Discovery of Innovative Polymers for Next-Generation Gas-Separation Membranes using Interpretable Machine Learning

Jason Yang [a,†], Lei Tao [b,†], Jinlong He [b,†], Jeffrey R McCutcheon [c,d], and Ying Li [b,d*]

[a] Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States
[b] Department of Mechanical Engineering, University of Connecticut, Storrs, Connecticut 06269, United States
[c] Department of Chemical & Biomolecular Engineering, Center for Environmental Sciences and Engineering, University of Connecticut, Storrs, Connecticut 06269, United States
[d] Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, Connecticut 06269, United States

[†]Equal contribution

*Corresponding Author: Ying Li; Email: ying.3.li@uconn.edu; Tel: (860) 483-7110. Fax: (860) 486-5088.

ABSTRACT

Polymer membranes perform innumerable separations with far-reaching environmental implications. Despite decades of research on membrane technologies, design of new membrane materials remains a largely Edisonian process. To address this shortcoming, we demonstrate a generalizable, accurate machine-learning (ML) implementation for the discovery of innovative polymers with ideal separation performance. Specifically, multitask ML models are trained on available experimental data to link polymer chemistry to gas permeabilities of He, H_2, O_2, N_2, CO_2, and CH_4. We interpret the ML models and extract chemical heuristics for membrane design, through Shapley Additive exPlanations (SHAP) analysis. We then screen over nine million hypothetical polymers through our models and identify thousands of candidates that lie well above current performance upper bounds. Notably, we discover hundreds of never-before-seen ultrapermeable polymer membranes with O_2 and CO_2 permeability greater than 10^4 and 10^5 Barrer, respectively, orders of magnitude higher than currently available polymeric membranes. These hypothetical polymers are capable of overcoming undesirable trade-off relationship between permeability and selectivity, thus significantly expanding the currently limited library of polymer membranes for highly efficient gas separations. High-fidelity molecular dynamics simulations confirm the ML-predicted gas permeabilities of the promising candidates, which suggests that many can be translated to reality.

1. INTRODUCTION

Polymer membranes are a flexible, processable, and inexpensive platform to provide a myriad of separations that fill critical roles in climate change mitigation (carbon capture) and resiliency (water treatment). For gas separations, polymer membranes have been widely used in the separation of mixtures in many industrial processes, including oxygen enrichment[1], biogas purification[2], and post-combustion carbon capture[3]. In particular, carbon capture processes are garnering increased attention to reduce emissions to the environment, and membrane technologies offer known advantages such as high energy efficiency and operational simplicity due to flexibility and scalability[4–6]. In post combustion, pre-combustion, and oxy-combustion, CO_2/N_2, CO_2/H_2, and O_2/N_2 separations are respectively important for environmental conservation.

During membrane-based gas separation, a gas mixture is typically driven through a membrane by pressure where separation is achieved through differences in individual gas permeabilities[7]. The performance of membrane processes is determined by the membrane's permeability for a specific gas species, P_i, where i specifies the gas type. The membrane permeability is calculated from Fick's law of diffusion, $|J_i| = P_i \Delta p / l$, where J_i is the flux of gas i, and Δp is the pressure drop across a membrane of thickness l. When comparing the permeability of gas A with that of gas B, another performance measure is the membrane's selectivity between two gases, α, which is defined as $\alpha = P_A / P_B$. An ideal membrane for a given binary gas separation would have high permeability and high selectivity. Increasing gas permeability and selectivity in these membranes would allow for more efficient industrial processes by increasing the process throughput, reducing energy costs, and achieving a purer product[8–10]. However, there is a well-known permeability-selectivity trade-off for polymer gas-separation membranes[7], which is defined by the Robeson upper bound[11]. Over time, advancements in polymer designs have pushed the Robeson upper bound from 1991 values to updated 2008 values (and most recently 2015 values for O_2/N_2 separations and 2019 values for CO_2/CH_4 and CO_2/N_2 separations[12,13]) that reflect improved membrane performance. Identifying new materials that break this upper bound has driven and continues to drive materials discovery efforts for membranes[14–17].

Remarkably, in the decades of technological development in the membrane science field, design of new membrane materials has been, and remains, a largely trial-and-error process, guided by experience and intuition[18]. Current approaches generally involve tuning chemical groups to increase affinity and solubility towards a desired gas, or incorporating greater free volume to increase overall diffusivity[19]. When assembling a new polymer, typically a desired enhancement is targeted (i.e., higher CO_2 affinity, higher overall permeability, aging resistance etc.) and a chemical group that is likely to achieve that enhancement is incorporated into the polymer chemistry[20–23]. For achieving higher permeability, polymers of intrinsic microporosity (PIMs) have been extensively studied during the last two decades[24,25]. PIMs generally enhance fractional free volume via inefficient chain packing to increase permeability, while simultaneously stiffening the polymer backbone and improving solubility selectivity[25–27]. Efforts to design improved chemistries for PIMs generally involve tuning the contortion group, increasing steric frustration via modifications to side chains, or further stiffening the polymer backbone[28–31]. Still, many of these studies remain limited to an Edisonian approach, unable to identify or utilize big-picture rules of chemistry-property relationships in polymer membranes.

Further complicating matters, synthesis of new polymeric materials and subsequent testing of permeance and selectivity is a time-consuming, expensive, and incomplete process that can miss high-performance candidates. Molecular modeling approaches, such as Monte Carlo/molecular dynamics (MC/MD) simulations, can reasonably predict a polymer membrane's gas permeabilities without costly experiments[32–35]. However, even these high throughput molecular simulations are too computationally expensive to explore the vast chemical space of polymers on the order of 10^6~10^{10}. By contrast, simplified approximations to predict gas permeability for a given membrane are low cost but inaccurate. Most simply, group contribution methods sum together the gas permeability contribution of each chemical moiety in a polymer, but they do not necessarily consider connectivity and cannot expand into new classes of polymers[36]. Permeability can also be calculated via diffusivity based on the polymer's free volume and the solution-diffusion model of gas transport using various theoretical models, but these theories are incomplete[37–39]. In short, there is no efficient and accurate predictive model for gas permeability based on polymer-membrane chemistry.

Machine learning (ML) is a promising data-centric approach for prediction of gas permeabilities by learning a functional model based on polymer chemistry[40,41]. ML methods using chemical inputs have been successfully applied to accurately predicting many polymer properties including glass transition temperature[42–44], thermal conductivity[45], dielectric constants[46], organic photovoltaic properties[47,48], and transport properties[49–51]. The primary challenge for learning a generalizable ML model is training on robust and diverse data, which requires compiling multiple databases with the most recent literature values and imputing missing values[40]. While Barnett et al. have trained accurate ML models that link polymer chemistry to gas permeability[50], their training set notably lacks PIMs, and they only screened a limited chemical space of 11,000 existing homopolymers. Therefore, ML-approaches would benefit from considering an expanded chemical space, such as additional training data on PIMs. Overall, ML-directed molecular design of polymer membranes still faces significant challenges in the following aspects: (1) How can we define an appropriate chemical space to explore the molecular design of high-performance polymer membranes? (2) Even if ML models can be established for the gas permeability prediction of polymer membranes, how can we achieve a physical understanding of how membrane chemistry affects gas separations? (3) Can we exceed the Robeson upper bound simultaneously for separations of different gas pairs, such as O_2/N_2, CO_2/CH_4, CO_2/N_2, and H_2/CO_2?

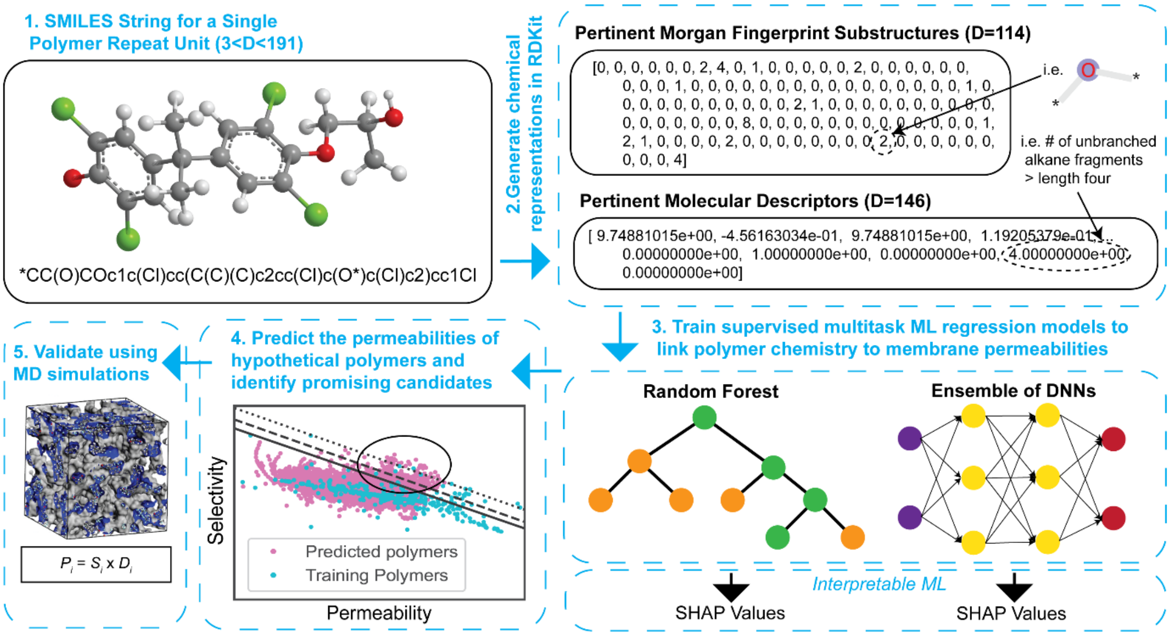
1. SMILES String for a Single Polymer Repeat Unit (3<D<191)
CC(O)COc1c(Cl)cc(C(C)(C)c2cc(Cl)c(O)c(Cl)c2)cc1Cl
2.Generate chemical representations in RDKit
Pertinent Morgan Fingerprint Substructures (D=114)
[0, 0, 0, 0, 0, 0, 2, 4, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4]
i.e.
O
*
*
Pertinent Molecular Descriptors (D=146)
[9.74881015e+00, -4.56163034e-01, 9.74881015e+00, 1.19205379e-01, ... 0.00000000e+00, 1.00000000e+00, 0.00000000e+00, 4.00000000e+00, 0.00000000e+00]
i.e. # of unbranched alkane fragments > length four
3. Train supervised multitask ML regression models to link polymer chemistry to membrane permeabilities
Random Forest
Ensemble of DNNs
Interpretable ML
SHAP Values
SHAP Values
4. Predict the permeabilities of hypothetical polymers and identify promising candidates
Selectivity
Permeability
Predicted polymers
Training Polymers
5. Validate using MD simulations
$P_i = S_i \times D_i$

Fig. 1 Workflow for ML-assisted discovery of innovative polymer membranes with ideal gas-separation performance, e.g., beyond the traditional Robeson upper bound. (1) We begin with the SMILES string for each polymer's repeating unit and its associated gas permeabilities for model training. (2) The molecule's relevant fingerprint substructures (MFFs) and molecular descriptors are extracted, which are used as chemical inputs for ML model training. The two representations have a dimensionality *D* on the order of 100. (3) Multitask RFs and ensembles of DNNs are trained to predict gas permeabilities, and physical insights can be extracted from the models using their SHAP values. (4) The models are used for high-throughput permeability prediction of hypothetical polymers with unknown permeabilities but known chemistries in a significantly expanded chemical space. (5) Finally, high-fidelity MD simulations are performed to verify the membrane permeabilities/selectivities of top polymer candidates.

