Chemical Physics Reviews

REVIEW ARTICLE | DECEMBER 17 2024

Transcend the boundaries: Machine learning for designing polymeric membrane materials for gas separation

Special Collection: AI and Machine Learning in Chemical and Materials Science

Jiaxin Xu ; Agboola Suleiman ; Gang Liu ; Renzheng Zhang ; Meng Jiang ; Ruilan Guo ; Tengfei Luo

Chem. Phys. Rev. 5, 041311 (2024)
https://doi.org/10.1063/5.0205433

Chemical Physics Reviews REVIEW pubs.aip.org/aip/cpr

Transcend the boundaries: Machine learning for designing polymeric membrane materials for gas separation

Cite as: Chem. Phys. Rev. **5**, 041311 (2024); doi: 10.1063/5.0205433
Submitted: 26 February 2024 · Accepted: 14 November 2024 ·
Published Online: 17 December 2024

Jiaxin Xu,[1] Agboola Suleiman,[2] Gang Liu,[3] Renzheng Zhang,[1] Meng Jiang,[3] Ruilan Guo,[2] and Tengfei Luo[1,2,a)]

AFFILIATIONS

[1]Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA
[2]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA
[3]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA

Note: This paper is part of the CPR Special Topic on AI in Chemical and Materials Science.
[a)]**Author to whom correspondence should be addressed:** tluo@nd.edu

ABSTRACT

Polymeric membranes have become essential for energy-efficient gas separations such as natural gas sweetening, hydrogen separation, and carbon dioxide capture. Polymeric membranes face challenges like permeability-selectivity tradeoffs, plasticization, and physical aging, limiting their broader applicability. Machine learning (ML) techniques are increasingly used to address these challenges. This review covers current ML applications in polymeric gas separation membrane design, focusing on three key components: polymer data, representation methods, and ML algorithms. Exploring diverse polymer datasets related to gas separation, encompassing experimental, computational, and synthetic data, forms the foundation of ML applications. Various polymer representation methods are discussed, ranging from traditional descriptors and fingerprints to deep learning-based embeddings. Furthermore, we examine diverse ML algorithms applied to gas separation polymers. It provides insights into fundamental concepts such as supervised and unsupervised learning, emphasizing their applications in the context of polymer membranes. The review also extends to advanced ML techniques, including data-centric and model-centric methods, aimed at addressing challenges unique to polymer membranes, focusing on accurate screening and inverse design.

TABLE OF CONTENTS

I. INTRODUCTION

Membrane-based gas separation has received significant attention in industrial applications, including natural gas sweetening, hydrogen

separation, and direct carbon dioxide capture.[1–4] This interest stems from its advantages over conventional thermal separation processes, such as a smaller carbon footprint, reduced spatial requirements, and significantly lower thermal demands due to the absence of phase change.[3,5,6] These inherent features distinguish membrane-based processes from established yet thermally intensive gas separation methods like cryogenic distillation and pressure swing absorption.[3,7–10] The membrane material is one of the core components of the entire membrane separation process.[4] Critical factors influencing a membrane's gas separation properties include permeability, selectivity, resistance to harsh chemical environments, and mechanical and thermal properties, as well as configuration (such as hollow fiber, spiral wound, or plate and frame) and the overall system design. These diverse elements underscore the complex nature of membrane-based gas separation.[1,3,4,11–13] There are two main gas separation membrane material categories: inorganic and organic. Inorganic materials have limited applications due to the difficulty in making continuous and defect-free membranes, inherent brittleness causing mechanical issues, and high production costs.[14] Among organic membrane materials, polymer membranes are preferred, dominating membrane-based gas separations in industry due to their lower cost, greater robustness, and ease of fabrication and scalability.[3,12,14–18]

The gas transport mechanism in a membrane mainly depends on its material and the microscopic morphology, whether porous or dense. As shown in Fig. 1(a), three common gas transport mechanisms are facilitated transport, molecular-sieving, and solution-diffusion mechanisms. Facilitated transport is the primary mechanism in facilitated-transport membranes, which offer high selectivity by incorporating a carrier agent into the membrane.[19,20] Molecular sieving is dominant in porous inorganic membranes, such as zeolites, carbon molecular sieves (CMS), zeolitic imidazolate frameworks (ZIFs), and metal-organic frameworks (MOFs).[12,13,21,22] In this case, the membrane's pore size is similar to the gas molecule size, allowing smaller molecules to diffuse at higher rates. Finally, gas transport through dense polymer membranes mostly follows the solution-diffusion mechanism.[23] This entails the dissolution of permeants at the surface of the nonporous membrane under high upstream pressure, followed by diffusion driven by a concentration gradient, and ends with desorption at the membrane surface on the low-pressure downstream side.

Polymer membrane materials exhibit different physical and chemical properties that align them with specific gas pairs for effective separation. The fundamental properties of permeability (P) and selectivity (α) are essential. P measures how fast the gas will pass through the membrane, and α indicates the membrane's capacity to discriminate between different gases. However, these two properties are inversely interconnected at most times. The P of a gas is the product of the sorption (S) and diffusion (D) coefficients, as follows:

$$P = S \times D, \tag{1}$$

S depends on the type of gas and its interface with the polymer matrix, while D relies on the migration of the gas molecules between the free volumes in the polymer. Free volume is an intrinsic property of a polymer matrix, which is an estimate of the unoccupied space within the polymer.[18,24,25]

Membrane-based gas separations pose unique challenges compared to other membrane applications like reverse osmosis, ultrafiltration, and pervaporation. This complexity arises from the minimal difference in kinetic diameters between molecules of the gases targeted for separation, which typically falls within the narrow range of 0.015–0.020 nm, as shown in Fig. 1(b).[27] An illustration of this challenge is the air separation process, where the size difference between the kinetic diameters of O_2 (0.346 nm) and N_2 (0.364 nm) is a mere 0.018 nm, i.e., approximately a 5.20% difference. Achieving effective separation under these conditions demands the precise and accurate design and synthesis of the molecular structure of polymer membranes.[23] Moreover, to further establish membranes in the gas separation industry, developing

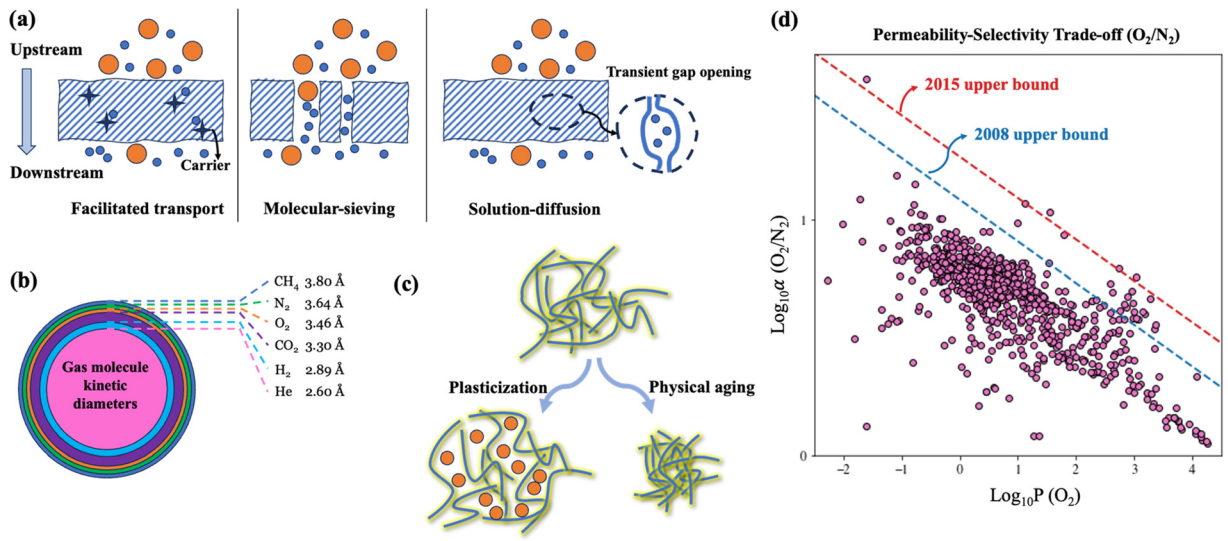

FIG. 1. (a) Three common gas transport mechanisms in membranes: facilitated transport, molecular-sieving, and solution-diffusion. (b) Kinetic diameters of common gas molecules in separation. (c) Schematics of plasticization and physical aging of polymer membranes. (d) Permeability-selectivity trade-off of O_2/N_2 data from the MSA database.[26] X-axis is the permeability of O_2 [$P(O_2)$], in a unit of log Barrer, and the y-axis is the selectivity of O_2 over N_2 [$\alpha(O_2/N_2) = P(O_2)/P(N_2)$]. Each dot represents one polymer. The blue dashed line is the 2008 upper bound, and the red dashed line is the 2015 upper bound.

tougher and higher-performance materials capable of withstanding harsh chemical environments is needed.[8] While polymer membranes have gained applications in gas separation, they still face three major challenges—permeability-selectivity trade-off, plasticization, and physical aging—that hinder broader applications [Figs. 1(c) and 1(d)].

The trade-off between permeability and selectivity, often restrained by the "upper bound," limits the performance of polymeric membranes. It was first reported by Robeson in 1991[28] and later theoretically validated by Freeman.[29,30] As shown in Fig. 1(d), this upper bound (see dashed lines), represented by a log–log plot of selectivity against permeability, has seen shifts over the years (2008, 2015, and 2019[31,32]), reflecting advancements in material discovery. However, these upper bound curves are derived based on pure gas permeation data and do not account for factors like plasticization and competitive sorption[33] under certain feed conditions, which can significantly influence permeability and selectivity.

Plasticization is another common issue when polymer membranes come into contact with highly soluble gases.[34–37] This leads to increased chain mobility and free volume due to the swelling of polymers, thus improving the diffusion coefficient and reducing the selectivity [Fig. 1(c)]. Among the gases of interest in gas separations, highly condensable gases like CO_2, H_2S, and C_3H_6, known for their high solubility, are frequently studied in relation to plasticization.[10,37,38] Other non-CO_2 causes of plasticization include impurities in the feed, conditioning effects, highly polymer-soluble penetrants.[39–41] A promising solution to this issue involves refining the microstructure of the polymer through methods like cross-linking, thermal rearrangement, carbonization, and functionalization.[42–44]

Physical aging occurs as the nonequilibrium polymer chains relax toward equilibrium, resulting in reduced free volume and lower permeability [Fig. 1(c)].[45] Most of the polymers currently used in gas separations are glassy polymers,[46] which have surplus free volume because of their non-equilibrium status.[45] The microstructure rearrangement toward equilibrium due to local segment motions decreases their free volume over time and increases their density. It also modifies the specific volume, enthalpy, entropy, and other physical properties of polymers.[47] Ultimately, physical aging causes a decrease in permeability and, in turn, an increase in selectivity. The loss in permeability and increase in density resulting from physical aging can be reversed by elevating the temperature of the polymer above its glass transition temperature (T_g).[31,48]

Various approaches, including polymer blends,[49] thermal methods,[50] UV cross-linking,[51,52] introduction of nanoparticles,[38,53] and the manipulation of macromolecular design,[7,31,54] are employed to control the free volume in polymers to address these challenges. For example, incorporating triptycene and pentiptycene into the backbone of the polyimides and polysulfones is useful to control the free volume in glassy polymers.[31,54] Other emerging polymers, such as the thermally rearranged (TR) polymers and polymers of intrinsic microporosity (PIMs), have also been studied extensively for gas separation because of their excellent performance.[55,56] However, PIMs exhibit a rapid performance loss due to physical aging, while TR polymers suffer from poor mechanical properties.[3,57] Mixed-matrix membranes (MMMs) represent another thriving field for gas separation within the polymer domain. These hybrid materials integrate high-performing inorganic materials, such as metals, zeolites, and MOFs, into polymers and can capitalize on the processability of polymeric membranes while addressing challenges linked to the selectivity-permeability trade-off, concurrently addressing the processing and fabrication difficulties associated with high-performing inorganic materials. Outperforming pristine polymers, MMMs owe their enhanced performance to the precise pore sizes and exact shape and geometry of the inorganic fillers, making them exceptional molecular sieves.[8] However, MMMs are yet to find commercial applications in gas separations due to persistent issues such as interfacial imperfections and the complexity of preparing thin, defect-free MMMs.[58]

Despite these progresses in polymeric membranes, meeting the demand for energy-efficient and sustainable gas separation requires continued innovation. This entails the precise design of polymers with exceptional separation performance. Traditional approaches, relying heavily on trial-and-error experiments, have time, cost, and efficiency limitations. Incorporating theoretical and empirical models can expedite the discovery and design of novel membrane materials. Researchers increasingly rely on sophisticated modeling techniques rooted in theoretical physics to gain insights into the underlying principles governing membrane performance, such as the group contribution-based methods and the graph theoretical approach.[17,59–66] These models allow for a deeper understanding of the intricate interplay of different factors influencing separation processes, guiding the design of membranes with improved selectivity and efficiency while saving development time. The amount of data accumulated from traditional experiments and computational and theoretical models has given rise to the adoption of machine learning (ML) techniques. The rapid development of ML has transformed various industries and scientific disciplines,[67,68] including materials science, where it is used to speed up the discovery and optimization of new materials.[69,70]

Within the dynamic landscape of materials science and ML, the emergence of polymer informatics stands out as a thriving field. Polymer informatics, using ML to uncover structure-property relationships, has grown rapidly with advances in polymer datasets and ML algorithms.[71–81] While ML can significantly speed up the exploration of vast chemical and structural spaces, its potential in designing and optimizing polymeric materials for specific applications, such as gas separation polymeric membranes, is still in its early stages. This review delves into different aspects of ML applications in polymer membranes for gas separation. We focus on three crucial components: polymer data, representation methods, and ML algorithms. By providing a comprehensive overview of the current landscape, challenges, and opportunities, this review seeks to advance polymer informatics for gas separation membranes and related domains.

II. MACHINE LEARNING FOR POLYMERIC GAS SEPARATION MEMBRANE DESIGN AND DISCOVERY

This section summarizes ML application for polymeric gas separation membranes based on the three primary components: data, representation, and algorithms. As illustrated in Fig. 2, these elements form a cohesive framework for applying ML to material design. Data are the foundational input, essential for training and validating the ML models. Polymer representations encapsulate intrinsic structural, chemical, and additional information into high-dimensional vectors, enabling ML to map them to the target properties (e.g., permeability). ML algorithms serve as the computational engine, uncovering hidden patterns and insights in polymer data to accelerate membrane design and discovery, and can further boost the acquisition of new data, creating an iterative process to enhance model accuracy and generalizability or reach certain design targets.

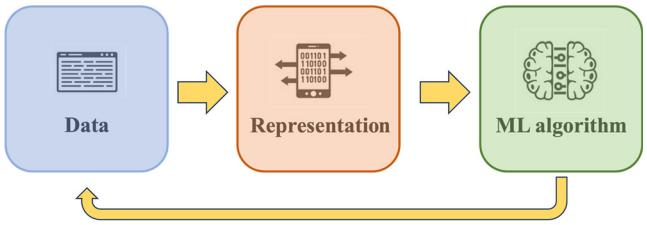

FIG. 2. Workflow of ML for polymeric gas separation membrane design and discovery, which is also used for materials in general.

A. Data

The bedrock of ML lies in the quality and quantity of the data it is trained on. In this subsection, we will discuss open-source or published datasets relevant to polymers, specifically related to gas separation. The data can be broadly categorized into experimental, computational, and synthetic datasets, depending on the annotation methods and how polymers are created. These datasets typically consist of labeled data (usually experimental and computational), which supports supervised learning, as well as unlabeled data (usually experimental and synthetic), which is valuable for unsupervised learning. Unlabeled data include synthesized polymers that are untested for gas permeability and synthetic polymer structures not yet verified. A summary of the databases covered in this section is listed in Table I.

1. Experimental datasets

PoLyInfo, a major polymer database,[82] offers diverse data for polymer properties. It includes around 29 000 polymers, including homopolymers (~18 700), copolymers (~7700), and polymer blends (~2500), with the primary data source being experiments documented in the academic literature. It covers information such as properties, chemical structures, processing methods, measurement conditions, monomers, and polymerization methods. It includes about 100 different kinds of properties, spanning from thermal and electrical to mechanical properties. Gas separation properties, such as permeability, solubility, and diffusivity coefficients for major industrial gases, are provided for ~1300 polymers (~800 are homopolymers, ~400 are copolymers, and ~100 are polymer blends). While PoLyInfo serves as a valuable resource for both labeled and unlabeled data in gas separation ML tasks and other polymer informatics applications, accessibility of the data remains a challenge due to its restrictions on data acquisition, limiting users to only web-based inquiries. Moreover, meticulous data cleaning is necessary for information obtained from PoLyInfo, including aggregating data for polymers with multiple entries from different sources, which can have significant variations.

Specifically for gas separation polymer applications, the Membrane Society of Australasia (MSA) provides a dataset, which can be used for ML tasks, known as the Polymer Gas Separation Membrane Database.[26] This database offers experimental data for ~1500 polymers spanning the years 1950 to 2018. Each major gas category (e.g., He, H_2, O_2, N_2, CH_4, and CO_2) has around 400–800 polymers with permeability coefficient records, though many have missing values. Additionally, other gases (e.g., C_2H_4, C_2H_6, C_3H_6, C_3H_8, C_4H_8, n-C_4H_{10}, CF_4, C_2F_6, and C_3F_8) are also represented. The dataset covers diverse membrane materials, including rubber and glassy polymers, CMS, and zeolites. Accessibility is facilitated through a directly downloadable table format, making it a good source of labeled data for model training.[91] However, it has several drawbacks. First, it has not been updated since 2018, excluding many high-performance polymers for gas separation developed in recent years, especially ladder polymers.[54,55,92,93] Second, polymer structural information is missing, requiring manual effort to obtain the structures for each polymer. Finally, data cleaning is needed, as homopolymers, copolymers, and composite materials are intermixed, and some polymers have multiple entries due to variations in synthesis and testing procedures (e.g., thermal treatment temperatures and aging times) across different data sources and labs. Unfortunately, these differences are not well-documented, posing a challenge for researchers in ensuring the accuracy and consistency of the dataset. Some works have expanded the MSA database by adding SMILES and additional polymer permeability data, enhancing its suitability for direct use in ML applications.[94,95]

In addition to the PoLyInfo and MSA, we introduce several other databases that offer extensive information on polymers and their properties. Like PoLyInfo, CHEMnetBASE[96] provides a database of detailed scientific and commercial information on over 1100 polymers, including transport properties like gas permeabilities. The Polymer Property Predictor and Database[97] offers predictions and experimental data on polymer properties such as Flory-Huggins Chi (χ), glass transition temperature (T_g), and binary polymer solution cloud point, which are useful for exploring polymer–polymer and polymer–solvent systems. The Materials Data Facility[98] is a platform for hosting and sharing materials data, including polymers. The Polymer Genome[76] is an ML-based platform for predicting polymer properties and retrosynthesis planning. However, the experimental and computational data used to train the models are inaccessible through the Polymer Genome website.

