MATERIALS SCIENCE

Designing exceptional gas-separation polymer membranes using machine learning

J. Wesley Barnett[1]*, Connor R. Bilchak[1]*, Yiwen Wang[1†], Brian C. Benicewicz[2], Laura A. Murdock[2], Tristan Bereau[3], Sanat K. Kumar[1‡]

The field of polymer membrane design is primarily based on empirical observation, which limits discovery of new materials optimized for separating a given gas pair. Instead of relying on exhaustive experimental investigations, we trained a machine learning (ML) algorithm, using a topological, path-based hash of the polymer repeating unit. We used a limited set of experimental gas permeability data for six different gases in ~700 polymeric constructs that have been measured to date to predict the gas-separation behavior of over 11,000 homopolymers not previously tested for these properties. To test the algorithm's accuracy, we synthesized two of the most promising polymer membranes predicted by this approach and found that they exceeded the upper bound for CO_2/CH_4 separation performance. This ML technique, which is trained using a relatively small body of experimental data (and no simulation data), evidently represents an innovative means of exploring the vast phase space available for polymer membrane design.

INTRODUCTION

Polymer membranes are used to effect a variety of gas separations (*1*–*6*) such as the removal of carbon dioxide from natural gas, oxygen from air, hydrogen recovery, and more recently in carbon capture. Separation performance is typically characterized by the membrane's permeability (P_i), i.e., the throughput of gas type *i*, and selectivity (α), the purity of the output stream. P_i is defined from Fick's law of diffusion, $|J_i| = P_i \frac{\Delta p}{\ell}$, where J_i is the flux of gas *i* and Δp is the pressure drop across a membrane of thickness ℓ. P_i is further decomposed into the product of a thermodynamic solubility constant and a diffusion constant, $P_i = D_i \times S_i$. The ideal selectivity, α, between two gases is the ratio of their permeabilities: $\alpha_{A/B} = \frac{P_A}{P_B} = \frac{D_A}{D_B} \times \frac{S_A}{S_B}$, where $\frac{D_A}{D_B}$ and $\frac{S_A}{S_B}$ are the diffusivity and solubility selectivities, respectively. While there has been an increased emphasis on the use of permeabilities and selectivities when gas mixtures (rather than pure gases) are used, the data on these systems are sparse, and hence, for the purposes of this work, we discuss pure gases.

While an optimal polymer membrane for a given gas pair should have both high permeability and high selectivity, these quantities are typically observed to be negatively correlated. This concept is demonstrated in a "Robeson plot" for a variety of polymers and gas pairs. The Robeson plot for CO_2/CH_4 separations is shown in Fig. 1; a multitude of other Robeson plots exist for different gas separations. These plots illustrate the empirically determined current best performance for a given separation as defined by the upper bound correlation (lines in Fig. 1) (*7*–*10*) Note that the upper bound evolves with time as scientists invent new materials so that while the slope of this line is apparently unchanged, the intercept increases with time. Thus, we use designations such as the 1991 upper bound (*10*) or the 2008 upper bound (*9*) to designate the temporal evolution of this observed trade-off relation. The challenge, therefore, in synthesizing next-generation polymer membranes is in designing materials that cross the current upper bound. These ideas have motivated the discovery of new classes of polymeric materials, e.g., thermally rearranged (TR) polymers (*11*, *12*) and polymers of intrinsic microporosity (*3*, *13*) with improved performance over conventional polymers.

Synthesizing and testing the vast number of possible polymer constructs and their potential chemical modifications with our currently available chemistry toolbox is an expensive and time-consuming proposition. Instead, several theoretical methods and models have been developed as a means to understand diffusion and solubility in polymeric materials, with the goal of permitting a more rational design of next-generation materials.

On the most basic level, gas permeability can be empirically predicted using group contribution methods, where polymer repeat units are decomposed into subunits and the estimated gas permeability contribution of each of these moieties is added together (*14*). This approach is only sensitive to the presence of various atoms/functional groups in a polymer backbone but does not necessarily take their connectivity into account. Further, these methods do not systematically evolve as newer classes of polymers are synthesized and measurement tools improve. Group contribution methods therefore represent a first step for predicting the gas transport properties of these polymeric materials. A more theoretically underpinned concept is that permeability, in the framework of the solution-diffusion model, can be predicted with knowledge of polymer free volume. This follows by relating the diffusion of a gas molecule of a known size with the amount of volume in the polymer that facilitates its motion. This idea has been developed to relate the "slope" of the upper bound line to the relative sizes of the gas molecules involved in a given separation (i.e., for a given Robeson plot) (*6*). However, this correlation is imperfect, and there is an incomplete understanding of the underpinning free volume concept (*15*, *16*). These free volume models have also been developed for estimating gas solubility in polymers. This concept is important for glassy polymers, which are known to swell and plasticize in the presence of CO_2, thereby markedly altering their gas solubility (*2*, *3*, *13*). While gas solubility in polymers can be elegantly derived from well-understood models