To tackle the above challenges, we demonstrate interpretable, supervised ML models that can accurately predict the He, H_2, O_2, N_2, CO_2, and CH_4 permeabilities of gas separation membranes based on polymer chemistry—as part of our ML-assisted discovery workflow outlined in **Fig. 1**. Our training data consists of polymer chemistry and experimental gas permeabilities from two large databases, PolyInfo[52] and Membrane Society of Australasia (MSA)[53], for hundreds of homopolymers, including PIMs. We utilize two representations for the polymer repeating unit, namely chemical descriptors as generated by RDKit[54] (listed in **Table S1** of Supporting Information) and the Morgan fingerprint with frequency (MFF)[55]. We impute missing permeabilities using multivariable imputation by chained equations (MICE)[56], and we then train multitask supervised ML models to establish synthesis-property relations for these polymer membranes. While various supervised ML models have been used in polymer informatics, including recurrent neural networks, support vector machines, gaussian processes, and others, we choose to focus our study to random forest (RF) regression and deep neural networks (DNN), which have demonstrated outstanding performance in our recent benchmark study[44]. Due to the high variance of DNNs, we perform bootstrap ensembling to further improve our predictions (achieving testing R^2 ~0.90). We also interpret our ML models by extracting feature importances using SHAP (SHapley Additive exPlanations)[57] analysis. Our analysis provides a chemical explanation for the well-known permeability-selectivity tradeoff in membranes, and many of the other physical insights that we draw are consistent with established membrane design principles. Using the trained ML models, we perform high-throughput screening of over nine million hypothetical polymers with unknown permeabilities, including many polyimides and ladder polymers that can be classified as PIMs. Thus, we identify thousands of promising polymers for gas separation membranes with desirable performance, which lie well above 2008 Robeson upper bounds. Finally, we perform high-fidelity MD simulations to confirm that the ML-predicted permeabilities of top-performing polymers are very accurate. Overall, our ML-assisted workflow is a promising method for the discovery of innovative polymers for next-generation gas-separation membranes to advance energy and environmental sustainability.

2. RESULTS

2.1 Datasets and chemical space under exploration

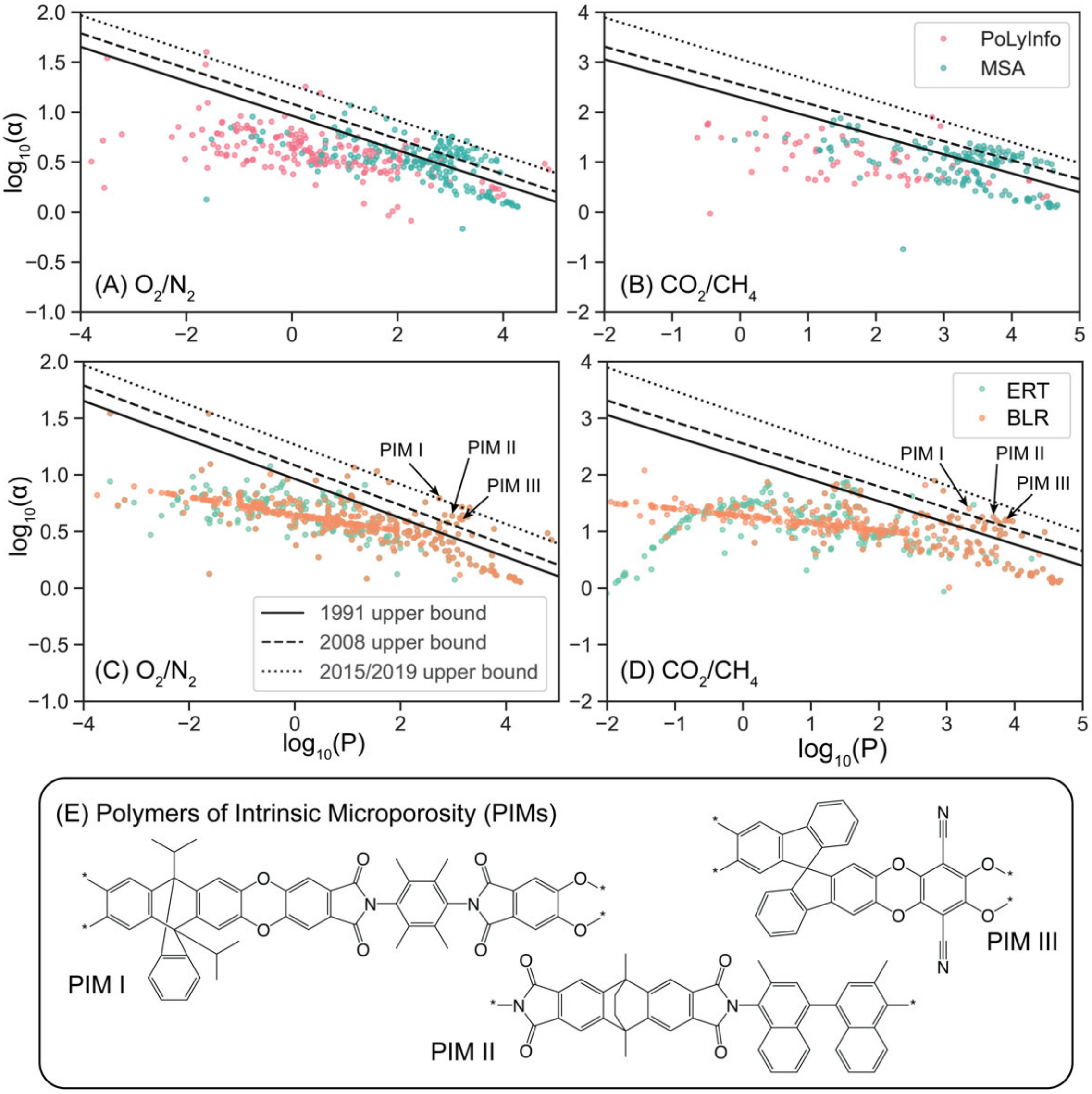

Fig. 2 Visualization of the permeability distribution of Dataset A, the training set. (a) O_2/N_2 and (b) CO_2/CH_4 Robeson plots for the raw data as obtained from the PoLyInfo and Membrane Society of Australasia (MSA) databases. (c) O_2/N_2 and (d) CO_2/CH_4 Robeson plots comparing the results of imputation using extremely randomized trees (ERT) vs Bayesian linear regression (BLR), with permeabilities averaged across entries that correspond to the same polymer. Units of permeability are Barrers. (e) The chemical structures of three existing examples of PIMs, with their performances identified in (c) and (d). Asterisks in chemical structures indicate connection points between repeating units.

Our training dataset, Dataset A, consists of 778 homopolymers (353 unique polymer chemistries), but not all entries have gas permeability data reported on all six gases under study: He, H_2, O_2, N_2, CO_2, and CH_4. Dataset A is manually collected from the PoLyInfo database (experimental data from before 2005) and is merged with data from the MSA database (beyond 2005). As shown in **Fig. 2(a-b)**, in general, the more recent MSA database contains polymers with higher permeability, e.g., CO_2 permeability greater than 10^3 Barrers, and entries that surpass the 2008 Robeson upper bound. **Fig. 2(c-d)** also show that there is not a significant difference when missing gas permeabilities are imputed via extremely randomized trees (ERT) vs Bayesian linear regression (BLR). In these plots, we identify several known PIMs, with their corresponding chemical structures shown in **Fig. 2(e)**. Many of these PIMs are ladder polymers, which have two connection points between consecutive monomers, such as PIM I and PIM II. Some of the polymers are polyimides, such as PIM III. Further visualizations of the results of the imputation process can be found in **Fig. S1** of Supporting Information.

	No. of Polymers	Permeabilities	Description	Source
Dataset A	778 (353 unique)	At least one gas known	Training Set	PoLyInfo[52] and MSA[56] Databases
Dataset B	995,799	Unknown	PI1M[58]	Hypothetical polymers generated from PoLyInfo through a recurrent neural network
Dataset C	8,205,087	Unknown	Polyimides	Hypothetical polyimides formed by known dianhydride and diamine/diisocyanate pairs from the PubChem[59]
Dataset D	1,124	Unknown	Ladder Polymers	Hypothetical polymers generated based on existing ladder polymers

Table 1 Summary of the datasets explored in this work. Dataset A is the training set, which contains polymers with known chemistries and permeabilities. Datasets B, C, and D contain hypothetical polymers with unknown permeabilities (used for screening and polymer discovery). They span three different chemical spaces: known polymers from PoLyInfo, polyimides, and ladder polymers, respectively.