2. Computational datasets

While experimental datasets are invaluable, their limitations, such as relatively small volumes (e.g., the MSA database collected ~1500 polymers' permeability information spanning around 70 years of effort in the literature) and uncertainties from varied experimental conditions, highlight the need for more standardized datasets. High-throughput computational simulation, such as molecular dynamics (MD) simulations, offers a reliable alternative to experimental data, as demonstrated in recent polymer informatics studies.[83,86,99–101]

A general workflow for calculating polymer properties using MD simulations is shown in Fig. 3(a). Using linear homopolymer as an example, the process starts by extracting structural information from a database to obtain the monomer structure, which serves as the building block of the polymer. Subsequently, the monomer is polymerized to the desired chain length, ensuring a sufficiently long polymer chain for simulation. The polymer chain is then replicated to generate an amorphous system, followed by optimization to obtain a more realistic polymer conformation, often verified by reproducing known properties such as density. In the final step, the optimized system undergoes a production run to calculate the target properties. The calculated properties are fed back into the database, gradually building a standardized resource for ML applications. However, this general workflow requires adjustments for other specific types of polymers. For instance, some glassy polymers like PIMs may need additional equilibration, such as the 21-step compression/relaxation scheme described by

TABLE I. Summary of datasets related to polymer membranes for gas separation.

Source	Properties				Data type			With SMILES	No. polymers	Note	Reference
	Solubility	Diffusivity	Permeability	FFV	Experimental	Computational	Hypothetical				
PoLyInfo	■	■	■		■			■	∼29 000	∼1300 polymers are related to gas separation	82
MSA			■		■				∼1500	Many missing values	26
Wang *et al.*				■		■			1683	Not openly available	83
Tao *et al.*				■		■		■	∼7900	6500 homopolymers and 1400 polyamides	84
PI1M							■	■	∼1 m	Trained on PoLyInfo database	85
Yang *et al.*							■	■	∼8 m	Include 1100 ladder polymers	86
OMG							■		∼12 m	17 polymerization rules	87
SMiPoly							■	■	169 347	19 monomer classes and 22 polymerization rules	88
Tiwari *et al.*							■	■	∼14.5 m	Seven types of polymer backbones and two small organic molecule datasets	89
Polymer expert							■		NA	Integrated in MedeA software	90

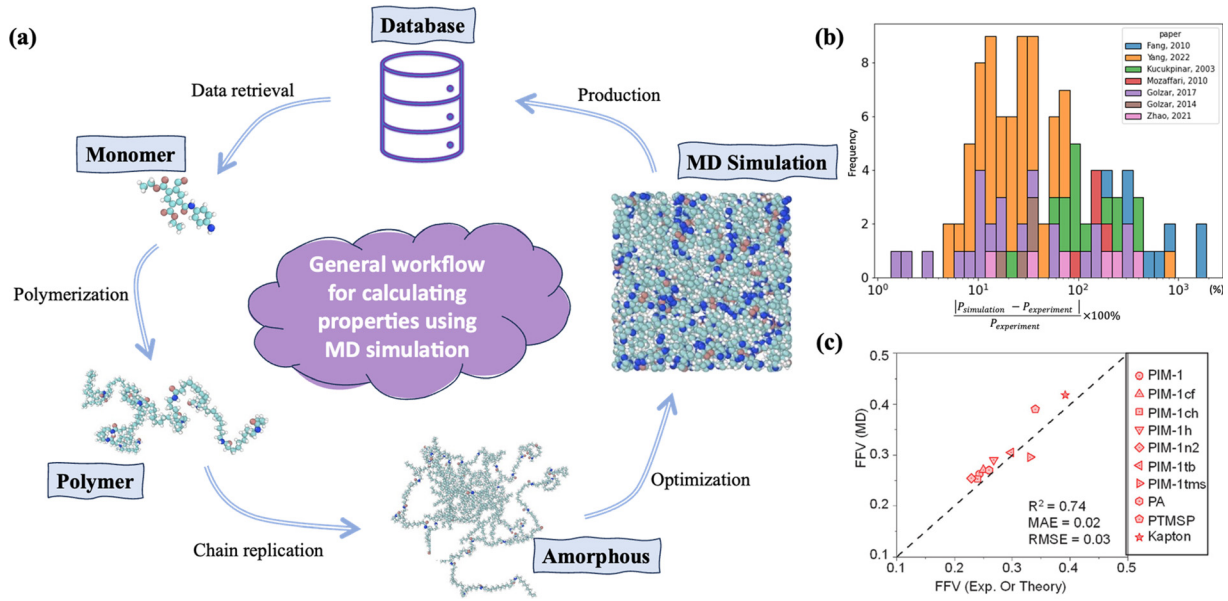

FIG. 3. A general workflow for calculating polymer properties using MD simulations and related works on gas separation polymer membranes. (a) The general workflow of building computational database of amorphous polymer properties using MD simulations. (b) Histogram of the absolute percentage error ($|P_{simulation} - P_{experiment}|/P_{experiment} \times 100\%$) collected and calculated from various studies on computational simulation of polymer permeabilities of different gases.[86,103–107] $P_{simulation}$ is the calculated permeabilities using MD simulation from different works, and $P_{experiment}$ is experimentally measured permeabilities compared in these works. (c) MD simulation results of polymer FFV from Tao *et al.* benchmarked against literature data points. Reproduced with permission from Tao *et al.*, J. Membr. Sci. **665**, 121131 (2023). Copyright 2023 Elsevier B.V.

Larsen *et al.*[102] For ladder polymers, the polymerization step must account for the double-stranded backbone. For copolymers, the polymerization step needs to include multiple monomer types in a specified sequence or ratio, such as random, block, or alternating. For polymer blends, an additional step is required to mix different polymer chains at the appropriate ratios, followed by optimization to ensure proper mixing and interaction between the different species.

However, MD simulations are computationally expensive for large polymer systems and gas permeability calculations, limiting their scalability for large gas permeability datasets. From a computational cost perspective, the workflow indicates that the calculation of permeability using MD simulation involves two key steps: (i) constructing and optimizing an amorphous polymer system and (ii) running property calculation, including solubility and diffusivity [Eq. (1)]. The computational burden arises from the need for a large polymer system, with sufficient atoms (considering chain length and the number of chains) to ensure an adequate representation of the real polymer, and the extended time required for optimization and property calculations. Furthermore, the accuracy of gas permeability calculated through simulations remains a challenge. Substantial uncertainties are inherent in simulations due to the randomness involved in the calculation process, such as initialization, polymerization process, gas molecule insertion procedures, and other parameters (e.g., molecular interaction potential) in the simulation setup. These factors, together with the propagation of uncertainty in Eq. (1), collectively contribute to the difficulty of accurately calculating gas permeabilities through computations. Figure 3(b) illustrates the absolute percentage error distribution across polymer permeability simulation.[86,103–107] Notably, over 86% of the simulations exhibit an absolute percentage error exceeding 10%, with more than 25% surpassing 100%. This underscores substantial computational uncertainties in comparison with experimental measurements. In addition to MD simulation, Monte Carlo simulations and transition-state theory are also utilized in the calculation of diffusivity, solubility, and permeability based on their effectiveness with different polymer and gas characteristics. The choice of forcefield, like PCFF (Polymer Consistent Force Field),[108] GAFF (General Amber Force Field),[109] and COMPASS (Condensed-phase Optimized Molecular Potentials for Atomistic Simulation Studies),[110] is also crucial to reasonably reflect the specific properties of the polymers under investigation.

In addition to permeability coefficients, fractional free volume (FFV) is an important feature that characterizes polymer microstructure and influences separation properties. The free volumes in polymer membranes are considered chain cavities resulting from chain packing, serving as diffusion paths for gas transport through the membrane. The FFV is used to quantify the ratio of free volume in polymer and is defined as

$$FFV = \frac{V - V_o}{V}, \tag{2}$$

where V is the specific volume of the polymer and V_o is the volume occupied by the polymer chains.[111–113] Determining FFV is more computationally feasible than permeability due to the absence of steps for calculating solubility and diffusivity, as well as considering interactions between gas molecules and the polymer matrix. This enables the development of relatively large simulation datasets for FFV of polymers. For example, Wang *et al.*[83] gathered 66 FFV data points from the literature

calculated using a combination of experimental data and Bondi's group contribution methods.[114] They then employed MD simulations to calculate the FFV of 1683 polymers, with verified MD protocols based on the 66 literature data. While the 66 data are accessible in their paper, the 1683 MD data are not openly available. After that, Tao *et al.*[84] utilized MD simulations, benchmarked against ten literature data points [as shown in Fig. 3(c)], to simulate over 6500 homopolymers and calculate the FFV. Additionally, they simulated over 1400 cross-linked polyamide systems using a multi-step cross-linking strategy.[115] Both the datasets of homopolymers and polyamides are openly accessible, providing a valuable supplementary source for ML models in the discovery of polymer membranes with high separation performance.

3. Synthetic datasets

In addition to already synthesized polymers, synthetic polymer structures are also valuable for ML screening in gas separation. Five representative works on synthetic polymer generation are illustrated in Figs. 4(a)–4(c). Ma *et al.*[85] introduced the PI1M database [Fig. 4(a)], a synthetic polymer dataset of approximately 1×10^6 entries. Learned from the SMILES (see Sec. II B Representation for more details on SMILES) of existing polymers in PoLyInfo[82] using a recurrent neural networks (RNN),[116] PI1M covers a chemical space similar to PoLyInfo but significantly populates the regions where the PoLyInfo data are sparse. However, the synthesizability of the synthetic polymers is not ensured, limiting their practical applicability in real-world applications.

Yang *et al.*[86] crafted two synthetic datasets tailored for membrane separation [Fig. 4(b)]. One dataset comprises around 8×10^6 synthetic polyimides, formed through the polycondensation of known dianhydride and diamine/diisocyanate pairs from PubChem.[117] The other one contains about 1100 synthetic ladder polymers by combining components from existing ladder polymers. There are also other rule-based generated polymers to address the difficulty of identifying synthetic routes for generated polymers, e.g., the Open Macromolecular Genome (OMG) dataset [Fig. 4(d)][87] and the SmiPoly dataset [Fig. 4(c)].[88] Both predefine a set of polymerization reaction rules to generate potentially synthesizable polymers. Similarly, Tiwari *et al.*[89] developed a simple backbone-fragment combination method to computationally construct polymers [Fig. 4(c)], generating 14 datasets from seven types of predefined polymer backbones and two small organic molecule datasets (the MOSES and the QM9 datasets). Polymer Expert[90] is another tool that is developed to generate novel polymers by starting with an initial set of existing polymer units and expanding the design space through systematic substitution of hydrogen with other fragments. This module is integrated into the MedeA software suite by Materials Design, Inc., to facilitate computational simulations. However, it is not open source.

B. Representation

Polymer representation is another critical aspect of ML-enabled polymer discovery and design, transforming intricate polymer

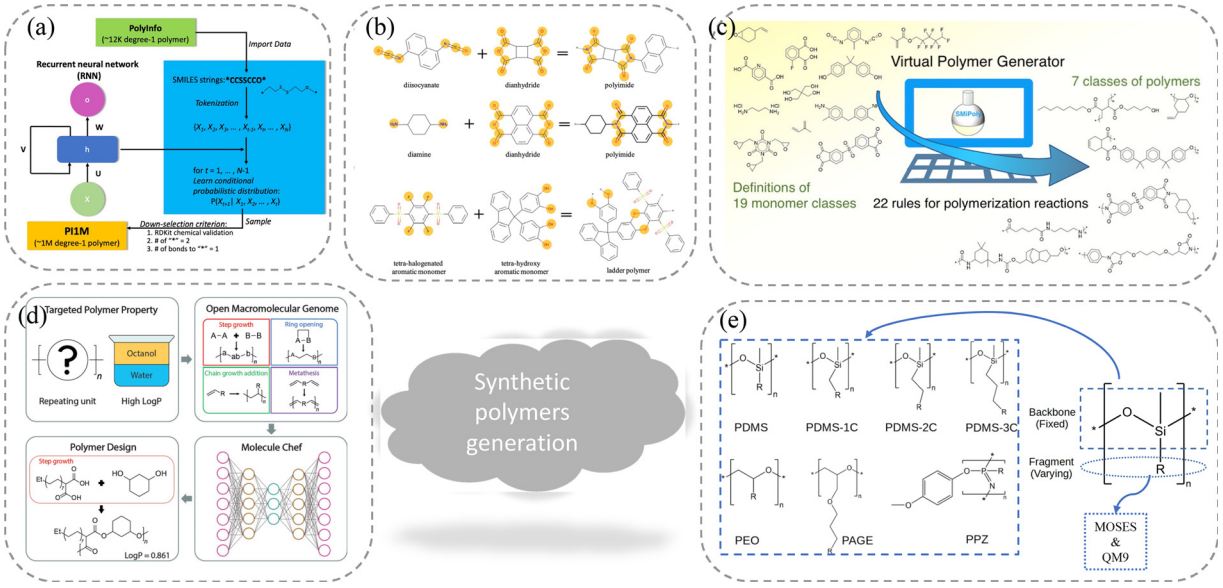

FIG. 4. Various methods for synthetic polymer generation for gas separation. (a) Creation of PI1M database, which consists of 1×10^6 synthetic polymers generated by an RNN trained on PoLyInfo polymer data. Reprinted with permission from Ma and Lo, J. Chem. Inf. Model. **60**, 4684–4690 (2020). Copyright 2020 American Chemical Society. (b) Synthetic polyimides and ladder polymers datasets tailored for membrane separation purposes. Polyimides are formed through the polycondensation of known dianhydride and diamine/diisocyanate pairs and ladder polymers are constructed by combining components from existing ladder polymers. Reprinted with permission from Yang *et al.*, Sci. Adv. **8**, eabn9545 (2022). Copyright 2022 Authors, licensed under a CC BY-NC 4.0 License/AAAS. (c) SmiPoly dataset creation which is based on 19 preselected monomer classes and 22 predefined polymerization rules. Reprinted with permission from Ohno *et al.*, J. Chem. Inf. Model. **63**, 5539–5548 (2023). Copyright 2023 Authors, licensed under a Creative Commons Attribution (CC BY) License. (d) Open Macromolecular Genome (OMG) dataset creation, which is based on four predefined polymerization rules (17 sub-rules). Reprinted (adapted) with permission from Kim *et al.*, ACS Polym. Au **3**, 318–330 (2023). Copyright 2023 American Chemical Society. (e) Synthetic polymer generation based on backbone-fragment combination using seven types of predefined polymer backbones and two small organic molecule datasets. Reprinted with permission from Tiwari *et al.*, J. Chem. Inf. Model. **63**, 5539–5548 (2024). Copyright 2024 American Chemical Society.

structures and other determining factors of polymer properties into simplified descriptors or high-dimensional vectors for tasks such as property prediction, similarity searching, and inverse design. The general workflow of polymer representation is summarized in Fig. 5. Here, we broadly classify polymer representation into two levels: chemistry-level and processing-level. The former focuses on the chemical structural information of the individual repeating units, while the latter typically includes parameters during processing, fabrication, and testing.

1. Chemistry-level polymer representation

The evolution of chemistry-level representation has seen a transition from traditional descriptors and fingerprints, typically predefined and offering a static representation of polymer chemistry, to techniques such as graph-based and language-based embeddings, which often dynamically learn to represent polymer chemistry in a supervised or self-supervised manner. This evolution signifies a shift in how polymers are analyzed and understood as ML techniques continue to progress. In this article, "representation" refers broadly to numerical vectors describing structural and/or compositional information of polymers. Representations are further categorized as "descriptor," "fingerprint," and "embedding." The former two are derived from static methods, while the latter is from dynamic learning.

Before generating representations of polymer structures, it is important to note that, for storage, retrieval, and identification, polymer structures are commonly represented as string notations. SMILES (Simplified Molecular Input Line Entry System) is the most popular method, as shown in Figs. 5(a) and 5(b), utilizing a simplified molecular string-line notation based on the principles of a molecular graph. SMILES represents a molecule as a sequence of characters, including letters, numbers, and symbols, each denoting specific atoms, bonds, and structures.[118] For polymers, SMILES often uses asterisks (*) to mark the connection points of repeat units, adapting the notation to represent the repeating nature of polymers. In the meantime, alternative string-based representation methods have been proposed to address some of its limitations. For instance, SELFIES (Self-referencing Embedded Strings)[119] is developed to overcome the issue of invalid or physically impossible molecule representation in SMILES [Fig. 5(a)]. Every SELFIES string can be converted into a valid molecular graph by utilizing a context-free grammar, providing a robust molecular representation. Another variant is BigSMILES,[120] specifically designed to represent larger and more complex polymers and macromolecules [Fig. 5(b)]. It introduces additional syntax elements to handle repeating units, copolymers, and other features typical of large molecules that are not addressed in the standard SMILES notation. Despite these advancements, SMILES remains widely used due to its simplicity and compatibility with open-source ML tools like RDKit.[121] The subsequent discussion in this section is based on the SMILES notation for polymers.

