Machine Learning in Nanoscale Thermal Transport

Yuhan Liu[1†], Sobin Alosious[2,3†], Jiahang Zhou[3], Meng Jiang[2,4], Tengfei Luo[2,3,5,6*]

[1]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, 46556, USA.

[2]Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN, 46556, USA.

[3]Department of Aerospace and Mechanical Engineering, 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;

†These authors contributed equally to this work.

Abstract

Thermal transport is a fundamental process underpinning a wide range of applications. Traditional experimental and computational methods have substantially advanced our understanding of nanoscale thermal transport. However, they continue to face challenges, such as the cost and speed associated with the experimental and computational methods, limiting their effectiveness for investigating complex nanoscale systems. Machine learning (ML) has emerged as a powerful approach in this domain, offering unique capabilities for processing extensive datasets, identifying intricate patterns, and designing novel materials. In this article, we provide an overview of recent advancements and applications of ML in nanoscale thermal transport, emphasizing polymeric systems along with interfaces and inorganic materials. Specifically, we review recent progress in property prediction, material design, atomistic simulations, and data analysis. We also highlight promising ML methods, including transfer learning, active learning, and physics-informed neural networks, which effectively address data scarcity and improve model accuracy. Finally, we present our perspective on emerging trends and future research directions, emphasizing their potential to guide the discovery and design of next-generation thermal materials and to unravel complex thermal phenomena.

Keywords: Nanoscale, Thermal Transport, Machine Learning (ML), Polymers, ML Interatomic Potentials

1 Introduction

Thermal transport at the nanoscale plays a vital role in diverse technologies, including nanoscale electronics, energy conversion systems, phase-change memory, thermal barriers, thermoelectric devices, photonic components, and biomedical applications.[1,2] As device dimensions approach nanometer scales, managing heat flow becomes increasingly challenging due to size effects, interface phenomena, and material heterogeneities.[3] For example, local hot spots in integrated circuits can degrade performance and reliability if heat is not efficiently dissipated.[4] In thermoelectric materials, low thermal conductivity (TC) is essential to enhance energy conversion efficiency.[5] A substantial portion of recent research, including the applications reviewed in this work, focuses on polymer-based systems because of their lightweight, flexible nature, and tunable properties. However, polymers generally exhibit low intrinsic TC, typically in the range of 0.1–0.5 W/mK, primarily due to their disordered molecular structure and weak inter-chain interactions that limit efficient phonon transport.[6] These diverse needs across applications, from efficient heat dissipation in nanoelectronics to thermal insulation in polymers and thermoelectric materials, highlight the importance of understanding and controlling heat transport mechanisms at the nanoscale.

Traditional modeling approaches for thermal transport provide essential physical insights but face limitations when dealing with nanoscale complexity. First-principles density functional theory (DFT) can yield accurate TC predictions. However, they are computationally expensive for large, complex, or disordered systems.[2] Classical molecular dynamics (MD) simulations can handle larger systems (including amorphous materials and interfaces).[7] However, their accuracy depends on the quality of interatomic potentials, and they cannot capture quantum effects.[8] Both DFT and MD cannot handle mesoscale structures encountered in practical applications (e.g., nanotransistors). At this scale, tools like Boltzmann transport equations (BTE) are accurate, but their numerical computational efficiency is low. Moreover, experimental or brute-force computational searches for materials with improved thermal properties are time-consuming. These challenges have motivated the integration of machine learning (ML) techniques into nanoscale thermal transport research.[9]

Data-driven ML offers an efficient way to model, predict, and design materials with targeted thermal properties. ML models can rapidly estimate TC from a material's composition or structure by learning

from existing data from experiments or simulations, enabling high-throughput screening to complement traditional calculations.[10] For example, ML models trained on physically meaningful descriptors have been successfully applied in a high-throughput screening framework to identify polymer chains with high TC, revealing structure-property relationships and accelerating the discovery of promising polymers for thermal management applications.[11] Additionally, ML-based interatomic potential can emulate near-DFT accuracy, allowing efficient simulations of phonon transport in systems that are otherwise difficult to treat with first-principles DFT methods.[12]

Recent demonstrations of the synergy between physics-based modeling methods (such as MD and DFT) and ML techniques (including regression models, neural networks, and symbolic regression) have spanned studies on polymers, interfaces, and inorganic nanostructures. In polymers, ML models have identified molecular structures that yield higher TC despite these materials' intrinsically poor heat transfer performance.[13] At material interfaces, which often dominate thermal resistance in nanostructured devices, ML models can predict thermal boundary conductance (TBC) and guide interface engineering to minimize resistance.[14] Similarly, ML models can quickly evaluate how defects or nanoscale architecture affect phonon transport in crystalline alloys and superlattices.[15,16]