Table 1 provides a summary of the training and screening datasets used in this work. Dataset B, PI1M, consists of polymers learned via a recurrent neural network (RNN) trained on simplified molecular input line entry system (SMILES) strings of existing polymers in PoLyInfo, as constructed by Ma and Luo[58]. Note that Dataset B covers an overlapping chemical space with Dataset A because the RNN model is also trained on the PoLyInfo database, but Dataset B significantly populates regions where PoLyInfo data are sparse[58]. Still, Dataset B mostly spans polymers similar to known polymers, which are generally not tailored for membrane separations. Thus, our motivation for constructing Datasets C and D is for rationally targeted exploration of the polymer design space, based on the established interest in PIMs within the membrane design community. Firstly, polyimides have garnered significant attention due to their superior permeability/selectivity tradeoff and high chemical and physical stability, largely due to a rigid aromatic backbone[21,60]. Thus, Dataset C is constructed as 8 million hypothetical polyimides formed by the polycondensation of known diamines/diisocyanates with dianhydrides from the PubChem library[59]. These 8 million hypothetical polyimides significantly expand the current chemical space of around 2000 polyimides in PoLyInfo. Ladder polymers adopt an alternative approach to stiffen the polymer backbone. These unique polymers have two-bond connections between repeating units and thus restricted rotation, except at a contortion site, which is often a spiro-center[61]. In our work, Dataset D contains hypothetical ladder polymers generated through the binary combinations of components of existing ladder polymers[62], supplemented by a RNN model. More details about the construction of these datasets are provided in **Fig. S2** of Supporting Information.

While we train ML models using both chemical descriptors and MFFs as inputs, for simplicity, we only screen new polymers using MFFs as model inputs. The feature spaces for fingerprints across the datasets studied in this work are visualized using uniform manifold approximation and projection (UMAP)[63] in **Fig. S3** of Supporting Information. In general, our training set, Dataset A, spans across the screening space of Datasets B, C, and D. Thus, our ML models can learn across a wide chemical feature space of interest. While Datasets B and C have more complete coverage due to the sheer number of samples, Dataset D only includes ladder polymers, which explains why they are more confined in the feature space.

2.2 Performance of ML models for gas permeability prediction

To quantify performance, we evaluate the accuracy and generalizability of our ML models, namely RFs and DNN ensembles trained on chemical descriptors and MFFs. For our supervised ML models, the metric of study is the R^2 correlation between the predicted and actual permeabilities on the training and test sets, as summarized in **Table 2**. We focus our analysis to models trained on the permeabilities imputed via BLR for consistency. Firstly, we find that the choice of ML model is more important than the choice of chemical features. The average test R^2 across all six gases for the RF is approximately 0.74 when trained on descriptors and very similar when trained on fingerprints. Similarly, the test R^2 values for the DNN ensembles are around 0.90 for both descriptors and fingerprints. We infer that fingerprints offer slightly better performance, which has been observed in the prediction of polymer glass transition temperature[42].

ML Model	He		H_2		O_2		N_2		CO_2		CH_4	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
RF (Descriptors)	0.96	0.73	0.96	0.74	0.96	0.75	0.96	0.74	0.96	0.75	0.96	0.74
DNN Ensemble (Descriptors)	0.85	0.87	0.87	0.88	0.89	0.89	0.90	0.90	0.88	0.90	0.89	0.89
RF (MFFs)	0.89	0.73	0.89	0.74	0.89	0.74	0.90	0.74	0.89	0.75	0.90	0.74
DNN Ensemble (MFFs)	0.88	0.91	0.88	0.90	0.90	0.92	0.90	0.91	0.89	0.90	0.89	0.88

Table 2 Summary of the performances of supervised ML models as scored by the R^2 value between the predicted and actual permeabilities. All ML models make multitask predictions for the six gas permeabilities of He, H_2, O_2, N_2, CO_2, and CH_4 and are trained on the data that is augmented using BLR imputation. The DNN ensemble models perform better than the RF models, and models trained on MFFs perform slightly better than models trained on molecular descriptors.

By contrast, we find that the choice of ML model has a significant impact on performance. RF learns a model with train R^2s of about 0.96 on descriptors and 0.90 on fingerprints, which reduce to test R^2s of about 0.74 for both inputs. In particular, the RF model seems to struggle to fit the datapoints with very low or very high permeabilities, as demonstrated in **Fig. S4** of Supporting Information by the points that have a high actual permeability but lie below the unit line. This would suggest that the RF does not prioritize fitting to the PIMs with high gas permeabilities in the training set, as PIMs make up a relatively small fraction of the training data and tend to have distinct chemistry compared to the rest of the training set.

On the other hand, the DNN ensemble learns a model with train R^2s of around 0.87 on descriptors and 0.89 on fingerprints, which generalizes very well to test R^2s of approximately 0.89 for both inputs. The similarity between train and test R^2s for the DNN ensembles suggests that the model is very generalizable and learns the underlying functional relationship between chemistry and permeability. In **Fig. S5** of Supporting Information, we note that the DNN ensemble generally predicts permeability reasonably well, though there are some outliers at low or high permeability. Uncertainty quantification of the DNN ensemble in **Table S2** of Supporting Information reveals that the ensemble of models performs better than the sum of its parts, as the average test R^2 for each individual model is only ~0.70. Overall, there is also around 10% average normalized variance in the predicted permeabilities across the 16 DNN models, which is quite high.

2.3 Physical insights from interpretation of ML models

Usually, ML models are treated as black boxes, which makes it challenging to understand any physical principles learned by the models. However, we find that obtaining SHAP values from our ML models on chemical descriptors and MFFs makes our models not only accurate but also interpretable. By extracting the most important chemical features that predict gas permeability, we draw physical insights into the molecular design of polymer membranes. Here we decide to focus our analysis on the DNN ensemble because of its better performance, but other model types can also be explained using the same method.

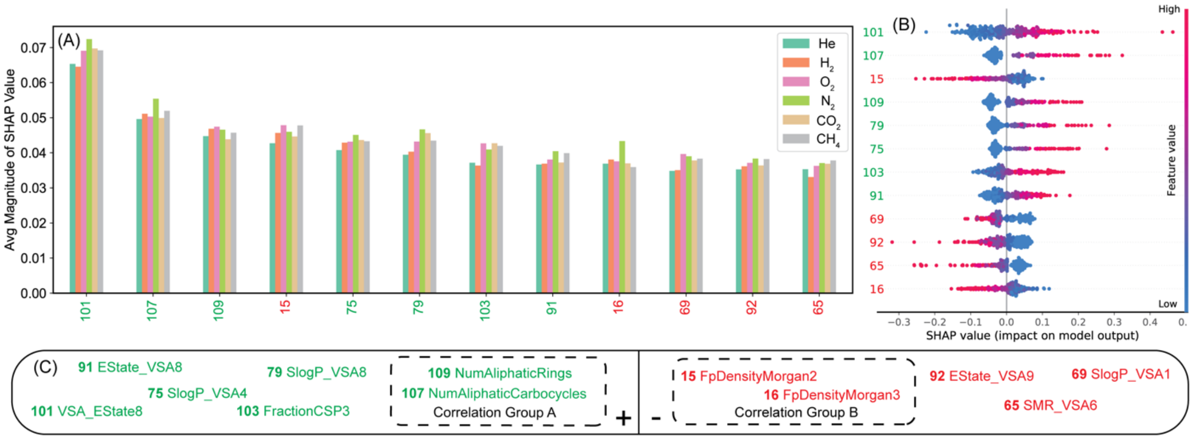

Fig. 3 Important molecular descriptors as identified using SHAP on the DNN-ensemble ML model trained on descriptors and BLR-imputed permeabilities. (a) Average SHAP importances for the top twelve descriptors on each of the six gas permeabilities (He, H_2, O_2, N_2, CO_2, and CH_4). (b) Impact of the top twelve descriptors on CH_4 permeability output. Each dot represents the impact of a particular sample in the training set. (c) Names of top descriptors, with highly correlated features circled. Green text signifies features that have positive effects on permeability, and red signifies a negative effect.

Fig. 3 summarizes the results of SHAP analysis on the DNN ensemble trained on chemical descriptors. **Fig. 3(a)** highlights the twelve most important chemical descriptors based on their average SHAP values–their relative impacts on the six gas permeabilities under study. A summary of the definitions of these descriptors is provided in **Table S3** of Supporting Information. The most important descriptor is VSA_EState8, which is a hybrid electronic state and van der Waals surface area (VSA) descriptor[64], based on precalculated surface area values derived from a list of functional groups. While some of these descriptors do not have obvious, intuitive physical meaning, the permeability of polymer membranes is determined by the solubility and diffusivity of gas molecules[34], which is affected by the electrostatic interactions and free volume elements respectively. Therefore, these identified chemical descriptors should play important roles in gas permeability of polymer membranes.

In **Fig. 3(b)**, we show how the values of each of the top descriptors impacts the model's CH_4 permeability prediction. If higher feature values result in more positive SHAP values, then the feature has a positive effect on permeability: the feature is directly correlated with gas permeability. On the other hand, if higher feature values result in more negative SHAP values, then the feature has a negative effect on permeability: the feature is inversely correlated with gas permeability. While we only show the impacts of features on CH_4 permeability output, we draw the same conclusions from the other gas permeability predictions, which produce almost identical SHAP impact plots (**Fig. S6** in Supporting Information), as all permeabilities are trained via the same multitask model.

Based on the correlation matrix between descriptor features in **Fig. S7** of Supporting Information, there are two main pairs of correlated features, which suggests that some of the top features do not have independent physical significance. Namely, descriptors 107 and 109 (Aliphatic Cycle Counts, Correlation Group A) are highly correlated with one another and relatively anti-correlated with features 15 and 16 (FpDensityMorgan, Correlation Group B). The steric space occupied by rings generally results in a lower molecular density, which explains the opposition between Group A and Group B. The features in Group A have a positive impact on gas permeability, while the features in Group B have a negative impact on gas permeability. This suggests that repeating units with more non-aromatic rings allow for larger free volume elements and lower densities, thereby higher gas permeabilities. This supports the emerging direction of polymer research on non-planar structures, such as, *kink*, *spiro*, *cardo* and pendant groups ($–CF_3$), bulky and flexible groups (–O–), or different spatial linkage configurations in polyimides for enlarging their microporosities[28,65].