Once represented by SMILES—whether as monomers, single repeated units, or n-mers[122]—various techniques can translate SMILES into polymer representations, using either traditional descriptors and fingerprints [Fig. 5(c)] or learned embeddings [Fig. 5(d)]. It should be noted that ML model accuracy shows a convergence pattern as a function of n, where n is the number of monomers used.[89,122]

a. Descriptors and fingerprints. Traditional polymer descriptors and fingerprints play an important role in characterizing and

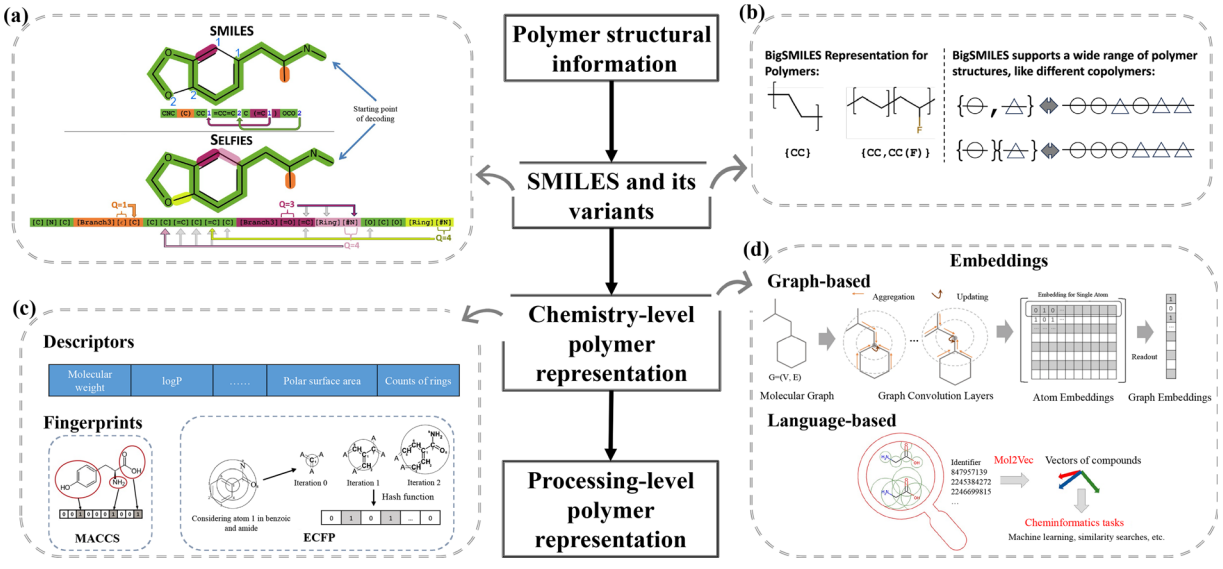

FIG. 5. General workflow of polymer representation. (a) Comparison between SMILES and SELFIES notations for expressing molecule structural information. Reproduced with permission from Krenn *et al.*, Mach. Learn. Sci. Technol. **1**, 045024 (2020). Copyright 2020 Authors, licensed under a Creative Commons Attribution 4.0 license. (b) BigSMILES illustration for representing larger and more complex polymer and macromolecular systems. (c) Examples of descriptors and fingerprints (MACCS and ECFP) to vectorize chemistry-level polymer information. Reprinted with permission from D. Rogers and M. Hahn, J. Chem. Inf. Model. **50**, 742–754 (2010). Copyright 2010 American Chemical Society. (d) Examples of graph-based and language-based (Mol2vec) embeddings to vectorize chemistry-level molecular information. Reproduced with permission from Jiang *et al.*, J. Cheminform **13**, 12 (2021). Copyright 2021 Authors, licensed under a Creative Commons Attribution 4.0 International License. Reprinted with permission from Jaeger *et al.*, J. Chem. Inf. Model. **58**, 27–35 (2018). Copyright 2018 American Chemical Society.

representing the chemical structures in cheminformatics and computational chemistry.[123,124] These methods reduce complex molecular structures into numerical or binary representations. Descriptors encompass a diverse set of quantitative parameters that capture the structural, topological, electronic, or thermodynamic properties of a polymer. Examples include molecular weight, logP (partition coefficient), polar surface area, and various connectivity indices.[123,125] On the other hand, molecular fingerprints are numerical or binary strings encoding the presence or absence of specific substructures. Each bit in a fingerprint corresponds to a predefined fragment or pattern, enabling efficient comparison and screening of large datasets. Commonly used fingerprints, as illustrated in Fig. 5(c), include the Molecular Access System (MACCS) keys,[126] the Morgan fingerprints (also known as Extended Connectivity Fingerprints, ECFP),[127] Daylight Fingerprints,[128] Topological Torsion Fingerprints,[129] etc.

b. Embeddings. These traditional approaches reply on predefined chemical rules to calculate descriptors or fingerprints. However, with the advancement of ML, particularly deep learning, there is a shift toward dynamically learning polymer representations directly from data, known as embeddings. Two prevalent methods for learning polymer embeddings are graph-based and language-based approaches [Fig. 5(d)].

Graph-based representations leverage the inherent structure of polymers, treating them as graphs where atoms serve as nodes and bonds as edges. This intuitive representation aligns well with graph ML techniques, like Graph Neural Networks (GNNs),[130] including Graph Convolutional Networks (GCN)[131] and Graph Isomorphism Networks (GIN).[132] By adopting this graph-centric perspective, powerful tools from network theory can be harnessed to learn embeddings that prove useful in predicting polymer properties[133–138] and generating novel polymer graph structures.[139]

On the other hand, language-based representations capitalize on the fact that polymers are initially stored in a SMILES format, which can be regarded as a "chemical language." The SMILES notation, with its encoded chemical rules or grammars, naturally lends itself to learning representations through Natural Language Processing (NLP)-based algorithms. In this approach, the chemical structure of polymers is treated as a "sentence," and functional groups or substructures are analogous to "words." NLP techniques, including sequence-to-sequence models[140] and transformers,[141] demonstrate proficiency in capturing sequential and contextual information in polymer structures, exemplified by the Polymer Embedding (PE),[85] which is based on Mol2Vec [Fig. 5(d)], an unsupervised learning approach to learn the representations of molecular substructures.[142] Moreover, they are also proven to be ideal for polymer generative tasks,[85] similar to graph-based representations. This linguistic analogy opens new avenues in polymer informatics, providing an effective tool for the analysis of large datasets of polymers using advanced language models.

However, the current chemistry-level representations of polymers still face challenges. First, using only the single repeating unit as input, they omit the degree of polymerization, which is a crucial factor affecting polymer properties. For example, polymers with identical composition usually exhibit higher melting temperatures when the degree of polymerization increases.[143] Second, there is no information describing polymer synthesis, fabrication, and measurement, which can influence the conformation and, hence, the properties of polymer materials. Interested readers are referred to Refs. 85, 122, and 144 for a detailed comparison between different polymer representation methods in polymer informatics tasks like density, glass transition temperature, and melting temperature.

2. Processing-level representation

Beyond chemistry-level representations, research also examines processing parameters like measurement and fabrication factors and their impact on membrane gas separation performance. Measurement-level parameters consider factors like temperature, pressures (on the feed side, permeate side, partial pressure for mixed gas tests), and feed gas flux. It typically focuses on a specific polymer, disregarding variations in polymer chemistry, and is crucial in membrane process design and control, as the membrane with the same chemistry and fabrication process can have different measured permeabilities under varying measurement conditions.[145,146] As for fabrication-level parameters, they primarily aim to optimize the membrane preparation step to achieve maximum performance for a given polymer, predominantly concentrating on polymer-based composite membrane materials. Key parameters include filler type and loading, type of solvent or boiling point of the solvent, concentration of the membrane casting solution, crosslinker concentration, catalyst concentration, membrane thickness, stirring time, synthesis time, and temperature.[147–152] Interested readers looking for more on processing-level representations of polymer membranes in gas separation are encouraged to refer to the review by Ricci *et al.*[81]

Nevertheless, establishing an extensive database of processing-level representations for polymer membranes is challenging, limiting these features to narrow datasets focused on one or a few polymers. This difficulty arises from the inconsistencies in experimental data due to varying fabrication techniques and measurement conditions across laboratories. The absence of a standardized protocol for recording information on polymer membrane fabrication and measurement is the primary cause of variations, complicating the integration of processing-level and chemistry-level representations. This hinders the enhancement of polymer membrane representation and subsequently limits the accuracy of ML predictions.

C. Machine learning algorithms

ML has revolutionized data analysis, prediction, and automated decision-making, becoming a powerful tool for the design and discovery of polymer materials for gas separation. In this subsection, as outlined in Fig. 6, we first briefly review fundamental ML concepts, including supervised and unsupervised learning, along with representative algorithms. Additionally, we summarize their applications in gas-separation polymer membranes. Given the scope of this review, for a more detailed introduction to fundamental ML techniques and their broader applications in materials and polymers, readers are encouraged to explore comprehensive reviews on polymer informatics, such as Refs. 69 and 153–155. Subsequently, we discuss advanced ML techniques and concepts tailored to specific challenges in polymer membranes for gas separation, i.e., accurate screening and inverse design. We focus on improving the model screening accuracy from data, model, and interpretability aspects. We also provide a summary of works (in Table II) that employ these techniques to navigate the extensive chemical and structural landscape of polymer membranes for gas separation.

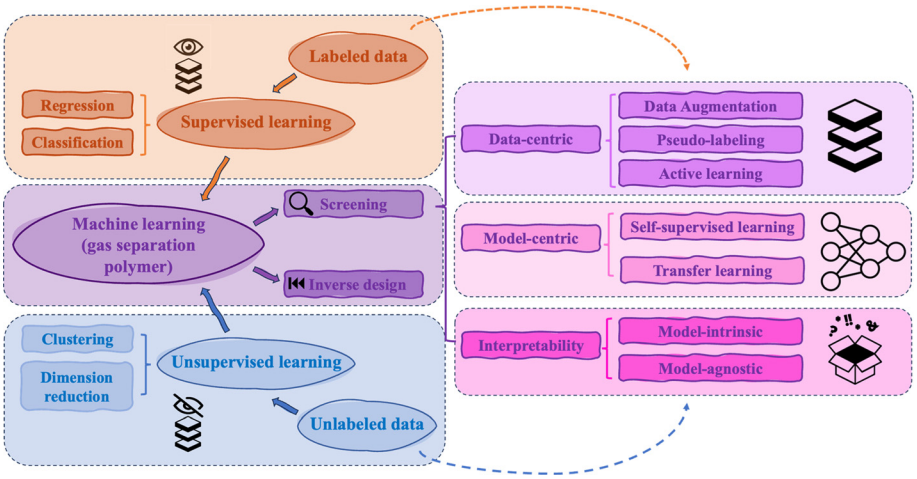

FIG. 6. An overview of fundamental machine learning (ML) concepts, encompassing supervised and unsupervised learning, alongside advanced ML techniques tailored to address specific challenges in polymer membranes for gas separation, including accurate screening and inverse design. Three critical aspects for accurate screening are highlighted: data-centric methods, model-centric methods, and model interpretability.

1. Traditional machine learning and applications

Traditional ML algorithms are usually categorized into supervised and unsupervised learning. Supervised learning entails mapping inputs to outputs based on labeled input-output pairs, commonly known as labeled data.[156] The model is trained to minimize a loss function, aligning the accuracy of predicted values with ground-truth values. Classification and regression are two key approaches in supervised learning. Classification predicts categorical values of the output, employing algorithms like logistic regression, decision trees, random forest (RF), neural networks (NN), and support vector machines.[157] Regression predicts continuous output quantities using algorithms such as Gaussian processes (GP), RF, and NN.[158] Due to the typical target output in gas separation being the permeability—a continuous variable—regression is the convention for ML in this context.

One of the pioneering works in this domain dates back to 1994. Wessling *et al.*[159] utilized an NN to model CO_2 permeability in 33 glassy polymers, using the infrared spectra of polymers as the input feature. Despite the limited data size, they achieved accurate predictions, highlighting the potential of ML in the QSPR analysis for polymeric membrane gas separation materials. Subsequent research endeavors have incorporated more labeled data, advanced polymer representations, and sophisticated ML algorithms to achieve enhanced prediction performance. For instance, Hasnaoui *et al.*[160] employed an NN to predict O_2, N_2, CO_2, and CH_4 permeability in 149 polymers, using 20 group contribution descriptors from Yampolskii *et al.*[170] and temperature as inputs. Zhu *et al.*[161] employed GP regression to predict permeability for He, H_2, O_2, N_2, CO_2, and CH_4, along with ideal selectivity for a dataset of 315 polymers. They used a hierarchical fingerprinting method, requiring only knowledge of the chemical structure of the polymer repeating unit (e.g., SMILES), as input features. Testing the model on a holdout dataset of 31 polymers revealed good performance on the major polymer classes and larger deviations for polymers from underrepresented classes. Following a similar procedure as Zhu *et al.*[161] Barnett *et al.*[162] utilized GP regression to create a gas permeability ML model, incorporating a topological, path-based fingerprint of the polymer repeating unit as the input feature. The model, trained on data for six gases and around 700 polymers, was used to predict the permeability values of around 11 000 unlabeled polymers. Experimental validation of two promising candidates for CO_2-related separation substantiated the alignment between the ML model predictions and real-world outcomes, demonstrating the great potential of ML to guide and inform experimental efforts in polymer research. Zhao *et al.*[163] developed a model for predicting gas separation performance with a specific focus on polyimide membranes using NN and the repeat unit structure. The dataset included 125 polyimides, and 20 descriptors were calculated using Yampolskii's group contribution fragments.[170] The model showed promising results on CO_2 permeability prediction, demonstrating applicability to polyimides and copolyimides. The model's simplicity makes it a valuable tool for guiding polyimide synthesis and structure screening. Meanwhile, selectivity is another important target variable defining membrane performance. A common practice is to derive selectivity from predicted permeabilities. However, Tiwari *et al.*[89] found that directly predicting selectivity yields greater accuracy by avoiding issues with uncertainty propagation.

In addition to using fingerprints or descriptors for polymer representation, more recently, researchers utilized representation learning from deep neural networks to learn the embeddings of polymers. For example, Wilson *et al.*[166] treated polymer structures as graphs and developed a multioutput GNN named PolyID to facilitate the efficient identification of high-performance bio-based polymers. PolyID enabled the discovery of bio-based poly(ethylene terephthalate) (PET) analogs with enhanced thermal and gas separation performance through the screening of over 22 000 polyester candidates. While one of the PET replacements was experimentally synthesized to validate the model performance, the validation focused solely on T_g prediction, with no validation for gas permeability prediction.

Unlike supervised learning, unsupervised learning deals with the input features without corresponding labels, also known as unlabeled data, making it suitable for clustering, dimensionality reduction, and associative rule mining.[171] Algorithms, such as K-means clustering, principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and autoencoders, are commonly used on unlabeled data. In polymeric gas separation membranes, these techniques are often used to visualize polymer representations.[171,172] For instance, Yang *et al.*[86] used UMAP and Wilson *et al.*[166] used PCA to visualize the chemical space (representation space) of their labeled and/or unlabeled polymer

TABLE II. Summary of chemistry-level ML works on polymer membranes for gas separation.

Work	Data						Generative	No. training data	Representation	ML model	Note	Reference
	He	H_2	O_2	N_2	CO_2	CH_4						
Wessling *et al.*					■			33	Infrared spectra	NN		159
Hasnaoui *et al.*			■	■	■	■		149	Group contribution descriptors, temperature	NN		160
Zhu *et al.*	■	■	■	■	■	■		315	Hierarchical fingerprint	GP		161
Barnett *et al.*	■	■	■	■	■	■		∼700	Topological, path-based fingerprint	GP	Successful experimental validation of two promising candidates	162
Zhao *et al.*					■			125	Group contribution descriptors	NN	Focus on polyimides	163
Yuan *et al.*	■	■	■	■	■	■		1378	N/A	MICE	Missing value imputation	164
Yang *et al.*	■	■	■	■	■	■		778	Morgan fingerprint with frequency	MICE, NN, RF		86
Liu *et al.*			■					595	Graph-based embedding	GNN	Data augmentation through rationale identification	138
Liu *et al.*			■				■	595	Graph-based embedding	GNN, Diffusion	Data augmentation through generative model	139
Liu *et al.*			■					595	Graph-based embedding	GNN	Pseudo-labeling	165
Wilson *et al.*			■					∼250	Graph-based embedding	GNN	Focus on biobased polymers	166
Kuenneth *et al.*	■	■	■	■	■	■		∼300–600	Language-based embedding	DeBERTa	Self-supervised learning, including co-polymers	167
Giro *et al.*					■		■	1169	Topological, geometrical and structural fingerprint	RF, SVM, etc.	MD calculated permeability	99
Basdogan *et al.*			■	■	■		■	780	ECFP4, MACCS	RF, GA		168
Xu *et al.*		■	■	■	■	■		∼500–800	Graph-based embedding, ECFP4, MACCS	GNN	Data augmentation and pseudo-labeling	169

datasets for gas separation, revealing the relationship between different datasets. These algorithms play a crucial role in deducing patterns and structures from unlabeled data, especially in polymers, where application-specific labels are usually limited.[72,155]

2. Beyond supervised and unsupervised learning

a. Accurate screening—the sparsity and imbalance in polymer membrane data. ML proves effective for material screening. However, annotating properties like gas permeability for polymers is costly and time-consuming,[71–73,85] limiting the size of available training data for accurate screening. While scaling up deep learning models has shown promise in discovering inorganic materials[173,174] and small molecules,[175] replicating similar success in polymers, especially for gas separation, is challenging due to limited data. This limitation arises from the typically higher complexity of polymer systems compared to small molecules and inorganic materials. To illustrate, PubChem[117] is a dataset hosting various molecular properties for drug discovery. For bioactivity and toxicity information, it has around 305×10^6 records. The Open Quantum Materials Database (OQMD)[174] consists of more than 1×10^6 density functional theory (DFT)-calculated thermodynamic and structural properties of inorganic compounds from the Inorganic Crystal Structure Database (ICSD).[176] On the contrary, datasets for gas separation polymers are notably smaller, as indicated in Table II, with only a few hundred entries. This substantial data size disparity poses a significant hurdle in training a generalizable ML model to accurately screen the large chemical space of gas separation membrane polymers.