In summary, integrating ML with conventional thermal transport study methods offers a powerful route to accelerate materials development and deepen our understanding of heat transfer at the nanoscale. This chapter highlights recent advances in applying ML techniques to nanoscale thermal transport in polymers, interfaces, and nanostructured materials, emphasizing how combining data-driven models with physics-based insight enables the design of materials with tailored thermal transport properties and help conventional thermal modeling tools to be more efficient.

2 Machine Learning Methods

2.1 Fundamental Principles of Machine Learning

ML techniques have shown promising applicability in addressing problems involving high-dimensional features and complex mappings, especially when large datasets are available. In thermal transport, one of the

primary goals of ML methods is to establish quantitative mappings between thermal transport properties and materials characteristics, such as structural features and other physical descriptors. Developing an ML-based model involves three essential components: datasets, descriptor selection, and learning algorithms.[17]

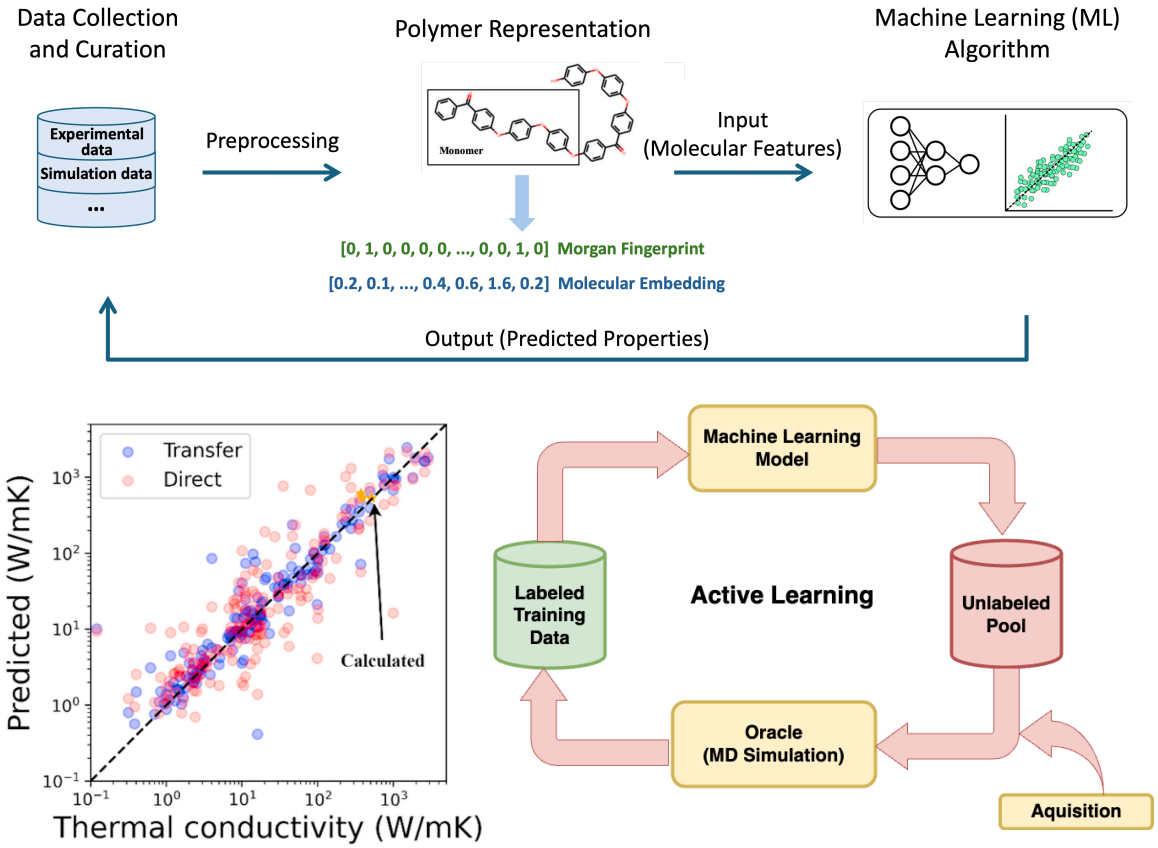

Fig. 1 (a) ML workflow establishing mappings from polymer fingerprints (input features) to target thermal properties. (b) Parity plot comparing TL-predicted TC values and high-fidelity TC data for three compounds ($GaBN_2$, $AlBN_2$ and B_2AsP). The error is quantified using the factor difference ($10^{|\log_{10}\kappa_{\text{pred}}-\log_{10}\kappa_{\text{DFT}}|}$). (Reprinted with permission from Liu et al.[18] Copyright 2022 Elsevier Ltd) (c) The schematic diagram of active learning (AL) strategy combined with high-throughput MD simulation to identify thermally conductive polymer blends.

