npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01681-8

Polymer design for solvent separations by integrating simulations, experiments and known physics via machine learning

Janhavi Nistane[1], Rohan Datta[2], Young Joo Lee[2], Harikrishna Sahu[1], Seung Soon Jang[1], Ryan Lively[2] & Rampi Ramprasad[1] ✉

This study guides the discovery of sustainable high-performance polymer membranes for organic binary solvent separations. We focus on solvent diffusivity in polymers, a key factor in quantifying solvent transport. Traditional experimental and computational methods for determining diffusivity are time- and resource-intensive, while current machine learning (ML) models often lack accuracy outside their training domains. To overcome this, we fuse experimental and simulated diffusivity data to train physics-enforced multi-task ML models, achieving more robust predictions in unseen chemical spaces and outperforming single-task models in data-limited scenarios. Next, we address the challenge of identifying optimal membranes for a model toluene-heptane separation, identifying polyvinyl chloride (PVC) as the optimal membrane among 13,000 polymers, consistent with literature findings, thereby validating our methodology. Expanding our search, we screen 1 million publicly available and 7 million chemically recyclable polymers, identifying greener halogen-free alternatives to PVC. This capability is expected to advance membrane design for solvent separations.

Separating organic solvents is essential in the chemical industry for producing fuels, chemicals, and other derived products. A notable example is the separation of aromatic compounds, such as toluene, from aliphatic compounds like n-heptane, which is vital for producing cleaner motor fuels with reduced aromatic content[1,2]. However, traditional separation methods such as extractive distillation or liquid-liquid extraction face significant challenges in separating such mixtures due to the close boiling points and similar physio-chemical properties of the solvent components[3]. Pervaporation (PV) presents a promising alternative by utilizing differences in the permeation rates of organic solvent molecules across the membrane for separation rather than relative volatility alone[4]. PV-based approaches also offer advantages such as enhanced safety, cost-efficiency, and reduced energy consumption compared to distillation-based methods[5–10]. In such separations, a liquid phase solvent mixture is introduced to one side of a polymer membrane, while the permeate exits from the opposite side in a vapor phase. Although polymeric membranes are widely used for such separations due to their low cost of fabrication and ease of scaling up[11–13], finding a suitable polymer that will achieve these separations effectively remains a challenge.

For successful solvent separations, one of the key membrane properties is solvent permeability, defined as the flux normalized by the membrane thickness and the driving force. The mass transport in the PV process follows the solution-diffusion mechanism, which states that the permeability P (Barrer) through a dense membrane is the product of diffusivity D (cm^2/s) and solubility coefficient S ($cc(STP).cc\ polymer^{-1}\ cmHg^{-1}$)[14].

$$P = D \cdot S \tag{1}$$

Knowledge of pure component diffusivity and sorption isotherms is essential for predicting membrane separation performance for complex mixtures[15]. Focusing on solvent diffusivity, we find that experimental data, typically obtained through gravimetric sorption or time-lag measurements, is limited and resource-intensive to expand[16,17]. Classical molecular dynamics (MD) simulations, while effective for calculating diffusivity, are constrained by high computational costs and the need for accurate force-field parameters[18–21]. In contrast, machine learning (ML)-driven methods offer a promising route allowing for rapid and accurate diffusivity predictions while effectively addressing the resource and scalability limitations of conventional experimental and computational techniques. Such data-driven methods have achieved remarkable strides in recent years, fundamentally reshaping the landscape of materials property predictions and the tailored design of materials with specific target characteristics[22–28]. While many ML models exist for predicting gas[29–32] and ionic diffusivity through polymers[33,34], as well as properties like polymer-solvent interaction

[1]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [2]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA. ✉e-mail: rampi.ramprasad@mse.gatech.edu

parameters[35,36], miscibility[36], swelling[37], and fractional free volume[38], models specifically for organic solvent diffusivity through polymers are still limited. In our recent work, we developed an ML model to predict solvent diffusion in polymers, which was used in conjunction with mass transport simulations to predict complex multi-component crude oil solvent permeation[15]. Despite these advances, current ML models for diffusivity encounter significant difficulties in extrapolating beyond the polymer-property space encompassed by their training data. This underscores a well-known limitation of ML models: their inability to "generalize" or reliably predict outcomes outside the training set. Therefore, the development of more robust and generalizable ML models is essential to ensure accurate, physically meaningful predictions as we explore new chemical spaces for membrane design.

In this work, we propose a two-fold methodology for enhancing the generalizability of current diffusivity ML models: multi-task (MT) learning and physics-enforced learning. In this context, MT learning involves training a model on multiple tasks, such as simultaneously learning from experimental and simulated data sources. While experiments provide the accurate ground truth, these datasets are limited and grow slowly. At the same time, lower-accuracy computational data can be generated at scale, exploring new chemical spaces beyond the reach of experimental datasets. Multi-task learning leverages the correlations between scarcely available but high-fidelity experimental data and the diverse but lower-fidelity computational data, thereby improving model generalizability, as recently demonstrated[29,39,40]. Thus, recognizing the power of multi-task learning, we augmented experimental diffusivity data with our in-house generated computational data and trained multi-task diffusivity models to improve prediction accuracy and generalizability.

To further enhance predictive performance, we leverage physics-enforced neural networks (PENN). While data-driven models can effectively fit observed data, their predictions may sometimes be physically inconsistent, especially when extrapolating beyond the training data, leading to poor generalization. To address this, it is crucial to incorporate fundamental physical laws and domain expertise into ML models. The wealth of physical and empirical observations, formulas, and axioms in literature can be exploited to provide informative priors to enhance the predictions of current ML models. Physics-informed machine learning leverages this prior knowledge, enabling more accurate predictions while increasing model interpretability[41–43], for applications such as material property prediction and design[44–47]. For example, Bradford et al. enhanced the accuracy of an ML model for predicting temperature-dependent ionic conductivity by incorporating the Arrhenius law, thereby improving the model's performance compared to those without such physics-based integration[46]. Building on these successes, we present a "physics-enforced" machine learning approach designed to improve generalizability, especially in data-limited scenarios. First, we use an empirical power law, developed in our prior work[15], to encode known correlations between solvent molar volume and diffusivity, capturing the slower diffusion of bulkier molecules and enabling accurate predictions for large solvents not included in the training set[48,49]. Second, we apply an Arrhenius-based relationship to model solvent diffusivity as a function of temperature, enabling extrapolation to higher temperatures common in industrial separations. We systematically compared and demonstrated the superiority of these physics-based multi-task models against single-task models that rely solely on limited experimental data.

Further, harnessing our generalizable ML diffusivity model, we addressed a key challenge in membrane discovery: identifying suitable membranes for specific solvent separations. Effective separations require not only high solvent permeability but also high permselectivity, which ensures the selective transport of the target species through the membrane. Ideal permselectivty is defined as:

$$\alpha_{AB} = \frac{P_A}{P_B} = \frac{D_A}{D_B} \cdot \frac{S_A}{S_B} \tag{2}$$

Here, P_A and P_B are permeabilities, D_A and D_B are diffusivities and S_A and S_B are solubility coefficients for pure solvents A and B. However, identifying membranes that possess both high solvent permeability and permselectivity remains challenging due to the inherent trade-off between these properties. In the gas separation community, researchers guide membrane discovery by maximizing permeability and permselectivity, typically illustrated through trade-off or Robeson plots, and continually seek new polymers that exceed established upper bounds for these properties[50–52]. This study addresses the gap in identifying optimal membranes for binary organic solvent separations by generating ML-predicted trade-off plots for 13,000 polymers, with a focus on toluene-heptane separations crucial for producing cleaner, aromatic-lean fuels. Using the solution-diffusion model (Equation: (1)), we estimated permeability and ideal permselectivity from ML-predicted diffusivity and solubility coefficients[15]. Notably, our ML predictions identify polyvinyl chloride (PVC) as an optimal choice polymer for this separation, aligning with its recognition as a membrane of choice in literature[53–56]. Despite PVC's notable separation performance, it is widely regarded as one of the most environmentally harmful plastics due to both its environmental impact and the fact that its monomer, vinyl chloride, is a potent carcinogen. Hence, we next aimed to find more sustainable non-halogenated alternatives to PVC. The initial screening of the 13,000 known polymers failed to identify any suitable PVC alternatives. To expand our search for sustainable and high-performance polymer alternatives, we screened across a space of virtually generated candidates. First, we leveraged an open-source dataset of 1 million candidates, "PI1M," produced by Luo et al. using a generative ML model trained on SMILES strings of existing polymers in PolyInfo database[57]. Additionally, we utilized a database of 7 million chemically recyclable ring-opening polymerization (ROP) polymers, previously created by Kern et al. by utilizing known monomers[58]. Leveraging these datasets of synthetically accessible and chemically recyclable polymers along with 13,000 known polymers, we propose halogen-free alternatives to PVC for toluene-heptane separation.

Results

Generation and validation of a computational dataset for solvent diffusivity

We established a high-throughput simulation protocol for calculating solvent diffusivity through polymers, employing classical molecular dynamics (MD) with the open-source LAMMPS package, as detailed in Fig. 1. First, polymer and solvent structures were generated using the open-source Polymer Structure Predictor (PSP)[59]. The polymer chains consist of approximately 150 atoms per chain, with the entire system totaling 4000–5000 atoms, a majority of them operating within a dilute solvent concentration regime, as can be seen in the Supplementary Fig. 1. GAFF2 was employed as the chosen force field. Next, a 21-step equilibration process was employed to ensure that polymers were equilibrated[60]. Afterward, all systems were subjected to an additional 10 ns equilibration in the NPT ensemble, followed by a 200 ns production run in the NVT ensemble[29]. The Nosé-Hoover thermostat and barostat, with a damping parameter of 100-time steps each, were employed, and a time step of 1 fs was used in all MD simulations. Post-simulation, the diffusivity of the solvent in the Fickian regime was estimated based on mean square displacement analysis for all atoms and averaged for that molecule. Thus the solvent diffusivity (D) is then calculated as:

$$D = \frac{1}{6N} \lim_{t\to\infty} \frac{d}{dt} \sum_{i=1}^{N_{gas}} \langle (r_i(t) - r_i(0))^2 \rangle \tag{3}$$

where N is the number of solvent molecules, t is the simulation time, $r_i(t)$ is the position vector of the solvent at time t, and $r_i(0)$ is the position vector at the initial time 0.