[1]Department of Chemical Engineering, Columbia University, New York, NY, USA. [2]Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC, USA. [3]Max Planck Institute for Polymer Research, Mainz, Germany.
*These authors contributed equally to this work.
†Present address: Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA.
‡Corresponding author. Email: sk2794@columbia.edu

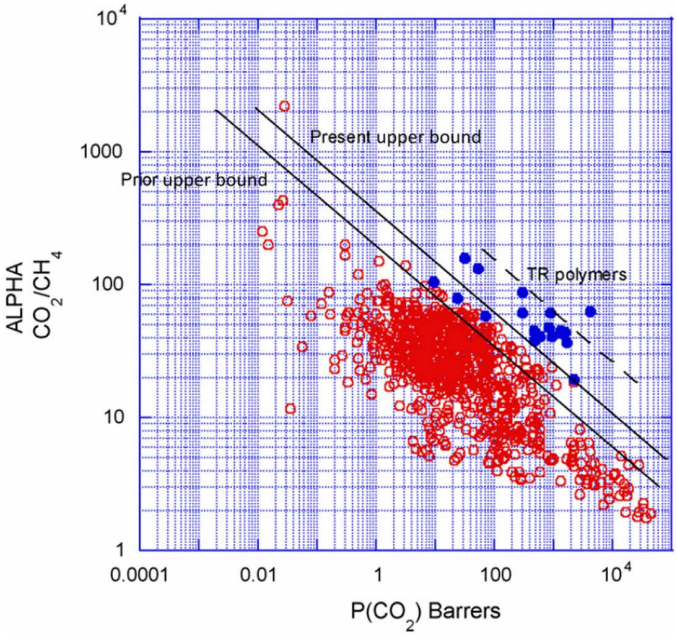

Fig. 1. Robeson plot of selectivity versus permeability for CO_2/CH_4 separations. The 1991 and 2008 Robeson upper bounds are shown as solid black lines. Each data points represents a single polymer; Robeson plots typically contain experimental data for >500 unique polymers. *P* is in units of Barrer (1 Barrer = 1 $\times 10^{-10}$ cm^3[STP] cm^2/cm^3 s cmHg). Reprinted from (*9*) with permission.

such as Sanchez-Lacombe theory extended to the nonequilibrium polymer glass state (such as the non-equilibrium lattice fluid model), these results often contain a number of unknown parameters to describe polymer-gas interactions and the extent of glassiness. This complicates a full, predictive understanding of the underlying phenomena. Other models, such as the dual-mode sorption, have also been found to qualitatively explain trends observed for gas solubility in glassy materials, but they are often limited to specific families of polymers. It is safe to say therefore that while there is good qualitative understanding of gas transport in glassy polymers, there is hardly any scientifically grounded, predictive models in this context. This concept is underpinned experimentally by the enormous scatter in the data shown on a Robeson plot, which represents not only our empirical understanding of gas transport but also the lack of design cues that can guide the synthesis of new materials. A means of rationally designing advanced membrane materials, without resorting to empirical experimentation, thus remains an open challenge.

Here we propose a different approach, which could eventually lead to the understanding of the underlying molecular processes, i.e., machine learning (ML) (*17*, *18*). In its current form, ML represents a class of statistical models that make predictions on properties based on a set of data, but without a detailed understanding of the underlying physics in these situations. These models are greatly dependent on the availability and accuracy of large sets of applicable data. Thus, when ML has been used for polymer property prediction, researchers have primarily focused on large sets of theoretically generated data ("Materials Project") (*19*–*21*). Other ML methods in the past have typically been applied to experimental datasets with less than 100 data points for any given property (*22*), which tends to limit the accuracy of the predictions of this exercise.

Our approach uses all the gas permeation data that we could find in the literature, i.e., typically 500 to 1000 polymers for each gas, to develop an ML model as outlined schematically in Fig. 2. While we have not chosen to curate these datasets to prevent user bias, this larger dataset appears to allow us to develop more reliable models. We train the ML algorithm using a training dataset (which is part of the available dataset) and test its predictions on the remaining polymers for which gas permeation data exist. This validated model can then predict the gas permeation behavior of a large body of polymers that have been synthesized to date (~11,000), but which have not been experimentally characterized in this context. Our ideas have some parallels to the group contribution methods discussed above, but with the advantage that we do not define the chemical building blocks ahead of time. Instead, we explore the polymers whose permeabilities have been measured by using a topological, path-based, fingerprinting method to describe the polymer backbone structure so that materials with previously unexplored chemistries can be easily added to the dataset as synthetic advances are made (*23*). Once the predictions on the 11,000 polymers have been made, we focus specifically on the polymers that are predicted to lie well above the upper bound, i.e., polymers particularly well suited for that separation but ones that have not been tested to date. We then experimentally validate the predicted values of P_{CO2} and P_{CO2}/P_{CH4} for these previously unexplored polymers. Thus, ML appears to be a powerful method to predict (and hence design) materials that are optimal for a given application, particularly with limited sets of experimental data.