The first step is dataset acquisition, comprising the target property data. Such datasets can be obtained from experiments, numerical simulations (e.g., MD and DFT), or available published data. The second step involves identifying appropriate descriptors - quantities that characterize the materials system and exhibit strong correlations with the property of interest.[19] Finally, ML algorithms are employed to discern underlying relationships between descriptors and target properties (Figure 1a).

Depending on their learning paradigms, ML algorithms can be classified into supervised, unsupervised, semi-supervised, and reinforcement learning (RL).[20] Supervised learning algorithms use labeled training datasets to learn a mapping function between input descriptors and output targets, enabling regression or classification tasks. Based on the nature of the output, if the model targets are discrete quantities, such as crystal structures or specific structural motifs, classification methods can be used to find the prediction function, whereas regression methods are applicable for continuous properties, such as glass transition temperature (T_g). In contrast, unsupervised learning, such as principal component analysis (PCA) and K-means clustering, is employed to uncover intrinsic data patterns and draw inferences without labeled outputs, facilitating dimensionality reduction and data clustering. Another category is semi-supervised learning, which bridges supervised and unsupervised learning, using a limited amount of labeled data to initially train a model and then refined iteratively using abundant unlabeled data. RL is a method that allows an agent to interactively learn through trial and error, optimizing its performance based on feedback from actions and experiences. Both conventional and deep learning methods[21] have been successfully applied to heat transfer, such as the prediction of TC and TBC. Common conventional ML methods are decision trees (DT), random forests (RF), support vector machines (SVM), and Gaussian process regression (GPR). Popular deep learning architectures include artificial neural networks (ANNs), graph neural networks (GNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs).

Before practical application, ML models must be evaluated on unseen datasets, known as test sets, to assess their generalization and extrapolation ability. Most thermal transport prediction tasks involve regression, commonly evaluated by metrics such as mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). These metrics quantify generalization error, with lower values indicating better model performance. However, they are sensitive to outliers. To mitigate data sensitivity issues, correlation metrics such as Pearson, Kendall, and Spearman coefficients have been employed to evaluate the model performance.[20] Additionally, model validation and optimization methods that rely on withholding subsets of data during training, ranging from a simple holdout, over k-fold cross-validation, leave-one-out cross-validation, Monte Carlo cross-validation, up to leave-one-cluster-out cross-validation, are also extensively used.[22]

A key challenge in nanoscale thermal transport is the limited availability of high-quality data points, often restricted to hundreds or thousands at most.[23] To overcome this limitation, recent advances in ML have emphasized techniques suited for small datasets, such as transfer learning (TL) and active learning (AL). Moreover, integrating physical laws explicitly into ML frameworks presents a promising strategy to enhance prediction accuracy and reliability, as detailed in subsequent subsections.

2.2 Transfer Learning and Multi-Fidelity Models

TL and multi-fidelity modeling have become increasingly important techniques in nanoscale thermal transport research, where high-quality data for properties like TC are often limited. TL enables models to learn from large datasets of proxy or lower-fidelity properties and then fine-tune on smaller high-fidelity datasets, thus improving predictive performance and enabling better generalization to new materials.[24] Multi-fidelity approaches combine information across data of varying accuracy levels, enhancing the model's robustness.[25]

Wu et al.[26] provided a detailed example of TL applied to polymer TC prediction. Their workflow consisted of two stages: first, a model was trained to predict the glass transition temperature (T_g) of polymers using a large dataset containing thousands of entries. Molecular descriptors were generated from the repeat unit structures of the polymers, and T_g was selected as the source property because it is both more widely available than TC and physically related to factors such as chain rigidity, segmental motion, and packing density, which also influence thermal transport. In the second stage, the parameters learned from the T_g model were transferred to initialize a TC prediction model, which was then fine-tuned using a much smaller dataset of experimentally measured polymer TCs. By leveraging the structural–property correlations captured in the T_g pretraining phase, the TL model significantly improved TC prediction accuracy compared to training from scratch. Using this approach, the authors identified and synthesized three polyimide-based polymers with room-temperature TCs 18–80% higher than typical amorphous polyimides, demonstrating that abundant T_g data can be effectively used to bootstrap accurate TC prediction when direct measurements are scarce. Building on leveraging large-scale data, Ju et al.[27] focused on lattice TC prediction in crystalline materials. By pretraining a neural network on low-order features and transferring the learned descriptors, their model

