

# Machine learning-based discovery of molecular descriptors that control polymer gas permeation

T. Shastry, M. R. Carbone

To be published in "Journal of Membrane Science"

March 2024

Computational Science Initiative  
**Brookhaven National Laboratory**

**U.S. Department of Energy**  
USDOE Office of Science (SC), Basic Energy Sciences (BES)

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

## **DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

# Machine Learning-Based Discovery of Molecular Descriptors that Control Polymer Gas Permeation

Tejus Shastry,<sup>1</sup> Yasemin Basdogan,<sup>2</sup> Zhen-Gang Wang,<sup>2</sup> Sanat K. Kumar,<sup>1,\*</sup> and Matthew R. Carbone<sup>3,†</sup>

<sup>1</sup>*Department of Chemical Engineering, Columbia University, New York, New York 10027, USA*

<sup>2</sup>*Division of Chemistry and Chemical Engineering,  
California Institute of Technology, Pasadena 91125, USA*

<sup>3</sup>*Computational Science Initiative, Brookhaven National Laboratory, Upton, New York 11973, USA*

(Dated: April 12, 2024)

While machine learning has found increasing use in predicting the properties of polymeric materials with only a knowledge of chain architecture, determining the molecular factors underpinning properties (“interpretable AI”) has remained less well explored. We show that encoding chain chemistry in commonly employed formats, e.g., binary-valued fingerprints, leads to uniqueness issues during the hashing process to save storage space. This is because the hashing algorithm can map several chemical moieties into the same bit. These issues carry over into the ML algorithms, especially for “inverse” design and interpretable AI, and cannot be avoided by changing the length of the fingerprint. Using MACCS key featurizations of monomer repeats resolves some of these issues, and we show that a few substructures consistently appear in top features for maximizing permeability across several gases and ML models. These are carbon-carbon double bonds (as in polyacetylenes) especially when they are associated with methyl groups (found in branching architectures). These results, derived from the limited data set of  $\sim 500$  polymers with experimental gas permeation data, are in agreement with physical insight and thus provide a robust foundation which could further enable study of these material classes through detailed experiments and simulations.

## KEYWORDS

Machine learning, membranes, polymers, gas transport, explainable AI

## I. INTRODUCTION

The future of a sustainable chemical industry crucially depends on new separation processes [1, 2]. However, knowledge gaps prevent the widespread adoption of emergent technologies that require clean hydrogen generation, carbon-neutral renewable natural gas, oxycombustion, and non-thermal chemicals purification. Polymer membranes offer a promising solution to address these needs. They are also one of the only unit operations with a clear pathway for integration into an electrified grid. Despite this pressing need, limited methods exist for the rational discovery of improved, temporally-stable, mechanically robust, and high-flux materials for selective separations of gas mixtures. Key separation figures of merit are the permeability of mixture species  $i$ ,  $P_i = D_i S_i$  ( $D_i$  and  $S_i$  are the gas diffusion and solubility coefficients, respectively) and the selectivity of component  $i$  over  $j$ ,  $\alpha_{ij} = P_i/P_j$ . Thus, the goal is to simultaneously maximize  $P_i$  and  $\alpha_{ij}$ .

Empirically testing a polymer formulation for gas transport properties is an arduous task, making high-throughput screening based on polymer structure highly

germane. Machine learning is particularly suited to this goal, as it has yielded promising results in several disciplines. For example, Bradford *et al.* successfully used chemistry-informed neural network models to screen solid polymer electrolyte formulations for applications in lithium-ion batteries [3]. Tao *et al.* developed several models to screen copolymers for optoelectronic properties, glass transition temperature, and more [4]. Arora *et al.* predicted phase behavior of block copolymers using random forest models, eliminating the need to empirically measure molecular parameters like Flory-Huggins interactions [5]. Specifically, in the context of gas transport properties, Barnett *et al.* successfully trained a Gaussian process regression model to predict single-gas permeabilities based on the structure of the polymer alone [6]. While this fitting exercise is useful in taking gas permeation data from a limited set of  $\sim 500$  polymers and generalizing it to the known  $\sim 11k$  homopolymers in a validated manner, it does not provide any mechanistic understanding of the factors critically controlling gas permeation. Further, there is no means to design polymers with a desired combination of permeabilities for a variety of gases.