RESULTS

We compiled a literature-based database of the diffusivities, solubilities, and permeabilities for six gases—methane (CH_4), carbon dioxide (CO_2), helium (He), hydrogen (H_2), nitrogen (N_2), and oxygen (O_2)—in a variety of polymers. The number of data points for each gas varied somewhat due to what was available in the literature, with a majority of the datasets having at least 500 polymers for each gas, as shown in Table 1; this represents a sizable portion of the polymers that are typically included in the most up-to-date Robeson plots. We then randomly split this dataset into one of two categories for each gas; one is used for training the ML model, while the other is initially withheld during training. The training datasets were ≈75% of our total database for each gas, which represented at least 250 polymers for each gas. We then apply the trained model to the remaining 25% of the polymers (test set) and use these data as verification of the model's accuracy. We found that the prediction of the ML model on these test datasets typically had an R^2 value of 0.8 or larger, although this correlation improves as the training dataset is made larger (see Table 1).

One challenge when creating ML models for evaluating physical properties is choosing appropriate descriptors to describe the materials being studied. Our first approach only included the number of each atom type in a repeat unit. However, this was found to be an ineffective means to properly model the experimental permeation data. Instead, we choose to use a fingerprinting method where the chemical connectivity in a polymer's repeating unit is represented numerically. Fingerprinting has a distinct advantage over traditional group contribution methods, where all of the possible building blocks must be defined a priori and remain static; fingerprinting methods are an inherently more dynamic representation because

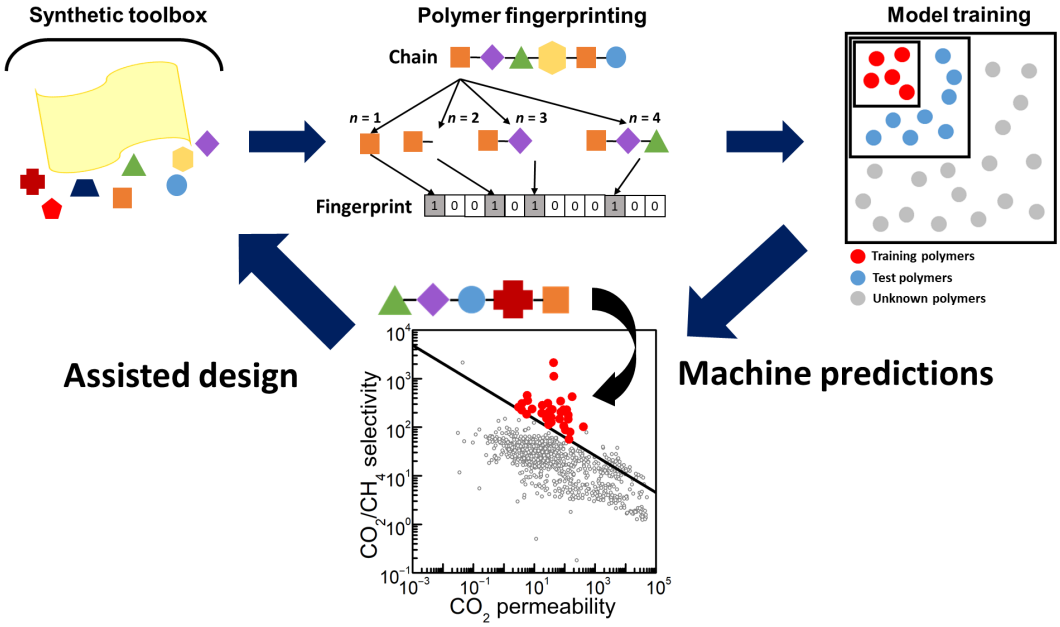

Fig. 2. Assisted design of high-performance polymer membranes. The large synthetic toolbox available for creating new polymers is simulated by translating the polymer into a binary "fingerprint," which is input to the ML algorithm. The model is trained with a random subgroup of polymers from our literature database and then tested against the remaining polymers. The model is then applied to a large set of literature data to discover high-performance polymers, thus facilitating machine-assisted design.

Table 1. Evaluation of model performance on hold-out test set. R^2 is the coefficient of determination. Sizes are the number of samples in the training set or test set for each gas. The ML package used was Scikit-learn.

Gas	Training size	Training R^2	Test size	Test R^2
N_2	514	0.986	172	0.847
O_2	523	0.985	175	0.903
H_2	324	0.985	109	0.827
He	282	0.980	94	0.799
CH_4	420	0.990	141	0.904
CO_2	471	0.986	158	0.875

they can evolve to include materials as they are synthesized. Further, they take into account the chemical connectivity between the different units. We transformed each polymer into a binary "fingerprint" using the Daylight-like fingerprinting algorithm as implemented in RDKit. This topological-based approach analyzes the various fragments of a molecule containing a certain number of bonds and then hashes each fragment to produce a binary fingerprint that computationally represents the molecule; this is shown schematically in Fig. 2. After a polymer's repeat unit was read into memory via a molfile, it was broken down into fragments containing between 1 and 7 units (represented for $n = 1$ to $n = 4$ in Fig. 2), and the structure was hashed into a fingerprint with 2048 bits of information to encode all of the possible connectivity pathways of the monomer. This process is repeated for each group in the molecule to generate the full fingerprint. Each bit was treated as a single feature in our model, which allows us to study the effects of various functional groups and their linkages on gas transport. Each monomer was connected to at least nine other identical repeat units to properly account for longer paths along the polymer backbone. This fingerprinting technique is the simplest representation of the polymer chemistry and structure that is sufficient to capture trends observed in the experimental data.