Since the diffusion coefficient D is ideally determined in the limit as t approaches infinity, simulations were run for sufficiently long periods to ensure that the diffusion properties do not vary significantly over time. We ensured that diffusivities are estimated in the Fickian regime by verifying that the slope of the log displacement-log time plot remains close to one. If

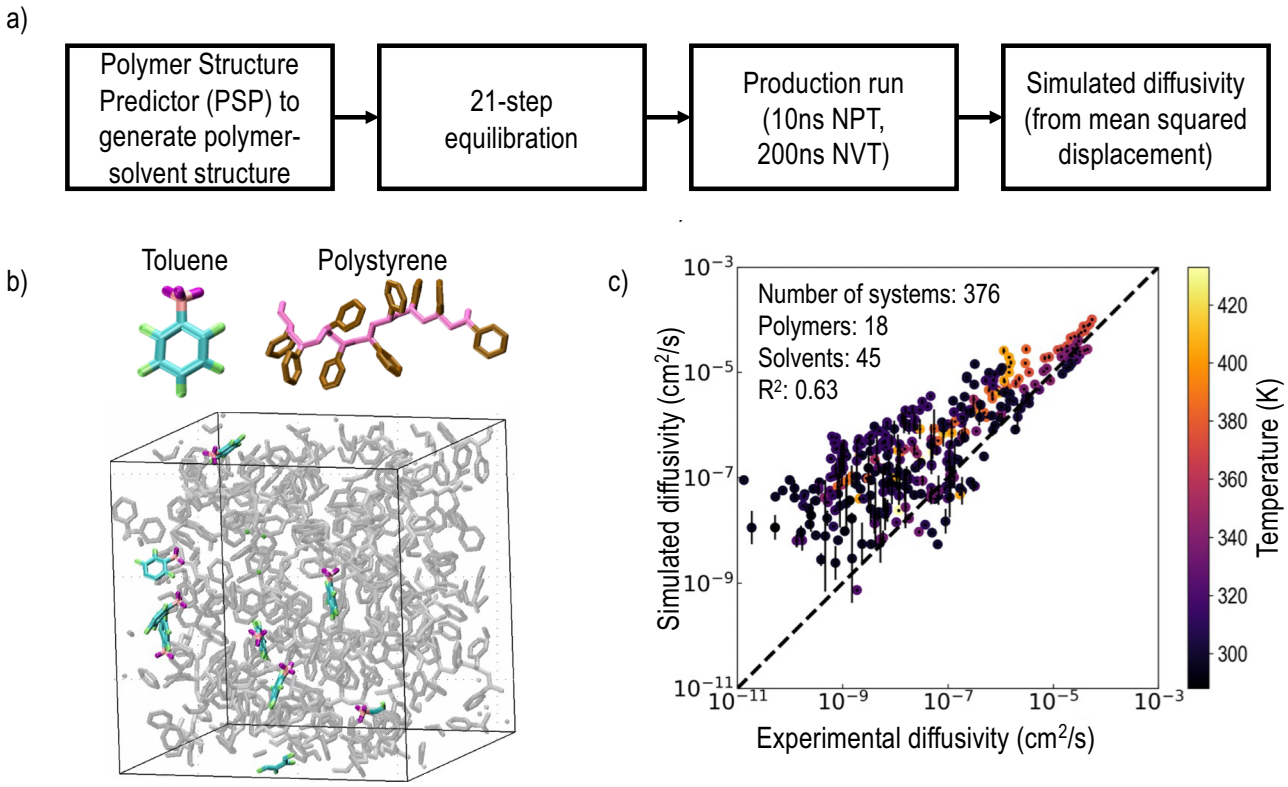

Fig. 1 | Simulation data generation and validation. a The simulation protocol is used to calculate the diffusivity of solvents within polymers. **b** A snapshot of a system used to measure polymer-solvent diffusivity. The diffusion of toluene molecules (colored) within polystyrene (grey) is observed through mean square displacement analysis. **c** Correlation between simulated and experimental diffusivity data:[62–75] The dashed black line represents the parity lines of optimal fit. The error bars represent the standard deviations obtained from diffusivity simulations.

the slope value is outside the range of 0.95-1.05, the diffusivity runs are excluded from subsequent calculations. Additionally, the standard block average method was used to estimate the simulated uncertainty from 10 blocks[61]. The diffusivity simulations were conducted as a function of solvent concentration and temperature. Using the simulation pipeline outlined above, diffusivity calculations were first conducted for polymer-solvent pairs with available experimental diffusivity data, as shown in Fig. 1. For 376 experimental systems across 18 polymers and 45 solvents, a coefficient of determination (R^2) of 0.63 was observed, indicating an acceptable correlation between experimental and simulated data[62–75]. Although the simulated results do not fully align with the experimental data due to inherent limitations in the force field and simulation protocol, the observed qualitative correlation is expected and sufficient for our objective of downstream multi-task machine learning models. After validating the correlation between experimental and simulated values for the 376 systems, the computational diffusivity dataset was expanded beyond the limited experimental space, resulting in a total of 623 systems, comprising 91 polymers and 69 solvents. This expansion of the simulated space is aimed to enhance the generalizability of the downstream multi-task ML model by broadening chemical space coverage, enabling it to learn from a more diverse dataset and make more accurate predictions beyond the constraints of the limited experimental data[29,39,40]. Additional details on the simulation protocol can be found in the Supplementary Information (sections 2 and 3).

Data augmentation

Building larger and more diverse datasets is essential for developing more effective models[76]. Hence, data augmentation is crucial for creating robust models in materials informatics, where data scarcity is a common challenge. The experimental diffusivity dataset was expanded in two key ways:

- **Activity-to-Concentration Conversion**: Previously, the literature-sourced experimental diffusivity data was recorded as a function of solvent activity[15]. In this work, additional data was collected, where diffusivity was recorded as a function of solvent concentration in the membrane[62–75], as visualized in Fig. 2. For the purpose of expanding the concentration-dependent datasets, we converted experimental activity-dependent data to concentration-dependent data using the Sorption Uptake ML model, as described in Section Sorption Uptake ML model. This step is essential for achieving consistency in datasets and units, especially when integrating with simulated data (recorded as a function of solvent concentration), as discussed in the following section.
- **Augmentation of Simulation Data**: The simulated diffusivity dataset, recorded as a function of solvent concentration in the membrane, was fused with the converted experimental concentration-dependent dataset using one-hot encoding, thereby enhancing the diversity of the data.

Together, these augmented datasets created a comprehensive, concentration- and temperature-dependent diffusivity dataset, as illustrated in Fig. 2, which was used to train the diffusivity ML model. Additional data analysis is provided in Supplementary Fig. 1.

ML model benchmarks for improved generalizability

To assess the impact of data fusion and physics-guided learning, we trained and compared both single-task and multi-task diffusivity models across various algorithms. The feature space incorporated Polymer Genome-derived polymer-solvent fingerprints[27,77] (described in Section Polymer and solvent fingerprinting), along with key descriptors such as solvent concentration (expressed as the weight fraction on a logarithmic scale) and

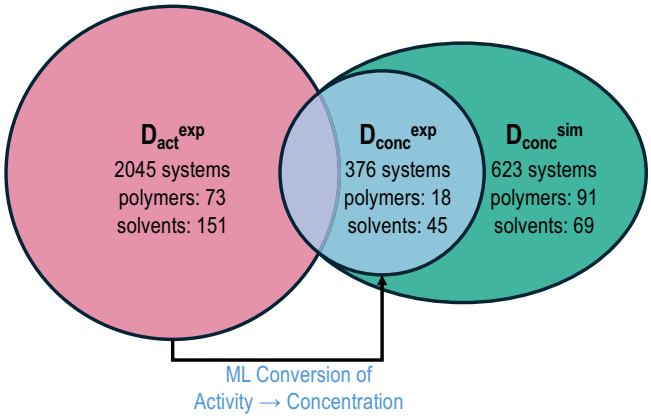

Fig. 2 | Expansion of the diffusivity dataset. The pink segment represents the experimental activity-dependent data, labeled as D_{act}^{exp}, which is converted to a concentration-dependent format using the Sorption Uptake ML model[15]. The blue circle denotes the experimental concentration-dependent data, D_{conc}^{exp} [62–75], while the green segment, D_{conc}^{sim}, includes concentration-dependent simulated data. These datasets collectively form a comprehensive concentration-dependent diffusivity dataset, comprising 3044 systems, 154 polymers, and 176 solvents, which serves as the foundation for developing the production diffusivity model.

