Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: S. Alosious, Y. Liu, J. Xu, G. Liu, R. Zhang, M. Jiang and T. Luo, *Digital Discovery*, 2026, DOI: 10.1039/D6DD00206D.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00206D

ADEPT–PolyGraphMT: Automated Molecular Simulation and Multi-Task Multi-Fidelity Machine Learning for Polymer Property Generation and Prediction

Sobin Alosious[1,2], Yuhan Liu[3], Jiaxin Xu[2], Gang Liu[4], Renzheng Zhang[2], Meng Jiang[1,4], Tengfei Luo[1,2,5,6*]

[1]Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN, 46556, USA.
[2]Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA.
[3]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, 46556, USA.
[4]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA.
[5]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA.
[6]Center for Sustainable Energy at Notre Dame (ND Energy), University of Notre Dame, Notre Dame, IN, 46556, USA.

*Corresponding author(s). E-mail(s): tluo@nd.edu;

Abstract

The discovery of polymers with targeted properties is challenged by the vast chemical design space and the limited availability of consistent, high-quality data across multiple properties. In this work, an integrated polymer informatics framework is presented that combines the Automated molecular Dynamics Engine for Polymer simulaTions (ADEPT) workflow with multi-task and multi-fidelity machine learning (PolyGraphMT). Polymer repeat units are represented as molecular graphs and processed using a graph neural network to learn structure–property relationships. Starting from SMILES representations for monomers, ADEPT automates the construction of atomistic models and the evaluation of their properties using molecular dynamics simulations and density

View Article Online
DOI: 10.1039/D6DD00206D

functional theory calculations. The simulation data are combined with curated experimental data and group contribution theory estimates to construct a unified dataset of approximately 62,000 polymer property values spanning 28 properties across thermal, mechanical, transport, electronic, optical, and structural properties. Using this dataset, inter-property correlations are analyzed, and multi-task learning strategies are evaluated for joint property prediction. The results show that multi-task models achieve performance comparable to single-task models in data-rich regimes and exhibit superior accuracy as training data become limited. In addition, fidelity-aware learning demonstrates improved agreement with experimental data for representative multi-fidelity properties. The trained models are further applied to large-scale property prediction for polymers in the PolyInfo database (∼13,000 polymers) and the PI1M virtual polymer library (∼1 million polymers), producing physically consistent property distributions across a broad chemical space. Overall, the proposed framework provides a structured approach for scalable prediction and screening of polymer properties across multiple property types and data fidelity levels.

Keywords: Polymer informatics, Multi-task learning, Multi-fidelity machine learning, Graph neural networks, High-throughput molecular simulations

1 Introduction

Polymers occupy a central role in modern materials science due to their exceptional chemical diversity and tunability across thermal, mechanical, transport, and electronic properties [1, 2]. This versatility underpins applications ranging from structural materials and membranes to energy storage, electronics, and biomedical devices [3–5]. At the same time, identifying polymers with targeted combinations of properties remains a fundamental challenge, as the accessible chemical space is effectively unbounded and experimental characterization is both time-consuming and property-specific [6, 7].

Data availability is a central requirement for materials informatics, as machine learning (ML) models rely critically on the quality, quantity, and diversity of training data [8, 9]. In inorganic materials science and small-molecule chemistry, large open databases derived from high-throughput first-principles calculations and curated experimental measurements, such as the Materials Project, AFLOW, OQMD, and QM9, have played a decisive role in advancing data-driven materials discovery [10–13]. In the polymer domain, dedicated databases such as PolyInfo and Polymer Genome have enabled important progress by compiling experimental measurements and first-principles data for selected properties [7, 14]. While these databases have enabled important progress in polymer informatics, the underlying data remain highly sparse and heterogeneous. For most polymers, only a limited number of properties are reported, often measured under different processing histories, molecular weights, and testing conditions.

Beyond database development, recent advances in polymer informatics have also focused on polymer representations, large-scale generative design frameworks, and foundation-model-inspired learning approaches. Several studies have emphasized the

View Article Online
DOI: 10.1039/D6DD00206D

importance of chemically meaningful polymer representations for enabling scalable ML workflows. In particular, the BigSMILES formalism introduced a stochastic extension of SMILES specifically designed to describe macromolecular structures, including copolymers and branched architectures, while preserving compatibility with text-based cheminformatics pipelines [15]. More recently, open polymer data infrastructures such as CRIPT have further highlighted the need for standardized and interoperable polymer data representations for community-scale informatics efforts [16]. At the same time, generative and language-model-based approaches have rapidly expanded the scope of AI-driven polymer design. Jackson and co-workers introduced the Open Macromolecular Genome (OMG), a large-scale polymer database and generative framework designed to enable synthetically accessible polymer discovery through reaction-aware generative modeling [17]. In parallel, Ramprasad and co-workers developed polymer language-model frameworks such as polyBERT, which treat polymer SMILES representations as a chemical language and enable fully machine-driven polymer informatics pipelines through transformer-based representation learning [18]. Earlier efforts from the same group also demonstrated the effectiveness of multi-task learning and multi-fidelity learning strategies for polymer property prediction using heterogeneous experimental and computational datasets [19, 20]. A variety of polymer representation strategies have additionally been proposed for polymer informatics, including periodicity-aware graph encodings [21], polymer fingerprinting and descriptor-based approaches [19], and simpler SMILES-derived molecular descriptors. While these representations differ in their treatment of polymer connectivity and periodicity, many have demonstrated competitive performance across polymer property prediction tasks. The primary focus of the present work is therefore not the development of a fundamentally new polymer representation scheme, but rather the integration of heterogeneous multi-property and multi-fidelity datasets within a unified graph-learning framework. Consequently, the overall workflow is expected to remain compatible with alternative polymer representation strategies. Together, these studies highlight the rapidly evolving ecosystem of polymer representations, databases, generative models, and transferable learning frameworks that are shaping modern polymer informatics.

In addition, recent studies have demonstrated the feasibility of high-throughput molecular dynamics (MD) workflows for generating polymer property data at scale [22, 23]. Hayashi et al. [23] introduced RadonPy, an open-source Python framework that enables fully automated, high-throughput all-atom MD simulations for polymer property prediction. The workflow integrates polymer structure generation from SMILES representations, force-field assignment, equilibration, equilibrium and nonequilibrium MD simulations, and automated trajectory post-processing within a unified pipeline, enabling systematic property generation directly from chemical structure.

More recently, Yoshida et al. [24] extended such automated simulation workflows toward the construction of PolyOmics, an omics-scale computational polymer database generated using fully automated MD pipelines. The database comprises physical property data for over 10^5 polymeric materials and includes thermal, mechanical, dielectric, and transport properties computed under standardized simulation protocols. PolyOmics further demonstrated that ultralarge computational polymer

View Article Online
DOI: 10.1039/D6DD00206D

datasets can be effectively leveraged for machine learning through simulation-to-real (Sim2Real) transfer learning, where models pretrained on computational data are fine-tuned using limited experimental measurements. The study additionally reported power-law scaling behavior in prediction accuracy with increasing dataset size, highlighting the potential of ultralarge simulation datasets as foundational resources for polymer informatics.

At the same time, it is well established that simulation-derived properties, particularly those obtained from classical MD, often exhibit systematic deviations from experimental measurements due to force-field limitations, finite-chain effects, incomplete equilibration, and the classical treatment of atomic vibrations [25, 26]. Nevertheless, MD predictions frequently preserve strong correlations with experimental trends for multiple polymer properties, including thermal conductivity (κ), specific heat capacity (C_p), density (ρ), and glass transition temperature (T_g), across chemically diverse polymer classes [27–30]. This suggests that simulation data can serve as informative lower-fidelity proxies for learning global structure–property relationships, even when systematic offsets relative to experiment remain present.

ML has emerged as a powerful framework for predicting polymer properties directly from chemical structure, enabling rapid evaluation of materials that would be infeasible to characterize experimentally or computationally at scale [8]. Single-task learning approaches, in which independent models are trained for individual properties, have demonstrated promising accuracy for targets such as T_g, elastic moduli, and dielectric response [31]. However, these models are inherently data-inefficient for sparsely sampled properties and do not exploit the strong correlations among polymer properties [32].

Multi-task learning addresses these limitations by learning shared representations across multiple tasks, enabling information transfer between related properties. Kuenneth et al. [19] demonstrated that multi-task neural networks trained on a dataset comprising 36 polymer properties for approximately 13,000 polymers consistently outperform single-task models, with error reductions of up to 20–30% for sparsely sampled properties. Similarly, Queen et al. [33] developed POLYMERGNN, a graph neural network framework for multi-task prediction using an experimentally curated dataset of over 240 synthesized polyesters, achieving predictive performance of $R^2 \approx 0.72$ for T_g and $R^2 \approx 0.70$ for intrinsic viscosity under five-fold cross-validation. Although this work focused exclusively on experimental data, it demonstrated the effectiveness of shared latent representations for learning correlated polymer properties.

Despite these advances, polymer property datasets remain inherently heterogeneous, combining experimental measurements with MD, density functional theory (DFT), and group contribution (GC) data that differ substantially in fidelity. Treating all data sources as equally reliable can bias learned models toward abundant but lower-fidelity data, while discarding computational data limits chemical coverage. Although multi-fidelity learning strategies have been explored in broader materials science contexts, their systematic integration with multi-task learning for polymer property prediction remains limited. [34–36]

The present work differs from existing Sim2Real transfer-learning frameworks in several important ways. PolyOmics primarily focuses on pretraining ML models on

View Article Online
DOI: 10.1039/D6DD00206D

large homogeneous MD datasets followed by fine-tuning on limited experimental data for downstream tasks [24]. In contrast, the framework developed here is designed around the joint integration of heterogeneous datasets spanning multiple physical domains, properties, and fidelity levels within a unified multi-task graph-learning framework. Rather than treating simulation and experiment through a sequential pretraining–fine-tuning paradigm alone, the present approach incorporates experimental measurements together with multiple computational data sources of differing physical accuracy and computational cost, including DFT, MD, and GC estimates, through fidelity-aware weighted optimization and task-grouped multi-property learning. Importantly, not all properties contain the same fidelity combinations, and the effective multi-fidelity coverage therefore remains strongly property-dependent. This formulation enables the model to exploit both cross-property correlations and complementary information across heterogeneous fidelities while remaining scalable to diverse polymer property domains.

In this work, we introduce ADEPT, an Automated molecular Dynamics Engine for Polymer simulaTions, for systematic generation of polymer properties from chemical structure, and integrate it with a newly developed multi-task, multi-fidelity ML framework (referred to here as PolyGraphMT) for large-scale polymer property prediction. ADEPT is designed for high-performance computing, enabling massively parallel, high-throughput property calculations using property-specific MD workflows, including both equilibrium and nonequilibrium methods. It combines atomistic MD and DFT calculations with curated experimental data and GC estimates to construct a heterogeneous dataset of approximately 62,000 data points spanning 28 properties across thermal, mechanical, transport, electronic, optical, and structural domains.

Using this dataset, we analyze inter-property correlations, evaluate task grouping strategies, and quantify the benefits of multi-task learning under varying data availability while accounting for data fidelity. The PolyGraphMT framework represents polymer repeat units as molecular graphs and employs graph neural networks to learn structure–property relationships across multiple properties and fidelity levels.

The integrated ADEPT–PolyGraphMT framework enables (i) simultaneous multi-property prediction, (ii) fidelity-aware learning across heterogeneous data sources, and (iii) scalable screening over large polymer spaces. The resulting models enable unified prediction and consistent screening across both experimental polymer databases ($\sim$ 13,000 polymers) and large virtual polymer libraries ($\sim$ 1 million polymers), which is not achievable with existing fragmented workflows. This establishes a scalable framework for data-driven polymer discovery by integrating property generation, model training, and large-scale screening. The overall framework integrates the ADEPT simulation engine with the PolyGraphMT learning architecture.

2 Methodology

2.1 ADEPT Workflow

Figure 1 presents an overview of the end-to-end methodology developed in this work. The workflow begins with polymer chemical structures represented using standardized repeat-unit SMILES containing explicit polymerization points, which provide a

View Article Online
DOI: 10.1039/D6DD00206D

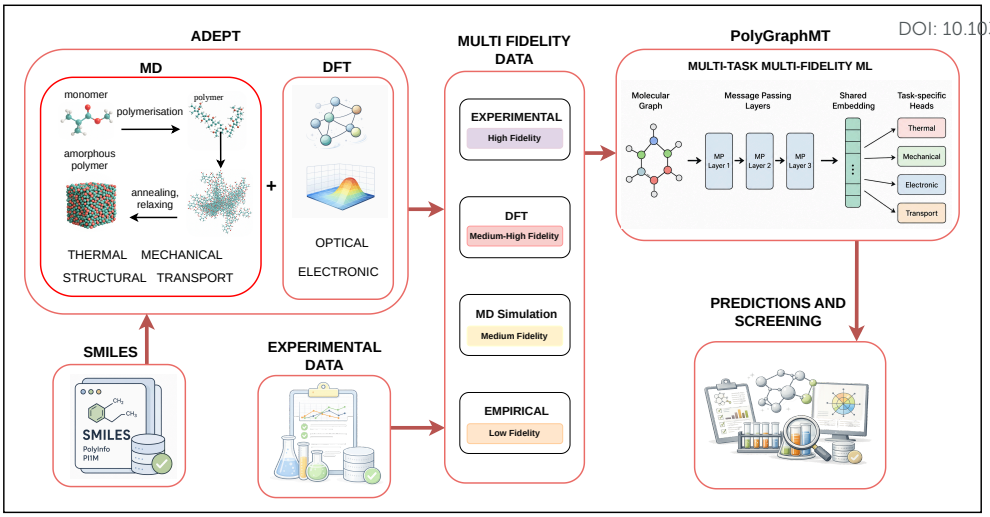

Fig. 1 Overview of the ADEPT and PolyGraphMT workflow. Polymer SMILES are processed through the ADEPT pipeline to generate thermal, mechanical, structural, transport, electronic, and optical properties using MD and DFT. These computational data are combined with curated experimental measurements to construct a unified dataset, which is used to train multi-task, multi-fidelity ML models for polymer property prediction and large-scale screening.

common input representation for all subsequent stages. These repeat units are used both for graph-based representation learning and for constructing finite linear-chain polymer models within the ADEPT simulation workflow. Consequently, the current framework does not explicitly represent branching, copolymer sequence distributions, polydispersity, or chain-length-dependent effects, which remain important limitations of the present approach. These structures are processed using the ADEPT framework to automatically generate atomistic polymer models, perform equilibration and production simulations, and compute a broad range of polymer properties using all-atom MD. DFT calculations are additionally employed to obtain selected electronic and optical properties that are not accessible from classical molecular simulations. To improve clarity, detailed methodologies, including polymer structure generation from SMILES, amorphous polymer construction and equilibration protocols, property calculations using MD simulations, and electronic and dielectric property evaluations from DFT, are provided in the Methods section (see Methods – ADEPT Workflow: Polymer Simulation and Property Generation). The simulation-derived properties are combined with experimentally measured data curated from the literature to construct a unified dataset spanning thermal, mechanical, transport, electronic, optical, and structural property domains. Due to the different origins of the data, the resulting dataset contains multiple levels of data fidelity, with experimental measurements treated as the highest fidelity, followed by DFT, MD simulations, and GC estimates. While experimental data from the literature will undoubtedly contain noise depending on the source of the data, they are the best reflection of true polymers in the real world. Any computational models have certain levels of approximations. This dataset is used to train

View Article Online
DOI: 10.1039/D6DD00206D

multi-task, multi-fidelity ML models that learn shared chemical representations across properties while accounting for systematic differences between experimental and computational data sources. The trained models are then applied to property prediction and large-scale screening of both experimentally reported polymers and virtual polymer libraries, enabling efficient exploration of polymer chemical space. The ADEPT workflow and the associated ML pipelines are released as open-source software, as described in the Data and Code Availability section.

