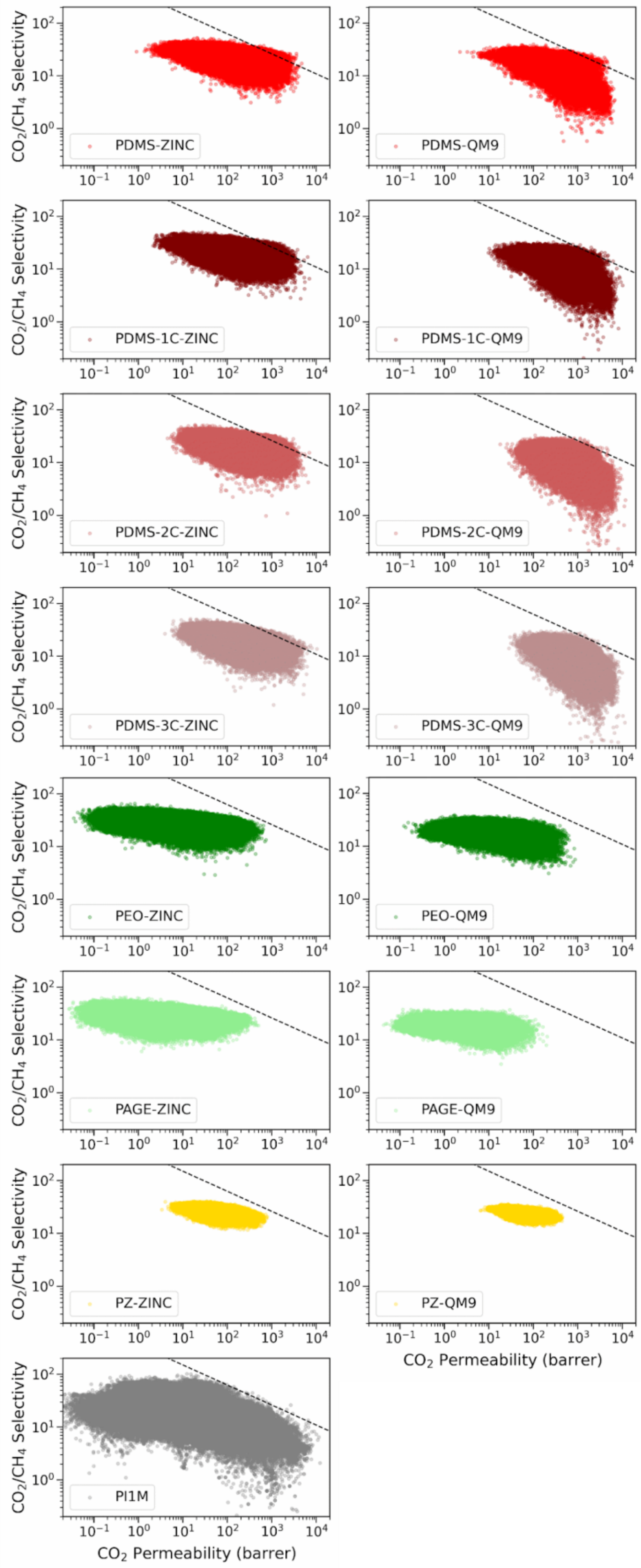

Figure 7. Plots of predicted CO_2/CH_4 selectivity versus CO_2 permeability for various polymers datasets.

CO_2/N_2 screening

The plots of CO_2/N_2 selectivity versus CO_2 permeability are displayed in Figure 8. Unfortunately, none of the generated polymers in this study were found to be above the Robeson bound. However, six polymers from the PI1M dataset were found to be above the Robeson bound. To broaden the search, the criteria for screening were lowered to include polymers with CO_2 permeability and CO_2/N_2 selectivity larger than 250 barrer and 25, respectively, as indicated by the yellow shade in the plots. Despite these changes, no generated polymers derived from ZINC and QM9 datasets in this study met these criteria. On the other hand, the PI1M dataset had 452 polymers that were in the specified region, with 444 of them being non-ring polymers.

It is important to note that the CO_2/N_2 selectivity model fitting shown in Figure 5 performed slightly poorly in predicting selectivities above 25, which was one of the goals of our screening task. One of the reasons for this result is the data imbalance; the number of polymers with selectivity data points above 25 (236) was smaller than the number of selectivity data points below 25 (484), causing the model to be biased against predicting larger selectivities.