enabled extrapolative predictions across over 60,000 compounds. The study emphasized that descriptors like phonon relaxation time and dipole polarizability play a critical role more than mechanical hardness alone in determining TC. Extending this concept further, Liu et al.[18] showed that TL enhances lattice TC predictions by integrating limited high-fidelity experimental and first-principles data with extensive low-fidelity data from phenomenological empirical models. Treating these as related tasks, their approach improved model accuracy by up to 23% in R^2 and reduced the average factor difference, a logarithmic metric that quantifies the typical order-of magnitude error between predicted and reference values by 30%. The transfer-learned model screened an extensive semiconductor database, identifying several candidates with room-temperature TC above 350 W/mK, confirmed by first-principles simulations. Predicted top candidates are illustrated in Figure 1b comparing TL predictions with high-fidelity data. Recent works show that TL and multi-fidelity approaches can significantly enhance ML model performance in predicting complex thermal properties by effectively combining large low-fidelity datasets with limited but crucial high-fidelity information.[28]

2.3 Active Learning

AL algorithms typically comprise three key components: a surrogate model to predict the target property, an acquisition function to identify the most informative next experiment, and an iterative update step that incorporates newly acquired data into the training set. Surrogate models, such as GPR, approximate expensive simulation results or experimental measurements, enabling rapid screening of candidates at reduced computational cost.[23] AL frameworks further accelerate discovery by strategically selecting the most informative data points to evaluate next, balancing exploitation of known promising regions and exploration of uncertain areas.[29] Kim et al.[30] applied GPR with AL to identify polymers with high T_g. Starting from a small dataset, their iterative strategy terminated after identifying 10 polymers exceeding the T_g threshold, with benchmarking showing superior performance over random selection.

Extending to TC, Xu et al.[31] combined high-throughput MD simulations with AL to explore 550,000 polymer blends. Starting from MD simulations of about 600 single-component polymers and 200 blends, an AL framework was used to explore the TC of approximately 550,000 unlabeled blends using a weighted sum representation method. Figure 1c shows a schematic diagram of their AL strategy. Their method revealed

that AL effectively accelerated the discovery of high-performance blends and revealed a positive association between TC improvement, the radius of gyration (R_g), and hydrogen bonding. Zhang et al.[32] used an AL framework to discover amorphous polymers with enhanced TC under mechanical strain. They trained a GPR model on MD simulation data and applied Bayesian optimization to iteratively select new candidates from the PoLyInfo database for further simulation. After several AL cycles, they identified ten strained polymers with TC above 1 W/mK and revealed structural features linked to high TC under strain. These studies demonstrate that combining surrogate modeling with AL enables efficient exploration of vast chemical spaces.

2.4 Physics-Informed Machine Learning

Unlike the purely data-driven ML, Physics-Informed Neural Networks (PINNs) are a type of ML models that incorporate physical laws directly into the training of neural networks. The physics laws are typically expressed as partial differential equations (PDEs) or ordinary differential equations (ODEs) together with boundary/initial conditions. PINNs were originally introduced by Raissi, Perdikaris, and Karniadakis,[33,34] and have emerged as a powerful framework for solving forward and inverse problems in computational physics. The framework is particularly effective in regimes where data are limited or hard to obtain.

PINNs were developed to address challenges faced by classical numerical methods such as finite difference and finite element methods. By incorporating governing PDEs/ODEs into the loss function and leveraging automatic differentiation (AD), physics is enforced in PINN models without the need for domain discretization. To be specific, derivatives are computed analytically using chain rule via AD, and the loss function is constructed by evaluating PDE/ODE residuals. Since the introduction of PINNs, many frameworks have been developed to include advanced training techniques,[35,36] modified network architectures,[37–39] and improved constraint-handling strategies.[40]