2.2 PolyGraphMT Workflow

2.2.1 Learning objective and data structure

To enable scalable prediction of polymer properties beyond direct molecular simulations, a multi-task, multi-fidelity ML framework was developed using data generated by the ADEPT workflow, along with complementary experimental and GC datasets. The objective of the model is to jointly learn multiple polymer properties while accounting for differences in data fidelity arising from experimental measurements, MD simulations, DFT calculations, and GC estimates. The implementation builds upon and substantially extends the `torch-molecule` framework [37], which was originally developed for single-task molecular property prediction.

Each polymer is represented by its repeat-unit chemical structure and may be associated with one or more target properties evaluated at different fidelity levels. Let $i \in \{1, \ldots, N\}$ index polymers, $p \in \mathcal{P}$ index target properties (for example, C_p, κ, and T_g), and $f \in \mathcal{F}$ index data fidelity levels corresponding to experimental, MD, DFT, or GC sources.

For a given polymer i, property p, and fidelity level f, the observed value is denoted as $y_{i,p}^{(f)}$. In practice, only a subset of all possible (i, p, f) combinations is available, leading to a sparse, heterogeneous supervision structure. The learning objective is to construct a predictive model that leverages shared information across related properties and fidelity levels, while treating experimental data as the highest-fidelity reference during model training and evaluation.

2.2.2 Molecular graph representation

Polymer repeat units were represented as molecular graphs constructed from SMILES strings. For each polymer i, the molecular graph is defined as $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$, where $\mathcal{V}_i$ denotes the set of atoms (nodes) and $\mathcal{E}_i$ denotes the set of covalent bonds (edges). Node features encode atomic identity and local chemical attributes, while edge features represent bond connectivity and bond order.

This graph-based representation preserves the chemical structure of the repeat unit and enables the use of message-passing neural networks to learn structure–property relationships directly from molecular connectivity [38].

The graph representation used for machine learning is constructed from standardized polymer repeat units rather than explicitly polymerized multi-chain structures. Consequently, message passing occurs only within the repeat-unit molecular graph and does not explicitly propagate across covalently connected repeat units in a full polymer chain. To reduce ambiguities associated with terminal capping or monomer

View Article Online
DOI: 10.1039/D6DD00206D

fragmentation, polymers are represented using repeat-unit SMILES containing explicit polymerization points denoted by *. Finite-chain atomistic polymer models are generated separately within the ADEPT workflow for MD simulations, with chain lengths selected to balance computational tractability and approximate convergence for high-throughput property calculations.

2.2.3 Graph neural network encoder

A shared graph neural network encoder was used to map each polymer graph $\mathcal{G}_i$ to a fixed-dimensional latent representation $\mathbf{z}_i \in \mathbb{R}^d$. This mapping is expressed as

$$\mathbf{z}_i = f_{\boldsymbol{\theta}}(\mathcal{G}_i), \tag{1}$$

where $f_{\boldsymbol{\theta}}$ denotes the encoder parameterized by weights $\boldsymbol{\theta}$.

The encoder comprises multiple graph convolution layers in which atomic feature vectors are iteratively updated through neighborhood aggregation. At each layer, atomic representations are updated using information from directly bonded neighbors, followed by nonlinear activation and normalization. After message passing, a global pooling operation aggregates the node-level representations into a single polymer-level embedding $\mathbf{z}_i$. This shared embedding serves as the common input for all downstream property prediction tasks and follows standard message-passing neural network implementations [38]. Additional implementation details of the network architecture and training configuration are provided in the Methods section (see Methods – PolyGraphMT Framework).

2.2.4 Property-specific prediction heads

For each target property p, a dedicated prediction head maps the shared latent representation $\mathbf{z}_i$ to a scalar prediction,

$$\hat{y}_{i,p} = g_p(\mathbf{z}_i), \tag{2}$$

where g_p denotes a property-specific multilayer perceptron with parameters that are independent across properties.

All prediction heads operate on the same encoder output $\mathbf{z}_i$, which enforces a shared representation across tasks while allowing property-dependent nonlinear transformations in the final layers. This design supports information sharing among related properties without requiring identical functional forms for different prediction targets [39].

2.2.5 Multi-fidelity supervision strategy

When a target property p is available at multiple fidelity levels for a given polymer, all corresponding observations are included during training. The model produces a single property prediction $\hat{y}_{i,p}$ from the shared latent representation, and this prediction is compared independently against each available fidelity-specific observation $y_{i,p}^{(f)}$ through the training loss.

View Article Online
DOI: 10.1039/D6DD00206D

This formulation incorporates heterogeneous data sources without introducing separate prediction branches or explicit bias-correction terms for individual fidelity levels. Differences in data fidelity are handled through loss weighting during optimization, allowing lower-fidelity data to contribute to representation learning while higher-fidelity experimental data exert a stronger influence during model fitting and evaluation. Detailed formulations of the loss function, normalization strategy, and training procedures are provided in the Methods section (see Methods – PolyGraphMT Framework: Multi-Task Multi-Fidelity Learning).

Accordingly, the present framework represents a weighted heterogeneous learning strategy rather than an explicit discrepancy-learning or fidelity-mapping architecture commonly used in parts of the multi-fidelity learning literature. The current implementation emphasizes shared representation learning across heterogeneous fidelity sources using a common property-specific prediction head and fidelity-aware loss weighting.

3 Results and Discussions

The results presented in this section are obtained using the ADEPT workflow, which enables end-to-end evaluation of polymer properties starting from repeat-unit SMILES representations [40]. For a given polymer structure, ADEPT automates polymer construction, force-field assignment, MD simulations, electronic-structure calculations, and data processing to generate a wide range of thermodynamic, mechanical, transport, electronic, dielectric, and structural properties in a consistent and scalable manner. Data generated using ADEPT, including MD and DFT results, are combined with experimentally measured values and GC estimates to construct a unified dataset of polymer properties. This integrated dataset spans multiple property classes and data fidelity levels, providing a physically meaningful and chemically consistent foundation for data-driven modeling. The resulting framework enables systematic validation of simulation-derived properties against experimental data and supports the development of multitask and multifidelity learning models based on a common structural representation. The following sections first evaluate the accuracy of MD-predicted properties and then examine how the combined experimental, simulation, and GC datasets are used to predict polymer properties using ML.

Table 1 summarizes the polymer property dataset used in this work, comprising approximately 62,000 data points spanning thermal, mechanical, transport, gas permeability, electronic/optical, and structural properties. The dataset integrates information from multiple sources, including experiments, MD, DFT, and GC estimates, and covers a wide range of property magnitudes with highly variable data availability across tasks. The thermal subset includes T_m, T_g, α_{T}, κ, and C_p, with C_p and T_g having the most extensive coverage and the most diverse fidelity composition. Mechanical properties (E, G, K, and ν) and transport properties (η and D) are obtained from MD simulations. Gas permeability properties (P_{He}, $P_{\mathrm{H_2}}$, $P_{\mathrm{CO_2}}$, $P_{\mathrm{N_2}}$, $P_{\mathrm{O_2}}$, and $P_{\mathrm{CH_4}}$) are derived from experiments and span multiple orders of magnitude. Electronic and optical properties computed at the DFT level include α, E_{HOMO}, E_{LUMO}, E_g, μ, and E_{total}, which are combined with MD-derived information to estimate bulk properties (n, ε, and ϵ_r). Structural and physical properties (R_g and ρ)

Table 1 Summary of polymer properties used in this work, including symbols, units, data sources, number of data points, and corresponding value ranges.

Property	Symbol	Unit	Source	Points	Data range
Thermal					
Melting temperature	T_m	K	Exp.	3671	[210.6–873.1]
Glass transition temperature	T_g	K	Exp./MD	7360	[134.1–768.1]
Thermal diffusivity	α_{T}	m^2/s	MD	799	$[3.5\times10^{-8}$–$8.3\times10^{-7}]$
Thermal conductivity	κ	W/m.K	MD/Exp.	2327	[0.002–1.59]
Specific heat capacity	C_p	J/ kg.K	Exp./MD/GC	13104	[439.2–2831.7]
Mechanical					
Young's modulus	E	GPa	MD	1012	[0.36–11.04]
Shear modulus	G	GPa	MD	1012	[0.12–4.2]
Bulk modulus	K	GPa	MD	1017	[0.72–10.9]
Poisson ratio	ν	–	MD	1012	[-0.17–0.48]
Transport					
Viscosity	η	Pa·s	MD	704	$[9.5\times10^{-5}$–0.11]
Diffusivity	D	cm^2/s	MD	700	$[2.19\times10^{-9}$–0.096]
Gas Permeability					
He permeability	P_{He}	Barrer	Exp.	466	[0.05–17800]
H_2 permeability	$P_{\mathrm{H_2}}$	Barrer	Exp.	511	[0.02–36800]
CO_2 permeability	$P_{\mathrm{CO_2}}$	Barrer	Exp.	756	$[1.2\times10^{-6}$–47000]
N_2 permeability	$P_{\mathrm{N_2}}$	Barrer	Exp.	798	$[1.6\times10^{-4}$–16600]
O_2 permeability	$P_{\mathrm{O_2}}$	Barrer	Exp.	807	$[7.0\times10^{-7}$–18700]
CH_4 permeability	$P_{\mathrm{CH_4}}$	Barrer	Exp.	683	$[4.1\times10^{-4}$–35000]
Electronic / Optical					
Polarizability	α	a.u.	DFT	2036	[1.9–70.3]
HOMO energy	E_{HOMO}	eV	DFT	2916	[-13.4– -6.84]
LUMO energy	E_{LUMO}	eV	DFT	2916	[-3.01–3.08]
Band gap	E_g	eV	DFT	2916	[5.4–16.7]
Dipole moment	μ	Debye	DFT	2916	[0.003–12.49]
Total electronic energy	E_{total}	eV	DFT	2916	$[-3.2\times10^{6}$– $-2.5\times10^{6}]$
Refractive index	n	–	MD+DFT	744	[1.02–1.84]
Dielectric constant	ε	–	MD+DFT	744	[1.06–12.7]
Permittivity	ϵ_r	–	MD+DFT	744	[9.4–113.1]
Structural / Physical					
Radius of gyration	R_g	Å	MD	2500	[9.17–75.22]
Density	ρ	g/cm^3	MD/Exp.	3643	[0.11–2.97]

provide direct links between polymer conformational statistics, packing behavior, and macroscopic properties. Overall, the large number of target properties, heterogeneous data fidelity, and strongly imbalanced data availability across properties (Table 1) motivate the adoption of a unified multi-task, multi-fidelity ML framework. Although the overall dataset spans 28 polymer properties across multiple experimental and

View Article Online
DOI: 10.1039/D6DD00206D

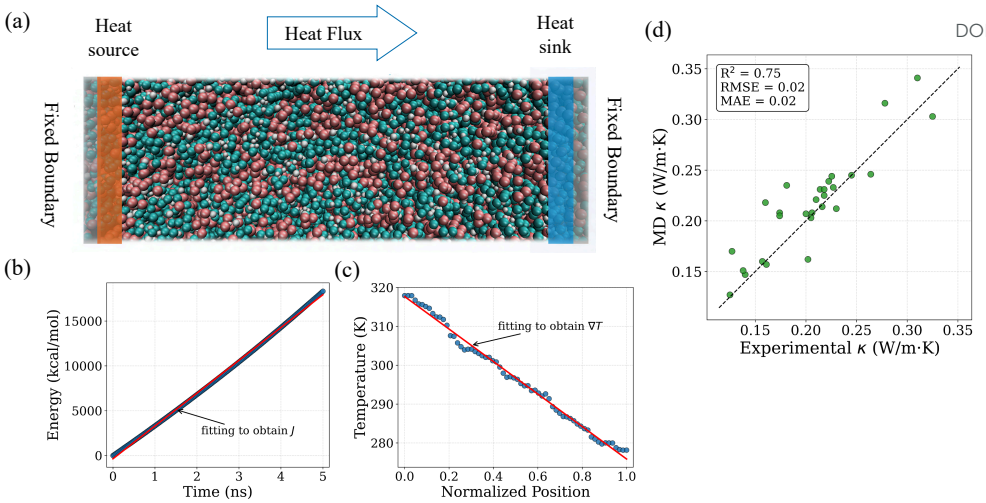

Fig. 2 Non-equilibrium molecular dynamics (NEMD) workflow and validation for thermal conductivity (κ) calculations. (a) Representative MD snapshot illustrating the NEMD setup, where a constant heat flux is imposed between hot and cold regions using thermostatted slabs, with boundary layers composed of fixed atoms. (b) Cumulative energy exchanged with the heat reservoirs as a function of time; a linear fit is used to obtain the heat flux J. (c) Steady-state temperature profile along the transport direction; the linear region is fitted to extract the temperature gradient ∇T. (d) Comparison of κ values obtained from MD simulations with experimental data, demonstrating good agreement between NEMD predictions and experiments.

computational sources, the fidelity coverage and overlap structure remain strongly property-dependent. Substantial multi-fidelity overlap is primarily available for properties such as specific heat capacity (C_p), glass transition temperature (T_g), thermal conductivity (κ), and density (ρ), whereas several mechanical properties are predominantly MD-derived and many electronic or optical descriptors are primarily obtained from DFT calculations. Even for representative multi-fidelity properties, only a subset of polymers contains both experimental and lower-fidelity computational data, while many polymers remain unique to a single fidelity source. A detailed property–fidelity availability matrix together with overlap statistics for representative multi-fidelity properties is provided in the Supporting Information (Table S1 and Fig. S1).