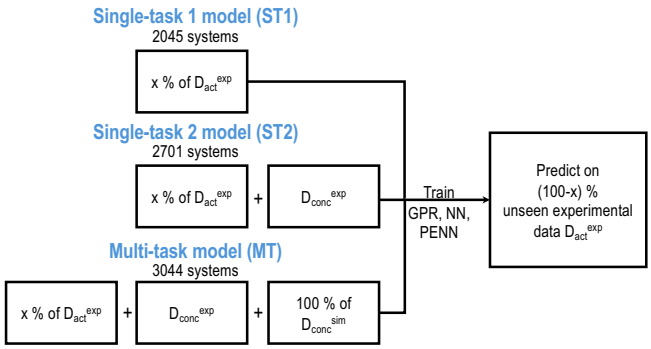

Fig. 3 | Models for the benchmarking. Single-task models (ST1 and ST2) were distinguished by the experimental diffusivity data they used for training- ST1 model is trained on x% D_{act}^{exp}, whereas the ST2 model is trained on x% $D_{act}^{exp} + D_{conc}^{exp}$. Here, x represents the percentage of the training set derived from D_{act}^{exp}. To ensure fair evaluation, any polymer in the test set is completely excluded from the training set. In contrast, multi-task models (MT) incorporated both experimental and simulated diffusivity data in their training process, consisting of x%$D_{act}^{exp} + D_{conc}^{exp} + D_{conc}^{sim}$. These models are trained using algorithms such as Gaussian Process Regression (GPR), Neural Networks (NN), and Physics enforced Neural Networks (PENN) models. The performance of these trained models is evaluated on the holdout (100-x)% D_{act}^{exp} data, with the best-performing models selected as the final diffusivity models for production.

temperature. Additionally, one-hot encoded selector vectors were used to distinguish between experimental and simulated data when both were present. Single-task models, trained solely on experimental data, were categorized as ST1 and ST2-based on the amount of data available, as shown in Fig. 3. In contrast, multi-task models combined experimental and simulated data through data fusion. By training on multiple correlated tasks, multi-task learning enables the models to recognize correlations between accurate experimental data and diverse simulated data, expanding prediction capabilities across a broader chemical space. We implemented these models using algorithms such as Gaussian Process Regression (GPR), Neural Networks (NN), and Physics-Enforced Neural Networks (PENN). While GPR and NN modes do not incorporate physical laws into training, the Physics-Enforced Neural Networks (PENN) models incorporate fundamental physical laws into the training process, ensuring that the model reproduces known physical behaviors. Two PENN models were developed. The PENN-1 model is based on an empirically observed relationship where the diffusivity of a solvent decreases with increasing molar volume due to its bulky nature[48,49]. This empirical law was initially designed in our previous work to make more accurate predictions for the large solvents (molar volume greater than 1000 cm^3/mol) found in real-world crude oil applications, which were absent from the more limited literature-based training dataset that was predominantly distributed around 250 cm^3/mol[15]. On the other hand, the Arrhenius equation is used to describe the temperature dependence of diffusivity in the PENN-2 model. By embedding the Arrhenius relationship[78,79] within the neural network, we propose improving its predictive accuracy, especially at temperatures outside the range of the training data.

More details about the model architecture are listed in Section Solvent diffusivity ML models.

The model's performance is evaluated using the Order of Magnitude Error (OME), which is essentially the Mean Absolute Error (MAE) computed on a logarithmic scale[15,29,44]. OME is expressed as :

$$\mathrm{OME} = \frac{1}{n}\sum_{i=1}^{n}|\log_{10}(y_i) - \log_{10}(\hat{y}_i)|$$

where y_i and $\hat{y}_i$ represent the actual and predicted values, respectively, and n is the number of data points. The test OME is calculated across various training set sizes using polymer-group splits. These polymer-based "group" splits ensure that test polymers are entirely excluded from training. To ensure statistical reliability, data was split into test-train sets using five different random seeds, with performance statistics reported in Fig. 4, comparing single-task models (ST1 and ST2) and multi-task models trained using GPR, NN, PENN architectures. As expected, test errors decreased as training size increased, eventually plateauing as the models approached their optimal performance.

First, we evaluate the performance of the models as a function of the training data by comparing the performance of single-task (ST1, ST2) and multi-task (MT) models. The ST1 model, trained on the smallest experimental dataset of 2045 polymer-solvent systems, showed the highest test error as a function of training set size (depicted in red in Fig. 4). The ST1 model serves as a baseline to assess the performance of more advanced models. The ST2 model, trained on 2421 experimental systems, was expected to show only modest performance improvements. As shown in green, ST2 models performed similarly to ST1, with slight gains in data-scarce scenarios (10% and 30% trainset sizes). More significant improvements were seen at trainset sizes of 50% and 70%, though the error plateaued at 70%, suggesting that the model was nearing optimal performance, with diminishing returns from additional data. In contrast, the multi-task diffusivity models (shown in blue), trained on a comprehensive dataset of 3,044 systems (combining both experimental and simulated data), significantly outperformed the single-task models. This was especially evident in data-scarce scenarios with as little as 10% trainset data. The multi-task model's enhanced generalization capabilities stem from leveraging diverse data and learning relationships between experimental and simulated data, underscoring the effectiveness of multi-task learning in improving predictive accuracy in data-limited environments.

Next, we analyzed the effect of different machine-learning algorithms on model performance. In the multi-task (MT) framework, within the data-scarce regime of 10% trainset size, Gaussian Process Regression (GPR) achieved the lowest averaged OME error (0.68), followed by PENN-2 (0.815), with PENN-1 (1.15) and the neural network (NN) model (1.12) showing comparable performance. This indicates that GPR models excel when training data is limited. Furthermore, we note that the limited performance of the PENN2 models in the low-data regime may be attributed to the insufficient coverage of the temperature range in the dataset, which constrains the model's ability to generalize effectively. Additionally, the complexity of the neural network architecture may contribute to this issue, as more sophisticated models typically require larger datasets to achieve

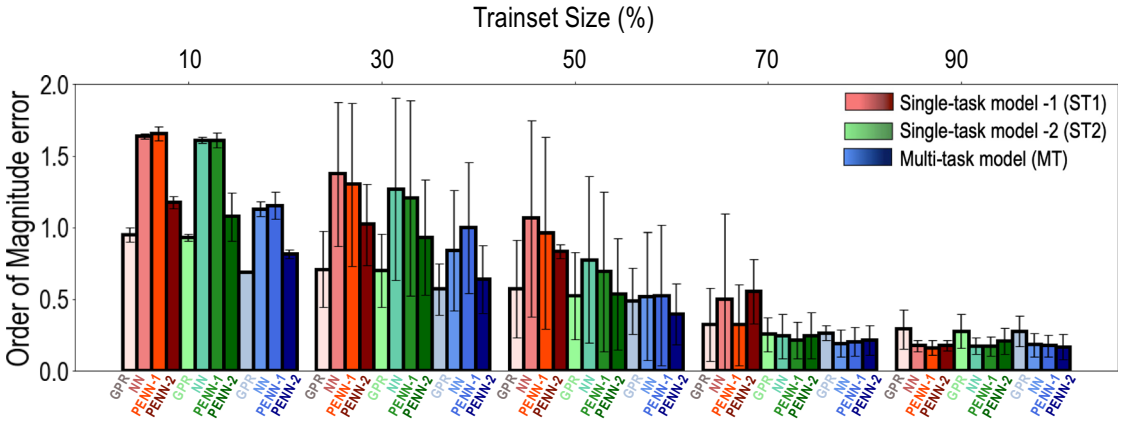

Fig. 4 | ML model benchmarking for accuracy and generalizability. This figure presents the test error (OME) as a function of the training set size (corresponding to x% of D_{act}^{exp}), averaged over five runs. MT models outperformed single-task (ST1 and ST2), especially with limited data. Physics-enforced PENN-2 models showed the lowest errors as the trainset size grew. MT models with physics-enforced learning improved robustness and generalizability.

robust generalization. This observation is consistent with previous studies, which have frequently linked performance limitations in such settings to overfitting[80–82]. However, as more data became available, particularly at 90% trainset size, neural networks significantly outperformed GPR, with errors of 0.179 for NN, 0.173 for PENN-1, 0.164 for PENN-2, and 0.27 for GPR. Notably, as trainset size increased, PENN-2 consistently outperformed other models, especially evident at 50% trainset size. While NN and PENN-1 performed competitively, PENN-2 demonstrated superior predictive power. As discussed in Section Solvent diffusivity ML models, the empirical solvent volume law (equation (4)) incorporated in PENN-1 is particularly effective for large molar volume solvents. However, due to the absence of these larger solvent molecules in the current test dataset (as direct diffusivity measurements were unavailable), the impact of this empirical law was diminished. Consequently, even with the inclusion of physics-based laws, PENN-1 did not reach the performance level of PENN-2. In summary, in data-scarce conditions, the model performance rankings were GPR > PENN-2 > PENN-1 ~ NN. As the training set size increased, this ranking shifted to PENN-2 > PENN-1 ~ NN > GPR. The trends observed in the MT model were similarly reflected in both the ST1 and ST2 models, exhibiting only minor deviations attributable to variability in the train-test splits.

Ultimately, this analysis illustrated that integrating multi-task models with physics-based approaches results in more accurate, robust, and generalizable models, even in scenarios with limited data. Consequently, our production-level diffusivity model was designed as a multi-task model built on a physics-enforced neural network architecture with Arrhenius temperature-dependence encoded (MT-PENN-2).

Comparative production level benchmark

In our previous work[15], we developed a physics-enforced single-task ML model (ST1-PENN-1), using the experimental dataset D_{act}^{exp}, which covers 2045 polymer-solvent systems. This model predicts polymer-solvent diffusivity based on polymer-solvent fingerprints and solvent activity. In our current work, we expanded this dataset by incorporating additional experimental and simulated data, resulting in a larger dataset of 3044 systems and trained multi-task models using physics-enforced methods (MT-PENN-2). The goal of this updated diffusivity model was to improve generalizability and robustness, enabling accurate predictions in polymer spaces previously unexplored by experiments. In this analysis, we compare the original model (ST1-PENN-1) with the updated model (MT-PENN-2).