Additionally, ML models for predicting gas permeability are generally fitted using log10 values, which can lead to larger prediction errors. Moreover, other properties of interest, such as the relative position of a polymer in the log–log plot of selectivity against permeability for a certain gas pair, are often observed less frequently above the satisfactory upper-bound threshold [Fig. 1(d)], creating an imbalanced nature in polymer data labels.[32] This imbalance tends to lead to a false negative problem (i.e., misclassifying upper-bound polymers as below the bound) in the virtual screening process, potentially biasing ML models toward polymers of lower interest [below the upper bound in Fig. 1(d)] and causing them to overlook promising candidates of excellent gas separation performance.

In this subsection, we review the methods developed in the ML community for addressing data sparsity and imbalance issues related to gas separation polymers and discuss their potential for enhancing virtual screening to identify high-performance materials. Following the three elements in the ML workflow in Fig. 2 (data, representation, and ML algorithm), we categorize the methods into two types: data-centric and model-centric methods. Data-centric methods prioritize enhancing both the quantity and quality of the data to improve the model performance, whereas model-centric methods focus on refining the learning of model parameters. Notably, the second element (polymer representation) can be integrated into the third element (the ML model) in advanced ML frameworks.

Data-centric methods. Data-centric methods, summarized in Fig. 7, concentrate on improving the training data quantity and quality to enhance model performance and screening accuracy. By integrating with model training, these methods allow for dynamical updates to the training set, continuously refining model performance. We categorize data-centric methods into three types: data augmentation (DA) (primarily using labeled data), pseudo-labeling (primarily using unlabeled data), and active learning (strategically selecting and labeling the most informative unlabeled data).

Data augmentation

To help ML model training, data augmentation (DA) approaches, inspired by techniques from fields like image augmentation,[177] focus on expanding training datasets by generating useful, albeit not necessarily realistic, examples. Common DA techniques can be categorized into perturbation-based and learning-based methods. Perturbation-based DA methods for polymers involve introducing small perturbations to the graph structure of polymers, as shown in Fig. 7(a), such as node feature masking,[178] edge dropping,[179] and subgraph replacement.[180] Sun *et al.*[180] introduced the concept of bioisosteres (subgraphs) replacement to ensure the chemical validity of augmented molecular graph, addressing concerns related to chemical integrity (some perturbation methods may make the augmented data chemically invalid) from some DA approaches like edge dropping.

Meanwhile, learning-based DA methods have emerged to avoid the heuristic search for optimal augmentation methods. For instance, as illustrated in Fig. 7(b), Liu *et al.* introduced a graph-based framework (Graph Rationalization enhanced by Environment-based Augmentations, GREA). This approach partitions the polymer graph into two parts: rationale (task-relevant subgraph) and environment (task-irrelevant subgraph). It then combines the rationale from one polymer with the environment of another in the latent space, thereby augmenting the limited training data.[138] Using GREA, an O_2 permeability dataset of 595 polymer membranes, primarily sourced from the MSA database, was augmented, achieving an average R^2 of 0.941 on the test set.

In addition to relying solely on the limited labeled data of polymer gas permeability prediction, leveraging large sets of unlabeled molecule or polymer data (e.g., from ZINC, QM9, and PoLyInfo databases) present a valuable resource to extract useful chemical knowledge (i.e., data distribution) to augment the labeled data. Building on this concept, as depicted in Fig. 7(c), Liu *et al.* developed a data-centric transfer framework (DCT) utilizing a generative diffusion model trained on a comprehensive unlabeled dataset to iteratively generate task-specific polymer examples, into which minimal sufficient knowledge from the unlabeled data was transferred. The generated data were used to augment the labeled dataset for training a more accurate polymer O_2 permeability prediction model. The results indicate a 17.7% decrease in prediction error on the test dataset compared to a vanilla GNN with no data augmentation and 14.6% compared to GREA, which relies solely on the training data for augmentation.[139]

Interested readers are referred to Refs. 181 and 182 for a more comprehensive introduction to graph DA. However, current approaches in DA primarily focus on augmenting the training set without directly addressing data sparsity and imbalance issues in polymer datasets. Future work in polymer tasks, e.g., prediction of gas permeability, should tailor DA approaches to generate training data points based on the specific data and label distribution within the task.

Pseudo-labeling

Pseudo-labeling, also known as self-training, is a semi-supervised learning approach that iteratively assigns pseudo-labels to unlabeled data and incorporates them into the labeled training set for model training.[183] The generalized workflow of pseudo-labeling is shown in Fig. 7(d). Liu *et al.*[165] utilized pseudo-labeling alongside DA techniques, proposing a Semi-supervised Graph Imbalanced Regression

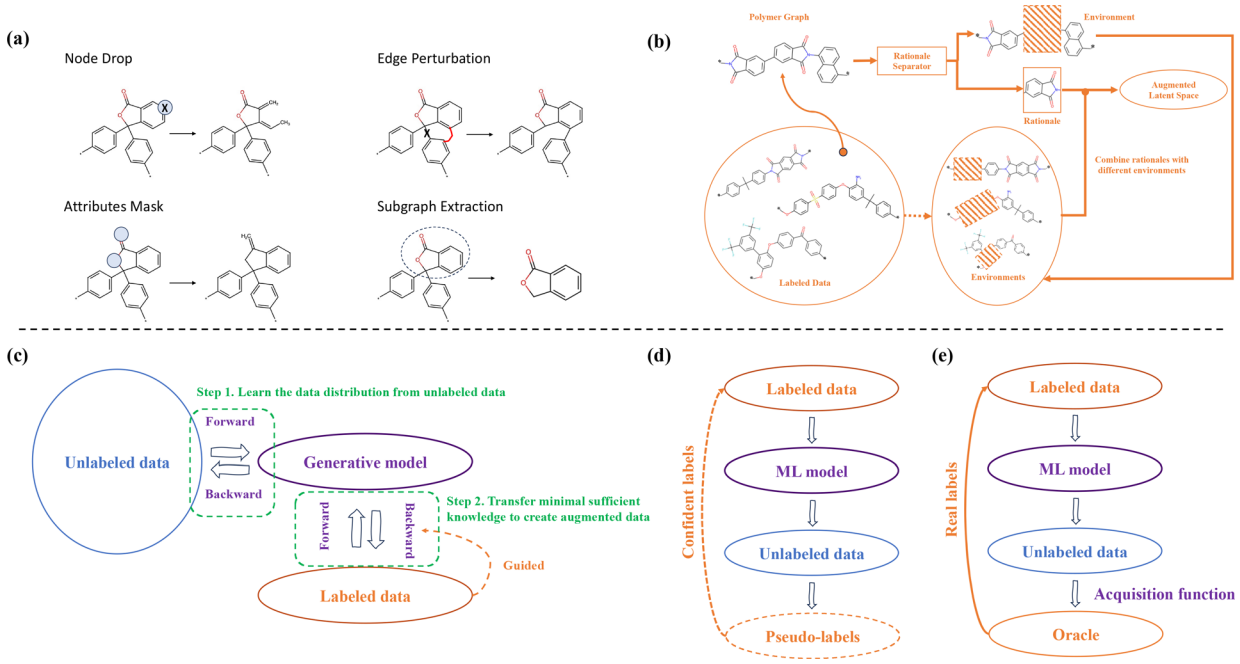

FIG. 7. Data-centric methods to enhance model performance and screening accuracy. (a) Small perturbations added to the structure of polymers to augment the data. (b) A graph rationalization enhanced by environment-based augmentation framework (GREA). It partitions the polymer graph into two parts: rationale (task-relevant subgraph) and environment (task-irrelevant subgraph) and then combines the rationale from one polymer graph with the environment from another in the latent representation space, thereby augmenting the limited training data. (c) A data-centric transfer (DCT) framework utilizing a generative diffusion model trained on a comprehensive unlabeled dataset to iteratively generate task-specific polymer examples for the labeled dataset. (d) The generalized workflow of pseudo-labeling. It iteratively assigns confident pseudo labels to the unlabeled data and incorporates them into the labeled training set for model training. (e) The general workflow of active learning, which mitigates model uncertainty by adding additional labeled data in the ML loop leveraging uncertainty estimation.

(SGIR) framework to address the training data sparsity and imbalance issue in the polymer permeability data. By enhancing the training data in the under-represented label areas, SGIR reduced average prediction error by 17.8% compared to vanilla GNN and 17.3% compared to GREA. Combining data augmentation (GREA) and pseudo-labeling (SGIR) techniques, Xu *et al.*[169] improved ML prediction accuracy on the small and imbalanced dataset of polymer gas permeability and validated their prediction by experimentally synthesizing two polymers with superior gas separation performance.

Missing value imputation can also be considered a form of pseudo-labeling in polymer permeability datasets, given that these datasets often include missing values. In many instances, a given polymer entry may possess limited gas permeability records, with some gases lacking recorded values. Yuan *et al.*[164] addressed this challenge by developing an ML model for imputing missing values in the datasets. They employed the multivariate imputation by chained equations (MICE) algorithm, incorporating Bayesian linear regression (BLR) and extremely randomized trees (ERT) for inference, to predict missing permeabilities for six common industrial gases. Building on this, Yang *et al.*[86] proposed an ensemble ML strategy for membrane materials discovery. They employed MICE to complete missing entries, then trained an ensemble of NN and RF models, and tested molecular descriptors and Morgan fingerprint with frequency as input features. Conducting high-throughput screening of over nine $\times$ 10^6 synthetic polymers, they identified ultra-permeable polymers validated through MD simulations. However, a potential issue of data leakage exists when training and testing with imputed data. Researchers need to be cautious when evaluating model in scenarios involving data imputation.

A key challenge in pseudo-labeling is defining a confidence score to assign pseudo-labels to confidently predictable labels.[183] Many studies have explored improving uncertainty estimation to aid the model in filtering out noise for reliable pseudo-labels.[184–186] However, this restricts pseudo-labeling to high-confidence labels, potentially overlooking a substantial number of high-uncertainty labels. Future work might consider integrating active learning (see next session) as a complementary approach to pseudo-labeling for handling high-uncertainty data. Moreover, like DA approaches, a sampling strategy for pseudo-labels is crucial to balance label distribution, rather than blindly adding examples to the training data. Without a thoughtful sampling strategy, there is a risk of exacerbating model bias using pseudo-labeling.

Active learning

Similar to pseudo-labeling, active learning [Fig. 7(e)] tries to improve model accuracy and generalizability by adding labeled data into the ML loop leveraging uncertainty estimation.[187–189] The key difference is that active learning focuses on selecting the most informative data, involving the trade-off between exploration and exploitation, and assigning real labels to them using oracle knowledge—an information source that provides accurate labels to the selected data. Although

active learning has not been employed to discover high-performance gas separation polymer membranes, its effectiveness has been demonstrated in other polymer discovery tasks, including polymers with high T_g, polymer blends with high thermal conductivity, peptide-binding polymers, among others.[190–194] Combining active learning with pseudo-labeling could be promising to harness the knowledge embedded in unlabeled polymer data and enhance the model prediction accuracy for gas separation performance.

Model-centric methods. In contrast to data-centric methods that prioritize improving the training data quantity and quality, model-centric methods concentrate on effective learning of model parameters (or data representation) through model structure design. There are two mains types: transfer learning (TL) and self-supervised learning (SSL). In TL, a training set usually comprises only labeled data for different tasks. As for SSL, the training set typically contains both labeled data and unlabeled data. Model-centric methods aim to create data-efficient models for tasks like representation learning and knowledge transfer. Typically, the labeled (and unlabeled) training data are assumed high-quality to solve the data imbalance and sparsity problems.

Transfer learning

TL is a promising framework for addressing data sparsity, which can leverage the information from a distinct but related "source" task to inform a model on a "target" task by transferring model parameters, as illustrated in Fig. 8(a).[195,196] It avoids the model parameter learning from scratch (i.e., a cold start) and reduces the amount of labeled data needed for a new task. In data-driven material research, TL has gained preliminary achievements. For instance, the transferability of metal-organic frameworks gas adsorption capacity is tested across different gas species and conditions, showing the great promise of TL when source and target tasks share common knowledge.[197] As for polymer research, Wu *et al.*[198] utilized features from well-populated proxy properties (T_g) to build a deep neural network on the thermal conductivity of polymers ($R^2 = 0.73$) with only 28 data points.

To assess the effectiveness of TL in predicting polymer gas permeability, we conduct a demonstrative cross-gas TL study using the N_2 and O_2 permeability data from the MSA database. Given the limited size of labeled data for both gases (~600) and recognizing potential relationships among different gas permeabilities of polymers, we hypothesize the existence of transferability between different gas permeabilities. Therefore, TL could enhance model performance, such as leveraging O_2 permeability as a source task to transfer knowledge and improve the accuracy of predicting N_2 permeability.

Herein, PE[85] is employed to map SMILES of polymers into a 300-dimensional continuous vector. A multilayer perceptron (MLP) with three hidden layers is constructed for the two data sets predicting the N_2 and O_2 permeabilities of polymers. These models, referred to as the proxy model if trained on O_2 data or the direct learning (DL) model if trained on N_2 data, do not involve knowledge transferred. Weights and bias for each hidden layer of the O_2 proxy model are saved after training on the corresponding whole dataset for subsequent cross-gas TL studies. Then, the proxy model trained on the O_2 data is further trained in the TL scheme for the N_2 dataset on the whole dataset (denoted as TL-whole, while DL models trained on the corresponding whole dataset are denoted as DL-whole). Prediction performance on N_2 permeability is evaluated in a fivefold cross-validation manner using mean squared error (MSE). To simulate common situations in polymer informatics where only very limited property data are available and to test the efficiency of TL under these circumstances, we train the O_2 proxy models using the TL scheme for N_2 data on 20 different sets (to eliminate overfitting) of 100 randomly selected N_2 data points, denoted as TL-100-full, and use the remaining N_2 data that are

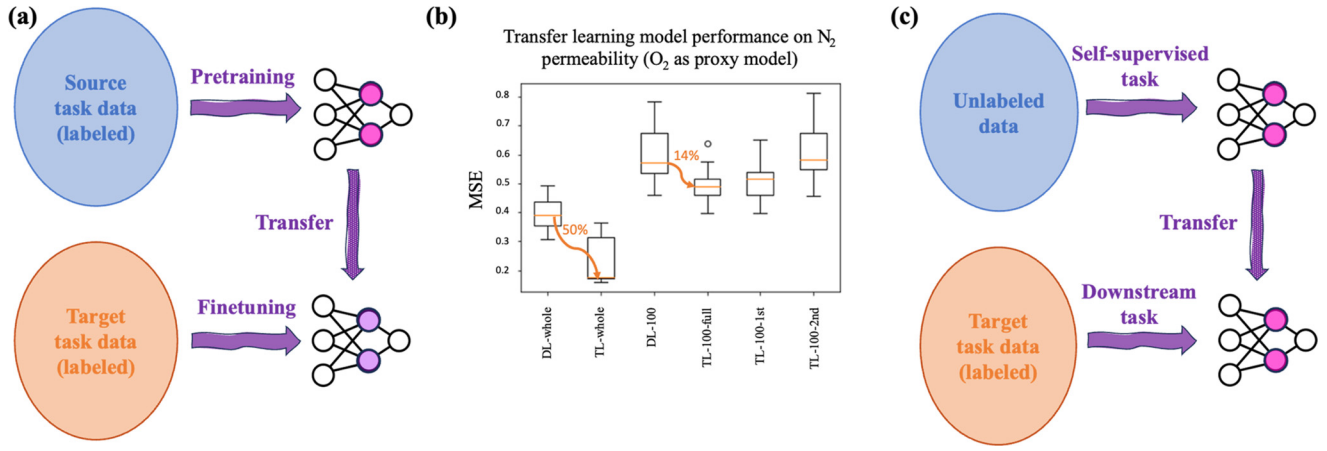

FIG. 8. Model-centric methods to enhance model performance and screening accuracy. (a) The framework of transfer learning. It leverages the information obtained from a "source" task to inform a model on a "target" task through the transfer of model parameters. Both the source and target tasks use labeled data. (b) Transfer learning performance on N_2 permeability prediction using O_2 permeability data to train a proxy model. All permeability values are in units of log10 Barrer. Prediction performance is evaluated in a fivefold cross-validation manner using mean squared error (MSE) and is visualized in box plot. The orange line in each box denotes the mean of MSE. "DL-whole" denotes the direct learning (DL) case using the whole N_2 data; "TL-whole" denotes the transfer learning (TL) case where we first use the whole O_2 data as the source task and then use the whole N_2 data to fine-tune the model parameters (three hidden layers); "DL-100" denotes the DL case only using 100 randomly sampled N_2 data to simulate the limited training data scenario; "TL-100-full" denotes the same TL case as "TL-whole," however only utilizing 100 randomly sampled N_2 data to fine-tune; "TL-100-1st" and "TL-100-2nd" denote the same TL case as "TL-100-full," however only transferring the 1st or 2nd hidden layer parameters from the source task. (c) The framework of self-supervised learning. It transfers knowledge from unlabeled data (self-supervised task training) to labeled data (downstream task or target task training) through model parameters.

not for training to test the model performance. MLP models directly trained on the same 100 N_2 data with the identical architecture as the O_2 proxy models are also constructed for comparison, denoted as DL-100. Moreover, to further validate the application of TL from O_2 permeability to N_2, an additional experiment is conducted. We only use the pre-trained parameters (weights and bias) in the first or second hidden layer and leave the rest randomly initialized, which are denoted as TL-100-1st and TL-100-2nd, respectively, in Fig. 8(b). The MLP is constructed using PyTorch with the rectified linear unit (ReLU) activation function and the Adam optimizer.[199]

As depicted in Fig. 8(b), the TL-whole model, transferred from O_2 to N_2, exhibits a substantial decrease in the median of MSE by over 50%, dropping from 0.390 (DL-whole) to 0.177 (TL-whole). This underscores the robustness of TL to enhance the prediction performance of MLP for polymer membrane gas permeability. Even in scenarios with only 100 selected data points for training, the TL scheme demonstrates a positive impact. The median of MSE decreases by more than 14%, going from 0.571 (DL-100) to 0.490 (TL-100-full). Therefore, TL proves effective in improving predictions when only limited training data are available. Compared to all hidden layers transferred case (TL-100-full), only transferring the information learned from the first hidden layer (TL-100-1st, MSE = 0.516) is lightly detrimental to the performance of TL, but still outperforms the DL case (DL-100). In contrast, when only the pre-trained parameters from the second hidden layer are transferred (TL-100-2nd), a similar or even slightly worse MSE (0.582) is observed compared to the DL-100 case. This indicates that the first hidden layer contains most of the useful knowledge learned from the proxy task, and the second hidden layer primarily serves as a complex nonlinear mixture of outputs from the first hidden layer, with little impact on the overall model performance.