Figure 2 presents the NEMD results used to validate the MD-derived κ values used in this work. A representative simulation snapshot showing the imposed heat-flow direction and boundary configuration is presented in Fig. 2(a), confirming the formation of a well-defined transport geometry without structural distortion [29]. The cumulative energy exchanged with the heat reservoirs shows a clear linear dependence on time (Fig. 2(b)), indicating the establishment of a stable steady-state regime from which the heat flux is reliably extracted. The corresponding temperature profile along the transport direction exhibits a well-defined linear region within the central conduction zone (Fig. 2(c)), with minimal curvature and noise, enabling robust estimation of the temperature gradient. The resulting κ values are directly compared with experimental measurements in Fig. 2(d), considering only data reported between 20 °C and

View Article Online
DOI: 10.1039/D6DD00206D

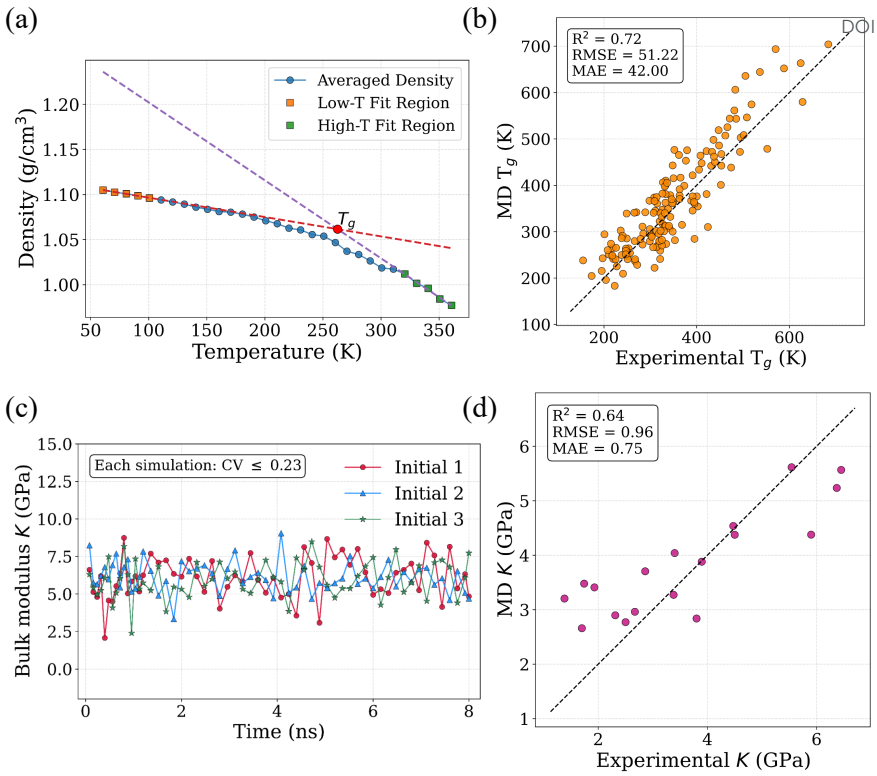

Fig. 3 Molecular dynamics workflows and validation for glass transition temperature (T_g) and bulk modulus (K) calculations. (a) Temperature-dependent density profile used to estimate T_g from the intersection of linear fits to the low- and high-temperature regimes. (b) Parity plot comparing MD-predicted and experimental T_g. (c) Time evolution of K obtained from independent simulations with different initial configurations, illustrating temporal fluctuations and convergence. (d) Parity plot comparing MD-predicted and experimental K.

30 °C to match the room-temperature MD simulation conditions. The MD predictions show good agreement with experiments, yielding an R^2 value of 0.75 and low absolute error. These results indicate that the MD workflow captures the dominant thermal transport trends across diverse polymer chemistries and provides physically consistent κ values suitable for subsequent ML analysis.

Figure 3(a–b) illustrates the MD-based evaluation of T_g. In MD simulations, T_g is identified from the temperature dependence of ρ and reflects changes in molecular mobility and thermal expansivity in amorphous polymers. Figure 3(a) shows the variation of ρ with temperature for a representative polymer. Two approximately linear regimes are observed, corresponding to the glassy and rubbery states. To estimate T_g, linear fits are performed separately for the low-temperature and high-temperature regions using different numbers of data points. Multiple fits are performed, and the final T_g is taken as the mean of the resulting intersection temperatures. If any fitted T_g value falls outside the initial temperature-scanning range, it is discarded. This approach captures changes in volumetric thermal expansion while reducing sensitivity

View Article Online
DOI: 10.1039/D6DD00206D

to the specific choice of fitting window. The parity plot in Fig. 3(b) compares MD-predicted T_g values with experimental data and shows good agreement, with an R^2 value of 0.72 and low prediction error across chemically diverse polymers. The estimated MD-derived T_g values remain sensitive to fitting-window selection, equilibration time, finite-time averaging, and the inherent limitations of simulation-based dilatometry approaches [41–44]. Although the present stepwise equilibration procedure reduces direct dependence on a continuous cooling rate, the resulting T_g estimates should still be interpreted as semi-automated simulation-derived approximations rather than exact thermodynamic transition temperatures.

Figure 3(c–d) presents the MD-based evaluation of the bulk modulus K. Figure 3(c) shows the time evolution of K obtained from simulations initiated from different configurations, illustrating temporal fluctuations around a stable mean. To validate the reliability of the MD predictions, polymers with experimentally reported K values were collected from PoLyInfo and other literature sources [45], considering only measurements conducted between 20 °C and 30 °C to match the room-temperature simulation conditions. The parity plot in Fig. 3(d) compares MD-predicted and experimental K values and shows reasonable agreement, with an R^2 value of 0.64 and low absolute error. The remaining deviations are attributed to variations in experimental conditions, polymer synthesis routes, and measurement techniques, which introduce noise into the reported modulus values. To assess the influence of polymer morphology on modulus evaluation, multiple independent initial structures were examined. The resulting variations in K were small, indicating that modulus predictions are largely insensitive to morphological randomness in the initial configurations [46]. To further reduce potential morphology-related bias, all reported MD-derived K values were averaged over simulations performed using three distinct initial structures for each polymer.

Although the finite-deformation approach reduces some of the convergence issues associated with fluctuation-based elastic-property calculations, the predicted elastic constants can still be influenced by morphology-dependent packing effects, residual internal voids, and finite-chain free-volume effects. Consequently, apparent agreement for individual elastic properties should not be interpreted as proof of a fully converged thermodynamic state point for all polymer systems. More rigorous structural convergence analysis and uncertainty propagation remain important future directions for large-scale automated polymer simulation workflows.

Figure 4(a–b) evaluates the accuracy of MD-predicted ρ through direct comparison with experimental data and illustrates the effect of bias correction. The parity plot in Fig. 4(a) shows a systematic deviation between MD predictions and experiments, despite a clear linear correlation across the sampled density range, indicating the presence of a global bias. The initial MD predictions yield a low R^2 value of 0.21. After applying a linear calibration using polymers common to both datasets, the agreement improves substantially, as shown in Fig. 4(b), with the R^2 increasing to 0.81. This corresponds to an improvement of approximately 286%, demonstrating a significant enhancement in agreement with experimental ρ values.

Figure 4(c–e) presents the corresponding analysis for the C_p. Figure 4(c) shows a representative enthalpy–temperature relationship obtained from MD simulations,

View Article Online
DOI: 10.1039/D6DD00206D

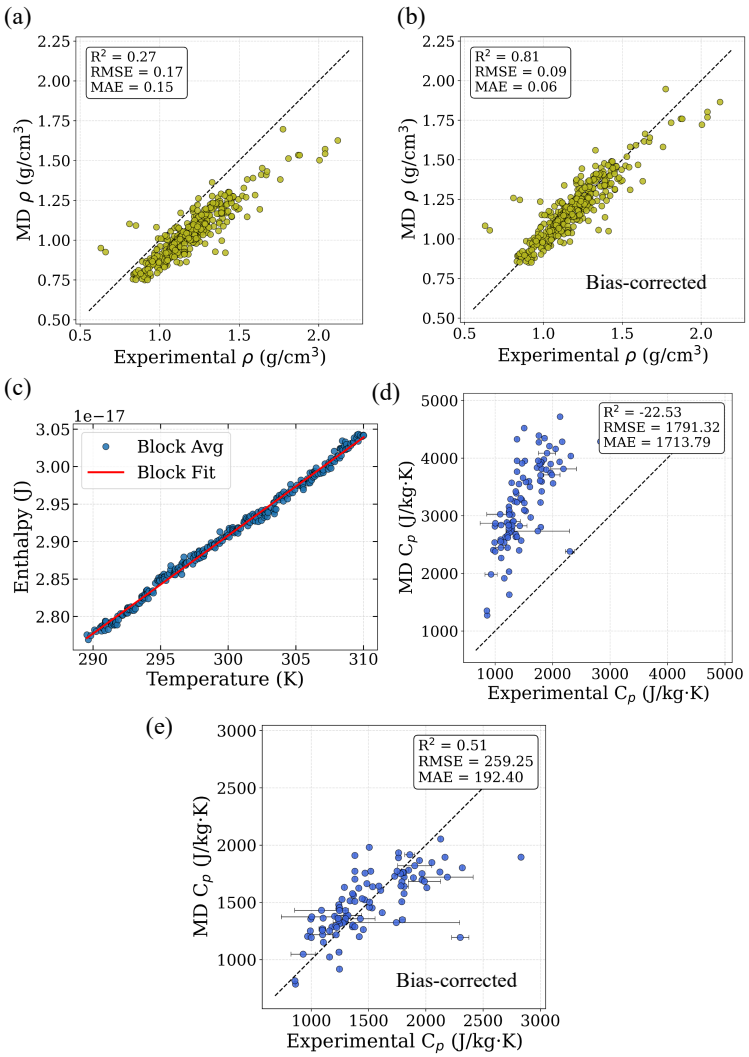

Fig. 4 Molecular dynamics prediction and bias analysis for density (ρ) and specific heat capacity (C_p). (a) Parity plot comparing MD-predicted ρ with experimental values, showing systematic deviation. (b) Density parity plot after bias correction, demonstrating improved agreement between MD and experimental ρ. (c) Representative enthalpy–temperature relationship obtained from MD simulations; block-averaged enthalpy values are fitted linearly to extract C_p. (d) Parity plot comparing MD-predicted C_p with experimental C_p, highlighting significant bias in raw MD predictions. (e) C_p parity plot after bias correction, showing improved correlation and reduced error relative to experimental data.

where block-averaged enthalpy varies linearly with temperature, allowing stable estimation of C_p from the slope. Direct comparison with experimental values reveals that MD-predicted C_p is systematically overestimated, as shown in Fig. 4(d), consistent

View Article Online
DOI: 10.1039/D6DD00206D

with previous studies based on classical MD simulations [23, 47, 48]. The overestimation of C_p arises from the classical treatment of vibrational degrees of freedom in MD simulations. In classical mechanics, all vibrational modes are fully excited according to the Boltzmann distribution, independent of vibrational frequency. In contrast, quantum mechanics predicts partial population of high-frequency vibrational modes at room temperature, following Bose-Einstein statistics. As a result, classical MD assigns excess vibrational energy to stiff bond-stretching and bond-bending modes, leading to inflated C_p values relative to experiments [25, 26]. Despite this systematic bias, MD-predicted C_p values exhibit a strong monotonic correlation with experimental data, indicating that MD captures the underlying structure–property trends. This behavior and its implications for data-driven correction and transfer learning have been analyzed in detail in our previous work [48]. The initial MAE of 1713.79 J/kg.K is reduced to 192.40 J/kg.K after applying a global linear correction, corresponding to an error reduction of approximately 89%. The corrected parity plot in Fig. 4(e) shows improved agreement with experiments, with reduced scatter and substantially lower prediction error.

The bias-corrected parity plots shown in Fig. 4e are included only as a diagnostic analysis to illustrate that the MD-derived C_p values preserve a strong systematic correlation with experimental trends despite the known classical overestimation of heat capacity. These corrected values were not used as inputs for model training or multi-fidelity learning. Instead, the multi-fidelity framework operates directly on the original heterogeneous datasets through fidelity-aware weighting without applying manual bias correction to the training labels. Importantly, MD-derived C_p values were not included in the joint multi-fidelity C_p learning framework because classical MD systematically overpredicts C_p due to the classical treatment of vibrational modes. Consequently, the fidelity-aware C_p analysis combines experimental and GC data rather than experimental and MD data. In contrast, for density (ρ), the original non-bias-corrected MD data were used directly within the multi-fidelity learning framework.