In **Fig. 4**, we perform the same type of feature importance analysis using SHAP values, for the DNN model trained on MFFs. Here, we highlight the most important chemical substructures in the prediction of gas permeability. As shown in **Fig. 4(a)**, the most important substructure overall is 2854, the methyl group. We believe that this feature facilitates permeability because it is hydrophobic and its shape contributes to steric frustration between polymer chains. Similarly, the quaternary carbon connected to an aliphatic ring (substructure 2168) contributes to increasing permeability, which supports our findings above. The DNN model also learns that the number of connection bonds, substructure 1781, is correlated with gas permeability, because many high-permeability PIMs are ladder polymers with four connection points per repeating unit, as opposed to two for a typical polymer. The correlation matrix between chemical substructures (**Fig. S9** of Supporting Information) suggests that most of the important substructure features are independent of one another. However, substructures 1781, 1432, and 822 are highly correlated and all have a positive relation to gas permeability (**Fig. 4(b)**). Upon closer examination, we find that substructure 1781 is contained within substructure 1432, which is contained in substructure 822. Substructures 1432 and 822, two double-bonded carbons connected to an aromatic ring, define polyacetylenes, which demonstrate some of the highest permeabilities among non-porous polymers in gas separations[66]. By contrast, polar groups generally have negative contributions to gas permeability, as shown in **Fig. 4(b)**. For example, double-bonded oxygens (799 and 2706), ethers (1519), and nitrogen atoms (2906) are all inversely correlated with gas permeability. Since most gas molecules are non-polar, the presence of these polar groups generally reduces the solubility of gases, which explains the negative effect.

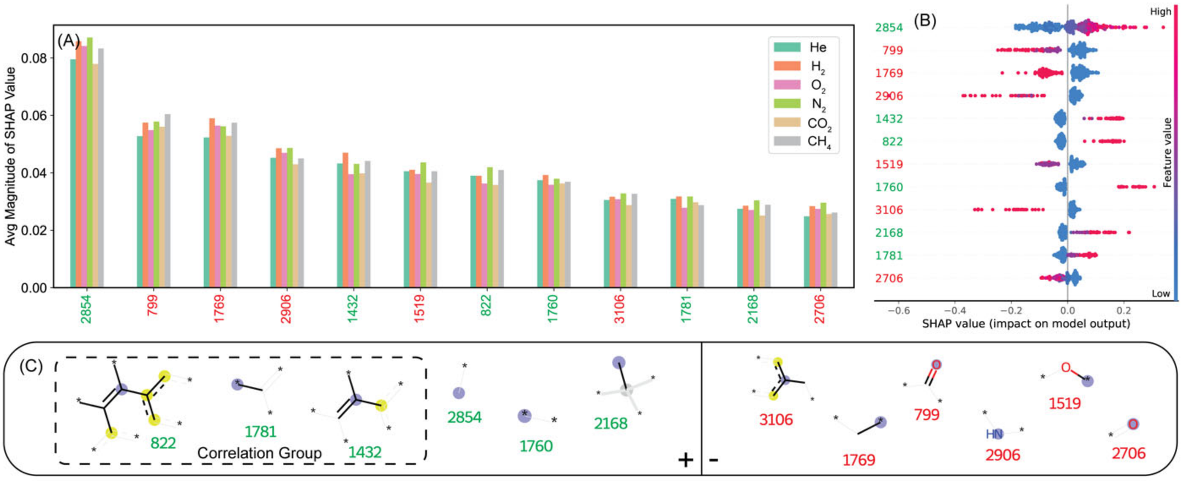

Fig. 4 Important molecular substructures as identified using SHAP on the DNN-ensemble ML model trained on MFFs and BLR-imputed permeabilities. (a) Average SHAP importances for the top twelve substructures for each of the six gas permeabilities (He, H_2, O_2, N_2, CO_2, and CH_4). (b) Impact of top twelve substructures on CH_4 permeability output. Each dot represents the impact of a particular sample in the training set. (c) Illustration of top substructures, with correlated features circled. Green text signifies features that have positive effects on permeability and red signifies a negative effect. In the substructure drawings, blue highlights the central atom in the environment, yellow indicates aromatic atoms, and gray indicates aliphatic ring atoms.

However, in **Fig. 4(a)**, our ML models show that these groups tend to have a greater negative impact (measured by SHAP value) on N_2 and CH_4 permeability, compared to O_2 and CO_2, which explains why the presence of these groups in polyimides, ladder polymers, and poly(ethylene oxides)[19] can increase selectivity by widening the permeability difference between certain pairs of gases, which is desirable for gas separations. This supports a known heuristic in membrane design, that CO_2 selectivity can be increased via increased CO_2 solubility by incorporating oxygen atoms into polymer membranes[19,23]. Across the board for substructures, SHAP values tend to be higher for N_2 and CH_4 compared to O_2 and CO_2 (**Fig. 4(a)**), which suggests that incorporating chemistries that increase permeability (i.e. methyl groups) is likely to come at the cost of selectivity. Our ML models thus elucidate a chemical basis for the permeability/selectivity tradeoff: chemical features that increase permeability are likely to do so to a greater extent for molecules that are less permeable (N_2 and CH_4), but chemical features that reduce permeability are also likely to impact these molecules to a greater magnitude–thereby increasing selectivity for the more permeable gas (O_2 and CO_2). Achieving high permeability and selectivity thus becomes a balancing act. This unique understanding is unlocked from the ability of ML to learn complex patterns in data.

2.4 Discovery of high-performance polymers and validation through MD simulations

After training our RF and DNN ensemble ML models, we use the models based on MFFs for high-throughput screening and discovery of high-performance polymers for gas separations. We choose the ML models using MFFs for simplicity due to their slightly better performance and lower memory requirements. We calculate MFFs for millions of hypothetical polymers in Datasets B, C and D, which span a wide and relevant chemical space. These inputs are then passed through the RF and DNN-ensemble models in a feed-forward manner. The predicted permeabilities for the DNN model, broken down by dataset, are plotted for O_2/N_2, CO_2/CH_4, CO_2/N_2, and H_2/CO_2 separations in **Fig. 5**. Similarly, the predictions for the RF model are visualized in **Fig. S10** of Supporting Information.

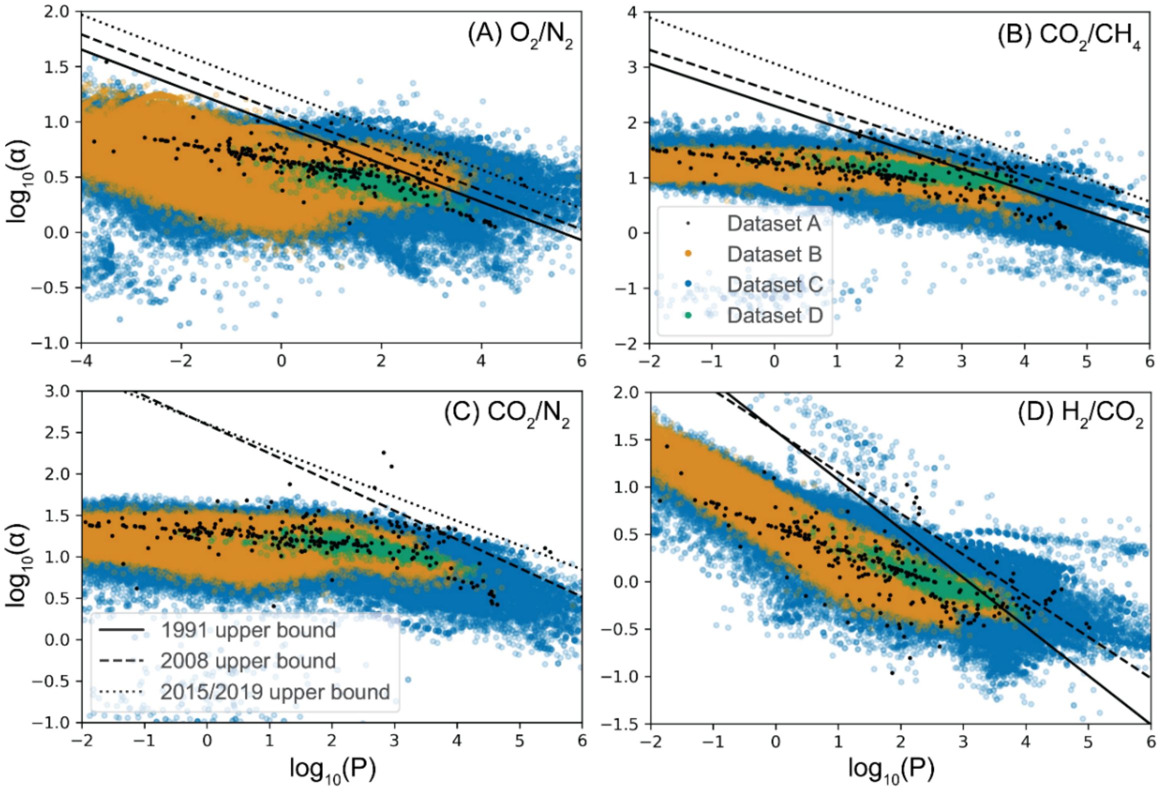

Fig. 5 Visualization of predicted permeabilities for hypothetical polymers in Datasets B, C, and D, based on the ensemble of DNNs trained on MFFs with BLR-imputed permeabilities. The training dataset (Dataset A) is overlayed on the predicted permeabilities. The data is visualized for (a) O_2/N_2, (b) CO_2/CH_4, (c) CO_2/N_2, and (d) H_2/CO_2 separations, with thousands of promising polymers lying at or above the Robeson upper bounds. Units of permeability are Barrers.

Broadly, we find that the RF model predicts permeabilities in a much narrower space than the DNN ensemble, which explains its lower R^2 values on the test set and further supports the observation that the DNN ensemble is more accurate and generalizable. Predicted permeabilities for each screening dataset lie in their expected region in the permeability-selectivity space, which further supports the accuracy of our ML models. Namely, both models predict permeabilities close to existing Robeson upper bounds for polymers in Dataset D, which consists entirely of ladder polymers (a subclass of PIMs). Similarly, Dataset C consists of polyimides (including many PIMs), and their permeability predictions span a space that includes polymers below and above the Robeson upper bound, reflecting the dataset's diversity. However, Dataset B corresponds to mostly polymers with low permeability and selectivity. We believe that this can be explained by the fact that PoLyInfo is a broad database that contains many polymers that are not suitable for gas separation applications, and Dataset B is populated from existing polymers in PoLyInfo.

Most promisingly, the DNN model predicts thousands of polymers from Dataset C to be above the 2008 Robeson upper bound, for O_2/N_2, CO_2/CH_4, CO_2/N_2, and H_2/CO_2 separations, which is summarized in **Table 3**. We further find that the DNN ensemble trained on MFFs not only generalizes but also extrapolates. We discover a class of hypothetical polymers in Dataset C with never-before-seen ultrahigh CO_2 permeability (greater than 10^5 Barrer) and a class of polymers with ultrahigh O_2 permeability (greater than 10^4 Barrer)—even though our training set only contains 12 polymers with O_2 permeability greater than 10^4 Barrer and only 2 polymers with CO_2 permeability greater than 10^5 Barrer.