Many tools exist that attempt to explain the predictions of ML algorithms. Xu *et al.* statistically quantified individual feature impacts on the distributions of hygroscopicity, thermal expansion, and tensile modulus [7]. Wellawatte *et al.* developed universal counterfactual explanations for molecule-based ML models, spanning representations like SMILES, SELFIES, and more [8]. Further improvements to this approach included natural language explanations of the molecular features [9]. Gao *et al.* applied Shapley additive explanations (SHAP) to ultrafiltration membranes to identify correlations between

---

\* sk2794@columbia.edu

† mcarbone@bnl.gov

membrane properties and performance [10]. Ideal tools for interpretability should be model-agnostic in the sense that they do not constrain users to specific models. Two major contenders that fit these criteria are SHAP values and importance by permutation. The Shapley value is a concept derived from Game Theory and represents the average marginal contribution of one player after all possible combinations have been considered [11]. Importance by permutation, on the other hand, measures the increase in the prediction error of a model after permuting a given feature over all values it can take [12]. Shapley values and importance by permutation give similar trends for the data we examine here.

In this study, we examine a database covering the past 60 years of publications on gas transport properties of polymer membranes (see Table I) and compare features of top importance to model predictions via permutation and Shapley value techniques. We also discuss the crucial role played by the representation of data in machine learning tasks, focusing in particular on bit collisions prevalent in common fingerprinting techniques. The results underscore the need for careful choice in representation when inverting or analyzing structure-activity relationships, as the model may perform perfectly well in the property prediction, but it may be conflating several phenomena in the process, thus casting doubt on any “learning” the models may have done.

## II. METHODS AND RESULTS

### A. Data Collection

TABLE I. Summary of datasets.

Gas	Train size	Test size	Total
H <sub>2</sub>	309	78	387
He	293	74	367
CO <sub>2</sub>	484	121	605
O <sub>2</sub>	530	133	663
N <sub>2</sub>	519	130	649
CH <sub>4</sub>	414	104	518

The database of gas transport properties studied herein is an expanded version of that used by Barnett *et al.* [6], covering roughly 60 years of empirical data in polymer membranes for gas separations (Table I). The data were sourced from any available publication from several journals, such as *ACS Macromolecules* and the *Journal of Membrane Science*. To compile this database, tools including ChemDraw and the NIH’s Optical Structure Recognition (OSRA) were employed to obtain SMILES strings from images or manually drawn chemical structures of repeat units. Among available data with both permeability values and chemical structures provided, the only criterion used to select data for

usage was having “clean” copolymer ratios, that is to say only those with ratios close to 25/75, 50/50, or similar. This was done to provide reliable periodic units to represent the overall polymer. There have been a few methods developed to address the repeating nature of polymer structure within ML applications, including periodic graph models [13] or the recent BigSMILES representation [14]. We cap the two ends of the monomer structure with hydrogen atoms to create a consistent data set. With this approach we were able to eliminate the inconsistencies in the data that might have arisen from the polymer molecular weight and different chain end decorations. We believe this is a reasonable assumption since there is experimental literature that suggests the chain end decorations of polymers does not significantly affect the polymer membrane’s performance for gas separations [15–19]. These data were then used to create training and testing sets for the ML study. We note that, due to limited data availability, temperature and pressure values were not used as model inputs. Though these variables certainly have an impact on permeability, it is possible for machine learning models to “average out” their effects, learning general trends to compensate for the lack of quality data. Learning curves from Barnett *et al.* [6] suggest that some of our machine learning models are at the lower end of the data requirement ( $\sim 400$  polymers) as-is. Thus, introducing temperature and pressure as additional variables may render prediction impossible with acceptable uncertainties. Our only option is thus to omit these variables while accepting the inherent uncertainties that such an approximation brings.

### B. Representations

Our previous paper [6] has shown that the minimum level of detail required for gas permeation prediction is the local chemical structure of the monomer. Information such as chain length and tacticity seem to play less important roles. Our approach thus examines different methods for fingerprinting these monomer structures.

Fingerprinting has become a vital tool in structure-property prediction given its convenience for regression models. Unlike more complicated representations, such as natural language or graph, fingerprints are fixed-length input vectors, allowing for seamless compatibility with simpler, more interpretable models (e.g. random forests). Two of the more common algorithms include the Daylight-like RDKit topological fingerprint (RDKit fingerprint) and the Morgan fingerprint/Extended Connectivity Fingerprint (ECFP). Generally, these methods will encode molecules atom-by-atom, assigning some numeric identifiers to each atom, then iterating several times over larger and larger radii to incorporate neighboring atoms and bonds. At the end, this information is (one-way) hashed and the result is a binary-valued vector of predefined length [20, 21]. Unfortunately, it is now well understood that there are many instances in which the way a

substructure is hashed yields duplicate encodings, such as the one illustrated in Figure 1. Since fingerprinting is usually an instance-based operation, seemingly random substructures will yield the same value in the same index of the fingerprint. Given this, it is somewhat surprising that we have had success in using ML to relate polymer structure to properties. Regardless, this creates substantial issues for the interpretability and inverse design considerations which are the focus of the current work.