Consider a PDE of the form:

$$\mathcal{N}[u(x)] = f(x), \quad x \in \Omega, \tag{1}$$

with boundary or initial conditions:

$$\mathcal{B}[u(x)] = g(x), \quad x \in \partial\Omega. \tag{2}$$

The $\mathcal{N}[\cdot]$ is a differential operator (may include spatial and temporal derivatives) that characterizes the governing PDE, $\mathcal{B}[\cdot]$ is a boundary or initial condition operator, Ω is the spatial domain of interest, and $\partial\Omega$ is the boundary of the domain Ω. A PINN approximates the solution $u(x)$ using a neural network $u_\theta(x)$, where θ denotes the network parameters. The total loss function is expressed in Eq. 3,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{PDE}}\mathcal{L}_{\text{PDE}} + \lambda_{\text{BC}}\mathcal{L}_{\text{BC}} + \lambda_{\text{Data}}\mathcal{L}_{\text{Data}}, \tag{3}$$

where λ_{PDE}, λ_{BC}, and λ_{Data} are tunable weights. The $\mathcal{L}_{\text{PDE}}$ and $\mathcal{L}_{\text{BC}}$ are PDE residual and boundary/initial condition residual, respectively, as expressed in Eq. 4 and Eq. 5, and $\mathcal{L}_{\text{Data}}$ is data residual. Usually data are not needed if the complete physics is known. However, in inverse problems or in the scenarios where the physics is only partially available, data will be needed to guarantee the uniqueness of solution. The N_r and N_b denote the number of collocation points sampled in the interior domain Ω and on the boundary (or initial surface) $\partial\Omega$, respectively. Training a PINN is an optimization process in which neural network parameters θ are iteratively updated to minimize the loss function, thereby the governing equations and boundary/initial conditions can be satisfied.

$$\mathcal{L}_{\text{PDE}} = \frac{1}{N_r}\sum_{i=1}^{N_r}\left|\mathcal{N}[u_\theta(x_r^{(i)})] - f(x_r^{(i)})\right|^2, \tag{4}$$

$$\mathcal{L}_{\text{BC}} = \frac{1}{N_b}\sum_{i=1}^{N_b}\left|\mathcal{B}[u_\theta(x_b^{(i)})] - g(x_b^{(i)})\right|^2. \tag{5}$$

Once training is complete, the well-trained PINN serves as a non-linear function that maps any input variables (i.e., independent variables of the governing PDE) to their corresponding solutions (i.e., dependent variables of the governing PDE). Theoretically, a well-trained PINN can infer solutions across the entire input space, whereas the traditional numerical methods are limited to providing solutions only at mesh nodes or cell centers.

To solve PDEs in heat transfer problems, PINNs can enforce governing physics (e.g., Fourier's heat conduction equation, phonon BTE, etc) through loss functions. Take phonon BTE for example, Li et al.[41] used a PINN framework with two subnets to solve the parametrized nongray phonon BTE under small