Dataset C	Gas or Separation	No. of Polymers
Above 2008 Robeson upper bound	O_2/N_2	~80,000
	CO_2/CH_4	~3,000
	CO_2/N_2	~800
	H_2/CO_2	~10,000
Permeability above 10^4 Barrer	O_2	197
Permeability above 10^5 Barrer	CO_2	225

Table 3 Summary of the number of polymers with exceptional performance discovered from Dataset C. Dataset C contains about 8 million hypothetical polyimides formed by known dianhydride and diamine/diisocyanate pairs from PubChem[59].

We select several hypothetical polymers with high predicted permeability and selectivity across all four separations under study, and we validate their performance using MD simulations to calculate permeability. Details of intermediate values calculated during the simulation process can be found in **Table S5** of Supporting Information, while SMILES strings for the selected polymers are given in **Table S4**. Note that our MD simulation protocol has been benchmarked against experimental and simulation values for the gas permeabilities of PIM-1[61], with good agreement with the literature (**Table S7** of Supporting Information).

Both the RF and DNN model can identify polymers with high performance in Datasets C and D. The chemical structures of the selected polymers are drawn in **Fig. 6(a)**. We highlight some of the top substructures identified from SHAP analysis (**Fig. 4**) in these chemical structures, which corroborates our earlier conclusions. For instance, the higher permeability polymers tend to have more methyl groups (substructure 2854) and methyl groups attached to aliphatic rings (substructure 2168) to increase steric frustration. Meanwhile, double-bonded oxygens (substructure 2706) in the polyimide backbone help to maintain selectivity for gases such as O_2 and CO_2.

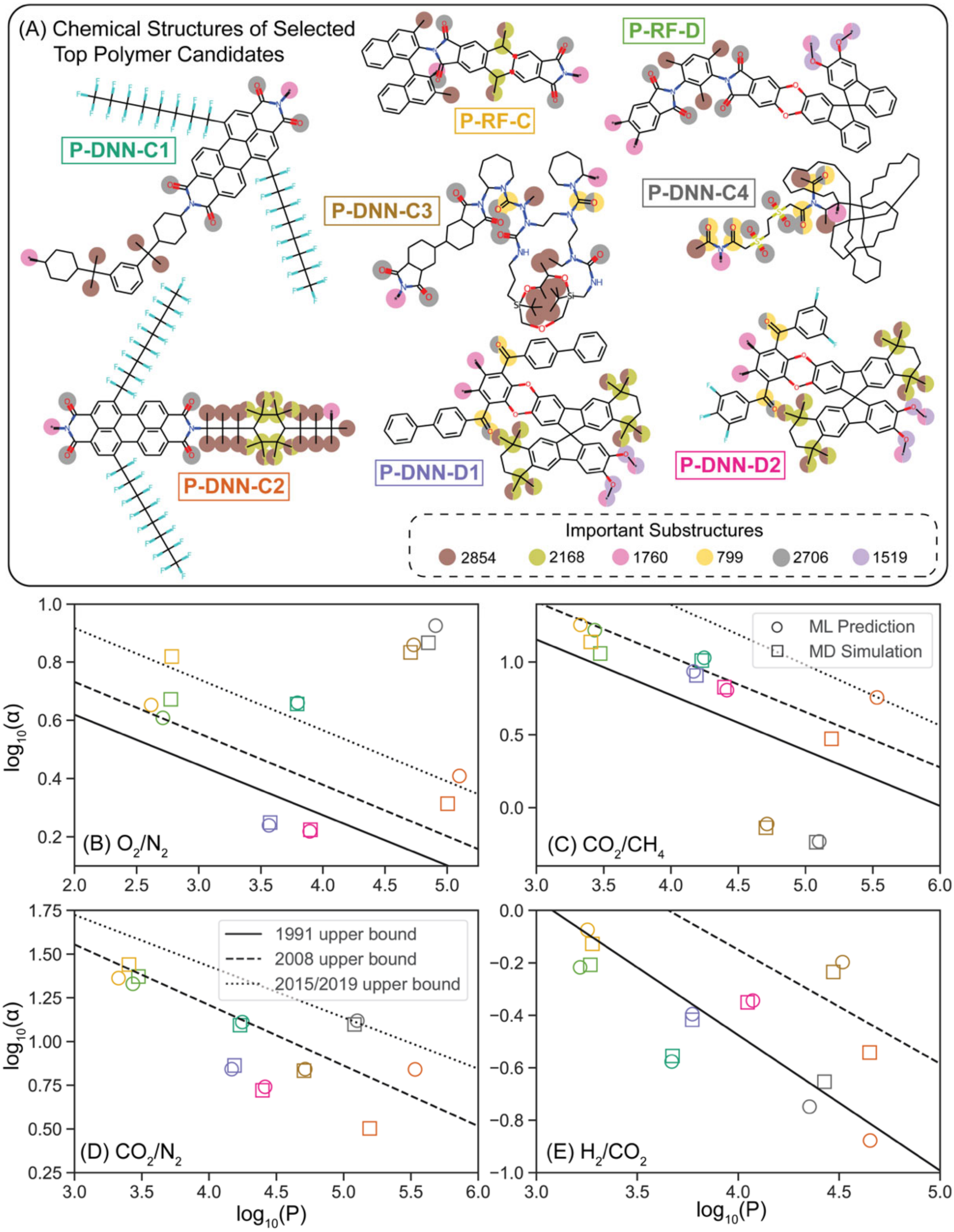

Fig. 6 Validation, using MD simulations, of the performance of selected top polymer candidates from the ML models trained on MFFs with BLR-imputed permeabilities. (a) Chemical structures for the selected polymer candidates with high performance. Important chemical substructures are highlighted in the molecules. Asterisks in chemical structures indicate connection points between repeating units. The predictions from ML models are shown as circles, while corresponding MD simulation values are shown as squares, for (b) O_2/N_2, (c) CO_2/CH_4, (d) CO_2/N_2, and (e) H_2/CO_2 separations. Units of permeability are Barrers. P-RF-C is identified from the RF model, from Dataset C. P-RF-D is identified from the RF model, from Dataset D. P-DNN-C1 through P-DNN-C4 are identified from the DNN ensemble model, from Dataset C. P-DNN-D1 and P-DNN-D2 are identified from the DNN ensemble model, from Dataset D.

As shown in **Fig. 6(b-e)**, ML-predicted performances lie very close to their respective MD-simulated performances for separations involving O_2, N_2, CO_2, CH_4, and H_2. Error ranges for permeability calculations from simulations and predictions from ML models are provided in **Table S6** of Supporting Information. In general, the DNN-ensemble model predictions differ less from the values given by MD simulations, compared to those of the RF model. While the permeability predictions tend to have larger error and uncertainty as the DNN model extrapolates to higher permeability values, our *experimentally validated* MD simulations confirm the predicted performances of these top candidates. Thus, thousands of polymers in our screening datasets with predicted permeabilities above the Robeson upper bound, or ultrahigh predicted permeabilities, could translate to real polymer membranes with exceptional separation performance.

Notably, P-DNN-C3 and P-DNN-C4, hypothetical polyimides in Dataset C, demonstrate O_2/N_2 selectivity significantly beyond the 2015 upper bound of existing known polymers, as predicted by our DNN-ensemble model and further validated by MD simulations. To our knowledge, these novel discoveries have the highest O_2/N_2 selectivities for their respective O_2 permeabilities, discovered to date. These hypothetical polyimides can each be formed through the polycondensation of a previously synthesized diisocyanate with a dianhydride, as shown in **Fig. S15-16** of Supporting Information.

Because of the multitask nature of the DNN-ensemble ML model, many polymers are predicted to perform well across several metrics. For example, P-DNN-C3 surpasses current upper bounds for O_2/N_2 and H_2/CO_2 separations, while P-DNN-C4 demonstrates exceptional performance for O_2/N_2 and CO_2/N_2 separations. However, these two polymers have poor CO_2/CH_4 selectivity. P-DNN-C1 performs near or above the 2008 Robeson upper bounds for O_2/N_2, CO_2/CH_4, and CO_2/N_2 separations. In another vein, P-DNN-C2 has both ultrahigh O_2 permeability and CO_2 permeability, while maintaining high selectivity. Similarly, P-DNN-C1 and P-DNN-C2 can each be formed through the polycondensation of a known diamine and a dianhydride, as given in **Fig. S13-14** of Supporting Information.

To further investigate the superior permeability of the selected top polymer candidates, we generate their realistic structural models and analyze their pore structures via molecular simulations in **Fig. S17** of Supporting Information. Details of our simulated polymerization algorithm are given in the **Methods** section (and additionally **Fig. S12-16** of Supporting Information). In comparison with PIM-1, our top candidates have more voids, enhanced microporosity, and larger pore radii. The pore size distribution of the top candidate polymers is wider and shifted to the right, further suggesting enhanced microporosity and permeability.

3. DISCUSSION

In this work, we demonstrate an accurate and cost-effective ML implementation that can effectively explore the ever-expanding design space for polymeric gas-separation membrane materials, by learning their synthesis-property relationships. Firstly, our study reveals that fixed chemical descriptors or fingerprints are both excellent representations for predicting gas permeabilities of polymer membranes. Corroborating our recent benchmark study on polymer glass-transition temperature[44], we conclude that the choice of chemical representation generally plays a limited role in each ML model's performance, as long as sufficient chemical substructures are captured. Additional features, such as microstructure, could be considered in future ML models, given the importance of microstructural characteristics such as FVEs in solution-diffusion transport theory of membranes[67]. Incorporation of such characteristics as input features has improved metal-organic framework adsorption prediction[68,69], compared to using solely chemical descriptors. These microstructural features could be efficiently calculated via MD simulations, as being demonstrated in this work. Alternatively, because high-throughput MD simulations can calculate gas permeabilities with reasonable accuracy, these simulations could also be used to augment the training set or be incorporated into active learning frameworks to reduce the uncertainty of ML models[71]. Nevertheless, we find that using fixed chemical features captures sufficient information to predict the gas permeabilities of the polymer membranes studied here.