An extensive convergence test with respect to different optimization procedures has been conducted in our previous work [29]. To further examine possible residual free-volume effects and equilibration-related density deviations, additional post-equilibration NPT validation simulations were performed for a representative subset of polymers initialized from the final equilibrated structures. Density, volume, and chain-size observables were monitored over the validation window, and mean-squared internal distance scaling was analyzed from unwrapped trajectories as a qualitative chain-statistics stability check. The corresponding analyses are provided in the Supporting Information, including density/volume stability plots, radius-of-gyration trajectories, internal-distance scaling behavior, and quantitative stability metrics (Figs. S4, S5, S6 and Table S2). Overall, the validation subset exhibited relatively small density drift, modest late-time volume fluctuations, and no obvious signatures of abrupt chain collapse or unstable late-time structural relaxation.

The dielectric-property workflow combines DFT-derived electronic polarizabilities with dipolar fluctuations obtained from classical MD simulations to estimate the static dielectric response [23]. Specifically, the electronic contribution is estimated through the Lorentz–Lorenz relation using isotropic polarizabilities [49], while the orientational

View Article Online
DOI: 10.1039/D6DD00206D

contribution is obtained independently from simulation-cell dipole fluctuations. The total dielectric constant is then approximated through an additive decomposition of electronic and dipolar contributions. We note that this workflow combines several approximations, including monomer-level electronic structure calculations, isotropic polarizability assumptions, classical MD dipole fluctuations, and additive partitioning of dielectric contributions. Consequently, the predicted dielectric and permittivity distributions, particularly for extreme high-value candidates identified within PI1M, should be interpreted cautiously and primarily as screening-level estimates requiring future validation. Nevertheless, the predicted refractive-index range obtained in this work ($n = 1.02$–1.84) remains physically reasonable for polymeric materials. The present framework therefore aims to provide scalable approximate dielectric-property estimation suitable for large-scale polymer screening rather than quantitatively definitive dielectric predictions for all polymer classes Also, even if the method may contain systematic errors, the relative ranking of the properties should still be validated in general.

The present workflow should not be interpreted as independently validating each simulated property in isolation, since several quantities remain thermodynamically coupled through density, volume, and related structural state variables. Systematic deviations in density can therefore propagate into derived quantities such as thermal diffusivity, refractive index, and dipolar dielectric response, while compensating errors may contribute to apparently good agreement for some observables. Consequently, agreement for an individual property does not necessarily guarantee that the simulated systems reproduce a fully self-consistent experimental thermodynamic state point. The present results primarily demonstrate the ability of the workflow to capture useful structure–property trends at scale, while more rigorous state-point validation, structural convergence analysis, and uncertainty propagation remain important directions for future work. Representative repeat-unit structures corresponding to selected polymers from the training and validation datasets are provided in the Supporting Information (Fig. S3) to aid interpretation of the reported property trends and representative examples discussed throughout the manuscript.

Having established the physical consistency of the MD-derived properties and quantified their systematic deviations relative to experiments, we now turn to data-driven polymer property modeling using ML. The validated MD, experimental, and GC datasets provide a complementary foundation for learning shared structure–property relationships across multiple targets. In the following sections, we examine how multi-task and multi-fidelity ML strategies leverage these datasets to improve predictive performance, particularly in data-scarce regimes.

Figure 5 presents a Spearman rank correlation heatmap of the polymer properties included in the analysis. Only property pairs with at least 30 shared data points are retained to ensure statistical reliability. The heatmap reveals clear blocks of correlated behavior within subsets of properties, along with weaker but non-negligible cross-property associations, highlighting the potential benefit of learning shared representations across multiple targets. Strong correlation patterns are observed among electronic and dielectric descriptors, including α, E_g, E_{HOMO}, E_{LUMO}, μ, n, ε, ϵ_r, and E_{total}, consistent with their shared dependence on electronic structure and polarization

View Article Online
DOI: 10.1039/D6DD00206D

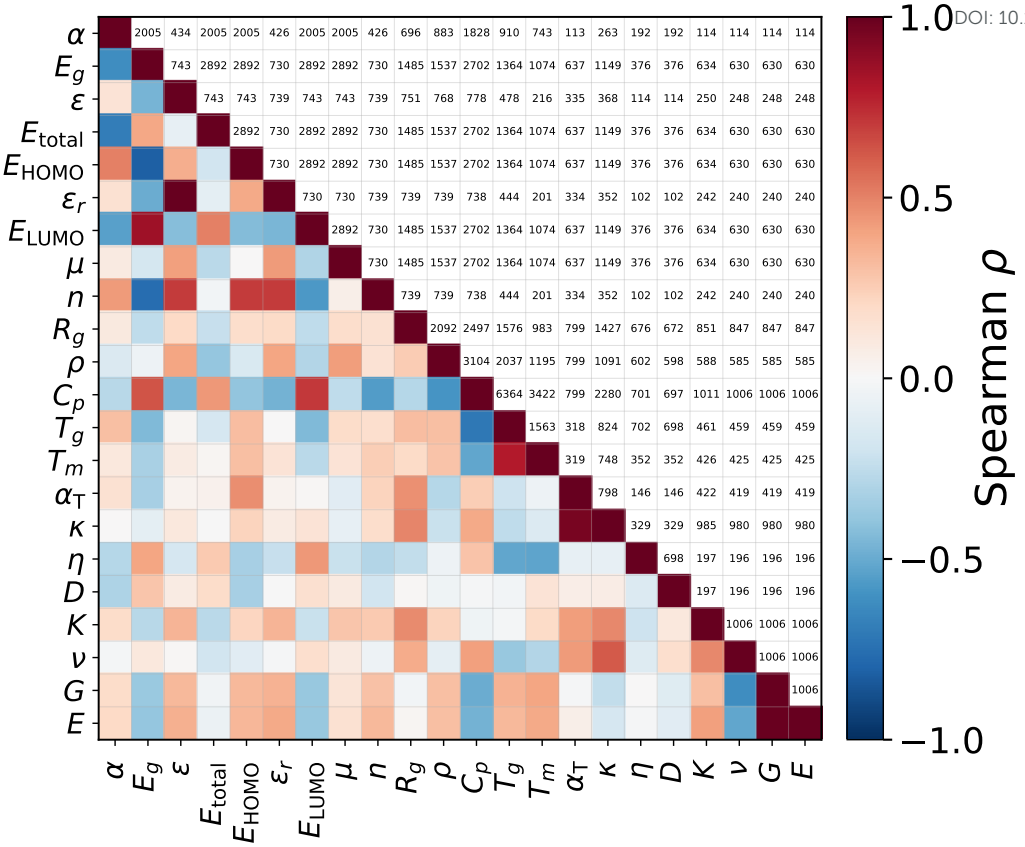

Fig. 5 Polymer property heatmap of Spearman correlation coefficients and data overlap counts. The lower triangle shows Spearman correlation coefficients, where red regions indicate positive correlations, blue regions indicate negative correlations, and white regions indicate negligible correlations. The upper triangle shows the number of overlapping polymer entries used to compute each property pair. Correlation coefficients are shown only for property pairs with at least 30 overlapping data points ($N \geq 30$).

response. Mechanical properties (E, G, K, and ν) also exhibit internally consistent correlations, reflecting their common sensitivity to stiffness and packing constraints in amorphous polymers. Thermal properties (C_p, T_g, T_m, α_T, and κ) show more heterogeneous correlations, indicating that while certain structural factors influence multiple thermal responses, each property retains distinct physical drivers. The heatmap further shows measurable cross-domain correlations between structural descriptors (ρ and R_g) and several thermal and mechanical properties, as well as selected electronic and dielectric quantities. These patterns suggest that polymer packing and conformational statistics encode shared information that can support prediction across different property classes, even when direct physical coupling is indirect. Guided

View Article Online
DOI: 10.1039/D6DD00206D

Table 2 Property groupings used in this work. Physical groups correspond to domain-based categorization, while correlation-based groups (G1–G4) are derived from Spearman correlation analysis and used for multi-task learning.

Group type	Group name	Properties
Physical	Electronic	α, E_g, ε, E_{total}, E_{HOMO}, E_{LUMO}, μ, n, ϵ_r
Physical	Thermal	C_p, T_g, T_m, α_{T}, κ
Physical	Mechanical	K, ν, G, E
Physical	Permeability	P_{He}, P_{H_2}, P_{CO_2}, P_{N_2}, P_{O_2}, P_{CH_4}
Physical	Other	η, D, ρ, R_g
Correlation	Group 1 (G1)	α, E_g, E_{HOMO}, E_{LUMO}, n, E_{total}
Correlation	Group 2 (G2)	C_p, E_{LUMO}, E_g, R_g, ρ, T_g
Correlation	Group 3 (G3)	T_g, T_m, η
Correlation	Group 4 (G4)	P_{He}, P_{H_2}, P_{CO_2}, P_{N_2}, P_{O_2}, P_{CH_4}, C_p, κ

by these observations, two complementary grouping strategies are adopted for multi-task learning, as summarized in Table 2. First, physical groupings are defined based on property domains (electronic, thermal, mechanical, permeability, and other). Second, correlation-based groups (G1–G4) are constructed by clustering properties that exhibit stronger Spearman rank correlations in Fig. 5. Although the correlation-based task groups used in the present study were selected using the observed correlation structure together with considerations of physical interpretability and sufficient data availability, the grouping process could also be automated using clustering algorithms applied to the property correlation matrix. For example, pairwise distances derived from correlation strength could be used together with hierarchical clustering or related graph-based partitioning methods to construct task groups in a fully data-driven manner. Such automated grouping strategies represent a promising direction for future extensions of the framework. These grouping schemes enable a systematic evaluation of whether correlation-aware task selection provides additional benefits beyond conventional domain-based multi-task learning, which is examined in the following section.

We now turn to data-driven polymer property modeling using multi-task and multi-fidelity ML. Instead of training independent models for each target property, the proposed framework learns a shared molecular representation across multiple properties while using property-specific prediction heads. This design enables information sharing among related targets and allows lower-fidelity data to support representation learning, while higher-fidelity data provides stronger supervision where available. The following results evaluate the impact of task grouping and joint learning on predictive accuracy, using single-task models as a baseline.

Figure 6 compares the scaled MAE obtained from single-task and multi-task ML models across different property groups and task configurations. In Fig. 6(a), electronic and dielectric properties generally show reduced error under multi-task learning, particularly when correlation-based groupings (G1 and G2) are used. In contrast, training a single model across all properties does not consistently improve performance, suggesting that indiscriminate task aggregation can yield limited gains or negative transfer for some electronic targets. Thermal properties (Fig. 6(b)) exhibit a more mixed response. Multi-task learning improves prediction accuracy for C_p, whereas only

View Article Online
DOI: 10.1039/D6DD00206D

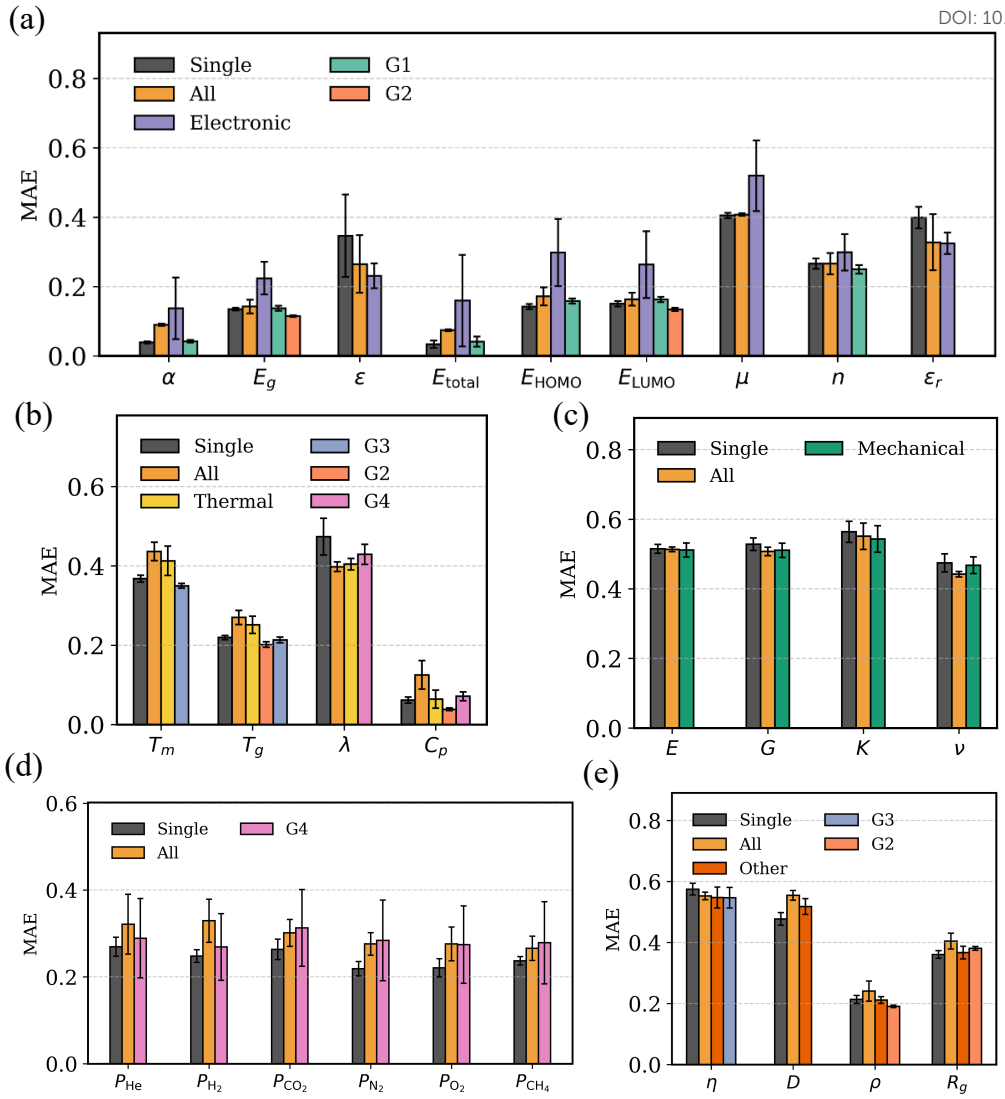

Fig. 6 Comparison of scaled mean absolute error (MAE) for different polymer properties obtained from single-task and multi-task learning models. The multi-task models are trained using different sets of target properties, including all properties, property-class-based groups (thermal, mechanical, electronic, permeability, and others), and correlation-based groups (Groups 1–4).