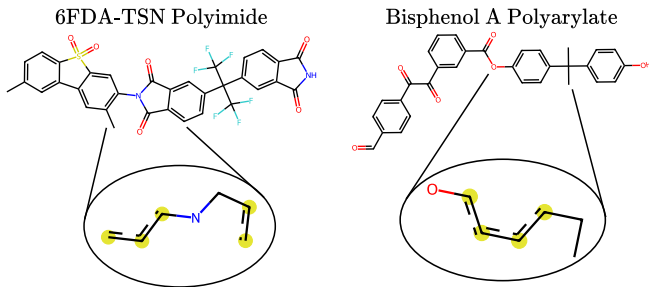


FIG. 1. Example of a bit collision (specifically bit 1062) between two RDKfingerprints (RDKit: Open-source cheminformatics, version 2023.3.1.; [rdkit.org](https://rdkit.org)), using a maximum radius of 3 bonds and a total length of 2048 bits

A sufficiently robust alternative to popular formats like RDKit’s implementation of the Daylight-like fingerprint or Morgan fingerprints is the MACCS key vectorization. These map specific substructures to individual indices (called keys). These provide an important benefit over hash-based methods by precluding bit collisions. Given that the major aim is to interpret individual indices’ contributions to permeability, it is crucial that each index is mapped uniquely to chemical motifs. These keys constitute a binary vector of 166 bits, each encoding a specified chemical characteristic independently, such as aromatic rings or specific branching structures. While this approach removes the occurrence of bit collisions, it comes with a slight downside in that each index has lower descriptive power. Whereas the RDKit or Morgan fingerprint would show exact fragments of the overall molecule, MACCS keys will only show atomic architecture in certain cases, often times leaving intermediate atoms as asterisks to denote “any atom.” It is left to the user to determine which specific portion of the molecule activated that key. Frequencies of each MACCS key within the database, as well as some example keys, are provided in Figure 2. Overall, this representation allows models to be trained faster and in a more interpretable way, ensuring a more consistent foundation for explainable AI and inverse design.

### C. Models

From the database, six gas-specific models were trained for prediction of permeability in a base-10 logarithmic scale, using squared loss as the objective. Additional tri-

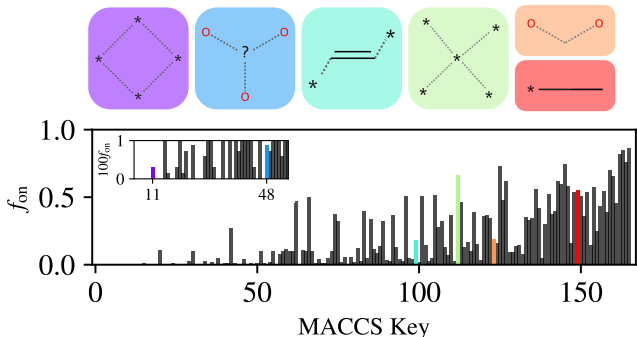


FIG. 2. MACCS Keys examples (top; keys 11, 48, 99, 112, 123 and 149), and the frequency of each key (with the examples highlighted) being “on” in the database ( $f_{\text{on}}$ ). Keys are given in SMARTS notation as in the MACCS master list, with asterisks denoting “any atom”, dashed lines “any bond”, and question marks representing heteroatoms.

als using Mean Average Percentage Error (MAPE) were conducted in hyperparameter tuning, though results were relatively similar. This is likely due to the target values all being in base-10 logarithm scale. Distributions of permeability values for each gas are provided in Figure 3. A wide variety of model architectures were surveyed, with early trials finding success with random forest regression. To maximize model accuracy, these trials were upgraded to gradient-boosted random forest models via XGBoost. All models were created using an 80/20 split for training and testing data, respectively, with the training set itself being split 80/20 for hyperparameter optimization. This subset was shuffled across the training set using 10-fold cross validation, yielding the models summarized below by Table II and the parity plots in Figure 4.