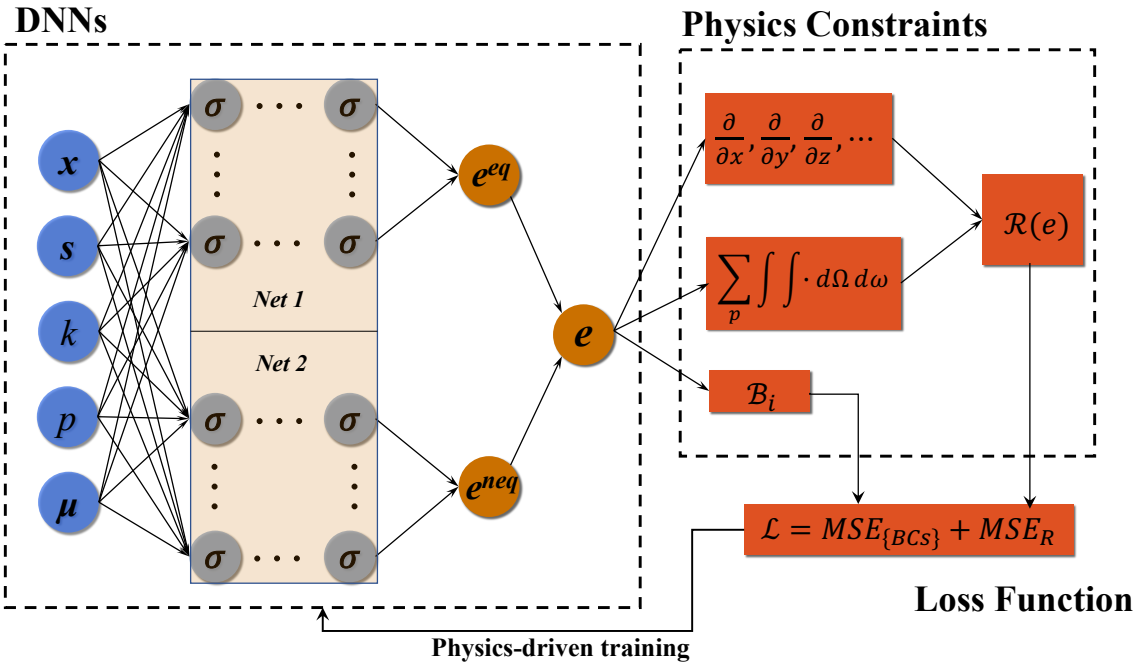

Fig. 2 Schematic of the PINN architecture for solving phonon BTE. Two deep neural networks (DNNs) are employed to approximate the equilibrium (e^{eq}) and non-equilibrium (e^{neq}) components of the phonon energy distribution, respectively. The inputs to the networks include the spatial coordinate vector $\mathbf{x}$, the directional unit vector $\mathbf{s} = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)$(θ is the polar angle and ϕ is the azimuthal angle), the wave number k, and the polarization index p. The symbol $\boldsymbol{\mu}$ represents additional parameters, which for example can be characteristic system length L, to enable parametric learning.

temperature difference, as shown in Figure 2. The framework is successfully validated on 1D cross-plane, 2D in-plane, 2D square, and 3D cuboid problems. Additionally, the Knudsen number is an extra input parameter into the model, enabling parametric learning and fast inference of solution under different Knudsen numbers. Later, Li et al.[42] extended this framework for solving the thermal transport problems with large temperature gradient, where the assumption of small temperature difference no longer holds and the equilibrium phonon energy distribution function cannot be linearized. This work features a pretrained shallow neural network that predicts the scaling coefficient that bridges local temperature and the equilibrium phonon energy. This enhanced framework enables accurate prediction for 1D to 3D thermal problems with any temperature gradient. Later, Zhou et al.[43] improved the PINN framework and adapt it to solve the transient phonon BTE. The time space is encoded into the neural network input layer, which allows the model to predict transient temperature change in 1D and 2D problems. To extend to model to handle coupled phonon and electron transport problems, Li et al.[44] used three subnets to approximate phonon and electron pseudo temperatures

and distribution functions separately. Two pretrained neural networks are used in the framework to predict scaling coefficient in the electron and phonon distribution functions.

3 Data Acquisition and Databases

Implementing ML in any domain requires relevant and sufficient data. In nanoscale thermal transport, data can come from experimental measurements, computational simulations, or increasingly from public databases that aggregate materials properties.

3.1 Experimental Databases

High-quality experimental databases are foundational in applying ML to thermal transport. They provide curated property data, synthesis conditions, and metadata in structured formats for model development and validation. Domain-specific resources in polymers, nanofluids, and thermoelectrics have enabled data-driven discovery, supporting model generalizability and benchmarking across diverse material classes. As highlighted in recent reviews, materials informatics frameworks integrating experiments and simulations with ML are accelerating the identification of materials with extreme TC and functional performance.[9] A leading example is PoLyInfo,[47] developed by NIMS, which contains over 500,000 curated experimental data points collected over two decades. It organizes polymers by constitutional repeating units and provides detailed structural, property, fabrication, and formation data. Designed for human and ML use, PoLyInfo enforces strict data curation, making it a cornerstone for polymer research and a model for managing complex polymer structures.

Beyond polymers, several other domain-specific experimental databases have been developed to support ML applications in areas such as nanofluids and thermoelectric materials. Mondejar et al.[48] compiled an open-access nanofluid database with 8,118 data records across 307 datasets, covering 13 base fluids and 19 nanoparticles. It organizes experimental data for five thermodynamic properties and two transport properties, offering easy access to information on nanofluid behavior. The database, hosted on the Dortmund Databank, also analyzes data consistency and challenges faced during collection.