Enhanced accuracy from multi-task and physics-enforced learning approaches. As shown in Fig. 4, MT-PENN-2 outperformed ST1-PENN-1, demonstrating superior prediction accuracy across a broader range of polymer chemistries. This success can be attributed to the expanded chemical space from a more diverse dataset (discussed in the following section) and the improved model architecture. It is important to note that ST1-PENN-1 was specifically designed using the PENN-1 architecture to predict diffusivity in large crude oil solvents. While PENN-1 is optimal for such predictions, the PENN-2 architecture excels in predicting solvent diffusivity for smaller organic solvents and new polymer chemistries. However, we note that since the original model of ST-PENN1 used the dataset D_{act}^{exp}, and this work converted it to a concentration-dependent format D_{conc}^{exp}, this is not a perfect comparison. This conversion was necessary to ensure consistent units for comparison with other models developed in this study. A direct comparison of models based on activity- or concentration-dependent data could only be made using a test set that overlapped these datasets, which was minimal (only 7 polymers and 84 data points), and this small representation of simple polymers (such as polyethylene) does not accurately reflect the generalizability of the models in the broader chemical space. As a result, we conducted an indirect comparison, and the findings clearly showed that the updated MT-PENN-2 models outperform the original model in prediction accuracy.

Enhanced generalizability due to expanded polymer chemical space. The principal component analysis (PCA) plot using Polymer Genome fingerprints in Fig. 5 demonstrates the broader polymer chemical space covered by the updated diffusivity model (MT-PENN-2) compared to the original model (ST1-PENN-1). The polymer chemical fingerprints are reduced to two dimensions using PCA, capturing the maximum variance and enabling a structured visual representation of the chemical space, wherein each marker represents a unique polymer. The gray markers represent the PCA projection of the fingerprints of a 13k known polymer space. The yellow markers represent the polymer representation for the "original" diffusivity model (ST1-PENN-1) built in our previous work, which was trained solely on experimental data D_{act}^{exp} [15]. In contrast, the red markers denoting the "updated" diffusivity model (MT-PENN-2) refers to the polymers in the more comprehensive dataset, incorporating $D_{act}^{exp} + D_{conc}^{exp} + D_{conc}^{sim}$, thus explaining the overlap and expansion in the chemical space.

Overcoming limitations in temperature-dependent-diffusivity predictions. The updated diffusivity model (MT-PENN-2) overcomes a key limitation of the original model by capturing temperature dependence. While the original model was limited to predicting diffusivity near room temperature, the updated model incorporates an Arrhenius-based temperature law, enabling it to extrapolate diffusivity behavior even when temperature-dependent diffusivity data is sparse. This is especially important since most available experimental data is concentrated around room temperature, while many separation applications occur at elevated

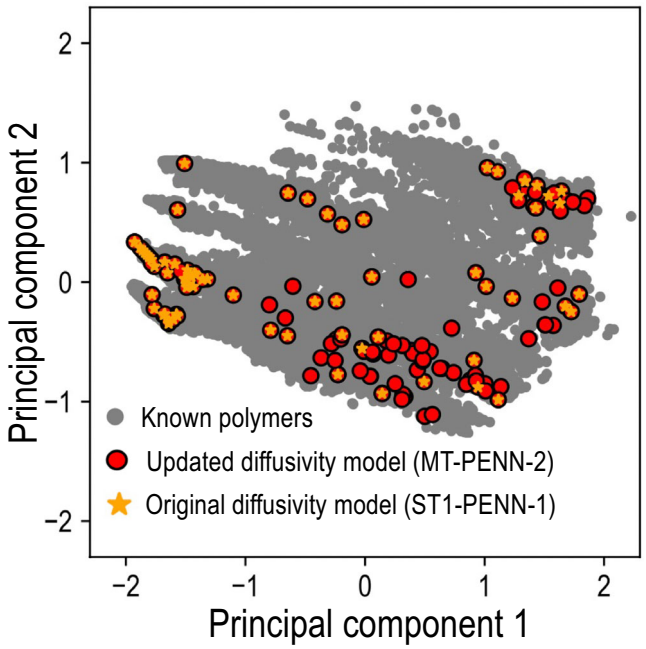

Fig. 5 | Expansion of the chemical space. The Principal Component Analysis (PCA) plot shows the expanded polymer chemical space covered by the updated diffusivity model (MT-PENN-2) compared to the original (ST1-PENN-1)[15]. Grey points represent the 13k known polymers in our database, yellow stars mark the polymers covered by the original model, and red points indicate the broader space covered by the updated model.

temperatures. Further details on this improvement have been visualized in Supplementary Fig. 6.

Design guidance for high-performance sustainable membranes

We now aim to identify the ideal pervaporation polymer membrane for separating toluene and heptane by maximizing permeability and ideal permselectivity. This separation is crucial for producing clean motor fuels. The selection of this solvent separation was additionally driven by the availability of extensive training data for this combination to ensure robust and reliable predictions. Toluene-heptane separation is particularly challenging because their physical differences, such as kinetic radius (toluene: 5.9 Å; heptane: 4.42 Å) and boiling point (toluene: 110.6 °C; heptane: 98.4 °C)—are relatively small[83,84]. As a result, membrane separations typically depend on exploiting differences in chemical structure and interactions, notably the aromatic nature of toluene versus the aliphatic character of heptane. To address this challenge and guide membrane design, we calculated single-component permeability as the product of ML-predicted diffusivity and solubility coefficients, along with ideal permselectivity as explained in Section Construction of ML based solvent-trade-off plots. This work presents the largest solvent trade-off plots to date, encompassing 13,000 known polymers, along with virtually generated 1 million publicly available polymers in the P1IM database[57] and 7 million chemically recyclable ring-opening polymerization (ROP) based candidates[58]. Details for this virtual polymer search space can be found in Section Membrane Design Search Space.

Rediscovery of known high-performance candidates. Focusing first on the known polymer space as shown in Fig. 6a (red data points), we present permeability-based trade-off plots, where halogenated polymers (like polyvinyl chloride PVC, denoted by A) emerged as top candidates, exhibiting both high predicted permeability and ideal permselectivity. The diffusivity-based trade-off plot in Fig. 6b, showed similar trends. Figure 6c focuses on solubility, where nitrogen-containing bulky aromatic compounds displayed high solubility and selectivity, likely due to their strong affinity for toluene.

Validating ML-generated trade-off plots is challenging due to limited benchmarking methods. PVC membranes and their composites are recognized for toluene-heptane separations via pervaporation[53–56]. In agreement with literature, our ML predictions independently identified PVC as a top-performing polymer, with a predicted toluene permeability of $10^{3.56}$ Barrer and ideal permselectivity of $10^{7.7}$. The low prediction uncertainties in ML-predicted diffusivity ($10^{-9\pm0.41}$ cm^2/s) and sorption uptake ($10^{0.74\pm0.07}$ mmol solvent/g polymer) further reinforce the model's confidence in its predictions, effectively ruling out potential statistical outliers. While the identification of PVC for toluene-heptane separation is not a novel discovery as PVC is well-established in the literature for this separation process, its re-discovery through our ML approach reinforces the validity of our methodology. Further validity is provided by Aouinti et al. who reports PVC's solubility parameter as 19.2 MPa$^{1/2}$, closely matching toluene's 18.2 MPa$^{1/2}$ and significantly differing from heptane's 15.1 MPa$^{1/2}$, implying PVC's higher affinity for aromatic toluene[54].

Searching the space of 13k known polymers for non-halogenated alternatives. Despite PVC's effective separation performance, it is considered one of the most environmentally harmful plastics; thus, identifying sustainable, non-halogenated alternatives is imperative[85,86]. To address this, we conducted a systematic screening of 13,000 known polymers to identify non-halogenated candidates with separation performance comparable to PVC. The initial screening was based on a moderately stringent criteria: ideal perm selectivity >$10^{5.5}$ and permeability >10^4 Barrer. To identify non-halogenated alternatives to PVC, we lowered the ideal permselectivity threshold and slightly increased the permeability threshold, as most non-halogenated candidates seemed to show lower ideal permselectivity. However, even the use of this moderately stringent criteria yielded no viable candidates from a dataset of 13k known polymers, underscoring the complexity of the problem-akin to finding a needle in a haystack. In response, we defined a relaxed criteria, such that ideal permselectivity >10^4 while maintaining the permeability threshold. This shift uncovered 46 candidates, out of which 31 are non-halogenated promising polymers that can be used for toluene-heptane separation. From this pool of polymers, we highlight Polymer B, which emerged as a sustainable candidate, being a known non-halogenated ring-opening polymer with a strong tendency for depolymerization. The rest of the non-halogenated alternative polymers (Supplementary Fig. 7) contain highly polar ester, carbonyl, ketone groups, aromatic or electronegative groups. These trends align with literature findings[53,54,83], as emphasized by Liu et al.[84], who reviewed 100 membranes for toluene-heptane separation and highlighted that those featuring electronegative groups and aromatic backbones-enabling π-π interactions-exhibit enhanced toluene affinity and separation performance.

Searching the space of chemically recyclable hypothetical polymers. In addition to exploring the known polymer search space to discover sustainable and efficient membranes, we extended our investigation to virtually generated datasets. To ensure reliable predictions, we excluded virtual polymers containing inorganic elements, such as Na, P, and Si, that were absent in the training data.