### D. Discussion

Analysis of machine learning models can take many forms. Some models have built-in feature importance tools that measure, for example, the reduction in mean squared error (MSE) associated with a given feature being used as the split criterion in a forest of decision trees, as in scikit-learn’s implementation of Random Forest Regression. The XGBoost toolkit adds more options like the total number of times a feature was used as the split criterion in the entire forest. For the purposes of this study, two model-agnostic techniques were chosen: Shapley analysis and permutation importance. We emphasize that these methods are nontrivial, and any features determined important are not necessarily either the most or least frequent among the database.

Shapley value feature contributions are taken as the average difference between prediction values over multiple coalitions wherein that feature is set to a “background” (non-contributing) value. In the cases where each index of the representation uniquely corresponds to

TABLE II. XGBoost Results, trained on base-10 logarithm values, 10-fold CV. Mean values are provided plus or minus two standard deviations.

Gas	Train $R^2$	Train MSE	Test $R^2$	Test MSE
H <sub>2</sub>	0.939 ± 0.005	0.060 ± 0.005	0.764 ± 0.153	0.226 ± 0.171
He	0.958 ± 0.005	0.034 ± 0.004	0.755 ± 0.214	0.187 ± 0.164
CO <sub>2</sub>	0.889 ± 0.026	0.184 ± 0.046	0.654 ± 0.382	0.519 ± 0.624
O <sub>2</sub>	0.949 ± 0.007	0.082 ± 0.010	0.723 ± 0.190	0.420 ± 0.27
N <sub>2</sub>	0.950 ± 0.004	0.086 ± 0.007	0.706 ± 0.162	0.484 ± 0.196
CH <sub>4</sub>	0.884 ± 0.015	0.257 ± 0.025	0.724 ± 0.310	0.500 ± 0.252

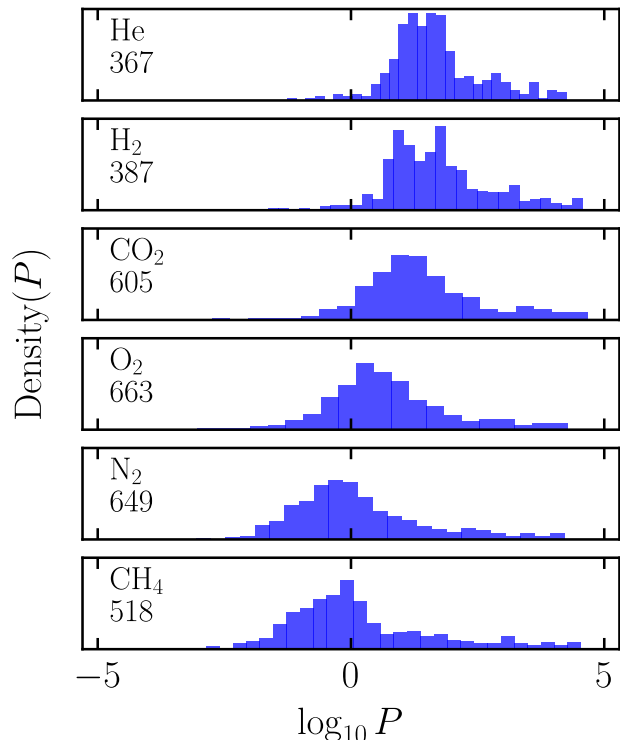


FIG. 3. Permeability distributions for each gas-specific database. Note the log-scaled  $x$ -axis. Total amount of data in that gas’s dataset is shown below the gas type.

a specific chemical substructure, the SHAP approach can be considered akin to group contribution models, which have a well-developed history in the field of polymer science [22, 23]. As an alternative to SHAP values, feature importance can be calculated by permutation of values on a per-feature basis, measuring the impact on model fits. For features vital to the training of a model, permuting the feature value at random should cause large changes in the model’s score. Conversely, unimportant features should have little to no impact. For the database studied herein, the features of the MACCS key vector were shuffled 10 separate times and the average values were taken as their importance by permutation. The top five contributing features for both the Random Forest Regres-