After training our model [which uses the Gaussian process regression (GPR) method] on each gas's permeability dataset (see Materials and Methods), we used both cross-validation in the training set and a hold-out test set to evaluate model performance. While Table 1 includes data from relatively large train set sizes, we have systematically varied the size of this initial training set—we find that the mean squared errors only begin to decrease for train sizes larger than ~400 and that the mean square error of the model (see Supplementary Materials) decreased monotonically as this size is increased. This explains why previous efforts, which typically used 100 polymers in their ML studies, were less insightful. Our choice of large train sizes thus reflects our goal to have a more generally applicable ML-derived model. Despite the varying amount of test data for each gas, each model performed similarly well with mean squared errors on the order of 2 to 4 Barrer (1 Barrer = 1×10^{-10} cm^3[STP] cm^2/cm^3 s cmHg; correlation curves for each gas are provided in the Supplementary Materials). Overall, we were satisfied with the test set performance and retrained the models on the full dataset to be used in predictions on new polymers never before tested. We then downloaded 11,325 molfiles from the National Institute for Materials Science (NIMS) Materials database (which represents a large repository of previously synthesized polymers) and apply the ML model to these polymers to predict their gas transport performance (*24*). Only a few structures (≈1.5%) in this prediction dataset were also in our full training set, meaning that the vast majority of polymers in the NIMS database that we predict represent new gas transport data, with no known experimental data.

One of the challenges in using ML modeling for property prediction is associating these predictions with physically meaningful quantities. This is the focus of much current research. Our model, which uses a fingerprinting method, makes it difficult to point to a specific set of physical quantities that are important in the prediction of gas permeabilities, such as free volume descriptors of the polymer chain. However, by examining the higher-performing materials—those which are above the upper bound—and their common characteristics, we are able to gain insight into what physical quantities are important for enhancing gas permeability and selectivity. We can also analyze the chemical structure of these high-performance materials to discern design motifs that are expected to give the best performance. Figure 3 (A and B) shows the learned gas transport data of the polymers in the NIMS database for O_2/H_2 and CO_2/CH_4, plotted in the Robeson plot format. Representative data used for model training are also shown.

Almost all predicted selectivities/permeabilities remain just below the Robeson 2008 upper bound line for the O_2/N_2 and CO_2/CH_4 gas pairs. However, more than 100 polymers are significantly above the 2008 upper bound for the CO_2/CH_4 gas pair. The polymers that are above this bound have several common characteristics. Of the 11,325 polymers in the dataset, polysulfides accounted for only 7.00%; however, they made up most (53.00%) of the polymers that crossed the CO_2/CH_4 2008 upper bound. In addition, the percentage of polysulfones (5.30% total, 18.00% above the upper bound) and polyimides (17.65% total, 35.00% above the upper bound) have a larger share in the upper bound–breaking group. Aromatic polyethers consisted of 30.78% of the total prediction dataset, but only 21.00% of the upper bound–breaking group; similarly, polyvinyls consisted of 13.7% of the total dataset but only 1% were above the upper bound. This implies that these functional groups are typically linked to suboptimal membrane performance (additional statistical analysis of the polymer classes in the CO_2/CH_4 Robeson plot is shown in the Supplementary Materials). The upper bound–breaking polymers were further analyzed by creating a two-dimensional histogram for group pairs. It was found that 18.00% belonged to both the polysulfone and polyimide classes, and 17.00% belonged to both the polysulfone and polyether classes. Thus, it was observed that materials containing a sulfur group, an oxygen along the backbone, and/or nitrogen rings performed the best in this context (*25*). Thus, our models for this gas pair seem to point to physically meaningful chemistries that can be used to enhance gas separations and may be further used in the future to identify strategies that have not been experimentally studied.

We focused our attention on two polymers predicted to lie well above the upper bound for CO_2/CH_4 separations (SDs are from the GPR). These two polymers are identified in the NIMS database (*24*) as poly[(1,3-dioxoisoindoline-2,5-diyl)sulfonyl(1,3-dioxoisoindoline-5,2-diyl)-1,4-phenyleneoxy-1,4-phenylene] (ID: P432092) and poly[(1,3-dioxoisoindoline-2,5-diyl)sulfonyl(1,3-dioxoisoindoline-5,2-diyl)-1,4-phenylenemethylene-1,4-phenylene] (ID: P432095). Their locations on the CO_2/CH_4 Robeson plot, as well as the structure of their repeat units, are shown in Fig. 4. Both of these polymers are polyimides containing sulfone groups; in addition, P432092 contains an aromatic ether linkage; each of these groups is highlighted during our analysis of the ML data as being related to high CO_2/CH_4 selectivities.