However, the limited size of data in the source task may still impede model generalization. Considering larger and more diverse small molecule datasets, such as QM9,[200] holds promise in this endeavor,[198] though it may introduce task-irrelevant noise for polymer studies. Exploring further avenues, including empirical observations[201] and theoretical analyses on the efficiency of transfer learning from molecules to polymers, presents promising directions for future research.

Self-supervised learning

SSL transfers knowledge from unlabeled data to labeled data through model parameters, as shown in Fig. 8(c). SSL methods involve manually constructed predictive and contrastive tasks on the unlabeled data.[178,202–204] In graph ML for polymers, predictive tasks include masked atom attribute prediction[178] and masked subgraph prediction.[202] In contrast, contrastive tasks[203,204] entail perturbing polymer graph structures to generate positive pairs and minimize the representation distance between these positive pairs while maximizing the distance between negative pairs from different polymers. It encourages the model to learn a meaningful polymer representation where similar instances (positive pairs) are clustered together, and dissimilar instances (negative pairs) are separated.

Kuenneth *et al.*[167] introduced an end-to-end machine-driven polymer informatics pipeline for predicting polymer membrane gas permeabilities and various polymer properties, featuring polyBERT, a polymer embedding tool inspired by NLP. The polyBERT model was trained using a predictive manner of SSL, which predicts the masked tokens of 100×10^6 synthetic SMILES of unlabeled polymers. The trained polyBERT was then connected to downstream multitask property predictors, mapping the embedding to various properties, including gas permeability. The approach demonstrated a two-orders-of-magnitude speed improvement over existing fingerprint methods while maintaining accuracy.

However, most SSL methods encounter challenges in cross-domain knowledge transfer.[139,205] The first challenge is that the unlabeled data for SSL are typically small molecules or synthetic polymers, whereas the target tasks pertain to real polymers. Second, SSL tasks are hand-crafted (e.g., the masked tokens prediction method in polyBERT), usually differing from the downstream tasks of polymer property predictions. These differences between the unlabeled and labeled data as well as between the SSL tasks and downstream tasks may lead to potential knowledge gap in the cross-domain transfer. A third challenge is an underexplored question of whether the pre-trained model from SSL still exhibits label imbalance bias when transferring model parameters to the target tasks. Therefore, to effectively leverage SSL for polymer gas permeability screening, one should consider collecting specific, larger-scale high-quality polymer datasets[85,86] and designing SSL tasks relevant to polymer properties. Researchers must also carefully examine potential model bias toward to majority group in the labeled datasets.

Interpretable models. While accuracy is crucial in virtual screening, researchers also prioritize understanding the rationale behind model predictions. The ability to interpret the result of the model not only drives scientific discovery but also, in turn, helps improve overall model performance. As a result, endeavors have been made to enhance the transparency and interpretability of advanced ML techniques, facilitating their application in polymer gas separation as well as broader polymer research. The model interpretability can be model-intrinsic if it is inherently embedded within the model architecture or parameters, or model-agnostic if it offers interpretability without relying on the internal details of a particular model.

ML models, such as logistic regression, linear regression, and RF, offer intrinsic interpretability compared to NN models,[206] providing nuanced physical insights alongside predictions. For instance, Fig. 9(a) shows the intrinsic feature importance of RDKit descriptors for RF models predicting T_g and T_m (melting temperature) of polymers.[206] Though not exceptionally accurate in prediction because of the model limitation, this type of feature importance analysis is straightforward to implement and offers useful preliminary intuition on the global feature importance across the dataset.

In the meantime, recent efforts have seen an increased exploration of intrinsic interpretable deep learning models, showcasing a greater advantage in prediction accuracy and local interpretability (i.e., for individual datapoint). For example, Wilson *et al.*[166] used a GNN (PolyID) and bond contribution aggregation to predict properties for biobased polymers, including gas permeabilities. The relative bond contribution to the target property (using T_g as an example) prediction learned through model training can be visualized and compared on the specific polymers of interest, as shown in Fig. 9(b). The identified important substructures are further validated by MD simulation. Similarly, atom contributions can also be calculated and visualized. Liu *et al.*[138] utilized DA and rationale identification to identify representative subgraphs supporting and explaining GNN predictions on polymer gas permeabilities [Fig. 9(c)]. Atoms highlighted in

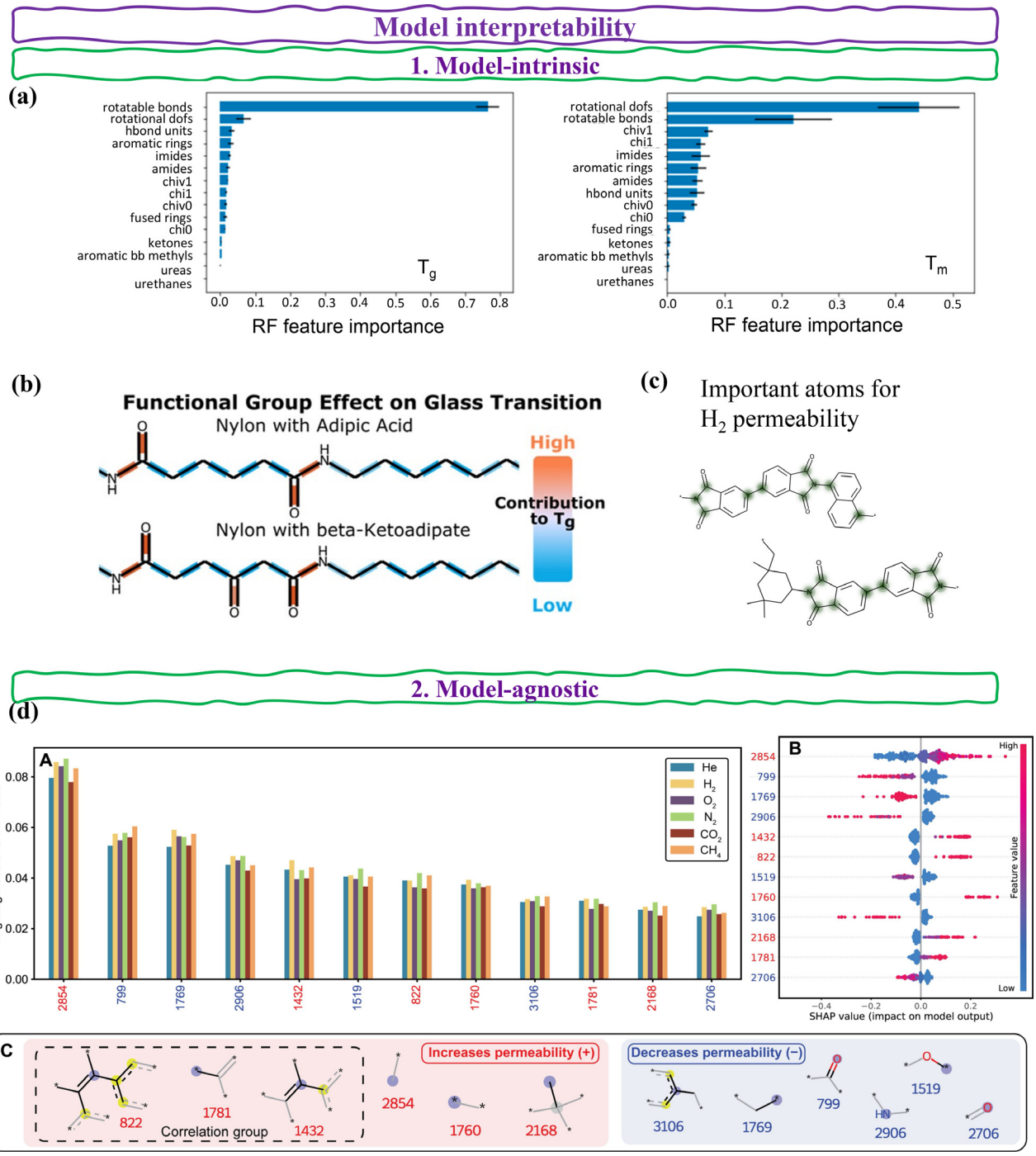

FIG. 9. Interpretable models: model-intrinsic and model-agnostic interpretability. (a) Example of feature importance obtained from intrinsic explainable models, e.g., random forest (RF). The RF is trained on T_g and T_m data of polymers represented by RDKit fingerprints. Reproduced with permission from Lee *et al.*, Polymers **13**, 3653 (2021). Copyright 2021 Authors, licensed under a Creative Commons Attribution (CC BY) License. (b) Visualization of relative bond contributions to the target property (T_g) prediction learned through model training. Reproduced from C. Kuenneth and R. Ramprasad *et al.*, Nat. Commun. **14**, 4099 (2023). Copyright 2023 Authors, licensed under a Creative Commons Attribution (CC BY) License. (c) Visualization of rationale atoms based on GNN predictions on polymer gas permeabilities. Atoms highlighted in green indicate a higher probability being classified as rationales, thus contributing more to gas permeability predictions. Reprinted with permission from Xu *et al.*, Cell Rep. Phys. Sci. **5**, 102067 (2024). Copyright 2024 Elsevier. (d) Feature importance analysis of gas permeability prediction models using the model-agnostic SHAP method. Reprinted with permission from Yang *et al.*, Sci. Adv. **8**, eabn9545 (2022). Copyright 2022 Authors, licensed under a CC BY-NC 4.0 License.

green indicate a higher probability of being classified as rationales, thus contributing more to the model prediction. However, these intrinsic interpretable deep learning models offer interpretability only at an individual data level, lacking global insights across the entire dataset.

In addition to the model-intrinsic interpretation, various model-agnostic interpretation approaches can enhance the interpretability of ML models, providing valuable insights at local and/or global levels. Examples include visualizations like partial dependence plots[207] and individual conditional expectation,[208] as well as techniques like LIME (Local Interpretable Model-agnostic Explanations)[209] and SHAP (Shapley Additive exPlanations).[210] These methods provide flexible tools for interpreting the decision-making processes of ML models. In ML for polymer membrane gas permeability prediction, Yang *et al.*[86] used SHAP for feature importance analysis [Fig. 9(d)], yielding insights consistent with experimental trends.

b. Inverse design. In addition to accurate screening, generative models also play a crucial role in achieving inverse design for gas separation polymers. Generative models for polymers generally fall into two categories: fragment-based and deep learning-based. Fragment-based approaches create a database of polymer substructures and functional groups, which are then combined in different ways to construct novel polymer structures. On the other hand, deep learning-based methods model the probabilistic distribution of polymer data, enabling inverse polymer designs conditional on desirable properties.[211] Generally, deep learning-based methods encode the high-dimensional chemical space of polymers into a continuous latent space, from which new polymers can be decoded/generated.

For fragment-based generative models, genetic algorithm (GA) is a prominent method for polymer inverse design in gas separation. It operates through an iterative evolution process [Fig. 10(a)], starting with a "gene pool" of polymer structural fragments to create the initial parent polymers. Crossover and mutation operations on segments then generate offspring polymers. Kim *et al.*,[212] shown in Fig. 10(a), utilized GA to design novel polymers with high bandgap and high T_g. Similarly, Basdogan *et al.*[168] employed an ML-driven GA to design polymer membranes for CO_2 separation from N_2 and O_2. They first utilized an RF model to predict gas permeabilities and selectivities using literature data for CO_2, N_2, and O_2, together with polymer fingerprinting methods (ECFP and MACCS). Then, GA was applied with the RF

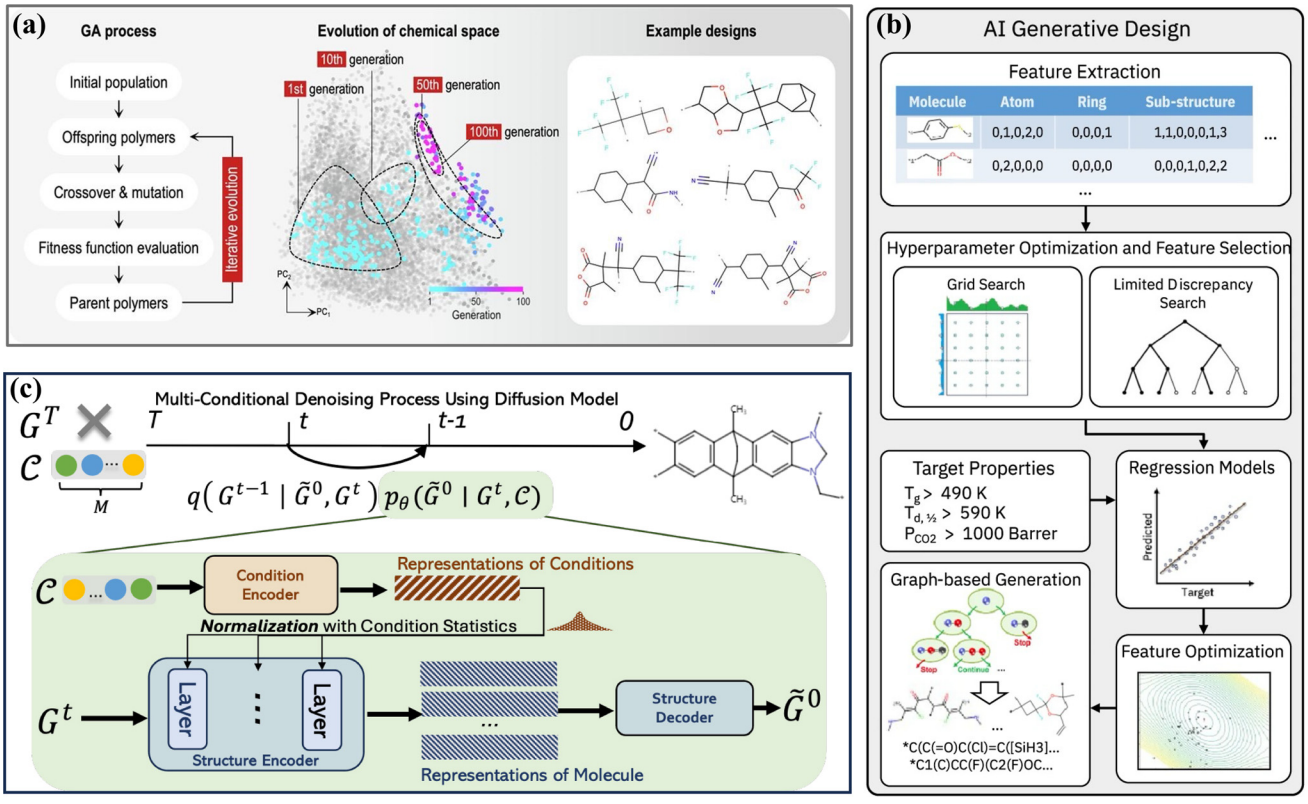

FIG. 10. Examples of inverse design for gas separation polymer membranes. (a) An example of genetic algorithm (GA) for polymer inverse design. Reprinted with permission from Kim *et al.*, Comput. Mater. Sci. **186**, 110067 (2021). Copyright 2020 Elsevier B.V. All rights reserved. (b) An example of graph generation algorithm through combination of fragments for novel polymer generation and design. Reproduced with permission from Giro *et al.*, npj Comput. Mater. **9**, 133 (2023). Copyright 2023 Authors, licensed under a Creative Commons Attribution (CC BY) License. (c) A multi-conditional diffusion guidance framework (Graph DiTs) designed for both polymers and small molecules generation and optimization. Reproduced with permission from Liu *et al.*, arXiv:2401.13858 (2024). Copyright 2024 Authors, licensed under a Creative Commons Attribution (CC BY) 4.0 International License.

model to design new polymers, optimizing for separation performance under constraints. The approach identified promising polymers for CO_2/N_2 and CO_2/O_2 separations, showcasing versatility in constrained optimization for polymer design. However, it is noteworthy that no experimental validation was conducted in this study and the synthesizability of proposed polymers has not been carefully considered.

In addition to GA, graph generation algorithm through combination of fragments (subgraphs) is another effective fragment-based approach for novel gas separation polymer membrane design. This process involves four steps: molecular feature encoding, target property prediction, feature search and optimization, and graph structure generation.[213] Giro *et al.*[99] presented a fully automated computational discovery process for polymer membranes in CO_2 separation based on graph generation algorithm, as illustrated in Fig. 10(b). The process couples graph generation polymer design with MD simulation of post-combustion CO_2 filtration, showing quantitative agreement between CO_2 permeability predictions from ML models and MD simulations on the generated polymers. However, while validated through MD simulation, no experimental validation was conducted, and selectivity was not considered, only focusing on CO_2 permeability, which lacks practical influence.

Deep learning-based generative methods have initially found application and validation in small molecules generation and design.[214] Representative algorithms include RNN,[85] variational autoencoders (VAEs), generative adversarial networks (GANs), normalizing flows (NFs),[215] and diffusion models.[216] Many successful examples of small molecule design have been reported using deep learning. For instance, GraphRNN with reinforcement learning,[217] Markov molecular sampling,[218] junction tree variational autoencoder with Bayesian optimization,[219] and long short-term memory networks on SMILES with Hill climbing.[220] While most molecular inverse design and optimization approaches can be adapted for polymers and gas separation because of the similarity shared by small molecules and polymers, very limited endeavors have been reported thus far.

Notably, Liu *et al.*[221] introduced a multi-conditional diffusion guidance framework (Graph DiTs) designed for both polymers and small molecules. As illustrated in Fig. 10(c), the Graph DiT employs a transformer-based architecture to encode numerical and categorical conditions (such as permeabilities for different gases) and to learn molecular representations. Leveraging a structure decoder for denoising from the learned and conditioned molecular representation, Graph DiT has shown success in an inverse polymer design task for O_2/N_2 gas separation. The study demonstrates the generation of polymers that meet multi-property constraints, showcasing the potential of deep learning in designing polymer membranes for gas separation.