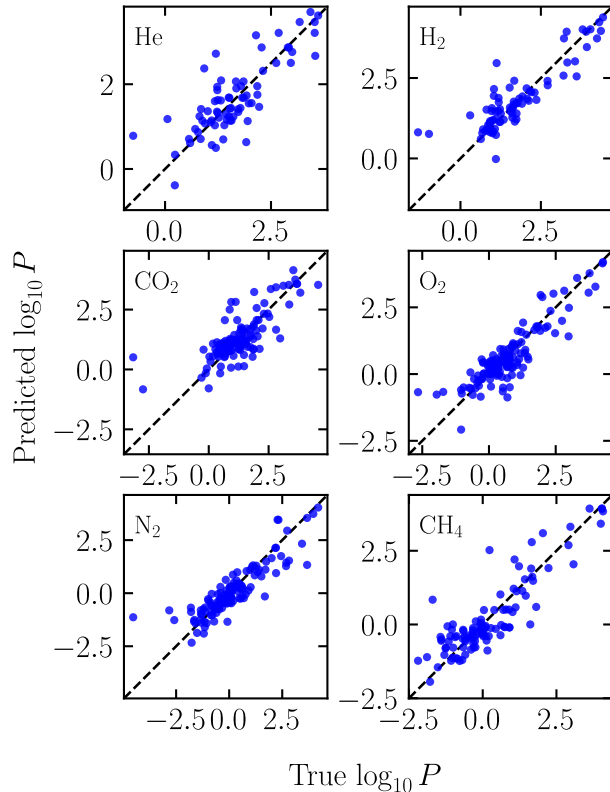


FIG. 4. XGBoost model performance parity plots, evaluated on the testing sets and resolved by gas type.

sor (RFR) and XGBoost models are summarized in Table III, with feature 1 having the highest average SHAP value over the training set. Features with an asterisk are negatively correlated with predicted permeability.

The top two features are largely consistent across the two models and six gases examined. Every model marked MACCS key 99 as the most important, which corresponds to a carbon-carbon double bond, specifically outside of aromatic rings. These findings are drawn from polyacetylene compounds which only constitute 20% of the training set. This emphasizes our statement above that our results are non-trivial and do not simply point to the most (or least) prevalent sample motifs. This re-

TABLE III. Feature importance for different gases and ML models: SHAP RFR/SHAP XGBoost (Permutation Importance RFR/Permutation Importance XGBoost). Features marked with an asterisk correspond to negative correlations with predicted permeability.

Gas	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
He	99/99 (99/99)	141/141 (149/149)	149/149 (112/141)	106/106 (141/162)	112/112 (106/112)
H <sub>2</sub>	99/99 (99/99)	141/141 (141/141)	164/74 (106/74)	106/112 (164/112)	74/164 (74/164)
CO <sub>2</sub>	99/99 (99/99)	141/141 (141/141)	112/112 (112/112)	106/146* (106/82)	159*/106 (82/157)
O <sub>2</sub>	99/99 (99/99)	141/141 (141/82)	106/112 (82/141)	159*/82* (159/112)	112/146* (106/159)
N <sub>2</sub>	99/99 (99/99)	141/159* (141/159)	106/141 (106/141)	159*/106 (159/106)	144/146* (144/74)
CH <sub>4</sub>	99/99 (99/99)	141/141 (141/154)	159*/154* (159/161)	106/161* (106/141)	154*/146* (74/159)

sult, further, is in agreement with molecular sieving principles, which dictate that stiff bonds along the polymer backbone disrupt three-dimensional packing, leading to higher free volumes and, by extension, higher permeability [24]. Keys 149 and 141 also appear frequently, both corresponding to methyl groups. While these features alone do not have an obvious role in gas transport, they tend to appear in conjunction with key 99 frequently, so it may be a case of importance by association. Considering the database as a whole, the probability of having keys 99 and 141 is roughly 20% and 55%, respectively. However, among those polymers with key 141, only about 25-30% have key 99. Conversely, those polymers with key 99 have a 75-80% chance of having key 141 as well. Therefore, we can assume that methyl groups alone do not notably impact membrane permeability, but can lend further contributions when present in conjunction with the polyacetylene backbone. The more likely scenario is based on the fact that the methyl substructures appear alongside branching architectures (keys 112 and 74), implying the methyl groups mostly boost permeability via branching off the main backbone. Steric hindrance has been shown to provide a similar disruption in packing structure [2, 25–28], leading to increased free volume and higher permeability.

While the permutation importance results indicate which features are most influential, they do not point to feature impact on their own. The Shapley values, taken in combination with the permutation importance, can provide a more holistic view of a feature’s impact on the overall membrane permeability. We note that, of the features discussed above, those that stiffen the polymer chains or otherwise disrupt the 3-D packing, can increase the permeability, as expected. The features correlated with lower permeability are mostly flat chain architectures with minimal branching and presence of highly electronegative atoms. For example, features 146 and 159 both correspond to presence of oxygen atoms in different amounts.

Going beyond single groups, we now look in-depth at the top two features for the CO<sub>2</sub> permeation model, keys 99 (carbon-carbon double bond) and 141 (>2 methyl groups) in Figure 5. Polymers with neither feature comprise the curve with the lowest average permeability.