First, we analyzed the PI1M dataset, comprising 1 million hypothetical polymers, to identify PVC alternatives by applying the moderately stringent screening criteria (ideal permselectivity > $10^{5.5}$, permeability > 10^4 Barrer). This screening yielded 152 viable polymers, including 74 non-halogenated candidates. As expected, we observe a significant increase in the number of viable candidates with the expansion of chemical space, outlining the importance of such generative design approaches for polymer design. Notably, Polymer C exhibited excellent separation performance, comparable to PVC, with slightly higher permeability, as shown in Fig. 6a. Furthermore, Polymer D emerged as a promising sustainable, non-halogenated alternative. Additionally, applying the relaxed criteria yielded 1243 candidates, of which 920 were non-halogenated. Although a large number of non-halogenated polymers were identified, the chemical space of these polymers

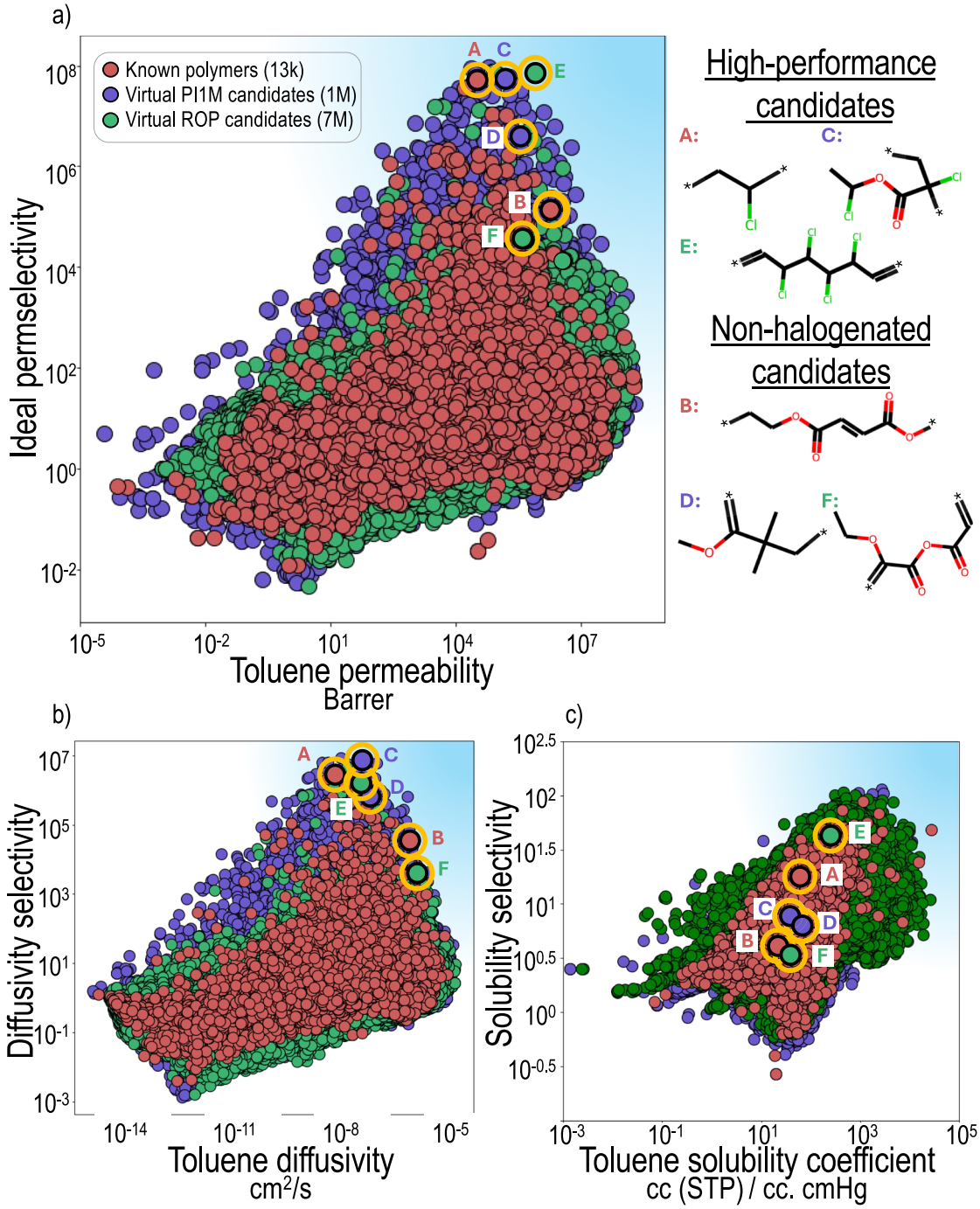

Fig. 6 | ML predicted trade-off plots. a Permeability, (**b**) diffusivity, and (**c**) solubility coefficients, along with their corresponding ideal selectivities for toluene/heptane separations across 13,000 known polymers (red), 1 million virtual PI1M (purple)[57], and 7 million virtual ROP polymers (green)[58]. The blue region in the top right highlights the target area of high transport properties and ideal selectivity. Focusing on high predicted toluene permeability and ideal permselectivity, among known polymers, Polymer A (polyvinyl chloride, PVC) emerges as the top machine learning-predicted candidate, aligning with its experimentally validated use in this application. Polymer B (a known polymer), a suggested sustainable alternative, demonstrates slightly lower performance. Within the virtual space, Polymer C (PI1M database) and E (ROP database) are identified as promising candidates with predicted performance exceeding that of PVC. Additionally, Polymers D (PI1M database) and F (ROP database) are virtual non-halogenated alternatives that offer comparable separation performance while representing more sustainable options compared to toxic halogenated membranes.

closely resembles that of the polymers in PolyInfo[87], and such polymers may not necessarily be optimized for recycling or represent more sustainable options.

Therefore, recognizing the need for more sustainable membranes, we further explored a dataset of 7 million synthetically accessible ring-opening polymerization (ROP) based polymers with an affinity for chemical recycling. Using the same moderately stringent criteria, we identified 9 candidate polymers, one of which was non-halogenated, denoted as polymer F. We also highlight Polymer E, which is predicted to exhibit slightly better separation performance than PVC. Although Polymer E is halogenated, it demonstrates a strong potential for depolymerization through ROP. Next, while screening for sustainable non-halogenated candidates within this virtual space, we noted that 2.1 million of the polymers were halogenated, significantly narrowing the search space for sustainable alternatives. By

using the relaxed screening criteria, we uncovered an additional 114 candidate polymers, of which 48 were non-halogenated (Supplementary Fig. 7). We note that the lower number of candidates passing the screening criteria in the ROP dataset, compared to the PI1M dataset-despite the ROP dataset being nearly seven times larger-may be attributed to its more limited chemical diversity. In contrast, the PI1M dataset benefits from a broader range of chemistries, providing a more diverse pool of potential candidates. We note that while leveraging a virtual space of 8 million polymers-compared to the known space of 13,000-led to the identification of more polymers meeting the specified criteria, the total number of successful candidates remained relatively small, underscoring the inherent complexity and challenges of polymer design. These results point to the fundamental limitations in achieving an optimal permeability-selectivity trade-off, which is primarily influenced by polymer packing and structural factors[88].

Further, we highlight a discrepancy arising from the ML-based overestimation of ideal diffusivity selectivity and perm-selectivity for the toluene-heptane system. Experimentally, the pervaporation perm-selectivity for a 50/50 multi-component mixture of toluene-heptane at 56 °C is approximately 10.1, significantly lower than the ML-predicted values (Please refer to the Supplementary Information, section 8 for detailed derivations of the formulas and explanations)[55]. This discrepancy underscores a fundamental limitation of ML predictions that rely on single-component transport behavior, and they fail to account for solvent-solvent interactions present in multi-component systems. Previously, we noted that single-component transport behavior is insufficient for describing multi-component systems[15], likely leading to this discrepancy. Furthermore, to assess the validity of the ML predictions, we compared PVC predictions using the published and validated PENN-1 model[15] alongside our PENN-2 models, finding that the predictions closely align, with values for toluene at $10^{-8.33}$ cm^2/s and $10^{-8.23}$ cm^2/s, and for extremely low values for heptane at $10^{-14.01}cm^2/s$ and $10^{-14.68}cm^2/s$. Thus after eliminating the possibility of invalid ML predictions as the cause of the observed discrepancy, we conduct a more thorough investigation. Aouinti et al. similarly observed the extremely sluggish transport of pure heptane in a modified PVC membrane in their experimental studies, to the extent that flux values could not be measured[89]. Additionally, they reported a significant swelling degree of 49% for PVC in toluene. Based on this, we hypothesize that toluene, due to its strong affinity for PVC, may alter the polymer's structure by inducing swelling and plasticization, thus creating a more conducive pathway for heptane diffusion. This hypothesis is supported by prior studies; for example, Mathias et al. attributed the loss in diffusivity selectivity to membrane plasticization and introduced the concept of "cohort diffusion"[90]. Here, friction-induced diffusion coupling effects cause faster molecules to slow down and slower molecules to speed up, leading to a reduction in diffusion selectivity. Additionally, Lee et al. observed a loss in diffusivity selectivity when the solubility difference between the polymer and solvents drops below a critical threshold (e.g., $\delta = 8\ MPa^{1/2}$), also attributed to polymer plasticization and swelling[91]. These works provide plausible explanations for why the observed selectivities are often much lower than ideal selectivities[92]. In summary, the discrepancy in the ML over-predictions of diffusivity selectivities can be attributed to the absence of solvent-solvent interactions and the effects of polymer swelling. Taking these over-predictions for ideal selectivity into account, we caution against drawing quantitative conclusions from ML predictions for multi-component systems; instead, these models should be used for qualitative insights, such as identifying PVC and other halogenated polymers as strong candidates for this separation.

In addition, we analyzed outlier ML predictions and corresponding prediction uncertainties; incorporating such uncertainty analysis enhances the credibility of ML predictions. By evaluating the standard deviation of predictions across 10 cross-validation (CV) models, we observed higher uncertainty in regions with low permeability and ideal perm-selectivity (Supplementary Fig. 8), indicating that these predictions should be interpreted with caution. Conversely, regions with high permeability and selectivity, which are the primary areas of interest for membrane design, exhibited much lower uncertainties, thus providing greater confidence for exploring new chemical spaces in this regime. In the future, exploring new polymer chemistries and validating the separation performance of these newly discovered polymers through multi-component transport simulations will lay the foundation for developing high-performance, sustainable polymers for solvent separation.