However, conducting comprehensive studies to evaluate the performance of different deep learning-based generative models on polymers remains a crucial avenue for future research. One notable challenge is the synthesizability of the generated polymers, a topic we will discuss more thoroughly in Sec. III.

III. CHALLENGES AND PERSPECTIVES

A. Data

1. Automatic data extraction

The scarcity of experimental data for gas separation polymeric membranes continues to pose significant challenges. Despite advancements in ML frameworks, as discussed in Sec. II, there is ongoing need for more training data to capture the vast chemical space of polymers. While computational simulation methods, such as MD simulations, offer valuable insights and the potential to augment data coverage, accuracy issues persist. Integrating NLP techniques for data mining from literature has shown promise in addressing this challenge. Recent studies on polymer blends have effectively harnessed NLP techniques to extract relevant data, showcasing its potential.[222] Similarly, Shetty *et al.*[223] trained a language model, MaterialsBERT, to automatically extract material property data from polymer literature abstracts, yielding ∼300 000 records from ∼130 000 abstracts. The data offered insights into applications like fuel cells and polymer solar cells, demonstrating the feasibility of an automated literature-to-data pipeline.

2. Data standardization and robustness

The variability in permeability data of polymer membranes presents a challenge for the robust application of ML. Experimental data frequently display inconsistencies stemming from different fabrication techniques and measurement conditions across laboratories. This variability, even among polymers with identical chemistry (e.g., PIM-1[224]), underscores the need for a standardized protocol in polymer membrane fabrication and measurement. The introduction of autonomous robotic labs for experiments[225–228] could enhance standardization. However, it is crucial to carefully address potential challenges in autonomous membrane fabrication, such as ensuring uniformity in coating thickness, maintaining consistent environmental conditions, and optimizing the reproducibility of experiments. Additionally, the effects of plasticization and physical aging on permeability values further contribute to data variability. A given polymer can exhibit different permeability measurements over time, complicating the assessment of data robustness. In addition to data variability, ensuring the robustness or reliability of current databases is also crucial. Issues like reverse selectivity in specific pure gas permeability datasets (e.g., MSA dataset[26]) underscore the importance of thorough data cleaning and validation before feeding training data into ML models.

3. Possibility beyond linear homopolymers

Currently, ML research is predominantly centered on simple linear homopolymers, constrained by data availability and polymer chemistry descriptions. However, there is a notable opportunity for advancement by expanding this focus to include a wider variety of polymers, such as ladder polymers,[92,224,229] polymer blends,[230] copolymers,[135,231–233] and polymer-based composites like MMMs.[8,14,58] This holds the potential to unlock improved separation performance in polymeric membranes once more data are available. Adopting advanced chemical language of polymer chemistry, such as BigSMILES,[120] and integrating its capability into popular cheminformatics software like RDKit could help capture and analyze the macromolecular details and complexities inherent in polymers.

B. ML algorithms

1. Generation of synthesizable polymer structures

Current generative algorithms, both fragment-based (e.g., GA and graph generation algorithm) and deep learning-based (e.g., diffusion model), hold promise for inverse designing of polymer structures. However, synthesizability, a critical factor in the inverse design process,

necessitates more sophisticated considerations in these models. Introducing a comprehensive synthesizability score (e.g., SA score,[234] SCScore,[235] SYBA score,[236] and RAScore[237]) could enhance the practical utility of generated structures. Still, it only provides limited information on the synthesizability of polymers. Bridging the gap between algorithmically proposed polymer structures and successful laboratory synthesis requires addressing crucial factors such as potential polymerization paths and optimal experimental conditions, including reactants, solvents, and film formation.

One promising solution to enhance the synthesizability of generated polymers is to integrate ML-driven retrosynthesis planning with generative algorithms. Retrosynthesis planning generally falls into two categories: template-based and template-free. Template-based approaches rely on summarized/extracted reaction rules defining atom and bond changes during reactions.[238,239] In contrast, template-free methods, often utilizing deep learning such as sequence-to-sequence models, directly predict reactants based on information like SMILES.[240–242]

Chen *et al.* developed a data-assisted retrosynthesis planning tool for polymer synthesis, demonstrating a template-based approach to polymer retrosynthesis.[243] This method extracts templates from a dataset of 11 448 polymerization paths (involving 9748 homopolymers and 8921 reactant monomers) and uses similarity-based predictions to select optimal synthesis routes for new polymers. However, this approach only considers reactants and products and is limited to three polymerization types for homopolymer synthesis, neglecting crucial factors such as solvents, catalysts, and experimental conditions as well as broader polymer types (e.g., copolymers, ladder polymers) and other polymerization classes. Additionally, it may only be applicable when the extracted templates are effective for a new polymer, requiring continuous template updates.

Template-free methods, although potentially more versatile, may require numerous reactions for training to identify meaningful reaction patterns without templates.[240–242] However, as of now, no work has been reported on polymer retrosynthesis using template-free methods. Exploring the potential of large language models for polymer structure generation and optimization, while considering retrosynthesis planning represents an intriguing direction for future research.[244–246]

2. Multi-objective inverse design

Beyond single-target optimization, the challenge of polymer membrane design lies in incorporating multi-objective optimization and considering the multifunctionality of polymeric materials. While permeability is crucial for applying polymer membranes for gas separation, other properties, such as mechanical and thermal characteristics,[1] are also critical in real-world applications. The exploration of ML-driven multi-objective optimization can contribute to the holistic design of polymer materials.

C. Other limitations—aging, plasticization, and environmental considerations

Addressing aging and plasticization challenges is imperative for the long-term performance of gas separation membranes. ML approaches may offer insights into mitigating these issues, enhancing membrane durability and stability. Ongoing research should explore the role of ML in overcoming these limitations to ensure the continued advancement of polymeric membranes. On the other hand, the need for environmentally friendly polymer membranes is growing. Presently, the most extensively researched polymers for gas separation feature fluorine-containing moieties, which enhance processibility, free volume, and thermal stability due to the strong C–F bond.[247,248] However, concerns about the end-of-life impacts of fluorine-containing polymers on health and the environment are driving stricter regulatory standards.[249–252] This highlights the urgency to develop high-performance polymer membranes that not only address drawbacks like permeability-selectivity trade-off, plasticization, and physical aging, but also align with tightening environmental legislation. How ML could accelerate the design of more environmentally friendly high-performance polymer membranes remains to be explored.

IV. CONCLUSIONS

In conclusion, adopting ML in polymer informatics, especially for gas separation membranes, is a promising avenue that can address the pressing energy and environmental challenges. This review discussed the critical role of three primary components: high-quality polymer data, advanced representation methods, and robust ML algorithms. We examined diverse datasets, representation techniques, and algorithms employed in recent studies on polymer membranes for gas separation, providing a comprehensive review of the current research landscape. Despite notable advancements, challenges and opportunities persist in applying ML to gas separation polymers. Key issues include data sparsity, imbalance, reliability, polymer synthesizability, and the multi-objective optimization of polymer structures, alongside environmental considerations. As the field evolves, addressing these challenges will facilitate more efficient and reliable ML applications in the design and discovery of polymeric materials for gas separation, as well as broader implications for polymer and material informatics.

ACKNOWLEDGMENTS

T.L. would like to thank the National Science Foundation (Grant Nos. 2102592 and 2332270) and DOE (Grant No. DE-EE0009103). A.S. gratefully acknowledges the support from the Center of Environmental Science and Technology (CEST) Predoctoral Fellowship at the University of Notre Dame.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Jiaxin Xu: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Agboola Suleiman:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Gang Liu:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Renzheng Zhang:**

Data curation (equal); Formal analysis (equal); Writing – original draft (equal). **Meng Jiang:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Ruilan Guo:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Tengfei Luo:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]R. Sidhikku Kandath Valappil, N. Ghasem, and M. Al-Marzouqi, "Current and future trends in polymer membrane-based gas separation technology: A comprehensive review," J. Ind. Eng. Chem. **98**, 103–129 (2021).

[2]W. J. Koros and G. K. Fleming, "Membrane-based gas separation," J. Membr. Sci. **83**, 1–80 (1993).

[3]D. F. Sanders, Z. P. Smith, R. Guo, L. M. Robeson, J. E. McGrath, D. R. Paul, and B. D. Freeman, "Energy-efficient polymeric gas separation membranes for a sustainable future: A review," Polymer **54**, 4729–4761 (2013).

[4]P. Bernardo, E. Drioli, and G. Golemme, "Membrane gas separation: A review/state of the art," Ind. Eng. Chem. Res. **48**, 4638–4663 (2009).

[5]D. S. Sholl and R. P. Lively, "Seven chemical separations to change the world," Nature **532**, 435–437 (2016).

[6]R. W. Baker and B. T. Low, "Gas separation membrane materials: A perspective," Macromolecules **47**, 6999–7013 (2014).

[7]J. R. Weidman and R. Guo, "The use of Iptycenes in rational macromolecular design for gas separation membrane applications," Ind. Eng. Chem. Res. **56**, 4220–4236 (2017).

[8]M. Galizia, W. S. Chi, Z. P. Smith, T. C. Merkel, R. W. Baker, and B. D. Freeman, "50th anniversary perspective: Polymers and mixed matrix membranes for gas and vapor separation—a review and prospective opportunities," Macromolecules **50**, 7809–7843 (2017).

[9]Y. Wang, B. S. Ghanem, Z. Ali, K. Hazazi, Y. Han, and I. Pinnau, "Recent progress on polymers of intrinsic microporosity and thermally modified analogue materials for membrane-based fluid separations," Small Struct. **2**, 2100049 (2021).

[10]M. Liu, A. Seeger, and R. Guo, "Cross-linked polymer membranes for energy-efficient gas separation: Innovations and perspectives," Macromolecules **56**, 7230–7246 (2023).

[11]S. Adhikari and S. Fernando, "Hydrogen membrane separation techniques," Ind. Eng. Chem. Res. **45**, 875–881 (2006).

[12]J. Caro, M. Noack, P. Kölsch, and R. Schäfer, "Zeolite membranes—state of their development and perspective," Microporous Mesoporous Mater. **38**, 3–24 (2000).

[13]E. Adatoz, A. K. Avci, and S. Keskin, "Opportunities and challenges of MOF-based membranes in gas separations," Sep. Purif. Technol. **152**, 207–237 (2015).

[14]T.-S. Chung, L. Y. Jiang, Y. Li, and S. Kulprathipanja, "Mixed matrix membranes (MMMs) comprising organic polymers with dispersed inorganic fillers for gas separation," Prog. Polym. Sci. **32**, 483–507 (2007).

[15]Z. Wang, D. Wang, S. Zhang, L. Hu, and J. Jin, "Interfacial design of mixed matrix membranes for improved gas separation performance," Adv. Mater. **28**, 3399–3405 (2016).

[16]G. Saracco, H. W. J. P. Neomagus, G. F. Versteeg, and W. P. M. van Swaaij, "High-temperature membrane reactors: Potential and problems," Chem. Eng. Sci. **54**, 1997–2017 (1999).

[17]Y. Yampolskii, "Polymeric gas separation membranes," Macromolecules **45**, 3298–3311 (2012).

[18]W. J. Koros, G. K. Fleming, S. M. Jordan, T. H. Kim, and H. H. Hoehn, "Polymeric membrane materials for solution-diffusion based permeation separations," Prog. Polym. Sci. **13**, 339–401 (1988).

[19]X. Hu, Y. Pang, H. Mu, X. Meng, X. Wang, Z. Wang, and J. Yan, "Synthesis and gas separation performances of intrinsically microporous polyimides based on 4-methylcatechol-derived monomers," J. Membr. Sci. **620**, 118825 (2021).

[20]C. Z. Liang, T.-S. Chung, and J.-Y. Lai, "A review of polymeric composite membranes for gas separation and energy production," Prog. Polym. Sci. **97**, 101141 (2019).

[21]H.-J. Schröter, "Carbon molecular sieves for gas separation processes," Gas Sep. Purif. **7**, 247–251 (1993).

[22]A. Phan, C. J. Doonan, F. J. Uribe-Romo, C. B. Knobler, M. O'Keeffe, and O. M. Yaghi, "Synthesis, structure, and carbon dioxide capture properties of zeolitic imidazolate frameworks," Acc. Chem. Res. **43**, 58–67 (2010).

[23]J. G. Wijmans and R. W. Baker, "The solution-diffusion model: A review," J. Membr. Sci. **107**, 1–21 (1995).

[24]A. S. Michaels, W. R. Vieth, and J. A. Barrie, "Diffusion of gases in polyethylene terephthalate," J. Appl. Phys. **34**, 13–20 (1963).

[25]A. S. Michaels, W. R. Vieth, and J. A. Barrie, "Solution of gases in polyethylene terephthalate," J. Appl. Phys. **34**, 1–12 (1963).

[26]See https://www.membrane-australasia.org for Membrane Society of Australasia, MSA (2024).

[27]N. Mehio, S. Dai, and D. Jiang, "Quantum mechanical basis for kinetic diameters of small gaseous molecules," J. Phys. Chem. A **118**, 1150–1154 (2014).

[28]L. M. Robeson, "Correlation of separation factor versus permeability for polymeric membranes," J. Membr. Sci. **62**, 165–185 (1991).

[29]B. D. Freeman, "Basis of permeability/selectivity tradeoff relations in polymeric gas separation membranes," Macromolecules **32**, 375–380 (1999).

[30]A. Y. Alentiev and Y. P. Yampolskii, "Free volume model and tradeoff relations of gas permeability and selectivity in glassy polymers," J. Membr. Sci. **165**, 201–216 (2000).

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

[32]L. M. Robeson, "The upper bound revisited," J. Membr. Sci. **320**, 390–400 (2008).

[33]W. J. Koros, "Model for sorption of mixed gases in glassy polymers," J. Polym. Sci. Polym. Phys. Ed. **18**, 981–992 (1980).

[34]M. S. Suleman, K. K. Lau, and Y. F. Yeong, "Plasticization and swelling in polymeric membranes in CO_2 removal from natural gas," Chem. Eng. Technol. **39**, 1604–1616 (2016).

[35]A. F. Ismail and W. Lorna, "Penetrant-induced plasticization phenomenon in glassy polymers for gas separation membrane," Sep. Purif. Technol. **27**, 173–194 (2002).

[36]J. E. Bachman, Z. P. Smith, T. Li, T. Xu, and J. R. Long, "Enhanced ethylene separation and plasticization resistance in polymer membranes incorporating metal–organic framework nanocrystals," Nat. Mater. **15**, 845–849 (2016).

[37]A. Bos, I. Pünt, H. Strathmann, and M. Wessling, "Suppression of gas separation membrane plasticization by homogeneous polymer blending," AIChE J. **47**, 1088–1093 (2001).

[38]W. F. Yong, K. H. A. Kwek, K.-S. Liao, and T.-S. Chung, "Suppression of aging and plasticization in highly permeable polymers," Polymer **77**, 377–386 (2015).

[39]D. Q. Vu, W. J. Koros, and S. J. Miller, "Effect of condensable impurities in CO_2/CH_4 gas feeds on carbon molecular sieve hollow-fiber membranes," Ind. Eng. Chem. Res. **42**, 1064–1075 (2003).

[40]L. S. White, T. A. Blinka, H. A. Kloczewski, and I. Wang, "Properties of a polyimide gas separation membrane in natural gas streams," J. Membr. Sci. **103**, 73–82 (1995).