Proceeding to those with either feature in isolation will provide an increase in permeability, with each feature adding a unique amount, but the largest permeability results from those polymers with both features simultaneously. Both of these features have been known to increase permeability by themselves, though the effect of methyl groups in isolation has been somewhat ambiguous, beyond the obvious disruption of 3D chain packing. Our results suggest that methyl groups occurring near carbon-carbon double bonds, rather than methyl groups alone, increase the packing frustration and give rise to higher permeability. Conversely, we note that certain features may interact destructively with regards to permeability. Features 99 and 106 (heteroatom on a single branch) show a different pattern, where the combined effect of both features results in a permeability distribution intermediate between having neither feature and having only one of the two features.

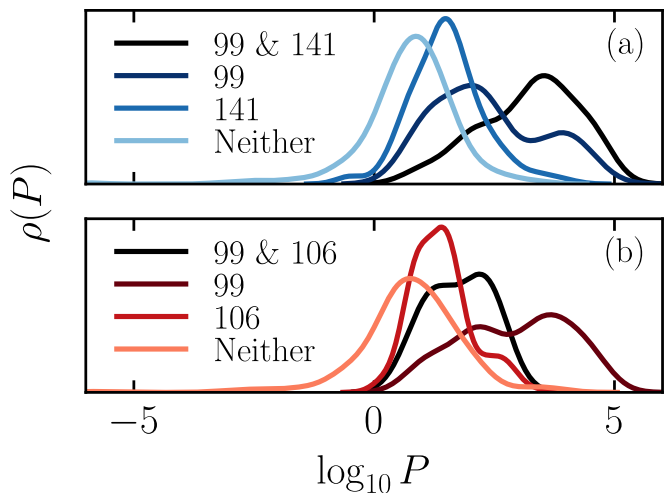


FIG. 5. Distribution of permeabilities for polymers containing neither Key 99 nor 141(106), only Key 99, only Key 141(106) and both Keys 99 and 141(106). Examples of (a) complementary (where both motifs being present enhances permeability) and (b) competitive (where both motifs being present reduces permeability) behavior is shown. Note that key 99 is a C-C double bond, 106 is any atom bonded to three heteroatoms, and 141 is a methyl group.

Representation of data was a central focus for this study. Since Morgan and other hash-based fingerprints adaptively set bit values in the fingerprint based on an instance (each individual molecule), they tend to result in multiple fingerprints having the same index corresponding to multiple substructures. In this scenario, any ML model built for property prediction may perform well enough, but the actual “learning” is questionable. If a particular index of the fingerprint is reported as highly important, it becomes a guessing game as to which instance of that index is the true influence. Here we point to recent work by Yang *et al.* [29] who used the Morgan fingerprint method to understand which groups controlled gas permeation. We arrive at the same conclusion that methyl groups significantly contribute to model predicted permeability, but with important context due to the MACCS key representation. Throughout the top 5 features per model, methyl group keys are nearly always accompanied by at least one branching-related key, usually either key 106 or 112. While Yang *et al.* hypothesizes the same correlation from their found methyl group impact, the Morgan algorithm could have obfuscated the exact instances under which the methyl groups contribute to predicted permeability. Since traditional SHAP analysis ignores feature dependencies, this could have gone unnoticed [9].

For this reason, we chose to encode repeat units as MACCS key vectors, which are robustly 1-to-1 mapping substructures to bits. However, there is another tradeoff in granularity taking this approach. Breaking molecules down to substructures large enough to maintain robust mappings causes a loss in connectivity of those substructures. Thus we know what substructures occur in a monomer but we do not have an unequivocal means of connecting them to form the monomer back. While reverse engineering will thus not give full details of how to assemble the substructures optimally, it will definitively report which substructures matter. Using MACCS keys therefore results in more mechanistic information that would be sufficient for standard genetic algorithms or other inverse design tools. Use of MACCS keys enables safe study of feature importance in the short term and establishes a solid foundation for more tailored inverse design projects down the line. Extending this database to a focus of substructure arrangement would pose one major hurdle. Since MACCS is not capable of tracking arrangements, a higher detail representation would need to be used. The more prescriptive the representation, the more data would be required to maintain or improve model performance. Without resorting to data imputation/infilling or other techniques that artificially fill in missing target data, this is currently a non-trivial task given the available data.