Na et al.[45] developed a public database containing experimentally synthesized thermoelectric materials and their measured thermoelectric properties to support data-driven discovery. Using this dataset, they

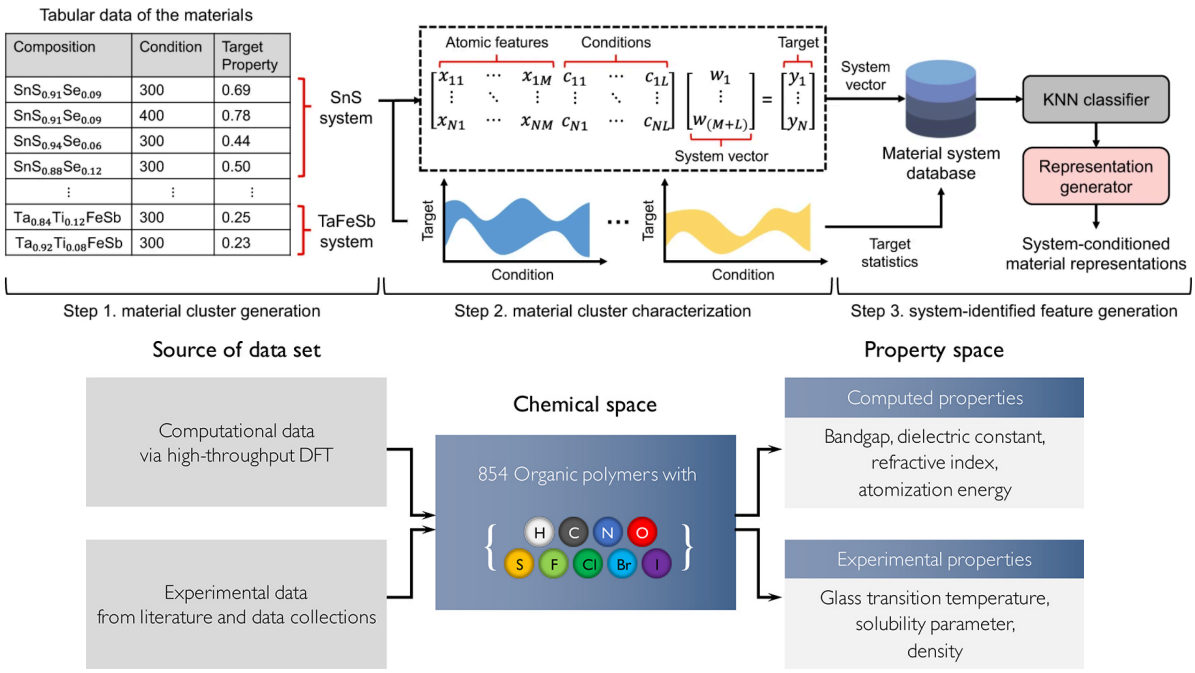

Fig. 3 (a) The overall process of SIMD generates the material representations for input tabular data of the materials. (Reproduced from Na et al.[45] under CC BY 4.0.) (b) Overview of Polymer Genome polymer data set used to develop property prediction models. (Reprinted with permission from Kim et al.[46] Copyright 2018 American Chemical Society)

built prediction models achieving R^2-scores greater than 0.9 for estimating thermoelectric properties from chemical compositions. They also introduced a new material descriptor, an identified material descriptor (SIMD), to improve ML extrapolation, significantly enhancing the R^2-score from 0.13 to 0.71 in predicting figure of merit (ZT) values for unexplored material groups. The overall process of SIMD in generating the material representations is shown in Figure 3a.

Lee et al.[49] introduced TEXplorer, a web-based platform for collecting and sharing thermoelectric materials data, including synthesis details, material characterization, transport measurements, and electronic structures from experiments and computations. The platform offers data upload, retrieval, visualization, post-processing, and ML-based property prediction tools. These databases offer essential foundations for thermal transport modeling by ensuring reproducibility, accelerating model training, and enabling rapid screening of material candidates using ML.

3.2 Simulation-Generated Databases

In addition to experimental resources, simulation-generated databases have become vital to ML in nanoscale thermal transport. These datasets, often derived from high-throughput first-principles calculations or MD simulations, offer expansive coverage of materials space at a relatively low cost compared to experimental methods. Several platforms have been developed to provide open access to large-scale simulation-derived data across polymers and inorganic materials. One of the most widely used platforms in this space is the Materials Project,[50] an open-access platform designed to accelerate materials discovery through high-throughput first-principles calculations. It provides an extensive database of computed structural, electronic, and energetic properties for over 33,000 inorganic compounds, accessible via web applications, APIs, and open-source analysis tools. The Materials Project enables rapid prototyping and data-driven exploration of new materials by combining large-scale computation, web-based tools, and community collaboration.

Focusing on polymers, Polymer Genome[46] is a data-powered informatics platform that accelerates polymer property prediction using surrogate ML models. It combines a curated dataset of computationally and experimentally measured properties for 854 polymers, represented through hierarchical fingerprints spanning atomic, molecular, and morphological features. The platform enables rapid property predictions such as bandgap, dielectric constant, glass transition temperature, solubility parameter, and density, with uncertainty estimates. An overview of the Polymer Genome dataset used for the development of property prediction models is shown in Figure 3b.