Although similar sulfur-containing polyimides have been tested for gas separations in general (*26*–*28*), CO_2/CH_4 selectivity has not been tested with these specific polymers. We synthesized both polymers and tested their CO_2/CH_4 transport performance to experimentally verify the ML data. The synthesized polymers were cast from solution into thin (≈30 μm) films via doctor-blading and tested using the well-known constant volume/variable pressure experimental technique with an upstream experimental pressure of ≈2 atm. The experimental results are plotted in relation to their predicted values in Fig. 4; the polymers exceed the 2008 Robeson upper bound for this gas pair as predicted by the ML model, and both P432092 and P432095 exhibit selectivities ~7 and 5.5 times, respectively, that of the upper bound at the same permeability value. Further, we find that the experimental and predicted data points are in relatively good agreement with each other (within the error of the prediction), indicating that the ML model may be used as a predictive tool in identifying previously unexplored polymers for gas separations.

DISCUSSION

The ML algorithm–based approach derives permeability predictions by using a detailed knowledge of the monomer structure and chemistry. We began with an approach that looked at only atoms, but found it to be insufficient; a description that includes connectivity within a monomer is found to be sufficient in terms of predicting

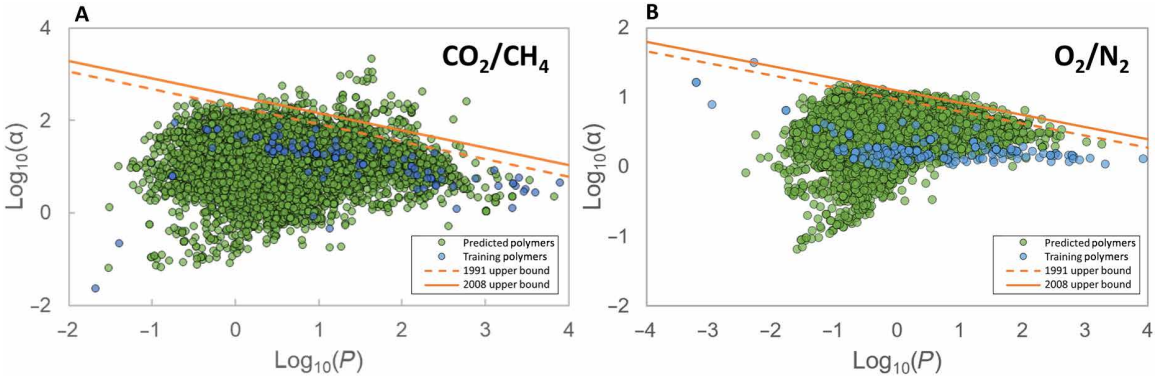

Fig. 3. Identification of polymer structures from machine-learning assisted design. Results of ML predictions on polymers in the NIMS database for (**A**) CO_2/CH_4 and (**B**) O_2/N_2 separations. A representative set of the training data is shown in blue for each polymer—note the relative sizes of the data used for training the models compared to that predicted using the ML algorithm. The 1991 and 2008 Robeson upper bounds are shown as dashed and solid lines, respectively.

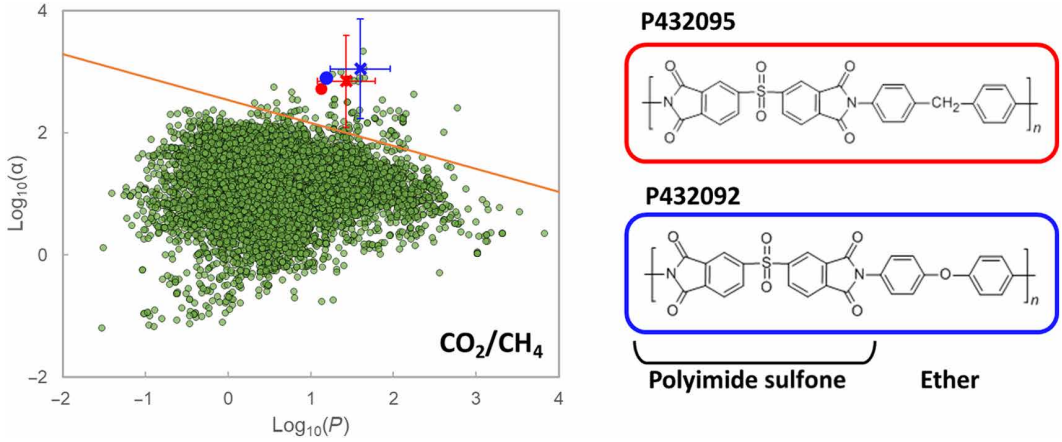

Fig. 4. Polymer candidates for advanced CO_2/CH_4 gas transport performance identified through ML and their experimental performance. CO_2/CH_4 Robeson plot showing learned permeability/selectivity data. The predicted locations for two potential high-performance polymers are marked with colored crosses—the measured experimental values are denoted with colored dots. The repeat units of both polymers contain functional groups identified through ML as being related to high-performance materials.