We additionally gain insight into how the choice of ML model affects performance. At the same time, we demonstrate that ensembling is a powerful technique for improving prediction accuracy while simultaneously quantifying uncertainty. Traditionally, RF models are thought to work better on small datasets, while deep learning is reserved for large training sets. But while decision trees are adequate for capturing simple relationships, neural networks can in principle approximate any function to arbitrary accuracy[72]. In our study, we demonstrate that deep learning can be effectively applied to small training datasets on the order of a few hundred training samples. We believe that our DNN method is accurate for two reasons. Firstly, a DNN that is deep enough will not overfit if it's in the "modern" interpolating regime[73]. Secondly, each DNN model, seeing limited data, captures complexities and nuances in the data, which results in individual predictions with high variance; however, the overall model generalizes well when predictions are averaged together via ensembling[74]. Importantly, training the 16 DNNs in our study and evaluating predictions for millions of samples is still computationally tractable from a cost standpoint. Though various other neural networks such as graph neural networks are garnering increased interest for certain molecular discovery and synthesis tasks[75], we do not observe notable performance gains from training graph convolutional, recurrent, or convolutional neural networks. We have reached a similar conclusion from our polymer informatics benchmark study on polymer glass transition[44]. In short, we believe that deep learning techniques, even standard multilayer perceptrons, have much broader applicability to small datasets of chemical features than previously assumed.

We further show that SHAP analysis can succinctly elucidate the impacts of input features, which erodes the paradigm that ML models are black boxes[76]. SHAP values can be calculated for nearly all supervised ML models, and we encourage future chemical and polymer informatics studies to take advantage of explainability in ML[77]. A recent study also used coloring of substructures when training a graph neural network for interpretable ML[78], which suggests that feature-importance analysis of ML models can be extended beyond fixed representations to learned chemical representations.

Our study of fixed feature importance solidifies many existing membrane design principles, but additionally offers unique, generalized guidance for the molecular engineering of new polymers for gas separations. Overall, SHAP analysis illuminates the chemical balancing act required for overcoming the permeability/selectivity tradeoff. Polymers must juggle (A) the number of bulky chemical moieties–i.e. methyl groups, aliphatic rings–that increase microporosity (permeability at the expense of selectivity) with (B) the number of polar groups–i.e. carbonyls, oxygens–that increase relative CO_2 and O_2 affinity (selectivity at the expense of permeability).

P-DNN-C3 and C4 are case studies into this balancing act. They achieve high permeability primarily through methyl groups and large aliphatic rings, which is a relatively underexplored strategy in membrane design. At the same time, these polymers attain unprecedented O_2/N_2 selectivity via the incorporation of polar groups, such as carbonyls and sulfonyls. By contrast, P-DNN-C1 and C2 each feature an inflexible polycyclic backbone and two trifluoromethyl containing side chains. Amazingly, they demonstrate that our DNN model learns the importance of bulky spherical groups (such as trifluoromethyl groups) for creating steric frustration[79], which has been recognized in the gas separation community as favoring higher gas permeability[25,80,81]. Restricted backbone mobility plus the presence of the bulky pendant groups disrupts polymer chain packing and leads to high fractional free volume and ultrahigh permeability, all while the polar polyimide backbone helps to maintain selectivity.

Differently, the discovered polymers from Dataset D utilize a rigid ladder-type backbone with a spirobifluorene (SBF) unit, like many other ladder-type PIMs[31]. The fused benzene rings in the SBF unit reduce the flexibility of the backbone around the spirocenter, and the two-bond ladder connections restrict the rotation ability of the backbone. The reduced chain flexibility may also prohibit chain motion to help resist physical aging[25]. To further increase permeability, P-DNN-D1 and P-DNN-D2 attach a fused tetramethyltetrahydronaphthalene (TMN) to the SBF unit, which incorporates additional aliphatic rings and methyl substituents[30].

Overall, the generalizable ML models presented here are capable of efficiently discovering promising polymers with high performance, with thousands of candidates lying beyond the 2008 Robeson upper bound[11]. Additionally, the ultra-high permeability polymers discovered in this work would allow for never-before-seen industrial gas separations with higher throughput while maintaining sufficient selectivity. Incredibly, the DNN model can extrapolate relatively accurately to high permeability predictions that it did not see in training. We believe that this amazing performance primarily arises from careful selection of diverse training samples and training with a neural network that can capture complexities but also generalizations through multitask parameter sharing and ensembling.

Our *experimentally validated* MD simulations of gas permeability confirm the ML predictions, which suggests that many of the polymer candidates discovered here can be translated to reality in experiments. As elaborated upon in **Fig. S2** of Supporting Information, each of the promising polyimides identified here have a well-defined cross-linking formation from existing PubChem chemicals, which makes their syntheses tractable. However, the difficulty of synthesizing complex polymers in a solution-processable manner should not be underestimated. Therefore, to facilitate the overcoming of this challenge, we have tabulated the thousands of promising polymers that we have identified and included them in the GitHub repository associated with this work (https://github.com/jsunn-y/PolymerGasMembraneML), which we encourage experimental and computational researchers to explore further. While our models consider membrane performance to be constant, future efforts should also take into account how aging, plasticization, and swelling can degrade membrane performance over time, which is an important consideration in membrane design[8,82].

Ultimately, we provide the membrane design community with many novel high-performance polymer candidates and key chemical features to consider when designing their molecular structures. Many of the concepts demonstrated here can likely be extended to other materials discovery and design tasks, such as polymer membranes for desalination and water treatment[14], high-temperature fuel cells[83], and catalysis[84]. With the continual improvement of ML techniques and an increase in computing power, we expect that ML discovery frameworks will only gain popularity and deliver increasingly substantial results in materials discovery for a wide range of applications[85].

4. METHODS

4.1 Calculation of chemical representations for polymers

The workflow for our ML method to learn synthesis-property relationships of gas separation membranes is shown in **Fig. 1**. In Step 1, the training set consists of the single repeating units of 353 unique homopolymers with at least one known gas permeability (among He, H_2, O_2, N_2, CO_2, and CH_4), as obtained from the PoLyInfo[52] and Membrane Society of Australasia[53] (MSA) databases. In the datasets, each polymer entry is identified based on its unique SMILES string, a notation for chemical structures that represents a molecule as a unique string of ASCII characters[86]. ML models for prediction of gas permeability from chemistry must utilize a descriptive and appropriate input to represent the polymer[44]. Thus, in Step 2, RDKit[54] is used to calculate two different chemical representations of each polymer's repeating unit: its molecular descriptors and its MFF[55].

First, 146 relevant chemical descriptors are calculated, which generally includes information such as number of certain atom types, presence/absence of features, and number of rings, among other physical descriptors that can be calculated from the atom types and connectivity. A list of the available descriptors in RDKit is provided in **Table S1** of Supporting Information. Thus, the important chemical features of each polymer repeating unit are identified. We additionally use RDKit to generate the MFF for each repeating unit chemistry. In short, the fingerprinting process consists of[42]: (1) assign each atom with an identifier, (2) update each atom's identifiers based on its neighbors, (3) remove duplicates, and (4) fold list of identifiers into a bit vector (a Morgan fingerprint). In our case, the chemical substructures considered are up to three units in radius, where each atom or bond is one unit, resulting in 3209 different substructures. In the fingerprint vector of length 3209, each bucket indicates if a certain substructure is present, and we minimize information loss by accounting for frequency if a substructure is present multiple times in a single repeating unit, known as the Morgan fingerprint with frequency (MFF)[42]. Finally, we shorten the fingerprint vector by only using the 114 most common substructures in the training set as input features. Unlike group-contribution methods, fingerprinting is dynamic and can evolve to include new chemical structures and connectivities between neighboring repeating units.

4.2 Training and interpretation of supervised ML models

The calculated molecular features are then used as inputs for multitask ML models and trained to learn gas permeabilities (He, H_2, O_2, N_2, CO_2, and CH_4), in Step 3. For each of our supervised ML models, training is based on the log of the permeability measured in Barrers and, for a given polymer, the permeability values are averaged across multiple literature sources, if available. Many polymer entries in our training database have missing data, where gas permeabilities are not available for all six gases under study. Yuan et al. have demonstrated effective imputation of missing gas permeability data using the MICE algorithm[56], if at least one gas permeability is available. We use their source code to impute missing gas permeabilities to augment our dataset. For MICE, we compare a linear predictive model, Bayesian linear regression (BLR), with a nonlinear predictive model, extremely randomized trees (ERT). By filling in missing gas permeabilities, imputation allows us to train multitask ML models. This improves our models via parameter sharing, which is physically reasonable, as permeabilities between different gases are related for a given membrane chemistry.

We train multitask RF and DNN models to predict gas permeabilities based on chemical descriptors and fingerprints, with 20% of the data reserved for the test set and the remaining 80% used for training, selected randomly. RFs reduce variance in decision trees by making regression predictions based on the average of many decision trees: during the growth of each tree, each new decision rule is made using only a random subset of datapoints and features[87]. RFs thus build generalizable models of non-linear relationships. We train each RF using 200 estimators, with training capped at the square root of the number of features for each decision tree and a max tree depth of 10. Additionally, we train dense multilayer perceptrons. These DNNs have 5 hidden layers with 64, 64, 32, 16, and 8 nodes, respectively; ReLU activation; and dropout of 0.1. The multitask models output 6 permeabilities for the 6 gases of study. The DNNs are trained using minibatch gradient descent with a batch size of 64, the Adam optimizer, and mean-squared-error loss.

Due to their density and complexity, deep learning models can be susceptible to particularly high variance. Especially, when trained on a small dataset, there is inherent stochasticity resulting from the network's initialization and the order of data processing during training. There exist many ways to quantify and reduce uncertainty for problems with chemical inputs[71,88]. One simple way to improve the predictive capacity of such models is through ensembling, or averaging together several models trained under different conditions. For example, given an ensemble of distinct models $\mathcal{E} = \{M_1, M_2, \ldots, M_n\}$ and inputs x, the ensemble prediction is given by the mean of all the model predictions

$$\bar{M}(x) = \sum_{M \in \mathcal{E}} \frac{M(x)}{n}$$

The uncertainty of the prediction can then be measured as the variance between model outputs.

$$U(x) = \sum_{M \in \mathcal{E}} \frac{\left(\bar{M}(x) - M(x)\right)^2}{n}$$

While there are many ways to perform ensembling, we choose to use bootstrapping, or training each model in the ensemble with a different random subset of the training data. First, we randomly select 20% of the data to be the holdout set, which is used for performance scoring. 16 independent models are trained, using 80% of the entries in the non-holdout set each time, selected at random. The training of our DNN ensemble on MFFs with BLR imputation of permeabilities is given in **Fig. S11** of Supporting Information.