modest or negligible changes are observed for T_m and T_g. This behavior is consistent with the partially distinct physical mechanisms underlying these thermal properties. In several cases, correlation-based groupings perform comparably to, or even better than, purely domain-based groupings, suggesting that statistically informed task selection can be beneficial. Mechanical properties (E, G, K, and ν) show relatively

View Article Online
DOI: 10.1039/D6DD00206D

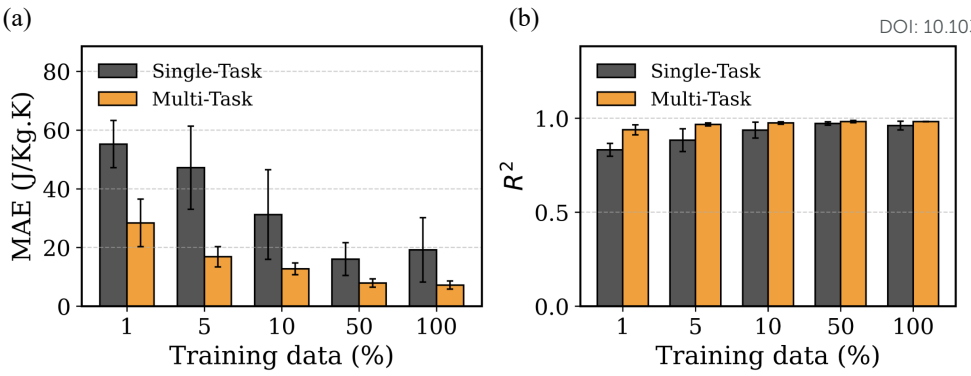

Fig. 7 Comparison of single-task and multi-task learning performance for specific heat capacity (C_p) prediction across different training data sizes. (a) Mean absolute error (MAE). (b) Coefficient of determination (R^2).

small differences between single-task and multi-task models (Fig. 6(c)). This indicates that the available MD data already provide adequate supervision for these targets, limiting the additional benefit of task coupling. Transport and structural properties (Fig. 6(e)) display selective improvements under multi-task learning, particularly for ρ and R_g, where shared structural information may contribute to improved predictions. Overall, Fig. 6 shows that the effectiveness of multi-task learning is strongly property-dependent and reflects the balance between positive transfer and task interference. Multi-task learning can reduce prediction error when task groupings capture meaningful statistical or physicochemical relationships, allowing the shared encoder to learn more transferable latent representations from limited data. This effect is particularly beneficial for sparsely sampled properties, where information from correlated tasks can improve generalization. However, when tasks exhibit weak correlations, substantially different data distributions, or large imbalances in dataset size and fidelity, joint optimization can introduce negative transfer, in which unrelated tasks interfere with one another and reduce predictive accuracy. Consequently, certain properties or task groupings remain better suited to single-task learning despite the shared-representation advantages of the multi-task framework. These observations support the use of correlation-aware task grouping and provide a basis for the multi-fidelity analysis presented in the following section. Comparison of R^2 for different polymer properties obtained from single-task and multi-task learning models is provided in the Supporting Information (Figure S2).

Figure 7 highlights the robustness of the multi-task learning framework relative to single-task models under varying training data availability. The analysis examines the effect of progressively reducing the fraction of training data while keeping the test set fixed. Training fractions of 100%, 50%, 10%, 5%, and 1% are considered, corresponding to approximately 10,000, 5,000, 1,000, 500, and 100 training data points, respectively.

When the full training dataset is used, single-task and multi-task models exhibit comparable performance in terms of both MAE and R^2, with the multi-task model

View Article Online
DOI: 10.1039/D6DD00206D

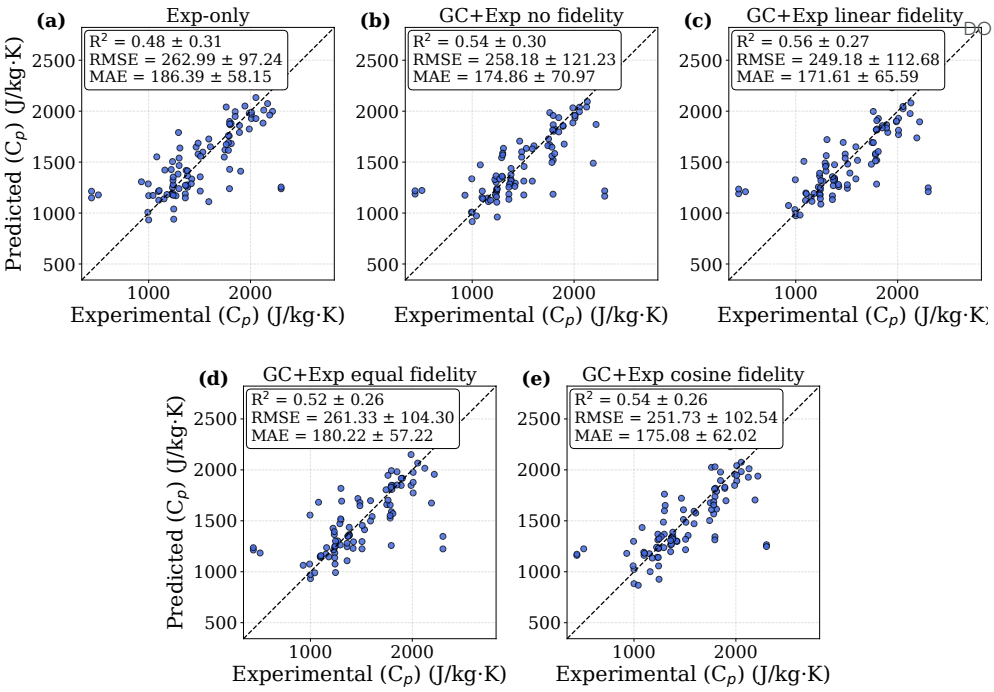

Fig. 8 Comparison of multi-fidelity training strategies for polymer C_p prediction evaluated on held-out experimental test sets. Parity plots compare experimental and predicted C_p values for (a) experimental-only training, (b) combined GC+experimental training without fidelity scheduling, (c) GC+experimental training with linear fidelity weighting, (d) GC+experimental training with equal fidelity weighting, and (e) GC+experimental training with cosine fidelity weighting. Reported metrics correspond to the mean ± standard deviation across multiple random seeds.

showing slightly lower MAE and marginally higher R^2 values (Fig. 7). This indicates that, in the data-rich regime, both models are capable of learning accurate structure–property relationships for C_p. As the training data size decreases, a consistent performance gap emerges across both error and correlation metrics. The multi-task model maintains lower MAE and higher R^2 values than the single-task model at all reduced training fractions. The difference becomes more pronounced at lower data availability, where the single-task model shows a clear degradation in predictive accuracy, while the multi-task model retains comparatively stable performance. This behavior is observed consistently in both MAE (Fig. 7(a)) and R^2 (Fig. 7(b)), indicating improved robustness of the multi-task model under limited training data. These results indicate that multi-task learning improves data efficiency by leveraging shared information across related tasks, leading to better generalization when training data are scarce. The observed trends support the use of multi-task learning for polymer property prediction in scenarios where high-fidelity experimental data are limited.

Parts of the present C_p analysis build upon our previous single-property transfer-learning study, which focused specifically on integrating MD, GC, and experimental

View Article Online
DOI: 10.1039/D6DD00206D

datasets for C_p prediction together with analysis of the systematic overprediction of C_p in classical MD simulations [48]. In the present work, these C_p-related discussions are included primarily to provide context for the fidelity-aware learning strategy. The primary new contribution of the current study is the extension from a single-property transfer-learning framework to a unified multi-task and multi-fidelity graph-learning framework spanning 28 polymer properties across multiple physical domains.

Figure 8 examines the influence of multi-fidelity training strategies on experimentally relevant property prediction using representative polymer properties with different experimental and lower-fidelity data availability. To systematically evaluate the role of lower-fidelity data and fidelity-weighting strategies, we compared experimental-only training together with multiple multi-fidelity formulations, including equal weighting, linear fidelity weighting, cosine fidelity weighting, and unscheduled combined-fidelity training. The corresponding parity plots for representative C_p models are shown in Fig. 8, while quantitative results for additional properties are summarized in the Supporting Information (Table S4).

For C_p, the experimental dataset is relatively sparse (118 experimental samples), while the GC dataset provides substantially broader chemical coverage (12,076 samples). Under these conditions, incorporating GC data generally improved experimentally evaluated prediction performance relative to experimental-only training. The experimental-only model yielded an average R^2 value of 0.48 ± 0.31 and an MAE of 186.39 ± 58.15 J/kg.K, whereas the multi-fidelity models achieved improved average performance, with the best results obtained using linear fidelity weighting ($R^2 = 0.56 \pm 0.27$, MAE $= 171.61 \pm 65.59$ J/kg.K). The remaining fidelity-weighting strategies produced comparable performance improvements relative to the experimental-only baseline. These results suggest that lower-fidelity GC data provide useful representation-learning benefits by expanding chemical coverage when experimentally measured data are limited.

The behavior observed for other properties further highlights the property-dependent nature of multi-fidelity learning. For density (ρ), where both experimental and MD datasets are relatively large and physically correlated, fidelity-aware learning produced more consistent improvements over experimental-only training, with cosine weighting yielding the best overall performance ($R^2 = 0.78 \pm 0.06$ compared with 0.70 ± 0.17 for experimental-only training). In contrast, for T_g, the experimental dataset already contains substantially more data than the corresponding MD dataset, and all fidelity strategies produced nearly identical performance ($R^2 \approx 0.89$), indicating limited additional benefit from lower-fidelity integration. Thermal conductivity (TC) remained challenging across all training strategies because of the extremely limited experimental dataset (65 samples), resulting in large variance and unstable generalization behavior regardless of the fidelity treatment.

Overall, these results indicate that the effectiveness of multi-fidelity learning depends strongly on the balance between experimental-data availability, lower-fidelity dataset coverage, and the degree of correlation between fidelity levels. The primary benefit of the multi-fidelity framework arises from the incorporation of broader lower-fidelity chemical information when experimental datasets are sparse, while differences among specific fidelity-weighting schedules remain comparatively modest for most

View Article Online
DOI: 10.1039/D6DD00206D

Table 3 Summary statistics of predicted polymer properties for the PolyInfo database (13000 real polymers) and the PI1M database (1 million virtual polymers).

Property	PolyInfo Database			PI1M Database		
	Min	Median	Max	Min	Median	Max
Thermal						
T_m (K)	207.87	509.66	880.45	203.78	447.25	918.02
T_g (K)	143.58	394.01	702.26	113.38	338.99	691.51
α_T (m^2/s)	5.0×10^{-8}	1.8×10^{-7}	5.6×10^{-7}	2.5×10^{-8}	1.8×10^{-7}	6.0×10^{-7}
κ (W/m.K)	0.11	0.26	0.49	0.14	0.26	0.83
C_p (J/kg.K)	437.43	1257.86	2560.19	551.00	1304.21	2796.52
Mechanical						
E (GPa)	1.46	5.26	10.77	0.82	5.18	10.16
G (GPa)	0.59	2.01	4.08	0.43	1.95	3.57
K (GPa)	1.20	4.90	8.61	1.24	4.84	9.95
ν	0.08	0.31	0.43	0.08	0.32	0.51
Transport						
η (Pa·s)	7.1×10^{-3}	1.1×10^{-2}	2.9×10^{-2}	6.4×10^{-3}	1.3×10^{-2}	2.9×10^{-2}
D (cm^2/s)	2.7×10^{-8}	3.0×10^{-4}	9.0×10^{-2}	1.3×10^{-8}	1.1×10^{-3}	2.9×10^{-1}
Gas Permeability						
P_{He} (Barrer)	0.55	14.93	26495.19	1.14	21.80	37246.60
P_{H_2} (Barrer)	0.56	17.83	38526.12	0.56	20.19	75825.66
P_{CO_2} (Barrer)	2.3×10^{-4}	9.81	61187.62	4.3×10^{-3}	14.80	54059.37
P_{N_2} (Barrer)	5.7×10^{-4}	0.46	18698.51	1.5×10^{-4}	0.86	40807.90
P_{O_2} (Barrer)	9.1×10^{-4}	1.64	23378.65	6.4×10^{-4}	2.17	30797.60
P_{CH_4} (Barrer)	1.7×10^{-3}	0.53	24869.14	5.6×10^{-4}	1.85	58958.58
Electronic / Optical						
α (a.u.)	2.10	45.39	93.94	0.25	40.04	89.77
E_{HOMO} (eV)	-12.96	-8.67	-6.74	-11.98	-9.37	-8.09
E_{LUMO} (eV)	-1.84	0.57	3.24	-2.97	1.03	3.17
E_g (eV)	6.57	9.15	15.48	7.55	10.91	16.70
μ (Debye)	0.02	2.90	6.92	0.01	2.97	8.24
E_{total} (eV)	-3.0×10^{6}	-8.9×10^{5}	-4.7×10^{4}	-4.3×10^{6}	-8.3×10^{5}	-4.1×10^{4}
n	1.26	1.55	1.74	1.25	1.52	1.91
ε	2.08	3.58	16.19	2.13	3.90	21.82
ϵ_r	13.77	24.88	78.64	13.54	25.65	86.77
Structural						
R_g (Å)	9.21	19.49	42.95	8.50	18.06	43.35
ρ (g/cm^3)	0.69	1.24	2.79	0.85	1.20	2.80

properties. Figure 6 and Figure 8 therefore address complementary objectives. Figure 6 evaluates shared representation learning within the multi-task framework, whereas Figure 8 focuses specifically on the influence of heterogeneous fidelity integration on experimentally relevant prediction performance.