[41]N. Tanihara, H. Shimazaki, Y. Hirayama, S. Nakanishi, T. Yoshinaga, and Y. Kusuki, "Gas permeation properties of asymmetric carbon hollow fiber membranes prepared from asymmetric polyimide hollow fiber," J. Membr. Sci. **160**, 179–186 (1999).
[42]M. Minelli, S. Oradei, M. Fiorini, and G. C. Sarti, "CO_2 plasticization effect on glassy polymeric membranes," Polymer **163**, 29–35 (2019).
[43]C.-C. Chen, W. Qiu, S. J. Miller, and W. J. Koros, "Plasticization-resistant hollow fiber membranes for CO_2/CH_4 separation based on a thermally crosslinkable polyimide," J. Membr. Sci. **382**, 212–221 (2011).
[44]N. R. Horn and D. R. Paul, "Carbon dioxide plasticization and conditioning effects in thick vs. thin glassy polymer films," Polymer **52**, 1619–1627 (2011).
[45]D. Dollimore, "Physical aging in amorphous polymers and other materials," Thermochim. Acta **54**, 242–243 (1982).
[46]B. W. Rowe, B. D. Freeman, and D. R. Paul, "Physical aging of membranes for gas separations," in *Membrane Engineering for the Treatment of Gases: Gas-separation Problems with Membranes* (The Royal Society of Chemistry, 2011), Chap. 3, pp. 58–83.
[47]T. M. Murphy, G. T. Offord, and D. R. Paul, "Fundamentals of membrane gas separation," in *Membrane Operations* (John Wiley & Sons, Ltd., 2009), pp. 63–82.
[48]B. W. Rowe, B. D. Freeman, and D. R. Paul, "Physical aging of ultrathin glassy polymer films tracked by gas permeability," Polymer **50**, 5565–5575 (2009).
[49]T. Visser, N. Masetto, and M. Wessling, "Materials dependence of mixed gas plasticization behavior in asymmetric membranes," J. Membr. Sci. **306**, 16–28 (2007).
[50]F. Y. Li, Y. Xiao, T.-S. Chung, and S. Kawi, "High-performance thermally self-cross-linked polymer of intrinsic microporosity (PIM-1) membranes for energy development," Macromolecules **45**, 1427–1437 (2012).
[51]Y. Liu, M. Ding, and J. Xu, "Gas permeabilities and permselectivity of photochemically crosslinked polyimides," J. Appl. Polym. Sci. **58**, 485–489 (1995).
[52]D. He, H. Susanto, and M. Ulbricht, "Photo-irradiation for preparation, modification and stimulation of polymeric membranes," Prog. Polym. Sci. **34**, 62–98 (2009).
[53]C. H. Lau, P. T. Nguyen, M. R. Hill, A. W. Thornton, K. Konstas, C. M. Doherty, R. J. Mulder, L. Bourgeois, A. C. Y. Liu, D. J. Sprouster, J. P. Sullivan, T. J. Bastow, A. J. Hill, D. L. Gin, and R. D. Noble, "Ending aging in super glassy polymer membranes," Angew. Chem. Int. Ed. **53**, 5322–5326 (2014).
[54]T. Corrado, Z. Huang, J. Aboki, and R. Guo, "Microporous polysulfones with enhanced separation performance via integration of the triptycene moiety," Ind. Eng. Chem. Res. **59**, 5351–5361 (2020).
[55]S. Bandehali, A. Ebadi Amooghin, H. Sanaeepur, R. Ahmadi, A. Fuoco, J. C. Jansen, and S. Shirazian, "Polymers of intrinsic microporosity and thermally rearranged polymer membranes for highly efficient gas separation," Sep. Purif. Technol. **278**, 119513 (2021).
[56]P. M. Budd, K. J. Msayib, C. E. Tattershall, B. S. Ghanem, K. J. Reynolds, N. B. McKeown, and D. Fritsch, "Gas separation membranes from polymers of intrinsic microporosity," J. Membr. Sci. **251**, 263–269 (2005).
[57]H. Sanaeepur, A. Ebadi Amooghin, S. Bandehali, A. Moghadassi, T. Matsuura, and B. Van Der Bruggen, "Polyimides in membrane gas separation: Monomer's molecular design and structural engineering," Prog. Polym. Sci. **91**, 80–125 (2019).
[58]M. Vinoba, M. Bhagiyalakshmi, Y. Alqaheem, A. A. Alomair, A. Pérez, and M. S. Rana, "Recent progress of fillers in mixed matrix membranes for CO_2 separation: A review," Sep. Purif. Technol. **188**, 431–450 (2017).
[59]J. Y. Park and D. R. Paul, "Correlation and prediction of gas permeability in glassy polymer membrane materials via a modified free volume based group contribution method," J. Membr. Sci. **125**, 23–39 (1997).
[60]D. Feldman, "Properties of Polymers, 3rd ed., edited by D. W. van Krevelen (Elsevier Science Publishers, Amsterdam, Oxford, New York, 1990), p. 875," J. Polym. Sci., Part B: Polym. Phys. **29**, 1654 (1991).
[61]M. Salame and S. Steingiser, "Barrier polymers," Polym.-Plast. Technol. Eng. **8**, 155–175 (1977).
[62]M. Salame, "Prediction of gas barrier properties of high polymers," Polym. Eng. Sci. **26**, 1543–1546 (1986).
[63]J. Bicerano, *Prediction of Polymer Properties*, 3rd ed. (CRC Press, Boca Raton, 2002).
[64]W. M. Lee, "Selection of barrier materials from molecular structure," Polym. Eng. Sci. **20**, 65–69 (1980).
[65]O. V. Malykh, A. Y. Golub, and V. V. Teplyakov, "Polymeric membrane materials: New aspects of empirical approaches to prediction of gas permeability parameters in relation to permanent gases, linear lower hydrocarbons and some toxic gases," Adv. Colloid Interface Sci. **164**, 89–99 (2011).
[66]*Materials Science of Membranes for Gas and Vapor Separation*, 1st ed., edited by Y. Yampolskii, I. Pinnau, and B. Freeman (Wiley, 2006).
[67]E. Mjolsness and D. DeCoste, "Machine learning for science: State of the art and future prospects," Science **293**, 2051–2055 (2001).
[68]M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**, 255–260 (2015).
[69]J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, and M. Lei, "Machine learning in materials science," InfoMat **1**, 338–358 (2019).
[70]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).
[71]L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, "Polymer informatics: Current status and critical next steps," Mater. Sci. Eng., R **144**, 100595 (2021).
[72]W. Sha, Y. Li, S. Tang, J. Tian, Y. Zhao, Y. Guo, W. Zhang, X. Zhang, S. Lu, Y.-C. Cao, and S. Cheng, "Machine learning in polymer informatics," InfoMat **3**, 353–361 (2021).
[73]K. Hatakeyama-Sato, "Recent advances and challenges in experiment-oriented polymer informatics," Polym. J. **55**, 117–131 (2023).
[74]C. Kuenneth, A. C. Rajan, H. Tran, L. Chen, C. Kim, and R. Ramprasad, "Polymer informatics with multi-task learning," Patterns **2**, 100238 (2021).
[75]D. J. Audus and J. J. de Pablo, "Polymer informatics: Opportunities and challenges," ACS Macro Lett. **6**, 1078–1082 (2017).
[76]C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, "Polymer genome: A data-powered polymer informatics platform for property predictions," J. Phys. Chem. C **122**, 17575–17585 (2018).
[77]S. Wu, H. Yamada, Y. Hayashi, M. Zamengo, and R. Yoshida, "Potentials and challenges of polymer informatics: Exploiting machine learning for polymer design," arXiv:2010.07683 (2020).
[78]M. M. Cencer, J. S. Moore, and R. S. Assary, "Machine learning for polymeric materials: An introduction," Polym. Int. **71**, 537–542 (2022).
[79]J. S. Peerless, N. J. B. Milliken, T. J. Oweida, M. D. Manning, and Y. G. Yingling, "Soft Matter Informatics: Current progress and challenges," Adv. Theory Simul. **2**, 1800129 (2019).
[80]S. S. Shukla, C. Kuenneth, and R. Ramprasad, "Polymer informatics beyond homopolymers," MRS Bull. **49**, 17–24 (2024).
[81]E. Ricci and M. G. D. Angelis, "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 (2024).
[82]See https://polymer.nims.go.jp/en/ for Polymer Database(PoLyInfo) - DICE:: National Institute for Materials Science.
[83]M. Wang and J. Jiang, "Accelerating discovery of high fractional free volume polymers from a data-driven approach," ACS Appl. Mater. Interfaces **14**, 31203 (2022).
[84]L. Tao, J. He, T. Arbaugh, J. R. McCutcheon, and Y. Li, "Machine learning prediction on the fractional free volume of polymer membranes," J. Membr. Sci. **665**, 121131 (2023).
[85]R. Ma and T. Luo, "PI1M: A benchmark database for polymer informatics," J. Chem. Inf. Model. **60**, 4684–4690 (2020).
[86]J. Yang, L. Tao, J. He, J. R. McCutcheon, and Y. Li, "Machine learning enables interpretable discovery of innovative polymers for gas separation membranes," Sci. Adv. **8**, eabn9545 (2022).
[87]S. Kim, C. M. Schroeder, and N. E. Jackson, "Open macromolecular genome: Generative design of synthetically accessible polymers," ACS Polym. Au **3**, 318–330 (2023).
[88]M. Ohno, Y. Hayashi, Q. Zhang, Y. Kaneko, and R. Yoshida, "SMiPoly: Generation of a synthesizable polymer virtual library using rule-based polymerization reactions," J. Chem. Inf. Model. **63**, 5539–5548 (2023).
[89]S. P. Tiwari, W. Shi, S. Budhathoki, J. Baker, A. K. Sekizkardes, L. Zhu, V. A. Kusuma, D. P. Hopkinson, and J. A. Steckel, "Creation of polymer datasets with targeted backbones for screening of high-performance membranes for gas separation," J. Chem. Inf. Model. **64**, 638–652 (2024).
[90]J. Bicerano, D. Rigby, C. Freeman, B. LeBlanc, and J. Aubry, "Polymer expert—A software tool for de novo polymer design," Comput. Mater. Sci. **235**, 112810 (2024).

[91]See https://research.csiro.au/virtualscreening/membrane-database-polymer-gas-separation-membranes/ for Virtual Screening of Materials, "Membrane Database - Polymer Gas Separation Membranes" (2012).
[92]H. W. H. Lai, F. M. Benedetti, J. M. Ahn, A. M. Robinson, Y. Wang, I. Pinnau, Z. P. Smith, and Y. Xia, "Hydrocarbon ladder polymers with ultrahigh permselectivity for membrane gas separations," Science **375**, 1390–1392 (2022).
[93]W. H. Lee, J. G. Seong, X. Hu, and Y. M. Lee, "Recent progress in microporous polymers from thermally rearranged polymers and polymers of intrinsic microporosity for membrane gas separation: Pushing performance limits and revisiting trade-off lines," J. Polym. Sci. **58**, 2450–2466 (2020).
[94]J. Yang (2024). "PolymerGasMembraneML (pgmML)," GitHub. https://github.com/jsunn-y/PolymerGasMembraneML
[95]Tiwari et al. (2023). "Polymer_backbones_paper_data," GitHub. https://github.com/sptiwari/polymer_backbones_paper_data
[96]See https://poly.chemnetbase.com for CHEMnetBASE (2024).
[97]See https://pppdb.uchicago.edu/ for Polymer Property Predictor and Database (2017).
[98]B. Blaiszik et al., "The materials data facility: Data services to advance materials science research," JOM **68**, 2045 (2016).
[99]R. Giro, H. Hsu, A. Kishimoto, T. Hama, R. F. Neumann, B. Luan, S. Takeda, L. Hamada, and M. B. Steiner, "AI powered, automated discovery of polymer membranes for carbon capture," npj Comput. Mater. **9**, 133 (2023).
[100]J. Choi, H. Kang, J. H. Lee, S. H. Kwon, and S. G. Lee, "Predicting the properties of high-performance epoxy resin by machine learning using molecular dynamics simulations," Nanomaterials **12**, 2353 (2022).
[101]R. Ma, H. Zhang, J. Xu, L. Sun, Y. Hayashi, R. Yoshida, J. Shiomi, J. Wang, and T. Luo, "Machine learning-assisted exploration of thermally conductive polymers based on high-throughput molecular dynamics simulations," Mater. Today Phys. **28**, 100850 (2022).
[102]G. S. Larsen, P. Lin, K. E. Hart, and C. M. Colina, "Molecular simulations of PIM-1-like polymers of intrinsic microporosity," Macromolecules **44**, 6944–6951 (2011).
[103]W. Fang, L. Zhang, and J. Jiang, "Polymers of intrinsic microporosity for gas permeation: A molecular simulation study," Mol. Simul. **36**, 992–1003 (2010).
[104]E. Kucukpinar and P. Doruker, "Molecular simulations of small gas diffusion and solubility in copolymers of styrene," Polymer **44**, 3607–3620 (2003).
[105]F. Mozaffari, H. Eslami, and J. Moghadasi, "Molecular dynamics simulation of diffusion and permeation of gases in polystyrene," Polymer **51**, 300–307 (2010).
[106]K. Golzar, H. Modarress, and S. Amjad-Iranagh, "Separation of gases by using pristine, composite and nanocomposite polymeric membranes: A molecular dynamics simulation study," J. Membr. Sci. **539**, 238–256 (2017).
[107]M. Zhao, C. Zhang, F. Yang, and Y. Weng, "Gas barrier properties of furan-based polyester films analyzed experimentally and by molecular simulations," Polymer **233**, 124200 (2021).
[108]H. Sun, S. J. Mumby, J. R. Maple, and A. T. Hagler, "An *ab initio* CFF93 all-atom force field for polycarbonates," J. Am. Chem. Soc. **116**, 2978–2987 (1994).
[109]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem. **25**, 1157–1174 (2004).
[110]H. Sun, P. Ren, and J. R. Fried, "The COMPASS force field: Parameterization and validation for phosphazenes," Comput. Theor. Polym. Sci. **8**, 229–246 (1998).
[111]R. P. White and J. E. G. Lipson, "Polymer free volume and its connection to the glass transition," Macromolecules **49**, 3987–4007 (2016).
[112]A. Thran, G. Kroll, and F. Faupel, "Correlation between fractional free volume and diffusivity of gas molecules in glassy polymers," J. Polym. Sci., Part B: Polym. Phys. **37**, 3344–3358 (1999).
[113]A. X. Wu, S. Lin, K. Mizrahi Rodriguez, F. M. Benedetti, T. Joo, A. F. Grosz, K. R. Storme, N. Roy, D. Syar, and Z. P. Smith, "Revisiting group contribution theory for estimating fractional free volume of microporous polymer membranes," J. Membr. Sci. **636**, 119526 (2021).
[114]A. Bondi, "van der Waals volumes and radii," J. Phys. Chem. **68**, 441–451 (1964).
[115]C. Jang, T. W. Sirk, J. W. Andzelm, and C. F. Abrams, "Comparison of cross-linking algorithms in molecular dynamics simulation of thermosetting polymers," Macromol. Theory Simul. **24**, 260–270 (2015).
[116]A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Phys. Nonlinear Phenom. **404**, 132306 (2020).
[117]S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "PubChem 2023 update," Nucl. Acids Res. **51**, D1373–D1380 (2023).
[118]D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988).
[119]M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation," Mach. Learn. Sci. Technol. **1**, 045024 (2020).
[120]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, "BigSMILES: A structurally-based line notation for describing macromolecules," ACS Cent. Sci. **5**, 1523–1531 (2019).
[121]See https://www.rdkit.org/ for RDKit (2024).
[122]L. Tao, V. Varshney, and Y. Li, "Benchmarking machine learning models for polymer informatics: An example of glass transition temperature," J. Chem. Inf. Model. **61**, 5395–5413 (2021).
[123]A. R. Katritzky, V. S. Lobanov, and M. Karelson, "QSPR: The correlation and quantitative prediction of chemical and physical properties from structure," Chem. Soc. Rev. **24**, 279 (1995).
[124]*Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology Vol. 672, edited by J. Bajorath (Humana Press, Totowa, NJ, USA, 2011).
[125]H. C. Patel, J. S. Tokarski, and A. J. Hopfinger, "Molecular modeling of polymers 16. Gaseous diffusion in polymers: A quantitative structure-property relationship (QSPR) analysis," Pharm. Res. **14**, 1349–1354 (1997).
[126]RDKit (2022). "RDKit MACCSkeys," GitHub. https://github.com/rdkit/rdkit/blob/aaded67a2ac3e1f329e98750c9841c424fa40d4e/rdkit/Chem/MACCSkeys.py
[127]D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model. **50**, 742–754 (2010).
[128]See https://www.daylight.com/ Daylight (2022).
[129]R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, "Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors," J. Chem. Inf. Comput. Sci. **27**, 82–85 (1987).
[130]J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," AI Open **1**, 57–81 (2020).
[131]S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," Comput. Soc. Networks **6**, 11 (2019).
[132]K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," arXiv:1810.00826 (2019).
[133]R. Gurnani, C. Kuenneth, A. Toland, and R. Ramprasad, "Polymer informatics at scale with multitask graph neural networks," Chem. Mater. **35**, 1560–1567 (2023).
[134]O. Queen, G. A. McCarver, S. Thatigotla, B. P. Abolins, C. L. Brown, V. Maroulas, and K. D. Vogiatzis, "Polymer graph neural networks for multitask property learning," npj Comput. Mater. **9**, 90 (2023).
[135]M. Aldeghi and C. W. Coley, "A graph representation of molecular ensembles for polymer property prediction," Chem. Sci. **13**, 10486–10498 (2022).
[136]J. Park, Y. Shim, F. Lee, A. Rammohan, S. Goyal, M. Shim, C. Jeong, and D. S. Kim, "Prediction and interpretation of polymer properties using the graph convolutional network," ACS Polym. Au **2**, 213–222 (2022).
[137]M. Zeng, J. N. Kumar, Z. Zeng, R. Savitha, V. R. Chandrasekhar, and K. Hippalgaonkar, "Graph convolutional neural networks for polymers property prediction," arXiv:1811.06231 (2018).
[138]G. Liu, T. Zhao, J. Xu, T. Luo, and M. Jiang, "Graph rationalization with environment-based augmentations," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD'22* (Association for Computing Machinery, New York, NY, USA, 2022), pp. 1069–1078.
[139]G. Liu, E. Inae, T. Zhao, J. Xu, T. Luo, and M. Jiang, "Data-centric learning from unlabeled graphs with diffusion model," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)* (Curran Associates Inc., Red Hook, NY, 2024), pp. 21039–21057.
[140]I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2014), Vol. 27.