Alongside the models trained in Step 3 of our workflow, we can perform explainable ML. To strengthen our physical understanding of how chemical features are linked to performance in gas separation membranes, our primary tool involves assessing SHAP values from each model[57]. In essence, the SHAP approach considers how well a model performs when each feature is neglected during training. By analyzing the quantitative impact of leaving out a feature on the model prediction, a feature importance can be assigned. Moreover, each sample's impact on the final model prediction can also be evaluated.

Once the ML models are trained and achieve good performance, we then screen over nine million hypothetical polymers (summarized in **Table 1**) to predict their gas permeabilities, in Step 4. Our screening predictions are then used to identify promising polymer candidates with high permeability and selectivity. The code and datasets for our ML implementation can be found at https://github.com/jsunn-y/PolymerGasMembraneML.

4.3 Permeability validation using MD simulations

In Step 5, to validate the gas permeabilities of selected polymeric membranes, all-atom MD simulations, using Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[89], are performed to calculate each gas's permeability as the product of its solubility and diffusivity[34]:

$$P_i = D_i \times S_i$$

The polymer consistent force field (PCFF)[90–92] is employed to describe the interatomic interactions of both polymer and gas, which has been widely used to calculate the mechanical properties, cohesive energies, heat capacities and elastic constants of organic polymers[91,93,94].

We construct the polymeric membrane models via the multi-step crosslinking of binary components in the polymers of interest, as most high-performance polymers are polyimides or ladder polymers. The reactive atoms are first assigned to each monomer, and 45 of each component are packed into a 3D-periodic amorphous cell. Geometry optimization and 5 annealing cycles of the packed system are carried out. The optimized structure is then cross-linked under the NVT ensemble within an initial cutoff distance of 4.5Å. Covalent bonds are formed between reactive atoms, and the cross-linked network is relaxed under the NPT ensemble for 1 ns. After that, the next cross-linking step continues with an increased cutoff distance of 0.5Å until the cross-linking degree reaches 90%. During the cross-linking process, extra hydrogen atoms are removed, and partial charges are updated to follow assignments from the force field and charge neutrality. The generated, cross-linked polymer structure is used for subsequent calculations. Details of the cross-linking results for selected ladder polymers and polyimides are presented in **Fig. S12-S16** of Supporting Information.

For solubility calculations, MD simulations are performed with a time step of 1.0 fs in the following sequence: (1) energy minimization, (2) 0.5 ns NVT-MD simulation at 600 K, (3) 0.5 ns NPT-MD simulation at 600K and 1 bar, (4) 5 thermal annealing cycles from 600 to 300K with a temperature interval of 50K at 1 bar, (5) 0.5 ns NPT-MD simulation at 300K and 1 bar, and (6) 0.5ns NVT-MD simulation at 300K. Finally, the solubility coefficients of relevant gases are evaluated at infinite dilution, which are equal to their Henry's constants[95].

Before simulations of diffusivity, gas molecules, such as H_2, CH_4, CO_2, O_2 or N_2 are inserted into the simulation box of the cross-linked polymer. The system is first equilibrated through a 21-step MD equilibration protocol[95]. The system is then equilibrated for 1 ns under the NVT ensemble at 300K, followed by 2 ns under the NPT ensemble at 300K and 1 atm. Production runs are then performed for a duration of 7 ns. The first 2 ns are used for equilibration and the remaining 5 ns for analysis. The diffusion coefficient of gas molecules in the cross-linked polymer is estimated by the mean squared displacement (MSD) defined as

$$\mathrm{MSD(t)} = \frac{1}{6N}\frac{d}{dt}\lim_{t\to\infty}\sum_{i=0}^{N}\langle|r_i(t) - r_i(0)|^2\rangle$$

where N is the number of gas molecules and $r_i(t)$ is the position of molecule i at time t. MSD is calculated from the ensemble average $\langle\cdots\rangle$ of the trajectory, and we use the multiple-origin method to improve the statistical accuracy. In addition, to account for molecular adsorption to the polymer membrane at the saturation state, we consider the diffusivity of different numbers of gas molecules: 5, 10, 20, 30, 40, 50, and 100. We find that using 20, 30, or 40 gas molecules result in similar diffusivities, which are averaged to give the calculated diffusivity. Finally, based on the solution-diffusion mechanism, gas permeability (P_i) in a polymer membrane can be expressed as the product of the diffusivity (D_i) and the solubility constant (S_i). As summarized in **Table S7** of Supporting Information, our benchmark study on the solubility, diffusivity, and permeability of five pertinent gases (H_2, N_2, O_2, CO_2, and CH_4) in a PIM-1 membrane agrees well with available experimental data and simulation results[33,61], which suggests that our MD model and method are physically reasonable.

Acknowledgements

We gratefully acknowledge financial support from the Air Force Office of Scientific Research through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Dr. Ming-Jen Pan) and the National Science Foundation (CMMI-1934829 and CAREER Award CMMI-2046751). Y.L. would like to express thanks for the support from 3M's Non-Tenured Faculty Award. Y.L. and J.R.M would like to thank the support from the National Alliance for Water Innovation (NAWI), funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), Advanced Manufacturing Office, under Funding Opportunity Announcement Number DE-FOA-0001905. This research also benefited in part from the computational resources and staff contributions provided by the Booth Engineering Center for Advanced Technology (BECAT) at UConn. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Defense. The authors also acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (Frontera project and National Science Foundation Award 1818253) and National Renewable Energy Laboratory (Eagle Computing System) for providing HPC resources that have contributed to the research results reported within this paper. We would also like to thank Dr. Mayur Ostwal for helpful comments on the manuscript.

Author Contributions

Y.L. and L.T. conceived the idea and supervised the research. J.Y. and L.T. collected and analyzed the data and implemented the ML models. J.H. and Y.L. developed and analyzed the molecular simulations. J.Y., L.T., J.R.M., and Y.L. contributed to the design of the project and data analysis. J.Y., L.T., and J.H. wrote the first draft of the manuscript, and all authors contributed to revising the manuscript.

Competing Interests

The authors declare no competing interests.

Supporting Information

The supporting information is available free of charge.

REFERENCES

1. Bhide, B. D. & Stern, S. A. A new evaluation of membrane processes for the oxygen-enrichment of air. II. Effects of economic parameters and membrane properties. *J. Membr. Sci.* **62**, 37–58 (1991).
2. Basu, S., L. Khan, A., Cano-Odena, A., Liu, C. & J. Vankelecom, I. F. Membrane-based technologies for biogas separations. *Chem. Soc. Rev.* **39**, 750–768 (2010).
3. Zhao, S. *et al.* Status and progress of membrane contactors in post-combustion carbon capture: A state-of-the-art review of new developments. *J. Membr. Sci.* **511**, 180–206 (2016).
4. Han, Y. & Ho, W. S. W. Polymeric membranes for CO2 separation and capture. *J. Membr. Sci.* **628**, 119244 (2021).
5. Bui, M. *et al.* Carbon capture and storage (CCS): the way forward. *Energy Environ. Sci.* **11**, 1062–1176 (2018).
6. Vitillo, J. G. Introduction: Carbon Capture and Separation. *Chem. Rev.* **117**, 9521–9523 (2017).
7. Freeman, B. D. Basis of Permeability/Selectivity Tradeoff Relations in Polymeric Gas Separation Membranes. *Macromolecules* **32**, 375–380 (1999).
8. Sanders, D. F. *et al.* Energy-efficient polymeric gas separation membranes for a sustainable future: A review. *Polymer* **54**, 4729–4761 (2013).
9. Park, H. B., Kamcev, J., Robeson, L. M., Elimelech, M. & Freeman, B. D. Maximizing the right stuff: The trade-off between membrane permeability and selectivity. *Science* **356**, eaab0530 (2017).
10. Merkel, T. C., Lin, H., Wei, X. & Baker, R. Power plant post-combustion carbon dioxide capture: An opportunity for membranes. *J. Membr. Sci.* **359**, 126–139 (2010).
11. Robeson, L. M. The upper bound revisited. *J. Membr. Sci.* **320**, 390–400 (2008).

12. Comesaña-Gándara, B. *et al.* Redefining the Robeson upper bounds for CO2/CH4 and CO2/N2 separations using a series of ultrapermeable benzotriptycene-based polymers of intrinsic microporosity. *Energy Environ. Sci.* **12**, 2733–2740 (2019).

13. Swaidan, R., Ghanem, B. & Pinnau, I. Fine-Tuned Intrinsically Ultramicroporous Polymers Redefine the Permeability/Selectivity Upper Bounds of Membrane-Based Air and Hydrogen Separations. *ACS Macro Lett.* **4**, 947–951 (2015).

14. Werber, J. R., Osuji, C. O. & Elimelech, M. Materials for next-generation desalination and water purification membranes. *Nat. Rev. Mater.* **1**, 1–15 (2016).

15. Galizia, M. *et al.* 50th Anniversary Perspective: Polymers and Mixed Matrix Membranes for Gas and Vapor Separation: A Review and Prospective Opportunities. *Macromolecules* **50**, 7809–7843 (2017).

16. Wang, S. *et al.* Advances in high permeability polymer-based membrane materials for CO2 separations. *Energy Environ. Sci.* **9**, 1863–1890 (2016).

17. Du, N., Bum Park, H., M. Dal-Cin, M. & D. Guiver, M. Advances in high permeability polymeric membrane materials for CO 2 separations. *Energy Environ. Sci.* **5**, 7306–7322 (2012).

18. National Academies of Sciences, Engineering, and Medicine. *A Research Agenda for Transforming Separation Science*. (The National Academies Press, 2019).

19. Liu, J., Hou, X., Park, H. B. & Lin, H. High-Performance Polymers for Membrane CO2/N2 Separation. *Chem. – Eur. J.* **22**, 15980–15990 (2016).

20. Corrado, T. J. *et al.* Pentiptycene-based ladder polymers with configurational free volume for enhanced gas separation performance and physical aging resistance. *Proc. Natl. Acad. Sci.* **118**, (2021).

21. Sanaeepur, H. *et al.* Polyimides in membrane gas separation: Monomer's molecular design and structural engineering. *Prog. Polym. Sci.* **91**, 80–125 (2019).