View Article Online
DOI: 10.1039/D6DD00206D

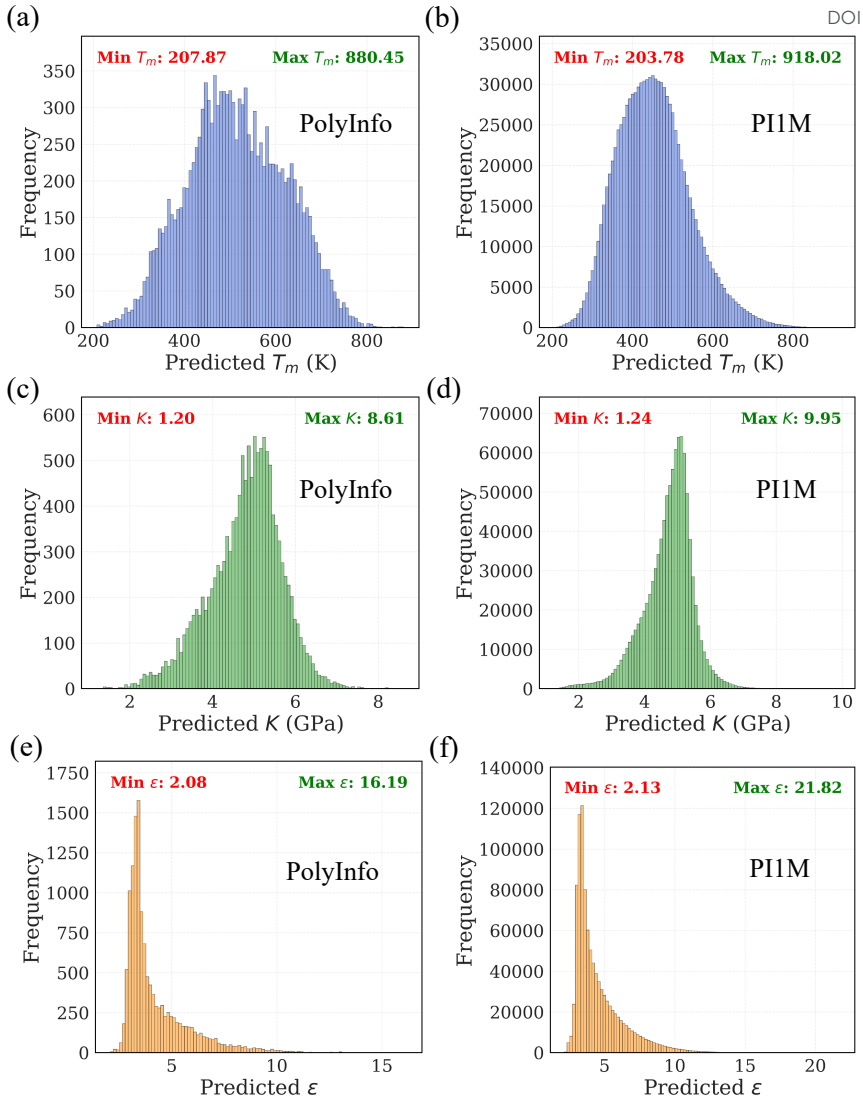

Fig. 9 Machine-learning-predicted property distributions for polymers in the PolyInfo and PI1M databases. (a,b) Histograms of predicted T_m for polymers in the PolyInfo and PI1M databases, respectively. (c,d) Histograms of predicted K for PolyInfo and PI1M polymers. (e,f) Histograms of predicted ε for PolyInfo and PI1M polymers. Minimum and maximum predicted values for each property are indicated in the corresponding panels.

Having established the predictive accuracy and robustness of the proposed ML models, we next apply them to large-scale polymer property prediction. Large-scale property prediction is performed for polymers in the PolyInfo database, comprising approximately 13,000 real polymers, and the PI1M [50] database, containing one million virtual polymers. In total, predictions were generated for 28 polymer properties,

View Article Online
DOI: 10.1039/D6DD00206D

corresponding to approximately 3.6×10^5 property values for the PolyInfo dataset and 2.8×10^7 property values for the PI1M virtual polymer library. For each target property, the model configuration that achieved the best validation performance in the preceding analyses is selected. Depending on the property, this corresponds to either a single-task model or a multi-task model with an appropriate task grouping. Table 3 summarizes the minimum, median, and maximum predicted values for all properties across both datasets. For the PolyInfo database, the predicted ranges for thermal properties such as T_m, T_g, κ, and C_p are consistent with experimentally reported values, with median predictions lying within physically expected regimes. Mechanical properties (E, G, K, and ν) exhibit relatively narrow distributions, reflecting the constrained mechanical response typical of amorphous polymers. Electronic and optical properties display broader ranges, particularly for α, μ, and ε, consistent with their sensitivity to chemical composition and electronic structure. Structural properties (R_g and ρ) show limited variation, reflecting packing constraints in polymer melts. For the PI1M database, the predicted medians for most properties remain comparable to those observed for PolyInfo, while the overall property ranges are generally broader. This widening reflects the substantially larger and more diverse chemical space represented in the virtual polymer set. Importantly, predicted values across all properties remain within physically reasonable bounds, indicating stable model behavior when extrapolated to a larger design space.

Figure 9 presents representative predicted distributions for selected properties, including T_m, K, and ε, for both PolyInfo and PI1M polymers. The PolyInfo distributions show a single dominant peak with a spread comparable to that observed in experimental datasets, whereas the PI1M distributions exhibit smoother profiles and extended tails due to the larger sample size. The minimum and maximum values shown are consistent with the summary statistics reported in Table 3 and further illustrate the broader coverage achieved for the virtual polymer library. Together, the results in Table 3 and Fig. 9 indicate that the trained models generate physically consistent property distributions at scale while preserving trends observed in experimentally characterized polymers. This capability enables systematic screening and comparative analysis of large polymer libraries across multiple property dimensions.

Taken together, the results indicate that combining physics-based simulations, experimental data, and ML within a unified framework enables consistent and scalable prediction of multiple polymer properties. Validation against experimental measurements provides confidence in the simulation-derived datasets, while correlation-aware multi-task learning and fidelity-weighted training improve predictive performance, particularly when experimental data are limited. The large-scale predictions for both real and virtual polymer libraries demonstrate the potential of the proposed framework for systematic evaluation of polymer properties across diverse chemical spaces.

4 Conclusion

In this work, an integrated polymer informatics framework is presented that combines physics-based data generation with data-driven modeling for the prediction of

View Article Online
DOI: 10.1039/D6DD00206D

polymer properties. The ADEPT workflow enables automated evaluation of polymer properties from repeat-unit SMILES, using MD and DFT to compute thermal, mechanical, transport, structural, and electronic descriptors consistently. When combined with experimentally curated measurements and GC estimates, this approach results in a heterogeneous dataset of approximately 62,000 property values spanning multiple property classes and data fidelity levels.

Using this dataset, inter-property correlations are quantified to assess the suitability of joint learning across targets. The results show that the effectiveness of multi-task learning depends on task-relatedness, with correlation-informed groupings providing more consistent improvements than uniform task aggregation. The analysis further indicates that multi-task learning improves data efficiency, with performance differences between single-task and multi-task models becoming more pronounced as the available training data are reduced. In addition, fidelity-aware training strategies improve predictive accuracy when combining experimental and computational data by balancing broad coverage from lower-fidelity sources with stronger supervision from experimental measurements.

The trained models are subsequently applied to large-scale property prediction for polymers in the PolyInfo database and the PI1M virtual polymer library. The resulting property distributions remain physically reasonable and reflect trends observed in experimentally characterized polymers, supporting the use of the models for systematic screening across extended chemical spaces.

Overall, the results demonstrate that integrating simulation-derived data, experimental measurements, and ML within a unified framework provides a practical approach for multi-property polymer prediction. The proposed methodology supports the scalable evaluation of both known and virtual polymers and provides a structured framework for polymer screening and analysis in settings where data availability and fidelity vary across properties.

5 Methods

5.1 ADEPT Workflow: Polymer Simulation and Property Generation

5.1.1 Polymer Structure Generation from SMILES

Polymer repeat units were defined represented by SMILES of monomers, with polymerization sites specified using isotope-labeled terminal hydrogens denoted as "*" to identify head and tail atoms. The present framework primarily focuses on linear homopolymers represented using standardized repeat-unit SMILES with explicit polymerization points. Branching, copolymer sequence distributions, polydispersity, tacticity, and chain-length-dependent effects are not explicitly represented in the current graph representation or automated simulation workflow. Consequently, the present framework should be interpreted as an approximate repeat-unit-level representation of polymer behavior rather than a complete description of all macromolecular structural complexities. Three-dimensional monomer geometries were generated using the ETKDG algorithm [51] and imported into the PySIMM framework for polymer

View Article Online
DOI: 10.1039/D6DD00206D

construction [52]. Polymer chains were built using a random-walk polymerization algorithm [53], with the chain length selected to yield approximately 600 atoms per polymer chain, ensuring comparable molecular weights across different polymers. During chain growth, terminal hydrogens were removed, and the final polymer chains were capped at both ends with methyl ($-CH_3$) groups. All polymer structures were parameterized using the General AMBER Force Field (GAFF2) force field [54]. GAFF2 was used for all MD simulations due to its broad chemical coverage and established parameterization for polymers; a detailed force-field sensitivity analysis for C_p calculations is reported in our previous work and is not repeated here[48]. Atomic partial charges were assigned using either the Restrained Electrostatic Potential (RESP) method [55] or Gasteiger charges [56]. RESP charges were obtained from gas-phase quantum chemical calculations performed on optimized monomer geometries, whereas Gasteiger charges were used for computationally efficient simulations when dielectric properties were not required. Different charge assignment schemes were used depending on the target property and computational requirements of the workflow. Gasteiger charges were adopted for most high-throughput MD calculations because of their low computational cost and scalability for large polymer datasets. In contrast, dielectric-property calculations employed RESP charges due to the stronger sensitivity of dipole fluctuations to the underlying electrostatic model. Consequently, the resulting dataset should be interpreted as a heterogeneous multi-source polymer property collection optimized for scalable informatics workflows rather than as a fully self-consistent thermodynamic reference dataset generated using a single electrostatic model throughout. After the force-field assignment, the polymer chains were energy-minimized and replicated to generate amorphous configurations consisting of six chains. These configurations were initially packed at a ρ of approximately 0.01 g/cm^3. The resulting amorphous polymer structures were written in formats compatible with LAMMPS [57] for subsequent MD simulations.

5.1.2 Amorphous Polymer Generation and Equilibration

Following polymer chain construction and force-field assignment, bulk amorphous polymer structures were generated and equilibrated using MD simulations in LAMMPS. Multiple polymer chains were randomly packed into a three-dimensional simulation cell at a low initial ρ and equilibrated using a multi-stage protocol comprising an initial relaxation and annealing.

Initial relaxation: Electrostatic interactions were initially disabled, and Lennard-Jones (LJ) interactions were truncated at a cutoff distance of 0.3 nm to reduce large forces arising from unfavorable contacts in the randomly packed configuration. The system was first equilibrated under an NPT ensemble at 100 K for 2 ps using a 0.1 fs time step. The temperature was then increased from 100 K to 1000 K over 1 ns under NVT conditions. This was followed by equilibration at 1000 K and 0.1 atm for 50 ps under NPT conditions, and a subsequent 1 ns NPT simulation during which the pressure was gradually increased from 0.1 atm to 500 atm using a 1 fs time step. SHAKE [58] constraints were applied throughout this stage to constrain covalent bond lengths and ensure numerical stability.

View Article Online
DOI: 10.1039/D6DD00206D

Annealing: Electrostatic interactions were re-enabled using the particle–particle particle–mesh (PPPM) Ewald summation method [59], and the LJ cutoff distance was increased to 0.800 nm. The system was equilibrated at 1000 K and 1 atm under NPT conditions for 2 ps using a 0.1 fs time step. The system was then cooled to 300 K at a rate of 140 K/ns while maintaining NPT conditions and SHAKE constraints. A final NPT simulation was performed at 300 K and 1 atm for 8 ns using a 1 fs time step to obtain a stable amorphous polymer configuration for subsequent property calculations.

5.1.3 Property Calculations from Molecular Dynamics Simulations

Unless otherwise specified, all polymer properties were computed from equilibrated amorphous configurations obtained after the annealing protocol described above. Production simulations were performed under NPT conditions at 300 K and 1 atm, and property values were obtained by averaging over statistically converged time intervals. Electronic properties were evaluated at the monomer level using DFT calculations. The following subsections describe the methodologies used for calculating structural, thermodynamic, transport, mechanical, electronic, and dielectric properties.

5.1.4 Structural Properties

Density (ρ). The mass density was computed as the ratio of the total system mass to the time-averaged simulation cell volume,

$$\rho = \frac{m}{\langle V \rangle}. \tag{3}$$

Radius of gyration (R_g). The radius of gyration was calculated for each polymer chain using the LAMMPS `compute gyration/chunk` command, with chains identified by molecule ID. This corresponds to the mass-weighted radius of gyration,

$$R_g = \sqrt{\frac{\sum_{k=1}^{N} m_k \left|\mathbf{r}_k - \mathbf{r}_{\mathrm{cm}}\right|^2}{\sum_{k=1}^{N} m_k}}, \tag{4}$$

where N is the number of atoms in the chain, m_k is the mass of atom k, $\mathbf{r}_k$ is its position, and $\mathbf{r}_{\mathrm{cm}}$ is the center-of-mass position of the chain. Reported R_g values correspond to averages over all polymer chains and statistically converged time intervals.

5.1.5 Thermal and Thermodynamic Properties

Glass transition temperature (T_g). The glass transition temperature was determined from density–temperature curves obtained from temperature-dependent NPT simulations. Linear regressions were performed separately for the low-temperature and high-temperature regions, and T_g was identified as the intersection of the corresponding fitted lines,

$$T_g = \frac{c_{\mathrm{high}} - c_{\mathrm{low}}}{m_{\mathrm{low}} - m_{\mathrm{high}}}, \tag{5}$$

View Article Online
DOI: 10.1039/D6DD00206D

where m and c denote the slopes and intercepts of the respective linear fits. To reduce sensitivity to the choice of fitting window, multiple fits were performed using different numbers of data points in each temperature regime. The final T_g value was taken as the mean of the resulting intersection temperatures. Any fitted T_g value falling outside the initial temperature-scanning range was discarded.