To address this issue, we employed a combination of a classifier model followed by regression models. The method and results are presented in the SI. Although the fitting of this method was slightly poor (with an R^2 of 0.53 for the test dataset), it did manage to screen 143 polymers derived from the ZINC and QM9 datasets that fell within the target region. In addition, 198 polymers were screened from the PI1M dataset. A full list of screened polymers from both the models is included in the SI.

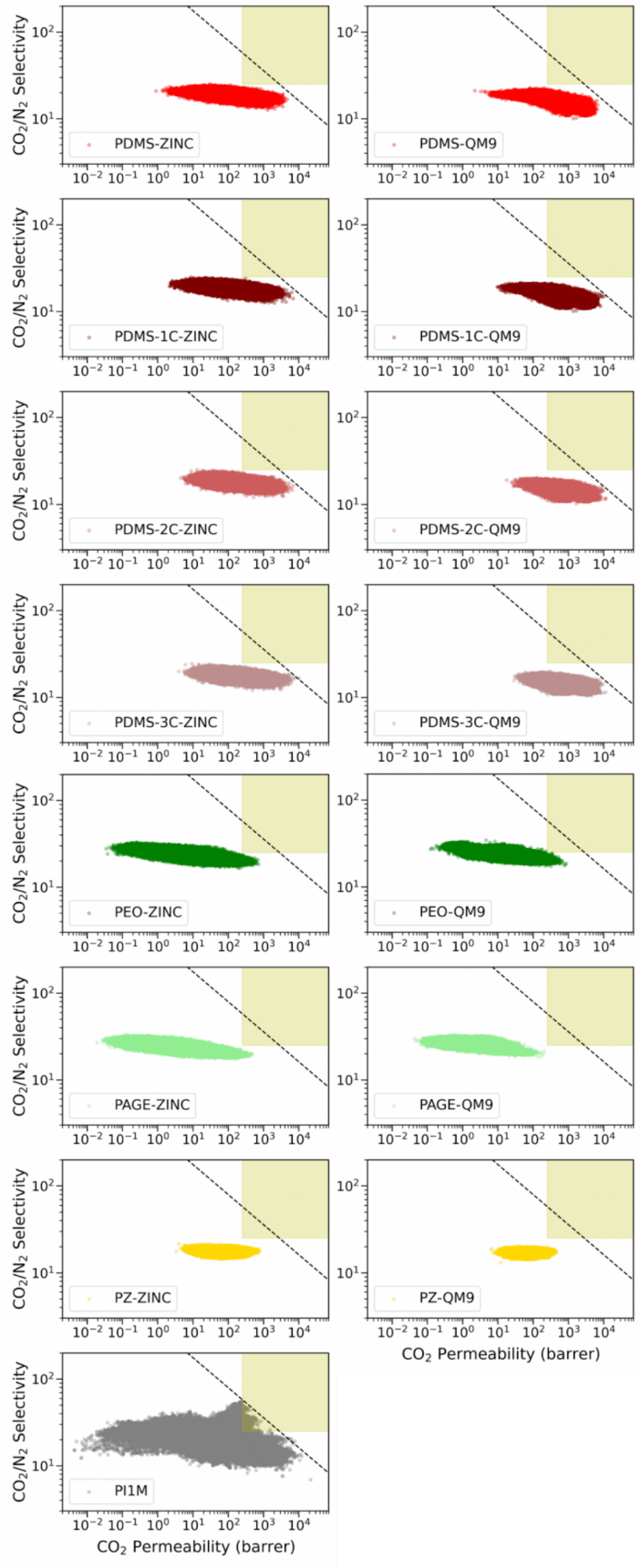

Figure 8. Plots of predicted CO_2/N_2 selectivity versus CO_2 permeability for various polymers datasets.

Conclusions

A simple approach was developed to synthesize polymers by combining the SMILES of a polymer backbone and a molecular fragment. By using seven different polymer backbones and two large molecular datasets (ZINC and QM9), 14 unique polymer datasets were generated. Additionally, an already available polymer dataset PI1M was also included in the study. The polymer datasets were found to be diverse, as demonstrated by their internal diversity scores.