Discussion

In conclusion, we proposed a methodology to guide the discovery of sustainable polymer membranes for a given organic binary solvent separation. Focusing on solvent diffusivity as a key parameter to understand solvent transport, we fused experimental and in-house generated simulated data to build robust multi-task and physics-enforced machine learning models. These models demonstrated enhanced prediction generalizability and accuracy in data-scarce regimes as compared to traditional single-task ML models. Next, we leverage these models to address the issue of identifying optimal membranes for a given binary solvent separation problem. For a case study of toluene-heptane separation, ML results indicate PVC as the optimal polymer for this separation, a finding consistent with the literature, thus affirming the validity of this methodology. We further proposed halogen-free alternatives to PVC by screening across a much larger chemical space of 8 million virtually produced candidates. This data-driven approach is scalable to other solvent systems and is thus expected to advance solvent separation technologies significantly. We acknowledge inherent assumptions and limitations, including the lack of consideration of solvent-solvent interactions and the need for further investigation into long-term membrane performance factors such as swelling, plasticization, and aging. Notably, despite these approximations, the ML-predicted trade-off plots represent a novel contribution, being the most significant of their kind to date, and will function as robust benchmarks that guide and expedite the discovery of membranes for solvent separations. We foresee that expanding the search space using virtual design algorithms, particularly those focused on promising candidates such as polymers with high intrinsic microporosity, will uncover new sustainable, high-performance membranes that are awaiting discovery.

Methods

Polymer and solvent fingerprinting

To numerically represent polymers and solvents, we used the hierarchical fingerprinting procedure that was developed in the past decade and referred to as the Polymer Genome fingerprinting[27,77]. The polymer and solvent SMILES were converted into numerical vectors to represent the chemical structure of the polymers and solvents. This fingerprinting scheme derives features from various hierarchical levels, including atomic-level, block-level, and morphological descriptors. Atomic-level fingerprints consider atomic triples (fragments of three contiguous atoms), while block-level fingerprints examine larger-scale blocks, such as benzene rings. Additionally, morphological descriptors encompass features like the shortest topological distance between rings, the fraction of atoms in side chains, and the length of the largest side chain. They also include quantitative structure-property relationship (QSPR) descriptors such as Van der Waals volume, surface area, and topological polar surface area (TPSA).

Sorption Uptake ML model

Sorption uptake refers to the amount of solvent absorbed per mass (or per volume) of the polymer matrix. However, experimental measurements of solvent sorption uptake, often performed using setups similar to those for diffusivity measurements, are notably time-intensive. To address this, we previously developed an ML model that accurately predicts the sorption uptake (or the specific uptake) of solvents in polymers[15]. We trained the "Sorption Uptake ML model" on a comprehensive dataset of experimental sorption values (expressed as mmol of solvent present per gram of polymer), covering 2275 systems of 46 polymers and 91 solvents. The neural network model comprises of an input layer, two hidden layers, and an output layer. This model utilized Polymer Genome-based[27,77] fingerprints to describe polymers and solvents (described in detail in Section

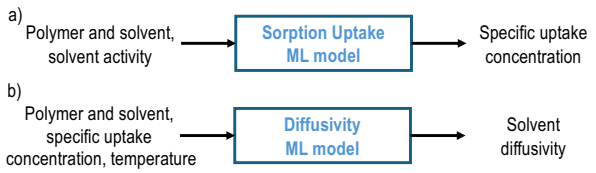

Fig. 7 | Sorption Uptake ML model. a This model enables prediction of specific uptake concentration as a function of solvent activity[15]. **b** The diffusivity ML model built in this work. Together, these two models are used to estimate solvent permeability through a polymer.

Polymer and solvent fingerprinting) along with solvent molar volumes, to predict specific uptake across a range of activities, as illustrated in Fig. 7, achieving a satisfactory predictive accuracy with a order of magnitude error or mean absolute error (MAE) in log scale of 0.13 mmol/g and an average R^2 of 0.93 across 10 runs on a 10% test split. Together with our diffusivity model, this sorption model provides a full description of solvent permeability through polymers. Previously, we used these models to predict complex mixture permeation by incorporating sorption and diffusivity data into a Maxwell-Stefan mass transport model. In this work, we leveraged the model in two ways: first, to expand the dataset, as discussed in Section Data augmentation, and second, to predict solubility coefficient values, as outlined in Section Solvent diffusivity ML models. More details about the model architecture are accessible in the Supplementary Section of our previous publication[15].

Solvent diffusivity ML models

- **Gaussian Process Regression (GPR)**: Gaussian process regression uses a Bayesian framework, wherein a Gaussian process is used to map inputs to the desired output property, diffusivity. GPR is widely used for smaller datasets, valued for its ability to deliver accurate predictions while quantifying uncertainty effectively. GPR models were built using Radial Basis Function (RBF) kernels and 5-fold cross-validation using scikit-learn libraries[93].
- **Neural Networks (NN)**: A fully connected neural network model was developed to predict the target property, consisting of an input layer, two hidden layers, and an output layer to predict the target property. The features polymer-solvent fingerprints, solvent concentration, temperature, and a selector vector for the data source are mapped to the target property. Each hidden layer consisted of a set of neurons that apply learned weights to the inputs, followed by a non-linear sigmoid activation function. This architecture, established in our previous work[15] and visualized in Supplementary Fig. 4a, was trained using backpropagation to minimize prediction errors. Given the small dataset size, 10 models were trained using 10-fold cross-validation, and an ensemble of these models was used for the final average prediction and to obtain the standard deviation of predicted values. Additionally, the dropout technique was implemented to prevent overfitting. All models were implemented using TensorFlow[94].
- **Physics enforced Neural Networks (PENN)**: The PENN model architecture is the same as that of neural networks, as described above, with an additional output layer to enforce physics (Supplementary Fig. 4b and 4c). For the PENN-1 model, polymer-solvent fingerprints, solvent concentration, temperature, and a selector vector for the data source are used as inputs. The molar volume of the solvent ($\hat{V}$) is incorporated into the output layer, where parameters A and B are determined through equation fitting to map the input features to the output. The empirical relationship was encoded using the equation:

$$\log_{10} D = A \cdot \log_{10} \hat{V} + B \tag{4}$$

where $\hat{V}$ (m^3/mol) represents the solvent molar volume, and A and B were the learned parameters.

In the PENN-2 model, the Arrhenius equation[78,79] is embedded within the neural network to predict solvent diffusivity. The PENN-2 model follows a similar architecture, with polymer-solvent fingerprints, solvent concentration, and a selector vector for the data source as inputs. However, temperature is included in the output layer. The encoded temperature dependence is:

$$D = D_0 \cdot \exp\left(-\frac{E}{RT}\right) \tag{5}$$

where D_0 (cm^2/s) represents the pre-exponential constant, E (J/mol) denotes the activation energy, R (8.314 J/mol-K) is the universal gas constant, and T (K) is the temperature. PENN models were developed using TensorFlow packages[94].

Construction of ML based solvent-trade-off plots

According to the simplified solution-diffusion theory (Equation (1)), which is accurate in the limit of dilute vapor streams, we calculated single-component permeability as the product of ML-predicted diffusivity and solubility coefficients[95]. Solubility coefficient (S) was determined by normalizing the ML-predicted specific uptake solvent concentration (C)[95] by the solvent vapor pressure (p_{sat}), as follows:

$$S = \frac{C}{p_{sat}} \tag{6}$$

Here, we utilize a simple approximation for the solubility coefficient, which imagines the concentration profile in the membrane as linear, with the feed-side driving force as the saturated vapor pressure p_{sat}, and the permeate side driving force as being essentially zero pressure (perfect vacuum). This situation roughly approximates an idealized pervaporation membrane separation.

The vapor pressure values for the solvents are obtained from PubChem database[96,97]. Additionally, polymer densities were determined using a separate ML model[27], allowing for the calculation of solubility in terms of volume. Moreover, the sorption uptake values served as concentration inputs for the updated production diffusivity model, and the final permeability was derived from the product of ML-predicted diffusivity and solubility coefficients. Similarly, predicted ideal permselectivity was calculated using Equation (2). It is important to note that this approach did not account for solvent-solvent interactions in a multi-component system, as our ML predictions were made for pure solvent systems. Therefore, caution must be exercised in cases of non-ideal solvent behavior. However, this approximation enabled the efficient screening of 13k known and 8M virtual polymers, enabling the construction of the largest solvent trade-off plots and providing a rapid framework for membrane design.

Membrane design search space

To find a suitable membrane for a given solvent separation, we explored the known chemical space consisting of 13,000 polymers. To expand the search for high-performance and sustainable membrane alternatives, we examined two virtual polymer databases. The first, developed by Luo et al.[57], is the open-source PI1M database, containing 1 million virtually generated polymers. These polymers were produced using a generative Recurrent Neural Network (RNN) ML model trained on the SMILES strings of existing polymers from the PolyInfo[87] database. This approach was designed to cover a chemical space similar to PolyInfo[87] while significantly enhancing data density in underrepresented regions. The second database comprises 7 million virtually generated polymers with strong potential for chemical depolymerization via ring-opening polymerization (ROP), created by Kern et al.[58]. This affinity for chemical recycling in ROP-based candidates stems from the favorable thermodynamics associated with these processes. This effort examined possible hypothetical polymer candidates that could be generated using existing

molecules, a concept referred to as Virtual Forward Synthesis (VFS). VFS utilizes 30,272,000 known commercial molecules and applies standard polymerization reaction procedures to generate hypothetical polymers. These candidates are then screened to identify sustainable options that meet ML-predicted criteria for high permeability and ideal permselectivity.