For each polymer, the temperature-scanning range was defined as a bounded interval around an initial estimate of T_g. This initial estimate was obtained from a separate multilayer perceptron model trained on experimentally reported T_g values. The MD temperature scan was then performed over a restricted window of ± 150 K around the estimated T_g, rather than across a broad global temperature interval. This procedure limits the required simulation window while ensuring adequate sampling of both glassy and rubbery regimes.

In the present workflow, T_g is estimated from the temperature dependence of density using independently equilibrated simulations at discrete temperature intervals rather than from a continuous cooling trajectory at a fixed cooling rate. An auxiliary MLP model trained on experimental T_g data is used only to guide the initial fitting-window selection and is not used directly as the final simulation label.

Specific heat capacity (C_p). The constant-pressure specific heat capacity was evaluated using both equilibrium MD (EMD) and non-equilibrium MD (NEMD) simulations under NPT conditions. The use of both approaches allows assessment of consistency while balancing accuracy and computational cost.

In the equilibrium approach, after sufficient equilibration, production simulations were performed at 300 K and 1 atm to sample thermodynamic fluctuations. The C_p was computed from enthalpy fluctuations using the fluctuation–dissipation relation [60],

$$C_p = \frac{\langle H^2 \rangle - \langle H \rangle^2}{k_B T^2 \, m}, \tag{6}$$

where H is the instantaneous system enthalpy, T is the absolute temperature, k_B is Boltzmann's constant, and m is the total system mass. Angle brackets denote ensemble averages over the equilibrated trajectory. This approach can be evaluated concurrently with other equilibrium properties, making it computationally efficient for large-scale screening studies.

In addition, a non-equilibrium approach based on the enthalpy–temperature relationship was employed. The system was first equilibrated at 290 K and 1 atm, followed by a controlled temperature ramp from 290 K to 310 K under NPT conditions. Average enthalpy values were recorded as a function of temperature, and C_p was obtained from the slope of a linear fit to the enthalpy–temperature curve [47],

$$C_p = \left(\frac{\partial H}{\partial T} \right)_P . \tag{7}$$

This non-equilibrium approach generally provides more stable estimates of C_p, but requires separate temperature-ramping simulations and therefore incurs additional computational cost; accordingly, all C_p values reported in this work were obtained using the NEMD method. A detailed comparison between EMD and NEMD

View Article Online
DOI: 10.1039/D6DD00206D

approaches for polymer C_p, including quantitative error analysis, is provided in our previous work [48], where NEMD was shown to offer improved numerical stability and was therefore adopted in the present study.

Thermal expansion coefficient (α_V). The volumetric thermal expansion coefficient was computed from temperature-dependent volume data according to

$$\alpha_V = \frac{1}{\langle V \rangle} \left(\frac{\partial \langle V \rangle}{\partial T} \right)_P , \tag{8}$$

and the linear thermal expansion coefficient was obtained as $\alpha_L = \alpha_V/3$.

Thermal conductivity (κ). Thermal conductivity was computed using a NEMD approach in which a steady-state temperature gradient was imposed along a designated transport direction [29]. After equilibration, localized hot and cold regions were maintained at fixed temperatures to induce a constant heat flux across the system. The simulation cell was divided into spatial bins along the transport direction, and the temperature profile was constructed from time-averaged kinetic temperatures within each bin.

The temperature gradient was extracted from the central conduction region by fitting the linear portion of the steady-state temperature profile, excluding bins adjacent to the thermostatted regions to minimize boundary effects. The heat flux was obtained from the cumulative energy exchanged with the thermal reservoirs and normalized by the cross-sectional area perpendicular to the transport direction. κ was then evaluated using Fourier's law,

$$\kappa = - \frac{J_q}{\mathrm{d}T/\mathrm{d}x}, \tag{9}$$

where J_q is the steady-state heat flux and $\mathrm{d}T/\mathrm{d}x$ is the temperature gradient along the transport direction.

To ensure reliable estimation of κ, only simulations exhibiting a linear increase in cumulative exchanged energy with time and a well-defined linear temperature gradient in the central region were retained for analysis. Reported κ values correspond to block-averaged means computed over the steady-state portion of the trajectory, with uncertainties estimated from the variance across independent time blocks.

Thermal diffusivity (α_T). The thermal diffusivity was computed from κ, ρ, and C_p as

$$\alpha_T = \frac{\kappa}{\rho \, C_p}, \tag{10}$$

using ρ and C_p obtained from NPT production simulations performed at 300 K and 1 atm.

5.1.6 Transport Properties

Self-diffusion coefficient (D). The self-diffusion coefficient was computed from the mean-square displacement using the Einstein relation,

$$D = \lim_{t \to \infty} \frac{1}{6t} \left\langle |\mathbf{r}(t) - \mathbf{r}(0)|^2 \right\rangle , \tag{11}$$

View Article Online
DOI: 10.1039/D6DD00206D

where $\mathbf{r}(t)$ denotes the position of a particle at time t, and angle brackets indicate an ensemble average over particles and time origins.

Viscosity (η). The shear viscosity was obtained from EMD trajectories using the Green–Kubo formalism,

$$\eta = \frac{\langle V \rangle}{k_B T} \int_0^\infty \langle P_{\alpha\beta}(0)\, P_{\alpha\beta}(t) \rangle \, \mathrm{d}t, \qquad \alpha \neq \beta, \tag{12}$$

where $P_{\alpha\beta}$ are the off-diagonal components of the pressure tensor. Ensemble averages were taken over independent time origins and Cartesian shear components.

5.1.7 Mechanical properties

The elastic properties of amorphous polymers were calculated using a finite-deformation stress-strain approach. Each equilibrated polymer cell was first converted from a cubic to a triclinic simulation box to enable shear as well as normal deformations. Symmetric finite strains of magnitude $\pm 2\%$ were applied along the six independent strain modes, comprising three normal and three shear components. Following each deformation, the system was simulated at 300 K using NVE integration coupled with a Langevin thermostat, consisting of a 250 fs brief equilibration followed by a 75 fs sampling segment with a 0.25 fs timestep, and the time-averaged stress tensor was recorded.

Within linear elasticity, the stress–strain relationship is given by

$$\sigma_i = \sum_{j=1}^{6} C_{ij}\, \varepsilon_j, \tag{13}$$

where σ_i and ε_j denote the stress and strain components in Voigt notation. The elastic stiffness coefficients C_{ij} were obtained from central finite differences of the stress response with respect to the applied strain, and the full 6×6 stiffness matrix was constructed.

For isotropic amorphous systems, the bulk modulus (K) and shear modulus (G) were computed using the Voigt–Reuss–Hill averaging scheme applied to the stiffness and compliance tensors. Young's modulus (E) and Poisson's ratio (ν) were then derived from K and G using standard isotropic relations [60],

$$E = \frac{9KG}{3K + G}, \tag{14}$$

$$\nu = \frac{3K - 2G}{2(3K + G)}. \tag{15}$$

To ensure statistical reliability, for each simulation, final mean values of elastic properties were averaged over 100 configurations evenly sampled from an 8 ns production trajectory. Reported values correspond to ensemble averages, with coefficients of variation (defined as the ratio of standard deviation to the mean) not exceeding 0.33.

View Article Online
DOI: 10.1039/D6DD00206D

5.1.8 Electronic and dielectric properties from DFT

Electronic descriptors were computed at the monomer level using DFT to obtain the static dipole polarizability (α), highest occupied molecular orbital energy (E_{HOMO}), lowest unoccupied molecular orbital energy (E_{LUMO}), electronic band gap (E_g), total electronic energy (E_{total}), and molecular dipole moment (μ). Monomer-based calculations were employed to enable scalable evaluation across large polymer libraries while retaining sensitivity to repeat-unit chemistry.

Monomer geometries were generated from repeat-unit SMILES using RDKit, where polymerization placeholders (*) were replaced with tritium tags ([^{3}H]) to preserve valence and identify connection sites [61]. Initial three-dimensional coordinates were generated using the ETKDGv2 algorithm with explicit hydrogens added prior to export in XYZ format [51].

All electronic-structure calculations were performed using Psi4 [62]. To obtain stable and physically reasonable geometries, a three-stage optimization protocol was applied: (i) HF/STO-3G pre-optimization, (ii) refinement at HF/6-31G, and (iii) final optimization using the range-separated, dispersion-corrected ωB97M-D3BJ functional with the 6-311+G(2d,p) basis set [63]. Geometry and self-consistent field convergence thresholds were progressively tightened at each stage.

The static dipole polarizability was evaluated at the optimized geometry using a finite-field approach. Single-point DFT calculations were performed under uniform electric fields of magnitude $\pm\delta$, with $\delta = 10^{-4}$ atomic units, applied independently along each Cartesian direction. The polarizability tensor was obtained from central finite differences of the induced dipole moments,

$$\alpha_{ij} \approx -\frac{\mu_i^{(+)} - \mu_i^{(-)}}{2\delta}, \tag{16}$$

and the isotropic polarizability was computed as $\bar{\alpha} = \frac{1}{3}\mathrm{Tr}\,\boldsymbol{\alpha}$ and reported in Å^3 [64].

Single-point electronic properties were then computed at the optimized geometries using ωB97M-D3BJ with density fitting and tight self-consistent field convergence. HOMO and LUMO energies were extracted from the converged orbital spectrum and reported in eV, and the electronic band gap was computed as $E_g = E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}$. Total electronic energies were reported in eV, and dipole moments were reported in Debye. For iodine-containing monomers, a mixed-basis treatment was employed in which LanL2DZ was applied to iodine while all other atoms used 6-311G(d,p) [65].

The electronic contribution to the dielectric response was estimated by first computing the refractive index (n) from the DFT-derived isotropic polarizability using the Lorentz–Lorenz relation [49],

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3} N\bar{\alpha}, \tag{17}$$

where N is the number density and $\bar{\alpha}$ is the isotropic polarizability. The electronic dielectric constant was then obtained as $\varepsilon_{\mathrm{el}} = n^2$.

View Article Online
DOI: 10.1039/D6DD00206D

The orientational, or dipolar, contribution to the static dielectric constant ($\varepsilon_{\mathrm{dip}}$) was obtained from MD simulations using dipole moment fluctuations,

$$\varepsilon_{\mathrm{dip}} = 1 + \frac{\langle M^2 \rangle - \langle \mathbf{M} \rangle^2}{3\,\varepsilon_0\,k_B T\,\langle V \rangle}, \tag{18}$$

where $\mathbf{M}$ is the total dipole moment of the simulation cell, ε_0 is the vacuum permittivity, k_B is Boltzmann's constant, T is the absolute temperature, and $\langle V \rangle$ is the time-averaged simulation cell volume. Angle brackets denote ensemble averages taken over the equilibrated trajectory. The total dipole moment was computed directly in LAMMPS using the `compute dipole` command, which evaluates the simulation-cell dipole moment from the instantaneous atomic charges and coordinates under periodic boundary conditions. Dipole fluctuations were obtained from the time series of the total dipole vector components and simulation-cell volume collected over the equilibrated MD trajectory.

The total static dielectric constant (ε) was calculated as

$$\varepsilon = \varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{dip}} - 1, \tag{19}$$

and the absolute permittivity was obtained as $\epsilon = \varepsilon\,\varepsilon_0$.

5.2 PolyGraphMT Framework: Multi-Task Multi-Fidelity Learning

5.2.1 Loss function and normalization

Model parameters were optimized by minimizing a composite loss function defined as

$$\mathcal{L} = \sum_{p \in \mathcal{P}} \sum_{f \in \mathcal{F}_p} \frac{w_{p,f}}{|\mathcal{I}_{p,f}|} \sum_{i \in \mathcal{I}_{p,f}} \left(\hat{y}_{i,p} - y_{i,p}^{(f)} \right)^2, \tag{20}$$

where $\mathcal{I}_{p,f}$ denotes the set of polymers for which property p at fidelity level f is available, $|\mathcal{I}_{p,f}|$ is the corresponding number of observed data points, and $w_{p,f}$ is a weighting factor used to balance the contributions from different properties and data fidelities.

The normalization by $|\mathcal{I}_{p,f}|$ ensures that properties or fidelity levels with larger datasets do not dominate the optimization objective. The weighting factors $w_{p,f}$ provide a mechanism to adjust the relative influence of experimental and computational data during training, reflecting their differing reliability without enforcing explicit bias correction. The present framework adopts a prescribed fidelity hierarchy in which experimental data are treated as the highest-fidelity reference, followed by DFT, MD, and GC data. This ordering is motivated by the overall physical accuracy and level of approximation associated with each method, while recognizing that the effective reliability may still vary across different properties. The fidelity weights used in the loss function are therefore prescribed based on the assumed relative reliability of the corresponding data sources rather than learned adaptively from the data. Importantly,

View Article Online
DOI: 10.1039/D6DD00206D

not all properties contain the same fidelity combinations. In particular, GC data are only used for C_p in the present work. Since classical MD systematically overpredicts C_p due to the classical treatment of vibrational modes, the fidelity-aware C_p analysis was restricted to GC and experimental data, where the lower-fidelity trends remain more consistent with the experimental target space. Consequently, MD and GC data are not simultaneously combined within the current C_p multi-fidelity training setup. Although the present framework uses fixed fidelity weights, future extensions could incorporate adaptive fidelity weighting or discrepancy-learning approaches in which the relative reliability of different fidelity sources is learned directly from the data.

Each target property was normalized independently using statistics computed from the training data prior to optimization. This normalization prevents properties with larger numerical magnitudes or different physical units from disproportionately dominating the loss function.