The datasets were then used as input to screen polymers for their applications in CO_2/CH_4 and CO_2/N_2 gas separation using membranes. To achieve this, ML models were trained using experimental data of gas permeabilities obtained from the literature. The ML models were used to predict the permeabilities and selectivities of the polymers in the generated datasets, helping to identify polymers that meet the criteria of being above the Robeson bound. This is a key characteristic of high-performance polymers in gas separation applications.

To obtain the selectivities, two methods were explored - one by dividing the predicted permeabilities, and the other by directly predicting selectivities from the ML models. However, the former method, which is commonly used in the literature, resulted in a large variance in selectivities due to error propagation. As a result, the latter method was used to report the screening results, ensuring more reliable and accurate results.

For the CO_2/CH_4 gas separation, the datasets that contain PDMS and its derivative backbones showed the highest number of polymers above the Robeson bound, with a total of 3025. In contrast, datasets containing PEO, PZ, and PAGE backbones show no polymers above the Robeson bound. The PI1M dataset demonstrated the broadest spread in the permeability-selectivity space due to its high internal diversity. Despite this, only 13 polymers were found to be above the Robeson bound, which was the key objective of this screening study. This highlights the importance of considering multiple datasets in the screening process to avoid missing the edge cases of interest.

For the CO_2/N_2 gas separation, only six polymers from the PI1M dataset were found to be above the Robeson bound. To broaden the search, the criteria for screening were lowered to include polymers with CO_2 permeability and CO_2/N_2 selectivity larger than 250 barrer and 25, respectively, as indicated by the yellow shade in the plots. Despite these changes, no polymers derived from the ZINC and QM9 datasets met the screening criteria. On the other hand, the PI1M dataset had 452 polymers that were in the specified region.

We observed that the CO_2/N_2 selectivity model performed slightly poorly in predicting larger selectivities, which was one of the goals of our screening task. One of the reasons for this result is the data imbalance for training, causing the model to be biased against predicting larger selectivities. To address this issue, we employed a combination of a classifier model followed by regression models. However, this method suffered from a lower accuracy.

Finally, it is important to acknowledge the limitations of this study. First, it only considered homopolymers and excluded block, ladder, and copolymers, some of which have proven to be effective in gas separation applications.[39,61–63] Second, the experimental gas permeability data used for ML model training may be flawed, leading to inaccurate models. This issue could be addressed by using data augmentation techniques for polymers.[64] Additionally, the ML model fittings are not perfect and result in rough polymer predictions, which should be validated through experiments or molecular simulations.

Despite these limitations, this study still provides a promising method for high-throughput screening of polymers for various applications. Additionally, materials designed in-silico may suffer from synthetic accessibility problems in the lab, for which various metrics have been proposed.[65–68] Finally, the created polymer datasets may not cover the entire chemical space. This issue can be addressed by utilizing inverse design methods,[21–27] and the polymer creation method proposed in this study can still be useful in these approaches.

DISCLAIMER

This project was funded by the U.S. Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Acknowledgements

Help from Ali Sekizkards, Victor Kusuma, Lingxiang Zhu and Shouliang Yi in assembling the training dataset is gratefully acknowledged. This work was performed in support of the U.S. Department of Energy's (DOE) Fossil Energy Transformational Carbon Capture Program.

References

(1) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

(2) Kell, D. B.; Samanta, S.; Swainston, N. Deep Learning and Generative Methods in Cheminformatics and Chemical Biology: Navigating Small Molecule Space Intelligently. *Biochem. J.* **2020**, *477* (23), 4559–4580. https://doi.org/10.1042/BCJ20200781.

(3) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(4) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323.

(5) *ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology | Chemical Research in Toxicology*. https://pubs.acs.org/doi/10.1021/acs.chemrestox.6b00135 (accessed 2023-02-08).

(6) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(7) *SIDER database of drugs and side effects | Nucleic Acids Research | Oxford Academic*. https://academic.oup.com/nar/article/44/D1/D1075/2502602 (accessed 2023-02-08).

(8) *MoleculeNet: a benchmark for molecular machine learning - Chemical Science (RSC Publishing)*. https://pubs.rsc.org/en/content/articlelanding/2018/SC/C7SC02664A (accessed 2023-02-08).