permeability. This approach ignores all higher-order polymer descriptors such as stereoregularity, polarity, and chain length. While we find that these variables are not required to gain reasonably accurate predictions of polymer properties, more sophisticated means of representing the polymer chain that can include these nuances may further increase the accuracy of the ML model and allow us to properly hone in on these more complex design cues. However, with current fingerprinting methods, there is no logical means by which appropriate descriptors can be defined for predicting an arbitrary property. How this choice should be made remains a topic of research. We also observe that other fingerprinting tools with similar complexity may have similar accuracy as the Daylight-like fingerprinting method we used here. An open question in the ML field is choosing the proper descriptor for a set application.

Our ML approach is designed with the specific goal of quickly characterizing gas permeabilities for an extremely large set of polymers and then a posteriori correlating high-performance materials with common functional groups and bond linkages; this allows us to determine which chemistries and structures are worth experimental observation. We emphasize that we do not relate these results to a molecular understanding of a polymer property as viewed through one of the many theoretical models available, e.g., for gas transport. In a similar vein, a number of past experimental work have focused on the effect of various polymer backbone properties on either solubility or diffusion, e.g., the effect of sulfur groups on CO_2 solubility and more polymer backbone stiffness on gas diffusion constants (*29*–*31*). Our approach focused specifically on predicting polymer permeabilities, as the available literature data that decomposes permeabilities into solubility and diffusion are less plentiful. The specific dependence of solubility and diffusivity on polymer structure can be potentially probed using this approach in the future provided a more complete database is available—this might allow us to probe the factors affecting solubility and diffusivity separately.

Our ML algorithm, as currently used, only tests against already synthesized polymers. A superior approach would be to include out-of-the-box polymer architectures in the algorithm and then imposing a "synthesizability" constraint as a means of selecting polymers for further study. However, practical implementation of this approach has not been determined, and it remains a topic of debate.

The approach presented above is easily amenable to an inverse design approach. Namely, we can design polymers with a desired combination of permeability and selectivity for a gas pair by using, e.g., a genetic algorithm to construct the optimal fingerprint vectors. This is ongoing work in our laboratory.

MATERIALS AND METHODS

Scikit-learn (*32*) was used to perform additional preprocessing and regression on the data. For each gas, the target value was the base-10 logarithm of the permeability in units of Barrer (1 Barrer = 1×10^{-10} cm^3[STP] cm/cm^2 s cmHg). The data for each gas were split into training and test sets, with 75% of the polymers and their permeabilities randomly placed in the training set and 25% placed in the test set (see Table 1 for exact numbers). After each feature was scaled by removing the mean and scaling to unit variance, GPR (*33*) was used to fit the training data. We used a kernel consisting of the sum of a radial basis function and a white noise term

$$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2}\left\|\frac{x_i - x_j}{l}\right\|_2^2\right) + \sigma_n^2 \delta(x_i - x_j) \quad (1)$$

where σ_f, σ_n, and l are optimized hyperparameters, with σ_f^2 the signal variance, σ_n^2 the noise variance, and l the length scale of the basis function. δ is the dirac delta function, and x_i and x_j correspond with samples i and j. Ten restarts with random initial hyperparameters were used for the optimization to avoid local maxima. Learning curves were created using 10-fold cross-validation on the training set. After fitting the training sets, the models were evaluated using the hold-out test sets. Last, we used the entire dataset to fit six models using GPR like we did with the training sets. We also performed an additional 10-fold cross-validation on the entire dataset, also producing learning curves. Results of the optimization for both the training set and the full dataset are shown in the Supplementary Materials.

The polymer P4320902 was synthesized using 2.3028 g of 4,4′-oxydianiline (11.5 mmol) and 4.1202 g of 5,5′-sulfonylbis (isobenzofuran-1,3-dione) (11.5 mmol). About 15 ml of dry dimethyl acetamide (DMAc) was added to each round bottom flask. Once monomers were dissolved, they were mixed together without

exposure to air and stirred at room temperature for 13.5 hours. The viscous solution was subsequently casted onto glass plates using a 10 mils doctor blade. The films were left untouched for 3 days to let the remaining DMAc evaporate, and a clear film was produced with a thickness of 30 ± 3 μm. The resulting poly(amic acid) was thermally converted to the polyimide structure by slowly heating the films to 250°C in an oven under a nitrogen atmosphere.