While there is no definitive metric to determine whether a dataset is big or small data, we can assume this database constitutes small data. Barnett *et al.* [6] found that the property prediction models’ mean squared error began to steeply decline after 400 polymers per gas-

specific model. The newly updated database has slightly more than this threshold for most models, the only exceptions being helium and hydrogen. Though this is definitely on the lower limit considering the model’s degrees of freedom, we have essentially exhausted the sources of empirical data. Without resorting to synthetic/augmented data of any sort, there are no clear options to expand the dataset. It is no coincidence that most works on this topic report the same 800-1,000 polymers from experimental papers. With this hard boundary in place, we resort to simpler regression models to mitigate overfitting as much as possible.

### III. CONCLUSION

Our results emphasize the importance of data representation when performing machine learning tasks, as well as the need for physical insight when interpreting models. We studied a minimally curated database of polymer repeat units spanning several decades of membrane research, fitted gradient-boosted random forest models for each gas of interest, and reported the top features correlated to high permeability. Despite a few necessary approximations to account for scarcity in data, the data-driven approach was still possible, with relatively performative models.

We find that using more predictable representations such as MACCS keys, a few substructures appear consistently in these top features. These substructures are unambiguous due to the absence of bit collisions, and agree with physical insight for gas transport through polymer membranes. Carbon-carbon double bonds, as in polyacetylenes, and methyl groups being the nearly unanimous top two features aligns with the expected disruption to 3-dimensional packing of repeat units, leading to higher free volume and permeability. In using MACCS keys, our models also make clear that methyl groups’ importance to permeability revolves around steric hindrance, as evidenced by their tendency to appear in top features alongside carbon-carbon double bonds and branching architectures.

While similar results can be gathered using hash-based fingerprints, the instance-based calculations of the hash may complicate interpretations. As multiple substructures fall under the same index, their impact on permeability may be convoluted with the others. Without rigorous attention to the analysis, it becomes easy to draw tenuous conclusions, further complicating more involved tasks like inverse design. We note that these bit collisions may not significantly impact simpler tasks like property prediction, but become critical issues for interpretability and design. Given that our models were built for exactly this application, the feature importance analysis herein serves as more of an audit of the “learning,” and a confirmation that robust mappings from chemical space to input space guarantee physical fidelity. Further work based on robust mappings like MACCS can be extended to a

wide range of goals, ranging from a relatively simple expansion to selectivity (maximize one permeability while minimizing the other) all the way up to more involved tasks like optimizing arrangement of substructures.

## MATERIALS AND METHODS

We used scikit-learn to preprocess the data and fit random forest regression models as well as XGBoost for further regression models. Model inputs were MACCS key vectors calculated using RDKit, and target values were base-10 logarithm values of permeability measured in Barrers. An 80/20 split was used for training and testing, respectively, with 10-fold cross-validation to obtain more characteristic model performance metrics. The XGBoost regression models were fit using a gridsearch over maximum depth, learning rate, L1 and L2 regularization ( $\alpha$  and  $\lambda$ ), and subsample ratios over training instances. Optimized model parameters can be found in Table IV. Interpretability results were carried out using the RDKit Draw module as well as Scikit-learn’s built-in random forest feature importance and the SHAP module.

## DATA AVAILABILITY

Data will be available at <https://doi.org/10.5061/dryad.5x69p8dbm> or avail-

able from authors upon request.

TABLE IV. Summary of model parameters. All models constructed using gbtree type.

Gas	Learning Rate	Max Depth	$\alpha$	$\lambda$	Subsample Ratio
H <sub>2</sub>	0.1	6	0.8	0.4	0.50
He	0.1	5	0.4	1.0	0.75
CO <sub>2</sub>	0.05	6	1.0	0.8	0.75
O <sub>2</sub>	0.1	6	0.0	1.0	0.50
N <sub>2</sub>	0.1	6	0.4	0.8	0.75
CH <sub>4</sub>	0.1	3	0.2	1.0	1.0

## ACKNOWLEDGEMENTS

Tejus Shastri acknowledges support from the Department of Energy through grant DE-SC-0008772. Sanat Kumar was funded by a grant from the King Abdullah University of Science and Technology. This material is also based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award Number DE-SC-0012704, and by Brookhaven National Laboratory, Laboratory Directed Research and Development grant no. 24-004.