Data Availability

The compiled experimental and our generated simulated data is publicly available on GitHub.

Code availability

The Polymer Structure Predictor (PSP) package to create simulation polymer structures is available free of charge on GitHub. The simulation scripts have also been made available on Github.

Received: 8 January 2025; Accepted: 2 June 2025;
Published online: 19 June 2025

References

1. Billard, P., Nguyen, Q., Leger, C. & Clement, R. Diffusion of organic compounds through chemically asymmetric membranes made of semi-interpenetrating polymer networks. *Sep. Purif. Technol.* **14**, 221–232 (1998).
2. Farshad, F., Iravaninia, M., Kasiri, N., Mohammadi, T. & Ivakpour, J. Separation of toluene/n-heptane mixtures experimental, modeling and optimization. *Chem. Eng. J.* **173**, 11–18 (2011).
3. Ribeiro, C. P., Freeman, B. D., Kalika, D. S. & Kalakkunnath, S. Aromatic polyimide and polybenzoxazole membranes for the fractionation of aromatic/aliphatic hydrocarbons by pervaporation. *J. Membr. Sci.* **390**, 182–193 (2012).
4. Villaluenga, J. G. & Tabe-Mohammadi, A. A review on the separation of benzene/cyclohexane mixtures by pervaporation processes. *J. Membr. Sci.* **169**, 159–174 (2000).
5. Khazaei, A., Mohebbi, V., Behbahani, R. M. & Ramazani, S. A. Energy consumption in pervaporation, conventional and hybrid processes to separate toluene and i-octane. *Chem. Eng. Process.-Process Intensif.* **128**, 46–52 (2018).
6. Li, Y. et al. Energy-saving and environmentally friendly pervaporation-distillation hybrid process for alcohol and ester recovery from wastewater containing three binary azeotropes. *Sep. Purif. Technol.* **281**, 119889 (2022).
7. Yong, W. F. & Zhang, H. Recent advances in polymer blend membranes for gas separation and pervaporation. *Prog. Mater. Sci.* **116**, 100713 (2021).
8. Shao, P. & Huang, R. Polymeric membrane pervaporation. *J. Membr. Sci.* **287**, 162–179 (2007).
9. Nagasawa, H. & Tsuru, T. Silica membrane application for pervaporation process. *Curr. Trends Futur. Dev. (Bio-) Membranes*, 217–241 (Elsevier, 2017).
10. Iravaninia, M., Mirfendereski, M. & Mohammadi, T. Pervaporation separation of toluene/n-heptane mixtures using a MSE-modified membrane: Effects of operating conditions. *Chem. Eng. Res. Des.* **90**, 397–408 (2012).
11. Aryafard, E., Rahmatmand, B. & Rahimpour, M. R. Application of computational fluid dynamics technique in pervaporation processes. *Curr. Trends Futur. Dev. (Bio-) Membranes*, 247–268 (Elsevier, 2022).
12. Lively, R. P. & Sholl, D. S. From water to organics in membrane separations. *Nat. Mater.* **16**, 276–279 (2017).
13. Thompson, K. A. et al. N-aryl–linked spirocyclic polymers for membrane separations of complex hydrocarbon mixtures. *Science* **369**, 310–315 (2020).
14. Wijmans, J. G. & Baker, R. W. The solution-diffusion model: a review. *J. Membr. Sci.* **107**, 1–21 (1995).
15. Lee, Y. J. et al. Data-driven predictions of complex organic mixture permeation in polymer membranes. *Nat. Commun.* **14**, 4931 (2023).
16. Reis, R., Vladimir Oliveira, J. & Nobrega, R. Diffusion coefficients in polymer-solvent systems for highly concentrated polymer solutions. *Braz. J. Chem. Eng.* **18**, 367–384 (2001).
17. Lee, J. K., Yao, S. X., Li, G., Jun, M. B. & Lee, P. C. Measurement methods for solubility and diffusivity of gases and supercritical fluids in polymers and its applications. *Polym. Rev.* **57**, 695–747 (2017).
18. Tsige, M. & Grest, G. S. Interdiffusion of solvent into glassy polymer films: A molecular dynamics study. *J. Chem. Phys.* **121**, 7513–7519 (2004).
19. Tsige, M. & Grest, G. S. Molecular dynamics simulation of solvent–polymer interdiffusion: Fickian diffusion. *J. Chem. Phys.* **120**, 2989–2995 (2004).
20. Wang, Z. et al. Polymer membranes for organic solvent nanofiltration: Recent progress, challenges and perspectives. *Adv. Membr.* **3**, 100063 (2023).
21. Müller-Plathe, F. Diffusion of water in swollen poly (vinyl alcohol) membranes studied by molecular dynamics simulation. *J. Membr. Sci.* **141**, 147–154 (1998).
22. Tran, H. et al. Design of functional and sustainable polymers assisted by artificial intelligence. *Nat. Rev. Mater.* **9**, 866–886 (2024).
23. Wu, C. et al. Rational design of all-organic flexible high-temperature polymer dielectrics. *Matter* **5**, 2615–2623 (2022).
24. Chen, L. et al. Polymer informatics: Current status and critical next steps. *Mater. Sci. Eng.: R: Rep.* **144**, 100595 (2021).
25. Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS macro Lett.* **6**, 1078–1082 (2017).
26. Tran, H., Shen, K.-H., Shukla, S., Kwon, H.-K. & Ramprasad, R. Informatics-driven selection of polymers for fuel-cell applications. *J. Phys. Chem. C.* **127**, 977–986 (2023).
27. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer Genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C.* **122**, 17575–17585 (2018).
28. Gurnani, R. et al. AI-assisted discovery of high-temperature dielectrics for energy storage. *Nat. Commun.* **15**, 6107 (2024).
29. Phan, B. K. et al. Gas permeability, diffusivity, and solubility in polymers: Simulation-experiment data fusion and multi-task machine learning. *Npj Comput. Mater.* **10**, 186 (2024).
30. Wessling, M. et al. Modelling the permeability of polymers: a neural network approach. *J. Membr. Sci.* **86**, 193–198 (1994).
31. Yuan, Q. et al. Imputation of missing gas permeability data for polymer membranes using machine learning. *J. Membr. Sci.* **627**, 119207 (2021).
32. Ricci, E. & De Angelis, M. G. A perspective on data-driven screening and discovery of polymer membranes for gas separation, from the molecular structure to the industrial performance. *Rev. Chem. Eng.* **40**, 567–600 (2024).
33. Khajeh, A. et al. Early prediction of ion transport properties in solid polymer electrolytes using machine learning and system behavior-based descriptors of molecular dynamics simulations. *Macromolecules* **56**, 4787–4799 (2023).
34. Ritt, C. L. et al. Machine learning reveals key ion selectivity mechanisms in polymeric membranes with subnanometer pores. *Sci. Adv.* **8**, eabl5771 (2022).
35. Nistane, J., Chen, L., Lee, Y., Lively, R. & Ramprasad, R. Estimation of the Flory-Huggins interaction parameter of polymer-solvent mixtures using machine learning. *MRS Commun.* **12**, 1096–1102 (2022).
36. Aoki, Y. et al. Multitask machine learning to predict polymer–solvent miscibility using Flory–Huggins interaction parameters. *Macromolecules* **56**, 5446–5456 (2023).
37. Xu, Q. & Jiang, J. Machine learning for polymer swelling in liquids. *ACS Appl. Polym. Mater.* **2**, 3576–3586 (2020).
38. Tao, L., He, J., Arbaugh, T., McCutcheon, J. R. & Li, Y. Machine learning prediction on the fractional free volume of polymer membranes. *J. Membr. Sci.* **665**, 121131 (2023).
39. Kuenneth, C. et al. Polymer informatics with multi-task learning. *Patterns* **2**, 4 (2021).