The polymer P4320902 was synthesized using 1.5862 g of 4,4′-methylenedianiline (8 mmol) and 2.8662 g of 5,5′-sulfonylbis (isobenzofuran-1,3-dione) (8 mmol). About 20 ml of dry DMAc was added to each round bottom flask. Once monomers were dissolved, they were mixed together without exposure to air and stirred at room temperature for 28 hours. The viscous solution was subsequently casted onto glass plates using a 5 mils doctor blade. The films were left untouched for 3 days to let the remaining DMAc evaporate, and a clear film was produced with a thickness of 30 ± 3 μm. The resulting poly(amic acid) was thermally converted to the polyimide structure by slowly heating the films to 250°C in an oven under a nitrogen atmosphere.

CO_2 and CH_4 permeabilities were measured using the constant volume/variable pressure technique. Sample films are mounted onto 47-mm brass discs with known inner diameter with epoxy to properly adhere the film to the brass. The brass functions as an impermeable mask to reduce the amount of film area necessary for the experiment. The films were supported by filter paper to avoid damage during the experiment. Films are loaded into the gas permeation cells' closed volume apparatus, where the pressure on either side of the film is monitored with electronic transducers. The sample and apparatus were then degassed under vacuum for a minimum of 12 hours, until the apparatus reached its maximum obtainable vacuum (≈15 mtorr). During this time, the apparatus is submerged in a water bath with a heater to regulate the cell temperature. The test apparatus downstream was then isolated from the vacuum, and the rate of pressure increase in the downstream chamber of known volume is measured to determine the "leak rate" of the sample. After the leak rate is measured for about 1 hour, the system is returned to ultimate vacuum to begin the experiment. Once ultimate vacuum is re-obtained, the downstream is again isolated from the vacuum and the upstream is charged with the desired penetrant gas. The upstream section was purged with the penetrant at least three times before starting the experiment to ensure that there is no competitive permeation between gases. The downstream pressure was monitored as a function of time as gas diffused through the film. The permeability of the gas, P_i, was calculated from the steady-state rate of pressure increase in the downstream

$$P = \frac{V_d \ell}{p_u ART}\left[\left(\frac{dp_d}{dt}\right)_{ss} - \left(\frac{dp_d}{dt}\right)_{leak}\right]$$

V_d is the downstream volume (calibrated using Burnett gas expansion with helium to within 0.001 cm^3), ℓ is the film thickness, p_u is the upstream pressure, A is the film area available for gas transport (defined by the brass disc inner diameter), R is the gas constant, and T is the experimental temperature (35°C).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/20/eaaz4301/DC1