5.2.2 Training protocol and hyperparameter optimization

Models were trained using mini-batch stochastic optimization on GPU-accelerated hardware. Data splits were defined at the SMILES level and kept fixed across all experiments to enable consistent comparison between single-task, multi-task, and multi-fidelity model configurations. All predictive performances reported in Figure 6 and related analyses were evaluated on held-out test datasets that were not used during either model training or hyperparameter optimization. For each property and learning configuration, the available data were partitioned into training, validation, and test sets at the polymer SMILES level to avoid data leakage between structurally identical polymers across splits. The validation set was used for hyperparameter optimization and model selection, while the final reported metrics were computed exclusively on the independent test set.

Hyperparameters, including learning rate, batch size, network depth, and regularization parameters, were optimized using Bayesian optimization implemented with Optuna [66]. Model selection was based primarily on validation mean absolute error (MAE), while RMSE and coefficient of determination (R^2) were additionally used for performance evaluation and comparison. For multi-task models, validation metrics were computed either as aggregated measures across tasks or with respect to a designated primary task, while the training objective itself remained unchanged. All reported model performances were averaged over five random seeds, and the variability across seeds was used to estimate uncertainty in the reported metrics.

5.2.3 Graph neural network architecture

PolyGraphMT employs a shared graph neural network (GNN) encoder with task-specific prediction heads for multi-task, multi-fidelity polymer property prediction. Molecular graphs are constructed from polymer repeat-unit SMILES using 35-dimensional atom features and 12-dimensional bond features. The shared encoder is implemented using a configurable message-passing architecture based on PyTorch Geometric, supporting GINE, GIN, and GCN convolutional backbones. In the present workflow, the Optuna hyperparameter optimization pipeline explored GINE and GIN variants.

View Article Online
DOI: 10.1039/D6DD00206D

The encoder consists of multiple message-passing layers in which node embeddings are iteratively updated through graph convolution, normalization, nonlinear activation, optional residual connections, and dropout. For the GINE backbone, edge features are explicitly incorporated during message passing, whereas GIN and GCN variants use only node features. Graph-level polymer embeddings are obtained through global pooling, with mean pooling used in the present study.

The number of message-passing layers, embedding dimensions, normalization strategy, activation functions, dropout values, and prediction-head architectures were optimized using Optuna based on validation performance and training stability. The hyperparameter search space included 3–7 message-passing layers, graph embedding dimensions ranging from 256 to 768, and prediction-head depths of 1–3 fully connected layers.

To enable simultaneous multi-property learning, all tasks share the same molecular graph encoder while maintaining separate task-specific multilayer perceptron (MLP) prediction heads. Each prediction head uses the same architectural template consisting of repeated linear layers and nonlinear activations followed by a task-specific output layer. Fidelity information is incorporated through learned fidelity embeddings, which are either concatenated with the graph embedding or applied through feature-wise linear modulation (FiLM) prior to property prediction. A summary of the key architectural hyperparameters and training configurations used in the PolyGraphMT framework is provided in Table S2 of the Supporting Information.

Acknowledgements

This work was supported by the Lucy Family Institute for Data & Society Postdoctoral Fellowship and National Science Foundation grants 2332270 and 2102592. The authors also acknowledge the Center for Research Computing at the University of Notre Dame for providing the necessary computing resources.

Declarations

Funding declaration. This work was supported by the Lucy Family Institute for Data & Society Postdoctoral Fellowship and National Science Foundation grants 2332270 and 2102592.

Competing interests. The authors declare no competing interests.

Data availability. The ADEPT workflow for automated polymer structure generation, molecular dynamics simulations, and DFT calculations is publicly available at https://github.com/sobinalosious/ADEPT. An archived version of the software used in this study is available through Zenodo (DOI: https://doi.org/10.5281/zenodo.20631234). The PolyGraphMT framework for multi-task and multi-fidelity polymer property prediction is publicly available at https://github.com/sobinalosious/PolyGraphMT. An archived version of the software used in this study is available through Zenodo (DOI: https://doi.org/10.5281/zenodo.20631261). The processed datasets, trained models, and scripts required to reproduce the reported results are

View Article Online
DOI: 10.1039/D6DD00206D

provided in the corresponding repositories. These repositories contain the data processing workflows, machine learning models, and analysis scripts necessary to reproduce the results reported in this study.

Ethics approval and consent to participate. Not applicable

Consent for publication. Not applicable

Materials availability. Not applicable

Author contributions. S.A. designed the study, developed the workflow, performed simulations, and implemented the machine learning models. Y.L., J.X., and R.Z. contributed to data generation and analysis. G.L. contributed to machine learning methodology and implementation. M.J. and T.L. supervised the research and contributed to the conceptual development of the study. All authors contributed to writing and reviewing the manuscript.

References

[1] Bicerano, J. *Prediction of polymer properties* (cRc Press, 2002).

[2] Sperling, L. H. *Introduction to physical polymer science* (John Wiley & Sons, 2015).

[3] Mike, J. F. & Lutkenhaus, J. L. Recent advances in conjugated polymer energy storage. *Journal of Polymer Science Part B: Polymer Physics* **51**, 468–480 (2013).

[4] Bujak, P. *et al.* Polymers for electronics and spintronics. *Chemical Society Reviews* **42**, 8895–8999 (2013).

[5] Lyu, S. & Untereker, D. Degradability of polymers for implantable biomedical devices. *International journal of molecular sciences* **10**, 4033–4065 (2009).

[6] Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS macro letters* **6**, 1078–1082 (2017).

[7] Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. Polyinfo: Polymer database for polymeric materials design. In 2011 International Conference on Emerging Intelligent Data and Web Technologies (2011). Pp. 22–29.

[8] Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3**, 54 (2017).

[9] Alosious, S., Jiang, M. & Luo, T. Computation and machine learning for materials: Past, present, and future perspectives: S. alosious et al. *MRS Bulletin* **50**, 1212–1224 (2025).

View Article Online
DOI: 10.1039/D6DD00206D

[10] Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* **1** (2013).

[11] Curtarolo, S. *et al.* Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science* **58**, 218–226 (2012).

[12] Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom* **65**, 1501–1509 (2013).

[13] Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **1**, 1–7 (2014).

[14] Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C* **122**, 17575–17585 (2018).

[15] Lin, T.-S. *et al.* Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science* **5**, 1523–1531 (2019).

[16] Walsh, D. J. *et al.* Community resource for innovation in polymer technology (cript): a scalable polymer material data structure (2023).

[17] Kim, S., Schroeder, C. M. & Jackson, N. E. Open macromolecular genome: Generative design of synthetically accessible polymers. *ACS Polymers Au* **3**, 318–330 (2023).

[18] Kuenneth, C. & Ramprasad, R. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature communications* **14**, 4099 (2023).

[19] Kuenneth, C. *et al.* Polymer informatics with multi-task learning. *Patterns* **2** (2021).

[20] Patra, A. *et al.* A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Computational Materials Science* **172**, 109286 (2020).

[21] Antoniuk, E. R., Li, P., Kailkhura, B. & Hiszpanski, A. M. Representing polymers as periodic graphs with learned descriptors for accurate polymer property predictions. *Journal of Chemical Information and Modeling* **62**, 5435–5445 (2022).

[22] Afzal, M. A. F. *et al.* High-throughput molecular dynamics simulations and validation of thermophysical properties of polymers for various applications. *ACS Applied Polymer Materials* **3**, 620–630 (2020).

[23] Hayashi, Y., Shiomi, J., Morikawa, J. & Yoshida, R. Radonpy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Computational Materials* **8**, 222 (2022).

[24] Yoshida, R. *et al.* Omics-scale polymer computational database transferable to real-world artificial intelligence applications. *arXiv preprint arXiv:2511.11626* (2025).

[25] McQuarrie, D. A. *Quantum chemistry* (University Science Books, 2008).

[26] Lukes, J. R., Li, D. Y., Liang, X.-G. & Tien, C.-L. Molecular dynamics study of solid thin-film thermal conductivity. *Journal of Heat Transfer* **122**, 536–543 (2000).

[27] Zhang, R., Xu, J., Zhang, H., Xu, G. & Luo, T. Active learning-guided exploration of thermally conductive polymers under strain. *Digital Discovery* **4**, 812–823 (2025).

[28] Xu, J. & Luo, T. Unlocking enhanced thermal conductivity in polymer blends through active learning. *npj Computational Materials* **10**, 74 (2024).

[29] Ma, R. *et al.* Machine learning-assisted exploration of thermally conductive polymers based on high-throughput molecular dynamics simulations. *Materials Today Physics* **28**, 100850 (2022).

[30] Buchholz, J., Paul, W., Varnik, F. & Binder, K. Cooling rate dependence of the glass transition temperature of polymer melts: Molecular dynamics study. *The Journal of chemical physics* **117**, 7364–7372 (2002).

[31] Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Scientific reports* **6**, 20952 (2016).

[32] Sha, W. *et al.* Machine learning in polymer informatics. *InfoMat* **3**, 353–361 (2021).

[33] Queen, O. *et al.* Polymer graph neural networks for multitask property learning. *npj Computational Materials* **9**, 90 (2023).

[34] Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science* **129**, 156–163 (2017).

[35] Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials* **7**, 77 (2021).

View Article Online
DOI: 10.1039/D6DD00206D

[36] Liu, Y., Alosious, S., Zhou, J., Jiang, M. & Luo, T. Machine learning in nanoscale thermal transport. *Annual Review of Heat Transfer* **28** (2025).

[37] Liu, G., Zhao, T., Xu, J., Luo, T. & Jiang, M. Graph rationalization with environment-based augmentations. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2022). Pp. 1069–1078.

[38] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning (ICML) (2017). Pp. 1263–1272.

[39] Caruana, R. Multitask learning. *Machine learning* **28**, 41–75 (1997).

[40] Liu, G. *et al.* Open polymer challenge: Post-competition report. *arXiv preprint arXiv:2512.08896* (2025).

[41] Suter, J. L. *et al.* Rapid, accurate and reproducible prediction of the glass transition temperature using ensemble-based molecular dynamics simulation. *Journal of Chemical Theory and Computation* **21**, 1405–1421 (2025).

[42] Hung, J.-H., Patra, T. K. & Simmons, D. S. Forecasting the experimental glass transition from short time relaxation data. *Journal of Non-Crystalline Solids* **544**, 120205 (2020).

[43] Banerjee, A., Iscen, A., Kremer, K. & Kukharenko, O. Determining glass transition in all-atom acrylic polymeric melt simulations using machine learning. *The Journal of Chemical Physics* **159** (2023).

[44] Zhang, S. & Webb, M. A. Asymmetric effects underlying dynamic heterogeneity in miscible blends of poly (methyl methacrylate) with poly (ethylene oxide). *Macromolecules* **59**, 2302–2314 (2026).

[45] Wypych, G. *Handbook of polymers* (Elsevier, 2022).

[46] Liu, Y., Xu, J., Zhang, R., Jiang, M. & Luo, T. Active learning-enabled multi-objective design of thermally conductive and mechanically compliant polymers. *arXiv preprint arXiv:2603.23494* (2026).

[47] Bhowmik, R., Sihn, S., Varshney, V., Roy, A. K. & Vernon, J. P. Calculation of specific heat of polymers using molecular dynamics simulations. *Polymer* **167**, 176–181 (2019).

[48] Alosious, S., Xu, J., Jiang, M. & Luo, T. A transfer learning framework integrating molecular dynamics and group contribution methods for predicting polymer specific heat capacity. *Polymer Chemistry* (2026).

[49] Lorentz, H. A. *The theory of electrons and its applications to the phenomena of light and radiant heat* Vol. 29 (GE Stechert & Company, 1916).

[50] Ma, R. & Luo, T. Pi1m: a benchmark database for polymer informatics. *Journal of Chemical Information and Modeling* **60**, 4684–4690 (2020).

[51] Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling* **55**, 2562–2574 (2015).

[52] Fortunato, M. E. & Colina, C. M. pysimm: A python package for simulation of molecular systems. *SoftwareX* **6**, 7–12 (2017).

[53] Theodorou, D. N. & Suter, U. W. Detailed molecular structure of a vinyl polymer glass. *Macromolecules* **18**, 1467–1478 (1985).

[54] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25**, 1157–1174 (2004).

[55] Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).

[56] Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **36**, 3219–3228 (1980).

[57] Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics* **117**, 1–19 (1995).

[58] Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics* **23**, 327–341 (1977).

[59] Hockney, R. W. & Eastwood, J. W. *Computer simulation using particles* (crc Press, 2021).

[60] Allen, M. P. & Tildesley, D. J. *Computer simulation of liquids* (Oxford university press, 2017).

[61] Landrum, G. Rdkit documentation. *Release* **1**, 4 (2013).

[62] Smith, D. G. *et al.* Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of chemical physics* **152** (2020).

[63] Mardirossian, N. & Head-Gordon, M. ωb97m-v: A combinatorially optimized, range-separated hybrid, meta-gga density functional with vv10 nonlocal correlation. *The Journal of chemical physics* **144** (2016).

DOI: 10.1039/D6DD00206D

[64] Buckingham, A. D. Permanent and induced molecular moments and long-range intermolecular forces. *Advances in chemical physics: Intermolecular forces* 107–142 (1967).

[65] Hay, P. J. & Wadt, W. R. Ab initio effective core potentials for molecular calculations. potentials for the transition metal atoms sc to hg. *The Journal of chemical physics* **82**, 270–283 (1985).

[66] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019). Pp. 2623–2631.

View Article Online
DOI: 10.1039/D6DD00206D

Data availability statement

The ADEPT workflow for automated polymer structure generation, molecular dynamics simulations, and DFT calculations is publicly available at GitHub (https://github.com/sobinalosious/ADEPT). An archived version of the software used in this study is available through Zenodo (DOI: 10.5281/zenodo.20631234).

The PolyGraphMT framework for multi-task and multi-fidelity polymer property prediction is publicly available at GitHub (https://github.com/sobinalosious/PolyGraphMT). An archived version of the software used in this study is available through Zenodo (DOI: 10.5281/zenodo.20631261).

The processed datasets, trained models, and scripts required to reproduce the reported results are provided in the corresponding repositories. These repositories contain the data processing workflows, machine learning models, and analysis scripts necessary to reproduce the results reported in this study.