40. Toland, A. et al. Accelerated scheme to predict ring-opening polymerization enthalpy: Simulation-experimental data fusion and multitask machine learning. *J. Phys. Chem. A* **127**, 10709–10716 (2023).
41. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
42. Kashinath, K. et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. R. Soc. A* **379**, 20200093 (2021).
43. Meng, C., Seo, S., Cao, D., Griesemer, S. & Liu, Y. When physics meets machine learning: A survey of physics-informed machine learning. *Mach. Learn. Comput. Sci. Eng.* **1**, 1–23 (2025).
44. Jain, A., Gurnani, R., Rajan, A., Qi, H. J. & Ramprasad, R. A physics-enforced neural network to predict polymer melt viscosity. *Npj Comput. Mater.* **11**, 42 (2025).
45. Pun, G. P., Batra, R., Ramprasad, R. & Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **10**, 2339 (2019).
46. Bradford, G. et al. Chemistry-informed machine learning for polymer electrolyte discovery. *ACS Cent. Sci.* **9**, 206–216 (2023).
47. Zhang, Z. & Gu, G. X. Physics-informed deep learning for digital materials. *Theor. Appl. Mech. Lett.* **11**, 100220 (2021).
48. Reynier, A., Dole, P., Humbel, S. & Feigenbaum, A. Diffusion coefficients of additives in polymers. i. correlation with geometric parameters. *J. Appl. Polym. Sci.* **82**, 2422–2433 (2001).
49. Othmer, D. F. & Thakar, M. S. Correlating diffusion coefficient in liquids. *Ind. Eng. Chem.* **45**, 589–593 (1953).
50. Robeson, L. M. The upper bound revisited. *J. Membr. Sci.* **320**, 390–400 (2008).
51. Robeson, L. M. Correlation of separation factor versus permeability for polymeric membranes. *J. Membr. Sci.* **62**, 165–185 (1991).
52. Comesaña-Gándara, B. et al. Redefining the Robeson upper bounds for CO_2/CH_4 and CO_2/N_2 separations using a series of ultrapermeable benzotriptycene-based polymers of intrinsic microporosity. *Energy Environ. Sci.* **12**, 2733–2740 (2019).
53. Aouinti, L., Roizard, D., Hu, G., Thomas, F. & Belbachir, M. Investigation of pervaporation hybrid polyvinylchloride membranes for the separation of toluene-n-heptane mixtures-case of clays as filler. *Desalination* **241**, 174–181 (2009).
54. Aouinti, L. & Roizard, D. Pervaporation of toluene-n-heptane mixtures with hybrid PVC membranes containing inorganic particles. *J. Earth Sci. Eng.* **5**, 473–481 (2015).
55. Tabbiche, A. & Aouinti, L. Preparation and characterization of nanocomposite membranes based on PVC/TiO_2 anatase for the separation of toluene/n-heptane mixtures via pervaporation. *Polym. Bull.* **80**, 643–666 (2023).
56. Friess, K. et al. Comparative study of sorption and permeation techniques for the determination of heptane and toluene transport in polyethylene membranes. *J. Membr. Sci.* **338**, 161–174 (2009).
57. Ma, R. & Luo, T. PI1M: a benchmark database for polymer informatics. *J. Chem. Inf. Modeling* **60**, 4684–4690 (2020).
58. Kern, J., Su, Y., Gutekunst, W. & Ramprasad, R. An informatics framework for the design of sustainable, chemically recyclable, synthetically-accessible and durable polymers. *arXiv preprint arXiv:2409.15354* (2024).
59. Sahu, H., Shen, K.-H., Montoya, J. H., Tran, H. & Ramprasad, R. Polymer structure predictor (PSP): a python toolkit for predicting atomic-level structural models for a range of polymer geometries. *J. Chem. Theory Comput.* **18**, 2737–2748 (2022).
60. Abbott, L. J., Hart, K. E. & Colina, C. M. Polymatic: a generalized simulated polymerization algorithm for amorphous polymers. *Theor. Chem. Acc.* **132**, 1–19 (2013).
61. Flyvbjerg, H. & Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **91**, 461–466 (1989).
62. Mamaliga, I., Schabel, W. & Kind, M. Measurements of sorption isotherms and diffusion coefficients by means of a magnetic suspension balance. *Chem. Eng. Process.: Process Intensif.* **43**, 753–763 (2004).
63. Hong, S.-U. Prediction of polymer/solvent diffusion behavior using free-volume theory. *Ind. Eng. Chem. Res.* **34**, 2536–2544 (1995).
64. Aminabhavi, T., Phayde, H., Ortego, J. & Vergnaud, J. Sorption/diffusion of aliphatic esters into tetrafluoroethylene/propylene copolymeric membranes in the temperature interval from 25 to 70° C. *Eur. Polym. J.* **32**, 1117–1126 (1996).
65. Prager, S. & Long, F. Diffusion of hydrocarbons in polyisobutylene[1]. *J. Am. Chem. Soc.* **73**, 4072–4075 (1951).
66. Doumenc, F., Guerrier, B. & Allain, C. Mutual diffusion coefficient and vapor- liquid equilibrium data for the system polyisobutylene+ toluene. *J. Chem. Eng. Data* **50**, 983–988 (2005).
67. Nunez, E. M., Myerson, A. S. & Kwei, T. Diffusion of benzene vapor in blends of poly (vinyl acetate) and poly (methyl acrylate). *Polym. Eng. Sci.* **31**, 1172–1175 (1991).
68. Vrentas, J. & Duda, J. Diffusion of small molecules in amorphous polymers. *Macromolecules* **9**, 785–790 (1976).
69. Aminabhavi, T. & Naik, H. Chemical compatibility study of geomembranes-sorption/desorption, diffusion and swelling phenomena. *J. Hazard. Mater.* **60**, 175–203 (1998).
70. Odani, H., Uchikura, M., Ogino, Y. & Kurata, M. Diffusion and solution of methanol vapor in poly (2-vinylpyridine)-block-polyisoprene and poly (2-vinylpyridine)-block-polystyrene. *J. Membr. Sci.* **15**, 193–208 (1983).
71. Hansen, C. M.*The three dimensional solubility parameter and solvent diffusion coefficient: Their importance in surface coating formulation* (1967).
72. Giacinti Baschetti, M., Piccinini, E., Barbari, T. A. & Sarti, G. C. Quantitative analysis of polymer dilation during sorption using ftir-atr spectroscopy. *Macromolecules* **36**, 9574–9584 (2003).
73. Rogers, C., Stannett, V. & Szwarc, M. The sorption, diffusion, and permeation of organic vapors in polyethylene. *J. Polym. Sci.* **45**, 61–82 (1960).
74. Galizia, M., Daniel, C., Guerra, G. & Mensitieri, G. Solubility and diffusivity of low molecular weight compounds in semi-crystalline poly-(2, 6-dimethyl-1, 4-phenylene) oxide: The role of the crystalline phase. *J. Membr. Sci.* **443**, 100–106 (2013).
75. Khinnavar, R. & Aminabhavi, T. Diffusion and sorption of organic liquids through polymer membranes. i. polyurethane versus n-alkanes. *J. Appl. Polym. Sci.* **42**, 2321–2328 (1991).
76. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *Npj Comput. Mater.* **4**, 25 (2018).
77. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106 (2015).
78. Rogers, C. Permeation of gases and vapours in polymers. In *Polymer permeability*, 11–73 (Springer, 1985).
79. Klopffer, M. & Flaconneche, B. Transport properties of gases in polymers: bibliographic review. *Oil Gas. Sci. Technol.* **56**, 223–244 (2001).
80. Safonova, A. et al. Ten deep learning techniques to address small data problems with remote sensing. *Int. J. Appl. Earth Observation Geoinf.* **125**, 103569 (2023).
81. Adadi, A. A survey on data-efficient algorithms in big data era. *J. Big Data* **8**, 24 (2021).
82. Du, S. S. et al. How many samples are needed to estimate a convolutional or recurrent neural network? *Adv. Neural Inf. Process Syst.* 31 (NeurIPS 2018).
83. Pakizeh, M., Karami, M., Kooshki, S. & Rahimnia, R. Advanced toluene/n-heptane separation by pervaporation: investigating the potential of graphene oxide (GO)/PVA mixed matrix membrane. J. Taiwan Inst. *Chem. Eng.* **150**, 105025 (2023).
84. Liu, H.-X., Wang, N., Zhao, C., Ji, S. & Li, J.-R. Membrane materials in the pervaporation separation of aromatic/aliphatic hydrocarbon mixtures-a review. *Chin. J. Chem. Eng.* **26**, 1–16 (2018).

85. Comaniță, E.-D. et al. Environmental impacts of polyvinyl chloride (PVC) production process. In *2015 E-Health and Bioengineering Conference (EHB)*, 1–4 (IEEE, 2015).
86. Wagoner, J. K. Toxicity of vinyl chloride and poly (vinyl chloride): a critical review. *Environ. Health Perspect.* **52**, 61–66 (1983).
87. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. *2011 International Conference on Emerging Intelligent Data and Web Technologies* 22–29 (2011).
88. 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).
89. Aouinti, L., Roizard, D. & Belbachir, M. PVC-activated carbon based matrices: A promising combination for pervaporation membranes useful for aromatic–alkane separations. *Sep. Purif. Technol.* **147**, 51–61 (2015).
90. Mathias, R. et al. Framework for predicting the fractionation of complex liquid feeds via polymer membranes. *J. Membr. Sci.* **640**, 119767 (2021).
91. Lee, Y. J. & Lively, R. P. A transition in diffusion behaviors of organic liquid mixtures in dense polymer membranes. *J. Membr. Sci.* **713**, 123346 (2025).
92. He, Z. & Wang, K. The 'ideal selectivity' vs 'true selectivity' for permeation of gas mixture in nanoporous membranes. In *IOP Conference Series: Materials Science and Engineering*, vol. 323, 012002 (IOP Publishing, 2018).
93. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122 (2013).
94. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems http://tensorflow.org/ (2015).
95. Lee, Y. M., Bourgeois, D. & Belfort, G. Sorption, diffusion, and pervaporation of organics in polymer membranes. *J. Membr. Sci.* **44**, 161–181 (1989).
96. National Center for Biotechnology Information. PubChem compound summary for CID 8900, heptane https://pubchem.ncbi.nlm.nih.gov/compound/8900#section=Vapor-Density. Retrieved October 29, 2024 (2024).
97. National Center for Biotechnology Information. PubChem compound summary for CID 1140, toluene https://pubchem.ncbi.nlm.nih.gov/compound/Toluene. Retrieved October 29, 2024 (2024).

Acknowledgements

The authors would like to thank the Office of Naval Research through a multidisciplinary university research initiative (MURI) for their funding support. We would also like to acknowledge Dr. Kuan-Hsuan Shen for their valuable support in building the simulation pipeline. We also extend a thank you to Dr. Lihua Chen for her guidance in the initial stage of the work. This research is supported in part through research cyber-infrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology and XSEDE/ACCESS for computational support through Grant No. TG-DMR080058N.

Author contributions

The work was conceived and guided by R.R. J.N. performed simulations, followed by training and evaluating the machine learning models. R.D. trained machine learning models and performed analysis. H.S. guided feedback on the manuscript. J.N., R.D., Y.J.L., H.S., S.S.J., R.L., and R.R. discussed the results and commented on the manuscript.

Competing interests

R.R. is a founder of Matmerize, Inc., a company specializing in materials informatics software and services. The other authors have no conflicts of interest to declare.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01681-8.

Correspondence and requests for materials should be addressed to Rampi Ramprasad.