REFERENCES AND NOTES

1. W. J. Koros, C. Zhang, Materials for next-generation molecularly selective synthetic membranes. *Nat. Mater.* **16**, 289–297 (2017).
2. R. W. Baker, B. T. Low, Gas separation membrane materials: A perspective. *Macromolecules* **47**, 6999–7013 (2014).
3. Y. Yampolskii, Polymeric gas separation membranes. *Macromolecules* **45**, 3298–3311 (2012).
4. T. C. Merkel, B. D. Freeman, R. J. Spontak, Z. He, I. Pinnau, P. Meakin, A. J. Hill, Ultrapermeable, Reverse-selective nanocomposite membranes. *Science* **296**, 519–522 (2002).
5. R. W. Baker, Future directions of membrane gas separation technology. *Ind. Eng. Chem. Res.* **41**, 1393–1411 (2002).
6. B. D. Freeman, Basis of permeability/selectivity tradeoff relations in polymeric gas separation membranes. *Macromolecules* **32**, 375–380 (1999).
7. H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, B. D. Freeman, Maximizing the right stuff: The trade-off between membrane permeability and selectivity. *Science* **356**, eaab0530 (2017).
8. L. M. Robeson, Q. Liu, B. D. Freeman, D. R. Paul, Comparison of transport properties of rubbery and glassy polymers and the relevance to the upper bound relationship. *J. Membr. Sci.* **476**, 421–431 (2015).
9. L. M. Robeson, The upper bound revisited. *J. Membr. Sci.* **320**, 390–400 (2008).
10. L. M. Robeson, Correlation of separation factor versus permeability for polymeric membranes. *J. Membr. Sci.* **62**, 165–185 (1991).
11. C. A. Scholes, B. D. Freeman, S. E. Kentish, Water vapor permeability and competitive sorption in thermally rearranged (Tr) membranes. *J. Membr. Sci.* **470**, 132–137 (2014).
12. D. F. Sanders, R. Guo, Z. P. Smith, Q. Liu, K. A. Stevens, J. E. McGrath, D. R. Paul, B. D. Freeman, Influence of polyimide precursor synthesis route and *ortho*-position functional group on thermally rearranged (TR) polymer properties: Pure gas permeability and selectivity. *J. Membr. Sci.* **463**, 73–81 (2014).
13. D. F. Sanders, Z. P. Smith, R. Guo, L. M. Robeson, J. E. McGrath, D. R. Paul, B. D. Freeman, Energy-efficient polymeric gas separation membranes for a sustainable future: A review. *Polymer* **54**, 4729–4761 (2013).
14. L. M. Robeson, C. D. Smith, M. Langsam, A group contribution approach to predict permeability and permselectivity of aromatic polymers. *J. Membr. Sci.* **132**, 33–54 (1997).
15. E. R. Hensema, M. H. V. Mulder, C. A. Smolders, On the mechanism of gas-transport in rigid polymer membranes. *J. Appl. Polym. Sci.* **49**, 2081–2090 (1993).
16. W. M. Lee, Selection of barrier materials from molecular-structure. *Polym. Eng. Sci.* **20**, 65–69 (1980).
17. D. J. Audus, J. J. De Pablo, Polymer informatics: Opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
18. A. L. Samuel, Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).
19. B. Kang, G. Ceder, Battery materials for ultrafast charging and discharging. *Nature* **458**, 190–193 (2009).
20. A. Van De Walle, G. Ceder, Automating first-principles phase diagram calculations. *J. Phase Equilib.* **23**, 348–359 (2002).
21. G. Ceder, Y.-M. Chiang, D. R. Sadoway, M. K. Aydinol, Y.-I. Jang, B. Huang, Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature* **392**, 694–696 (1998).
22. H. Hasnaoui, M. Krea, D. Roizard, Neural networks for the prediction of polymer permeability to gases. *J. Membr. Sci.* **541**, 541–549 (2017).
23. A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
24. Polyinfo. *https://polymer.nims.go.jp/index_en.html*.
25. S. A. Stern, Y. Mi, H. Yamamoto, A. K. St Clair, Structure permeability relationships of polyimide membranes. Applications to the separation of gas-mixtures. *J. Polym. Sci. Pol. Phys.* **27**, 1887–1909 (1989).
26. K. Tanaka, M. N. Islam, M. Kido, H. Kita, K.-I. Okamoto, Gas permeation and separation properties of sulfonated polyimide membranes. *Polymer* **47**, 4370–4377 (2006).
27. F. Piroux, E. Espuche, R. Mercier, M. Pineri, G. Gebel, Gas transport mechanism in sulfonated polyimides: Consequences on gas selectivity. *J. Membr. Sci.* **209**, 241–253 (2002).
28. C. R. Bilchak, E. Buenning, M. Asai, K. Zhang, C. J. Durning, S. K. Kumar, Y. Huang, B. C. Benicewicz, D. W. Gidley, S. Cheng, A. P. Sokolov, M. Minelli, F. Doghieri, Polymer-grafted nanoparticle membranes with controllable free volume. *Macromolecules* **50**, 7111–7120 (2017).
29. A. L. Khan, X. F. Li, I. F. J. Vankelecom, SPEEK/Matrimid blend membranes for CO_2 separation. *J. Membr. Sci.* **380**, 55–62 (2011).
30. W. J. Koros, Simplified analysis of gas polymer selective solubility behavior. *J. Polym. Sci. Pol. Phys.* **23**, 1611–1628 (1985).

31. A. Morisato, K. Ghosal, B. D. Freeman, R. T. Chern, J. C. Alvarez, J. G. De La Campa, A. E. Lozano, J. De Abajo, Gas separation properties of aromatic polyamides containing hexafluoroisopropylidene groups. *J. Membr. Sci.* **104**, 231–241 (1995).
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
33. C. E. Rasmussen, in *Advanced Lectures On Machine Learning: MI Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003 and Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, O. Bousquet, U. Von Luxburg, G. Rätsch, Eds. (Springer, 2004), pp. 63–71.

Acknowledgments
Funding: Financial support for this work was provided by the NSF Graduate Research Fellowship Program (grant DGE-16-44869 to C.R.B.), the DMREF Program (grant 1629502 to C.R.B., J.W.B., and S.K.K. and grant 1629052 to B.C.B. and L.A.M.), and the Emmy Noether Program of the Deutsche Forschungsgemeinschaft (to T.B.). S.K.K., J.W.B., and C.R.B. acknowledge that the majority of the work was funded by NSF through DMREF grant CTS-1629502. This project was started under a previous NSF grant (DMR-150730), which was continued as grant DMR-1929655. **Author contributions:** J.W.B., C.R.B., and Y.W. compiled the gas transport data. J.W.B. performed the ML simulations. J.W.B. and C.R.B. analyzed the ML data. C.R.B. performed gas transport experiments. L.A.M. synthesized the high-performance polymers. T.B. and S.K.K. designed the project. B.C.B. supervised the experimental verification of the ML results. S.K.K. wrote the paper in collaboration with all co-authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 9 September 2019
Accepted 2 March 2020
Published 15 May 2020
10.1126/sciadv.aaz4301

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

Science Advances

Designing exceptional gas-separation polymer membranes using machine learning

J. Wesley Barnett, Connor R. Bilchak, Yiwen Wang, Brian C. Benicewicz, Laura A. Murdock, Tristan Bereau and Sanat K. Kumar

Sci Adv **6** (20), eaaz4301.
DOI: 10.1126/sciadv.aaz4301