22. Wang, J. *et al.* Macromolecular Design for Oxygen/Nitrogen Permselective Membranes—Top-Performing Polymers in 2020—. *Polymers* **13**, 3012 (2021).

23. Lin, H. & Freeman, B. D. Gas Permeation and Diffusion in Cross-Linked Poly(ethylene glycol diacrylate). *Macromolecules* **39**, 3568–3580 (2006).

24. McKeown, N. B. Polymers of Intrinsic Microporosity (PIMs). *Polymer* **202**, 122736 (2020).

25. Corrado, T. & Guo, R. Macromolecular design strategies toward tailoring free volume in glassy polymers for high performance gas separation membranes. *Mol. Syst. Des. Eng.* **5**, 22–48 (2020).

26. Du, N. *et al.* Polymer nanosieve membranes for CO2-capture applications. *Nat. Mater.* **10**, 372–375 (2011).

27. Jimenez-Solomon, M. F., Song, Q., Jelfs, K. E., Munoz-Ibanez, M. & Livingston, A. G. Polymer nanofilms with enhanced microporosity by interfacial polymerization. *Nat. Mater.* **15**, 760–767 (2016).

28. Ghanem, B. S., Swaidan, R., Litwiller, E. & Pinnau, I. Ultra-Microporous Triptycene-based Polyimide Membranes for High-Performance Gas Separation. *Adv. Mater.* **26**, 3688–3692 (2014).

29. Ghanem, B. S., Swaidan, R., Ma, X., Litwiller, E. & Pinnau, I. Energy-Efficient Hydrogen Separation by AB-Type Ladder-Polymer Molecular Sieves. *Adv. Mater.* **26**, 6696–6700 (2014).

30. Rose, I. *et al.* Polymer ultrapermeability from the inefficient packing of 2D chains. *Nat. Mater.* **16**, 932–937 (2017).

31. Bezzu, C. G. *et al.* A spirobifluorene-based polymer of intrinsic microporosity with improved performance for gas separation. *Adv. Mater. Deerfield Beach Fla* **24**, 5930–5933 (2012).

32. Dutta, R. C. & Bhatia, S. K. Atomistic Investigation of Mixed-Gas Separation in a Fluorinated Polyimide Membrane. *ACS Appl. Polym. Mater.* **1**, 1359–1371 (2019).

33. Fang, W., Zhang, L. & Jiang, J. Polymers of intrinsic microporosity for gas permeation: a molecular simulation study. *Mol. Simul.* **36**, 992–1003 (2010).

34. Venable, R. M., Krämer, A. & Pastor, R. W. Molecular Dynamics Simulations of Membrane Permeability. *Chem. Rev.* **119**, 5954–5997 (2019).

35. Yi, S., Ghanem, B., Liu, Y., Pinnau, I. & Koros, W. J. Ultraselective glassy polymer membranes with unprecedented performance for energy-efficient sour gas separation. *Sci. Adv.* **5**, eaaw5459.

36. Robeson, L. M., Smith, C. D. & Langsam, M. A group contribution approach to predict permeability and permselectivity of aromatic polymers. *J. Membr. Sci.* **132**, 33–54 (1997).

37. Hensema, E. R., Hensema, E. R., Mulder, M. H. V., Smolders, C. A. & Smolders, C. A. On the mechanism of gas transport in rigid polymer membranes. *J. Appl. Polym. Sci.* **49**, 2081–2090 (1993).

38. Cohen, M. H. & Turnbull, D. Molecular Transport in Liquids and Glasses. *J. Chem. Phys.* **31**, 1164–1169 (1959).

39. Lin, H. & Yavari, M. Upper bound of polymeric membranes for mixed-gas CO2/CH4 separations. *J. Membr. Sci.* **475**, 101–109 (2015).

40. Chen, G. *et al.* Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges. *Polymers* **12**, 163 (2020).

41. Audus, D. J. & de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).

42. Tao, L., Chen, G. & Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2**, 100225 (2021).

43. Chen, G., Tao, L. & Li, Y. Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. *Polymers* **13**, 1898 (2021).

44. Tao, L., Varshney, V. & Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c01031.

45. Wu, S. *et al.* Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Comput. Mater.* **5**, 1–11 (2019).

46. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **6**, 20952 (2016).

47. Sun, W. *et al.* Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **5**, eaay4275 (2019).

48. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).

49. Wheatle, B. K., Fuentes, E. F., Lynd, N. A. & Ganesan, V. Design of Polymer Blend Electrolytes through a Machine Learning Approach. *Macromolecules* **53**, 9449–9459 (2020).

50. Barnett, J. W. *et al.* Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.* **6**, eaaz4301 (2020).

51. Liu, T. *et al.* Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning. *J. Mater. Chem. A* **8**, 21862–21871 (2020).

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

53. Thornton, A. W., Freeman, B. D. & Robeson, L. M. Polymer Gas Separation Membrane Database. *Membrane Society of Australasia*.

54. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Academic Press* (2013).

55. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

56. Yuan, Q. *et al.* Imputation of missing gas permeability data for polymer membranes using machine learning. *J. Membr. Sci.* **627**, 119207 (2021).

57. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *ArXiv170507874 Cs Stat* (2017).

58. Ma, R. & Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **60**, 4684–4690 (2020).

59. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).

60. Ghosh, A., Sen, S. K., Banerjee, S. & Voit, B. Solubility improvements in aromatic polyimides by macromolecular engineering. *RSC Adv.* **2**, 5900–5926 (2012).

61. Budd, P. M. *et al.* Gas separation membranes from polymers of intrinsic microporosity. *J. Membr. Sci.* **251**, 263–269 (2005).

62. Du, N., Guiver, M. D. & Robertson, G. P. Ladder polymers with intrinsic microporosity and process for production thereof. (2016).

63. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

64. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **18**, 464–477 (2000).

65. Kim, T. H., Koros, W. J., Husk, G. R. & O'Brien, K. C. Relationship between gas separation properties and chemical structure in a series of aromatic polyimides. *J. Membr. Sci.* **37**, 45–62 (1988).

66. Hu, Y., Shiotsuki, M., Sanda, F., Freeman, B. D. & Masuda, T. Synthesis and Properties of Indan-Based Polyacetylenes That Feature the Highest Gas Permeability among All the Existing Polymers. *Macromolecules* **41**, 8525–8532 (2008).

67. Wijmans, J. G. & Baker, R. W. The solution-diffusion model: a review. *J. Membr. Sci.* **107**, 1–21 (1995).

68. Pardakhti, M., Moharreri, E., Wanik, D., Suib, S. L. & Srivastava, R. Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Comb. Sci.* **19**, 640–645 (2017).

69. Moosavi, S. M. *et al.* Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **11**, 4068 (2020).

70. Venkatram, S. *et al.* Predicting Crystallization Tendency of Polymers Using Multifidelity Information Fusion and Machine Learning. *J. Phys. Chem. B* **124**, 6046–6054 (2020).

71. Soleimany, A. P. *et al.* Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).

72. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).

73. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci.* **116**, 15849–15854 (2019).

74. Ganaie, M. A., Hu, M., Tanveer*, M. & Suganthan*, P. N. Ensemble deep learning: A review. *ArXiv210402395 Cs* (2021).

75. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

76. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

77. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

78. Jiménez-Luna, J., Skalic, M., Weskamp, N. & Schneider, G. Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *J. Chem. Inf. Model.* **61**, 1083–1094 (2021).

79. Du, N. *et al.* Polymers of Intrinsic Microporosity Containing Trifluoromethyl and Phenylsulfone Groups as Materials for Membrane Gas Separation. *Macromolecules* **41**, 9656–9662 (2008).

80. He, Y. *et al.* Polymers with Side Chain Porosity for Ultrapermeable and Plasticization Resistant Materials for Gas Separations. *Adv. Mater.* **31**, 1807871 (2019).

81. Zhao, Y., He, Y. & Swager, T. M. Porous Organic Polymers via Ring Opening Metathesis Polymerization. *ACS Macro Lett.* **7**, 300–304 (2018).

82. Swaidan, R., Ghanem, B., Litwiller, E. & Pinnau, I. Physical Aging, Plasticization and Their Effects on Gas Permeation in “Rigid” Polymers of Intrinsic Microporosity. *Macromolecules* **48**, 6553–6561 (2015).

83. Li, Q., Jensen, J. O., Savinell, R. F. & Bjerrum, N. J. High temperature proton exchange membranes based on polybenzimidazoles for fuel cells. *Prog. Polym. Sci.* **34**, 449–477 (2009).

84. B. McKeown, N. & M. Budd, P. Polymers of intrinsic microporosity (PIMs): organic materials for membrane separations, heterogeneous catalysis and hydrogen storage. *Chem. Soc. Rev.* **35**, 675–683 (2006).

85. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

86. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

87. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

88. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **60**, 3770–3780 (2020).

89. Plimpton, S. Fast Parallel Algorithms for Short–Range Molecular Dynamics. *Journal of Computational Physics* **117**, 1–19 (1995).

90. Sun, H., Mumby, S. J., Maple, J. R. & Hagler, A. T. An ab Initio CFF93 All-Atom Force Field for Polycarbonates. *J. Am. Chem. Soc.* **116**, 2978–2987 (1994).

91. Sun, H., Ren, P. & Fried, J. R. The COMPASS force field: parameterization and validation for phosphazenes. *Comput. Theor. Polym. Sci.* **8**, 229–246 (1998).

92. Heinz, H., Lin, T.-J., Kishore Mishra, R. & Emami, F. S. Thermodynamically Consistent Force Fields for the Assembly of Inorganic, Organic, and Biological Nanostructures: The INTERFACE Force Field. *Langmuir* **29**, 1754–1765 (2013).

93. Bunte, S. W. & Sun, H. Molecular Modeling of Energetic Materials: The Parameterization and Validation of Nitrate Esters in the COMPASS Force Field. *J. Phys. Chem. B* **104**, 2477–2489 (2000).

94. Sun, H. COMPASS: An ab Initio Force-Field Optimized for Condensed-Phase ApplicationsOverview with Details on Alkane and Benzene Compounds. *J. Phys. Chem. B* **102**, 7338–7364 (1998).

95. Varshney, V., Patnaik, S. S., Roy, A. K. & Farmer, B. L. A Molecular Dynamics Study of Epoxy-Based Networks: Cross-Linking Procedure and Prediction of Molecular and Material Properties. *Macromolecules* **41**, 6837–6842 (2008).