(9) Tetko, I. V.; Engkvist, O. From Big Data to Artificial Intelligence: Chemoinformatics Meets New Challenges. *J. Cheminformatics* **2020**, *12* (1), 74. https://doi.org/10.1186/s13321-020-00475-y.

(10) Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U. F.; Tang, S.; Bi, J.; Chen, W.; Li, Y. Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges. *Polymers* **2020**, *12* (1), 163. https://doi.org/10.3390/polym12010163.

(11) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6* (10), 1078–1082. https://doi.org/10.1021/acsmacrolett.7b00228.

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

(13) Mannodi-Kanakkithodi, A.; Huan, T. D.; Ramprasad, R. Mining Materials Design Rules from Data: The Example of Polymer Dielectrics. *Chem. Mater.* **2017**, *29* (21), 9001–9010. https://doi.org/10.1021/acs.chemmater.7b02027.

(14) Wang, Y.; Xie, T.; France-Lanord, A.; Berkley, A.; Johnson, J. A.; Shao-Horn, Y.; Grossman, J. C. Toward Designing Highly Conductive Polymer Electrolytes by Machine Learning Assisted Coarse-Grained Molecular Dynamics. *Chem. Mater.* **2020**, *32* (10), 4144–4151. https://doi.org/10.1021/acs.chemmater.9b04830.

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

(16) *Polymer Database(PoLyInfo) - DICE :: National Institute for Materials Science*. https://polymer.nims.go.jp/en/ (accessed 2021-12-23).

(17) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci. Adv.* **2022**, *8* (29), eabn9545. https://doi.org/10.1126/sciadv.abn9545.

(18) O'Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115* (32), 16200–16210. https://doi.org/10.1021/jp202765c.

(19) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data* **2016**, *3* (1), 160012. https://doi.org/10.1038/sdata.2016.12.

(20) Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine Learning-Based Screening of Complex Molecules for Polymer Solar Cells. *J. Chem. Phys.* **2018**, *148* (24), 241735. https://doi.org/10.1063/1.5023563.

(21) Nigam, A.; Pollice, R.; Aspuru-Guzik, A. JANUS: Parallel Tempered Genetic Algorithm Guided by Deep Neural Networks for Inverse Molecular Design. *ArXiv210604011 Cs* **2021**.

(22) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. arXiv December 10, 2022. https://doi.org/10.48550/arXiv.1312.6114.

(23) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. arXiv June 10, 2014. https://doi.org/10.48550/arXiv.1406.2661.

(24) Kingma, D. P.; Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2018; Vol. 31.

(25) *Reinforcement Learning*. MIT Press. https://mitpress.mit.edu/9780262039246/reinforcement-learning/ (accessed 2023-02-09).

(26) Blum, C.; Roli, A. Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Comput. Surv.* **2003**, *35* (3), 268–308. https://doi.org/10.1145/937503.937505.

(27) Kim, S.; Schroeder, C.; Jackson, N. Open Macromolecular Genome: Generative Design of Synthetically Accessible Polymers. ChemRxiv January 24, 2023. https://doi.org/10.26434/chemrxiv-2023-3hn5r.

(28) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(29) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 1931. https://doi.org/10.3389/fphar.2020.565644.

(30) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), 140022. https://doi.org/10.1038/sdata.2014.22.

(31) Benhenda, M. ChemGAN Challenge for Drug Discovery: Can AI Reproduce Natural Chemical Diversity? *ArXiv170808227 Cs Stat* **2017**.

(32) Yang, H.; Xu, Z.; Fan, M.; Gupta, R.; Slimane, R. B.; Bland, A. E.; Wright, I. Progress in Carbon Dioxide Separation and Capture: A Review. *J. Environ. Sci.* **2008**, *20* (1), 14–27. https://doi.org/10.1016/S1001-0742(08)60002-9.

(33) Zhang, Y.; Sunarso, J.; Liu, S.; Wang, R. Current Status and Development of Membranes for CO2/CH4 Separation: A Review. *Int. J. Greenh. Gas Control* **2013**, *12*, 84–107. https://doi.org/10.1016/j.ijggc.2012.10.009.

(34) Iarikov, D. D.; Ted Oyama, S. Chapter 5 - Review of CO2/CH4 Separation Membranes. In *Membrane Science and Technology*; Oyama, S. T., Stagg-Williams, S. M., Eds.; Inorganic Polymeric and Composite Membranes; Elsevier, 2011; Vol. 14, pp 91–115. https://doi.org/10.1016/B978-0-444-53728-7.00005-7.