[141]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. ukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), Vol. 30.
[142]S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised machine learning approach with chemical intuition," J. Chem. Inf. Model. **58**, 27–35 (2018).
[143]P. J. Flory and A. Vrij, "Melting points of linear-chain homologs. the normal paraffin hydrocarbons," J. Am. Chem. Soc. **85**, 3548–3553 (1963).
[144]R. Ma, Z. Liu, Q. Zhang, Z. Liu, and T. Luo, "Evaluating polymer representations via quantifying structure–property relationships," J. Chem. Inf. Model. **59**, 3110 (2019).
[145]L. Wang, C. Shao, H. Wang, and H. Wu, "Radial basis function neural networks-based modeling of the membrane separation process: Hydrogen recovery from refinery gases," J. Nat. Gas Chem. **15**, 230–234 (2006).
[146]A. Shahsavand and M. P. Chenar, "Neural networks modeling of hollow fiber membrane processes," J. Membr. Sci. **297**, 59–73 (2007).
[147]A. Ghadimi, M. Sadrzadeh, and T. Mohammadi, "Prediction of ternary gas permeation through synthesized PDMS membranes by using principal component analysis (PCA) and fuzzy logic (FL)," J. Membr. Sci. **360**, 509–521 (2010).
[148]M. Rostamizadeh, M. Rezakazemi, K. Shahidi, and T. Mohammadi, "Gas permeation through H_2-selective mixed matrix membranes: Experimental and neural network modeling," Int. J. Hydrogen Energy **38**, 1128–1135 (2013).
[149]S. Ebrahimi, S. Mollaiy-Berneti, H. Asadi, M. Peydayesh, F. Akhlaghian, and T. Mohammadi, "PVA/PES-amine-functional graphene oxide mixed matrix membranes for CO_2/CH_4 separation: Experimental and modeling," Chem. Eng. Res. Des. **109**, 647–656 (2016).
[150]M. Fakhroleslam, A. Samimi, S. A. Mousavi, and R. Rezaei, "Prediction of the effect of polymer membrane composition in a dry air humidification process via neural network modeling," Iran. J. Chem. Eng. **13**, 73–83 (2016).
[151]R. Nasir, H. Suleman, and K. Maqsood, "Multiparameter neural network modeling of facilitated transport mixed matrix membranes for carbon dioxide removal," Membranes **12**, 421 (2022).
[152]S. A. Abdollahi and S. F. Ranjbar, "Modeling the CO_2 separation capability of poly(4-methyl-1-pentane) membrane modified with different nanoparticles by artificial neural networks," Sci. Rep. **13**, 8812 (2023).
[153]H. Yin, M. Xu, Z. Luo, X. Bi, J. Li, S. Zhang, and X. Wang, "Machine learning for membrane design and discovery," Green Energy Environ. **9**, 54–70 (2024).
[154]J. Wang, K. Tian, D. Li, M. Chen, X. Feng, Y. Zhang, Y. Wang, and B. Van der Bruggen, "Machine learning in gas separation membrane developing: Ready for prime time," Sep. Purif. Technol. **313**, 123493 (2023).
[155]P. Xu, H. Chen, M. Li, and W. Lu, "New opportunity: machine learning for polymer materials design and discovery," Adv. Theory Simul. **5**, 2100565 (2022).
[156]R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *2017 International Conference on Machine Learning and Data Science (MLDS)* (IEEE, 2017), pp. 37–43.
[157]I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies* (IOS Press, 2007).
[158]S. Ray, "A quick review of machine learning algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (IEEE, 2019), pp. 35–39.
[159]M. Wessling, M. H. V. Mulder, A. Bos, M. van der Linden, M. Bos, and W. E. van der Linden, "Modelling the permeability of polymers: A neural network approach," J. Membr. Sci. **86**, 193–198 (1994).
[160]H. Hasnaoui, M. Krea, and D. Roizard, "Neural networks for the prediction of polymer permeability to gases," J. Membr. Sci. **541**, 541–549 (2017).
[161]G. Zhu, C. Kim, A. Chandrasekarn, J. D. Everett, R. Ramprasad, and R. P. Lively, "Polymer genome–based prediction of gas permeabilities in polymers," J. Polym. Eng. **40**, 451–457 (2020).
[162]J. W. Barnett, C. R. Bilchak, Y. Wang, B. C. Benicewicz, L. A. Murdock, T. Bereau, and S. K. Kumar, "Designing exceptional gas-separation polymer membranes using machine learning," Sci. Adv. **6**, eaaz4301 (2020).
[163]M. Zhao, C. Zhang, and Y. Weng, "Improved artificial neural networks (ANNs) for predicting the gas separation performance of polyimides," J. Membr. Sci. **681**, 121765 (2023).
[164]Q. Yuan, M. Longo, A. W. Thornton, N. B. McKeown, B. Comesaña-Gándara, J. C. Jansen, and K. E. Jelfs, "Imputation of missing gas permeability data for polymer membranes using machine learning," J. Membr. Sci. **627**, 119207 (2021).
[165]G. Liu, T. Zhao, E. Inae, T. Luo, and M. Jiang, "Semi-supervised graph imbalanced regression," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD'23* (Association for Computing Machinery, New York, NY, USA, 2023), pp. 1453–1465.
[166]A. N. Wilson, P. C. St John, D. H. Marin, C. B. Hoyt, E. G. Rognerud, M. R. Nimlos, R. M. Cywar, N. A. Rorrer, K. M. Shebek, L. J. Broadbelt, G. T. Beckham, and M. F. Crowley, "PolyID: Artificial intelligence for discovering performance-advantaged and sustainable polymers," Macromolecules **56**, 8547–8557 (2023).
[167]C. Kuenneth and R. Ramprasad, "polyBERT: A chemical language model to enable fully machine-driven ultrafast polymer informatics," Nat. Commun. **14**, 4099 (2023).
[168]Y. Basdogan, D. R. Pollard, T. Shastry, M. R. Carbone, S. K. Kumar, and Z.-G. Wang, "Machine learning-guided discovery of polymer membranes for CO_2 separation with genetic algorithm," J. Membr. Sci. **712**, 123169 (2024).
[169]J. Xu, A. Suleiman, G. Liu, M. Perez, R. Zhang, M. Jiang, R. Guo, and T. Luo, "Superior polymeric gas separation membrane designed by explainable graph machine learning," Cell Rep. Phys. Sci. **5**, 102067 (2024).
[170]Y. Yampolskii, S. Shishatskii, A. Alentiev, and K. Loza, "Group contribution method for transport property predictions of glassy polymers: Focus on polyimides and polynorbornenes," J. Membr. Sci. **149**, 203–220 (1998).
[171]S. Naeem, A. Ali, S. Anam, and M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," Int. J. Comput. Digital Syst. **13**, 911–921 (2023).
[172]B. K. Tripathy, A. Sundareswaran, and S. Ghela, *Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization* (CRC Press, 2021).
[173]A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, "Scaling deep learning for materials discovery," Nature **624**, 80–85 (2023).
[174]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies," Npj Comput. Mater. **1**, 15010 (2015).
[175]N. Schapin, M. Majewski, A. Varela-Rial, C. Arroniz, and G. D. Fabritiis, "Machine learning small molecule properties in drug discovery," Artif. Intell. Chem. **1**, 100020 (2023).
[176]A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the inorganic crystal structure database (ICSD): Accessibility in support of materials research and design," Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater. **58**, 364–369 (2002).
[177]C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data **6**, 60 (2019).
[178]W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," arXiv:1905.12265 (2020).
[179]Y. Rong, W. Huang, T. Xu, and J. Huang, "DropEdge: Towards deep graph convolutional networks on node classification," arXiv:1907.10903 (2020).
[180]M. Sun, J. Xing, H. Wang, B. Chen, and J. Zhou, "MoCL: Data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD'21* (Association for Computing Machinery, New York, NY, USA, 2021), pp. 3585–3594.
[181]T. Zhao, W. Jin, Y. Liu, Y. Wang, G. Liu, S. Günnemann, N. Shah, and M. Jiang, "Graph data augmentation for graph machine learning: A survey," arXiv:2202.08871 (2023).
[182]K. Ding, Z. Xu, H. Tong, and H. Liu, "Data augmentation for deep graph learning: A survey," ACM SIGKDD Explor. Newsl. **24**, 61–77 (2022).
[183]D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning* (ICML, Atlanta, 2013), Vol. 3, p. 896.
[184]Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning* (PMLR, 2016), pp. 1050–1059.
[185]N. Tagasovska and D. Lopez-Paz, "Single-model uncertainties for deep learning," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32.

[186]A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2020), Vol. 33, pp. 14927–14937.
[187]B. Settles, "Active learning literature survey," Technical Report No. TR1648 (University of Wisconsin-Madison Department of Computer Sciences, 2009).
[188]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733 (2018).
[189]A. Tharwat and W. Schenck, "A survey on active learning: State-of-the-art, practical challenges and research directions," Mathematics **11**, 820 (2023).
[190]C. Kim, A. Chandrasekaran, A. Jha, and R. Ramprasad, "Active-learning and materials design: The example of high glass transition temperature polymers," MRS Commun. **9**, 860–866 (2019).
[191]A. Rakhimbekova, A. Lopukhov, N. Klyachko, A. Kabanov, T. I. Madzhidov, and A. Tropsha, "Efficient design of peptide-binding polymers using active learning approaches," J. Controlled Release **353**, 903–914 (2023).
[192]P. S. Ramesh and T. K. Patra, "Polymer sequence design via active learning," arXiv:2111.09659 (2021).
[193]S. Pruksawan, G. Lambard, S. Samitsu, K. Sodeyama, and M. Naito, "Prediction and optimization of epoxy adhesive strength from a small dataset through active learning," Sci. Technol. Adv. Mater. **20**, 1010–1021 (2019).
[194]J. Xu and T. Luo, "Unlocking enhanced thermal conductivity in polymer blends through active learning," npj Comput. Mater. **10**, 74 (2024).
[195]F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proc. IEEE **109**, 43–76 (2021).
[196]K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J. Big Data **3**, 9 (2016).
[197]R. Ma, Y. J. Colón, and T. Luo, "Transfer learning study of gas adsorption in metal–organic frameworks," ACS Appl. Mater. Interfaces **12**, 34041–34048 (2020).
[198]S. Wu, Y. Kondo, M. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, and R. Yoshida, "Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm," npj Comput. Mater. **5**, 66 (2019).
[199]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32.
[200]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 k molecules," Sci. Data **1**, 140022 (2014).
[201]G. Liu and M. Jiang, "Transfer learning with diffusion model for polymer property prediction" in Workshop on Machine Learning for Materials (2023).
[202]Z. Zhang, Q. Liu, H. Wang, C. Lu, and C.-K. Lee, "Motif-based graph self-supervised learning for molecular property prediction," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2021), vol. 34, pp. 15870–15882.
[203]Y. You, T. Chen, Y. Shen, and Z. Wang, "Graph contrastive learning automated," in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), pp. 12121–12132.
[204]D. Kim, J. Baek, and S. J. Hwang, "Graph self-supervised learning with accurate discrepancy learning," in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)* (Curran Associates, 2022), Vol. 35, pp. 14085–14098.
[205]R. Sun, H. Dai, and A. W. Yu, "Does GNN pretraining help molecular representation?," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS'22* (Curran Associates, Inc., Red Hook, NY, USA, 2024), pp. 12096–12109.
[206]F. L. Lee, J. Park, S. Goyal, Y. Qaroush, S. Wang, H. Yoon, A. Rammohan, and Y. Shim, "Comparison of machine learning methods towards developing interpretable polyamide property prediction," Polymers **13**, 3653 (2021).
[207]B. M. Greenwell, "pdp: An R package for constructing partial dependence plots," R J. **9**, 421 (2017).
[208]A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," J. Comput. Graph. Stat. **24**, 44–65 (2015).
[209]M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'16* (Association for Computing Machinery, New York, NY, USA, 2016), pp. 1135–1144.
[210]S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), Vol. 30.
[211]K. Sattari, Y. Xie, and J. Lin, "Data-driven algorithms for inverse design of polymers," Soft Matter **17**, 7607–7622 (2021).
[212]C. Kim, R. Batra, L. Chen, H. Tran, and R. Ramprasad, "Polymer design using genetic algorithm and machine learning," Comput. Mater. Sci. **186**, 110067 (2021).
[213]S. Takeda, T. Hama, H.-H. Hsu, V. A. Piunova, D. Zubarev, D. P. Sanders, J. W. Pitera, M. Kogoh, T. Hongo, Y. Cheng, W. Bocanett, H. Nakashika, A. Fujita, Y. Tsuchiya, K. Hino, K. Yano, S. Hirose, H. Toda, Y. Orii, and D. Nakano, "Molecular inverse-design platform for material industries," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'20* (Association for Computing Machinery, New York, NY, USA, 2020), pp. 2961–2969.
[214]W. Gao, T. Fu, J. Sun, and C. Coley, "Sample efficiency matters: A benchmark for practical molecular optimization," in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)* (Curran Associates, 2022), Vol. 35, pp. 21342–21357.
[215]C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, "Generative models for molecular discovery: Recent advances and challenges," WIREs Comput. Mol. Sci. **12**, e1608 (2022).
[216]C. Vignac, I. Krawczuk, A. Siraudin, B. Wang, V. Cevher, and P. Frossard, "DiGress: Discrete denoising diffusion for graph generation," arXiv:2209.14734 (2023).
[217]M. Popova, M. Shvets, J. Oliva, and O. Isayev, "MolecularRNN: Generating realistic molecular graphs with optimized properties," arXiv:1905.13372 (2019).
[218]Y. Xie, C. Shi, H. Zhou, Y. Yang, W. Zhang, Y. Yu, and L. Li, "MARS: Markov molecular sampling for multi-objective drug discovery," arXiv:2103.10432 (2021).
[219]W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," in *Proceedings of the 35th International Conference on Machine Learning* (PMLR, 2018), pp. 2323–2332.
[220]N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher, "GuacaMol: Benchmarking models for de novo molecular design," J. Chem. Inf. Model. **59**, 1096–1108 (2019).
[221]G. Liu, J. Xu, T. Luo, and M. Jiang, "Inverse molecular design with multi-conditional diffusion guidance," arXiv:2401.13858 (2024).
[222]Z. Liang, Z. Li, S. Zhou, Y. Sun, J. Yuan, and C. Zhang, "Machine-learning exploration of polymer compatibility," Cell Rep. Phys. Sci. **3**, 100931 (2022).
[223]P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang, and R. Ramprasad, "A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing," npj Comput. Mater. **9**, 52 (2023).
[224]Y. Wang, X. Ma, B. S. Ghanem, F. Alghunaimi, I. Pinnau, and Y. Han, "Polymers of intrinsic microporosity for energy-intensive membrane-based gas separations," Mater. Today Nano **3**, 69–95 (2018).
[225]A. A. Volk, R. W. Epps, D. T. Yonemoto, B. S. Masters, F. N. Castellano, K. G. Reyes, and M. Abolhasani, "AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning," Nat. Commun. **14**, 1403 (2023).
[226]J. A. Bennett and M. Abolhasani, "Autonomous chemical science and engineering enabled by self-driving laboratories," Curr. Opin. Chem. Eng. **36**, 100831 (2022).
[227]N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, and G. Ceder, "An autonomous laboratory for the accelerated synthesis of novel materials," Nature **624**, 86–91 (2023).
[228]E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," npj Comput. Mater. **8**, 84 (2022).
[229]Y. C. Teo, H. W. H. Lai, and Y. Xia, "Synthesis of ladder polymers: Developments, challenges, and opportunities," Chem. A Eur. J. **23**, 14101–14112 (2017).

[230]L. M. Robeson, *Polymer Blends: A Comprehensive Review* (Hanser, Munich Cincinnati, Ohio, 2007).
[231]C. Kuenneth, W. Schertzer, and R. Ramprasad, "Copolymer informatics with multitask deep neural networks," Macromolecules **54**, 5957–5961 (2021).
[232]L. Tao, T. Arbaugh, J. Byrnes, V. Varshney, and Y. Li, "Unified machine learning protocol for copolymer structure-property predictions," STAR Protoc. **3**, 101875 (2022).
[233]K.-H. Tu, H. Huang, S. Lee, W. Lee, Z. Sun, A. Alexander-Katz, and C. A. Ross, "Machine learning predictions of block copolymer self-assembly," Adv. Mater. **32**, e2005713 (2020).
[234]P. Ertl and A. Schuffenhauer, "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions," J. Cheminform. **1**, 8 (2009).
[235]C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, "SCScore: Synthetic complexity learned from a reaction corpus," J. Chem. Inf. Model. **58**, 252–261 (2018).
[236]M. Voršilák, M. Kolář, I. Čmelo, and D. Svozil, "SYBA: Bayesian estimation of synthetic accessibility of organic compounds," J. Cheminform. **12**, 35 (2020).
[237]A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist, and J.-L. Reymond, "Retrosynthetic accessibility score (RAscore)—rapid machine learned synthesizability classification from AI driven retrosynthetic planning," Chem. Sci. **12**, 3339–3349 (2021).
[238]J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, and H. Y. Ando, "Route designer: A retrosynthetic analysis tool utilizing automated retrosynthetic rule generation," J. Chem. Inf. Model. **49**, 593–602 (2009).
[239]C. W. Coley, W. H. Green, and K. F. Jensen, "RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application," J. Chem. Inf. Model. **59**, 2529–2537 (2019).
[240]B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande, "Retrosynthetic reaction prediction using neural sequence-to-sequence models," ACS Cent. Sci. **3**, 1103–1113 (2017).
[241]B. Chen, T. Shen, T. S. Jaakkola, and R. Barzilay, "Learning to make generalizable and diverse predictions for retrosynthesis," arXiv:1910.09688 (2019).
[242]K. Lin, Y. Xu, J. Pei, and L. Lai, "Automatic retrosynthetic route planning using template-free models," Chem. Sci. **11**, 3355–3364 (2020).
[243]L. Chen, J. Kern, J. P. Lightstone, and R. Ramprasad, "Data-assisted polymer retrosynthesis planning," Appl. Phys. Rev. **8**, 031405 (2021).
[244]A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh, and W. J. Peña Ccoa, "Do large language models know chemistry?," chemRxiv.
[245]K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White, and B. Blaiszik, "14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon," Digital Discovery **2**, 1233–1250 (2023).
[246]T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest, and X. Zhang, "What can large language models do in chemistry? A comprehensive benchmark on eight tasks," in *37th Conference on Neural Information Processing Systems (NeurIPS 2023)* (Curran Associates, 2023), pp. 59662–59688.
[247]F. W. Mercer and M. T. McKenzie, "Dielectric and thermal characterization of fluorinated polyimides containing heterocyclic moieties," High Perform. Polym. **5**, 97–106 (1993).
[248]A. Ghosh, S. K. Sen, S. Banerjee, and B. Voit, "Solubility improvements in aromatic polyimides by macromolecular engineering," RSC Adv. **2**, 5900–5926 (2012).
[249]N. J. Olsavsky, V. M. Kearns, C. P. Beckman, P. L. Sheehan, F. J. Burpo, H. D. Bahaghighat, and E. A. Nagelli, "Research and regulatory advancements on remediation and degradation of fluorinated polymer compounds," Appl. Sci. **10**, 6921 (2020).
[250]A. R. Bock and B. E. Laird, "PFAS regulations: Past and present and their impact on fluoropolymers" in *Perfluoroalkyl Substances* (The Royal Society of Chemistry, 2022), p. 634.
[251]R. Lohmann, I. T. Cousins, J. C. DeWitt, J. Glüge, G. Goldenman, D. Herzke, A. B. Lindstrom, M. F. Miller, C. A. Ng, S. Patton, M. Scheringer, X. Trier, and Z. Wang, "Are fluoropolymers really of low concern for human and environmental health and separate from other PFAS?," Environ. Sci. Technol. **54**, 12820–12828 (2020).
[252]I. T. Cousins, G. Goldenman, D. Herzke, R. Lohmann, M. Miller, C. A. Ng, S. Patton, M. Scheringer, X. Trier, L. Vierke, Z. Wang, and J. C. DeWitt, "The concept of essential use for determining when uses of PFASs can be phased out," Environ. Sci.: Processes Impacts **21**, 1803–1815 (2019).