- [1] D. S. Sholl and R. P. Lively, *Nature* **532**, 435 (2016), number: 7600 Publisher: Nature Publishing Group.
- [2] L. M. Robeson, *Journal of Membrane Science* **320**, 390 (2008).
- [3] G. Bradford, J. Lopez, J. Ruza, M. A. Stolberg, R. Osterude, J. A. Johnson, R. Gomez-Bombarelli, and Y. Shao-Horn, *ACS Central Science* **9**, 206 (2023), publisher: American Chemical Society.
- [4] L. Tao, J. Byrnes, V. Varshney, and Y. Li, *iScience* **25**, 104585 (2022).
- [5] A. Arora, T.-S. Lin, N. J. Rebello, S. H. M. Av-Ron, H. Mochigase, and B. D. Olsen, *ACS Macro Letters* **10**, 1339 (2021), publisher: American Chemical Society.
- [6] J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, *Science Advances* **6**, eaaz4301 (2020), publisher: American Association for the Advancement of Science.
- [7] X. Xu, W. Zhao, Y. Hu, L. Wang, J. Lin, H. Qi, and L. Du, *Journal of Materials Chemistry A* **10.1039/D2TA09272G** (2023), publisher: The Royal Society of Chemistry.
- [8] G. P. Wellawatte, A. Seshadri, and A. D. White, *Chemical Science* **13**, 3697 (2022), publisher: Royal Society of Chemistry.
- [9] H. A. Gandhi and A. D. White, *Explaining molecular properties with natural language* (2022).
- [10] H. Gao, S. Zhong, R. Dangayach, and Y. Chen, *Environmental Science & Technology* **10.1021/acs.est.2c05404** (2023), publisher: American Chemical Society.
- [11] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, *Nature Machine Intelligence* **2**, 56 (2020), number: 1 Publisher: Nature Publishing Group.
- [12] L. Breiman, *Machine Learning* **45**, 5 (2001).
- [13] E. R. Antoniuk, P. Li, B. Kailkhura, and A. M. Hiszpanski, *Journal of Chemical Information and Modeling* **62**, 5435 (2022), publisher: American Chemical Society.
- [14] T.-S. Lin, C. W. Coley, H. Mochigase, H. K. Beech, W. Wang, Z. Wang, E. Woods, S. L. Craig, J. A. Johnson, J. A. Kalow, K. F. Jensen, and B. D. Olsen, *ACS Central Science* **5**, 1523 (2019), publisher: American Chemical Society.
- [15] J. Liu, S. Zhang, D.-e. Jiang, C. M. Doherty, A. J. Hill, C. Cheng, H. B. Park, and H. Lin, *Joule* **3**, 1881 (2019).
- [16] J. Liu, G. Zhang, K. Clark, and H. Lin, *ACS applied materials & interfaces* **11**, 10933 (2019).
- [17] L. Hu, S. Pal, H. Nguyen, V. Bui, and H. Lin, *Journal of Polymer Science* **58**, 2467 (2020).
- [18] L. Huang, W. Guo, H. Mondal, S. Schaefer, T. N. Tran, S. Fan, Y. Ding, and H. Lin, *Macromolecules* **55**, 382 (2021).
- [19] G. Zhang, T. N. Tran, L. Huang, E. Deng, A. Blevins, W. Guo, Y. Ding, and H. Lin, *Journal of Membrane Sci-*

- ence **644**, 120184 (2022).
- [20] D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling* **50**, 742 (2010), publisher: American Chemical Society.
- [21] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, *Methods Virtual Screening*, **71**, 58 (2015).
- [22] L. M. Robeson, C. D. Smith, and M. Langsam, *Journal of Membrane Science* **132**, 33 (1997).
- [23] D. W. v. Krevelen† and K. t. Nijenhuis, *Properties of Polymers: Their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions* (Elsevier, 2009) google-Books-ID: bzRKwjZeQ2kC.
- [24] L. M. Robeson, B. D. Freeman, D. R. Paul, and B. W. Rowe, *Journal of Membrane Science* **341**, 178 (2009).
- [25] H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech, and B. D. Freeman, *Science* **356**, eaab0530 (2017), publisher: American Association for the Advancement of Science.
- [26] S. Matteucci, Y. Yampolskii, B. D. Freeman, and I. Pinnau, in *Materials Science of Membranes for Gas and Vapor Separation* (John Wiley & Sons, Ltd, 2006) pp. 1–47, section: 1 .eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/047002903X.ch1>.
- [27] B. D. Freeman and I. Pinnau, in *Polymer Membranes for Gas and Vapor Separation*, ACS Symposium Series, Vol. 733 (American Chemical Society, 1999) pp. 1–27, section: 1.
- [28] B. D. Freeman, *Macromolecules* **32**, 375 (1999), publisher: American Chemical Society.
- [29] J. Yang, L. Tao, J. He, J. R. McCutcheon, and Y. Li, *Science Advances* **8**, eabn9545 (2022), publisher: American Association for the Advancement of Science.