(35) Powell, C. E.; Qiao, G. G. Polymeric CO2/N2 Gas Separation Membranes for the Capture of Carbon Dioxide from Power Plant Flue Gases. *J. Membr. Sci.* **2006**, *279* (1), 1–49. https://doi.org/10.1016/j.memsci.2005.12.062.

(36) Liu, J.; Hou, X.; Park, H. B.; Lin, H. High-Performance Polymers for Membrane CO2/N2 Separation. *Chem. – Eur. J.* **2016**, *22* (45), 15980–15990. https://doi.org/10.1002/chem.201603002.

(37) Han, Y.; Ho, W. S. W. Polymeric Membranes for CO2 Separation and Capture. *J. Membr. Sci.* **2021**, *628*, 119244. https://doi.org/10.1016/j.memsci.2021.119244.

(38) Robeson, L. M.; Burgoyne, W. F.; Langsam, M.; Savoca, A. C.; Tien, C. F. High Performance Polymers for Membrane Separation. *Polymer* **1994**, *35* (23), 4970–4978. https://doi.org/10.1016/0032-3861(94)90651-3.

(39) Robeson, L. M. The Upper Bound Revisited. *J. Membr. Sci.* **2008**, *320* (1), 390–400. https://doi.org/10.1016/j.memsci.2008.04.030.

(40) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-Learning Predictions of Polymer Properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128* (17), 171104. https://doi.org/10.1063/5.0023759.

(41) Zhu, G.; Kim, C.; Chandrasekarn, A.; Everett, J. D.; Ramprasad, R.; Lively, R. P. Polymer Genome–Based Prediction of Gas Permeabilities in Polymers. *J. Polym. Eng.* **2020**, *40* (6), 451–457. https://doi.org/10.1515/polyeng-2019-0329.

(42) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv. 6* (20), eaaz4301. https://doi.org/10.1126/sciadv.aaz4301.

(43) Kahovec, J.; Fox, R. B.; Hatada, K. Nomenclature of regular single-strand organic polymers (IUPAC Recommendations 2002). *Pure Appl. Chem.* **2002**, *74* (10), 1921–1956. https://doi.org/10.1351/pac200274101921.

(44) Chen, X.; Engle, K. M.; Wang, D.-H.; Yu, J.-Q. Pd(II)-Catalyzed C–H Activation/C–C Cross-Coupling Reactions: Versatility and Practicality. *Angew. Chem. Int. Ed Engl.* **2009**, *48* (28), 5094–5115. https://doi.org/10.1002/anie.200806273.

(45) Hartwig, J. F. Carbon–Heteroatom Bond Formation Catalysed by Organometallic Complexes. *Nature* **2008**, *455* (7211), 314. https://doi.org/10.1038/nature07369.

(46) Tan, X.; Rodrigue, D. A Review on Porous Polymeric Membrane Preparation. Part II: Production Techniques with Polyethylene, Polydimethylsiloxane, Polypropylene, Polyimide, and Polytetrafluoroethylene. *Polymers* **2019**, *11* (8), 1310. https://doi.org/10.3390/polym11081310.

(47) Firpo, G.; Angeli, E.; Repetto, L.; Valbusa, U. Permeability Thickness Dependence of Polydimethylsiloxane (PDMS) Membranes. *J. Membr. Sci.* **2015**, *481*, 1–8. https://doi.org/10.1016/j.memsci.2014.12.043.

(48) Liu, S. L.; Shao, L.; Chua, M. L.; Lau, C. H.; Wang, H.; Quan, S. Recent Progress in the Design of Advanced PEO-Containing Membranes for CO2 Removal. *Prog. Polym. Sci.* **2013**, *38* (7), 1089–1120. https://doi.org/10.1016/j.progpolymsci.2013.02.002.

(49) Kargari, A.; Rezaeinia, S. State-of-the-Art Modification of Polymeric Membranes by PEO and PEG for Carbon Dioxide Separation: A Review of the Current Status and Future Perspectives. *J. Ind. Eng. Chem.* **2020**, *84*, 1–22. https://doi.org/10.1016/j.jiec.2019.12.020.

(50) Venna, S. R.; Spore, A.; Tian, Z.; Marti, A. M.; Albenze, E. J.; Nulwala, H. B.; Rosi, N. L.; Luebke, D. R.; Hopkinson, D. P.; Allcock, H. R. Polyphosphazene Polymer Development for Mixed Matrix Membranes Using SIFSIX-Cu-2i as Performance Enhancement Filler Particles. *J. Membr. Sci.* **2017**, *535*, 103–112. https://doi.org/10.1016/j.memsci.2017.04.033.

(51) Zhou, Z.; Jiang, Z.; Chen, F.; Kuang, T.; Zhou, D.; Meng, F. Research Progress in Energy Based on Polyphosphazene Materials in the Past Ten Years. *Polymers* **2023**, *15* (1), 15. https://doi.org/10.3390/polym15010015.

(52) Orme, C. J.; McNally, J. S.; Klaehn, J. R.; Stewart, F. F. Mixed Substituent Ether-Containing Polyphosphazene/Poly(Bis-Phenoxyphosphazene) Blends as Membranes for CO2 Separation from N2. *J. Appl. Polym. Sci.* **2021**, *138* (15), 50207. https://doi.org/10.1002/app.50207.

(53) Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models, 2022. https://github.com/molecularsets/moses (accessed 2022-10-29).

(54) *selfies/examples/vae_example/datasets at master · aspuru-guzik-group/selfies · GitHub*. https://github.com/aspuru-guzik-group/selfies/tree/master/examples/vae_example/datasets (accessed 2023-04-29).

(55) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52* (11), 2884–2901. https://doi.org/10.1021/ci300261r.

(56) *Membrane Database - Polymer Gas Separation Membranes*. Virtual Screening of Nanoporous Materials. https://research.csiro.au/virtualscreening/membrane-database-polymer-gas-separation-membranes/ (accessed 2022-11-04).

(57) A. W. Thornton, B. D. Freeman and L. M. Robeson. *Polymer Gas Separation Membrane Database (2012).*

(58) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer Informatics: Current Status and Critical next Steps. *Mater. Sci. Eng. R Rep.* **2021**, *144*, 100595. https://doi.org/10.1016/j.mser.2020.100595.

(59) *RDKit*. https://www.rdkit.org/ (accessed 2021-11-03).

(60) Michieletto, D.; Nahali, N.; Rosa, A. Glassiness and Heterogeneous Dynamics in Dense Solutions of Ring Polymers. *Phys. Rev. Lett.* **2017**, *119* (19), 197801. https://doi.org/10.1103/PhysRevLett.119.197801.

(61) Embaye, A. S.; Martínez-Izquierdo, L.; Malankowska, M.; Téllez, C.; Coronas, J. Poly(Ether-Block-Amide) Copolymer Membranes in CO2 Separation Applications. *Energy Fuels* **2021**, *35* (21), 17085–17102. https://doi.org/10.1021/acs.energyfuels.1c01638.

(62) Corrado, T. J.; Huang, Z.; Huang, D.; Wamble, N.; Luo, T.; Guo, R. Pentiptycene-Based Ladder Polymers with Configurational Free Volume for Enhanced Gas Separation Performance and Physical Aging Resistance. *Proc. Natl. Acad. Sci.* **2021**, *118* (37), e2022204118. https://doi.org/10.1073/pnas.2022204118.

(63) Wang, Y.; Ma, X.; Ghanem, B. S.; Alghunaimi, F.; Pinnau, I.; Han, Y. Polymers of Intrinsic Microporosity for Energy-Intensive Membrane-Based Gas Separations. *Mater. Today Nano* **2018**, *3*, 69–95. https://doi.org/10.1016/j.mtnano.2018.11.003.

(64) Lo, S.; Seifrid, M.; Gaudin, T.; Aspuru-Guzik, A. Augmenting Polymer Datasets by Iterative Rearrangement. ChemRxiv November 30, 2022. https://doi.org/10.26434/chemrxiv-2022-hxvcc.

(65) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.

(66) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.

(67) *SYBA: Bayesian estimation of synthetic accessibility of organic compounds | Journal of Cheminformatics | Full Text*. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00439-2 (accessed 2023-02-09).

(68) *Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning - Chemical Science (RSC Publishing)*. https://pubs.rsc.org/en/content/articlelanding/2021/sc/d0sc05401a (accessed 2023-02-09).