To further expand the design space, Ma et al.[51,52] developed PI1M, a benchmark database containing approximately 1 million virtual polymers generated to support ML research in polymer informatics. A generative model was trained on around 12,000 polymers from PolyInfo and used to create the expanded dataset. They introduced a polymer embedding (PE) representation and demonstrated its effectiveness in regression tasks for properties like density, T_g, T_m, and dielectric constants. PI1M was shown to cover similar chemical space as PolyInfo while significantly filling previously sparse regions, providing a valuable resource for future polymer informatics studies. Building upon this, Xu et al.[53] developed the POINT^2 database, which integrates PI1M and other datasets into a comprehensive benchmark with diverse polymer

representations and ML models, further enhancing property predictions, uncertainty estimation, and polymer synthesizability.

For inorganic materials, the Open Quantum Materials Database (OQMD)[54] contains over one million compounds calculated through high-throughput DFT studies, including both experimentally known and hypothetical structures. It supports materials discovery efforts worldwide and increasingly integrates ML-based projects to accelerate exploration. The database emphasizes open public access and develops universal querying protocols to align with Findable, Accessible, Interoperable, and Reusable (FAIR) data principles. OQMD remains a significant platform for data-driven materials research and innovation. However, building high-fidelity simulation-generated databases presents several challenges. Ensuring accuracy often requires extensive convergence testing, careful choice of exchange-correlation functionals, and validation against experimental benchmarks, which can be computationally expensive. For MD-based datasets, force field or potential selection strongly influences fidelity, and transferability across diverse chemical systems remains non-trivial. Achieving broad chemical coverage while avoiding data imbalance is another hurdle, as over-representation of certain chemistries can bias ML models. Additionally, consistent data curation, metadata annotation, and adherence to FAIR principles are essential to ensure reproducibility and interoperability across platforms. These challenges underscore the importance of standardized workflows, automated error detection, and community-driven validation efforts to maintain the reliability of large-scale computational materials databases.

3.3 Feature engineering

Feature engineering generally includes data representation and feature selection, both critical determinants of ML model performance. Appropriate data representation (also termed *descriptors* or *fingerprints*) is essential for ensuring model accuracy and interpretability. Collected chemical, structural, or material-property data should be converted into machine readable descriptors, serving as input features for ML algorithms. Insufficient descriptors may hinder model convergence, while redundant features can cause overfitting. Thus, effective material descriptors typically meet several requirements:[55] completeness, uniqueness, descriptiveness and efficiency.

Molecular structures are commonly represented as string notations, which are human-readable and machine-friendly, facilitating subsequent processing. The Simplified Molecular Input Line Entry System (SMILES)[56] is the most widely used format, which encodes molecular structures as sequences of characters denoting atoms, bonds, and substructures. Several extensions have been developed to address limitations of standard SMILES. For example, Self-referencing Embedded Strings (SELFIES)[57] resolve issues related to invalid or physically unrealistic molecular representations inherent in SMILES. In polymer representation, SMILES employs asterisks (*) to indicate the connection of repeating units, reflecting polymers' repetitive macromolecular structure. BigSMILES,[58] a polymer-specific variant, introduces additional syntax to accurately capture polymer features such as branching and copolymerization.

Fingerprinting intrinsically depends on the application context, capturing essential attributes of materials through either handcrafted or automatically generated methods. Hand-crafted descriptors are constructed based on explicit domain knowledge and predefined rules. For polymer systems, these include macroscopic properties such as molecular weight distributions, crystallinity, and chain morphology, as well as chemical group contributions[59] and fragment-based fingerprints. Traditional fingerprints encode structural features as binary vectors indicating the presence or absence of substructures, which effectively correlate polymer structures with thermal properties such as T_g and TC. Each bit in a fingerprint corresponds to a predefined fragment, facilitating efficient comparison and screening of large datasets. Common fingerprints include Morgan fingerprints (Extended Connectivity Fingerprints, ECFP),[60] Molecular Access System (MACCS) keys, Topological Torsion fingerprints,[61] and Daylight fingerprints.

Similarly, handcrafted descriptors for TBC include physical descriptors, such as acoustic velocities, Debye temperatures and densities; chemical descriptors, such as interfacial binding energy and lattice mismatch; and process descriptors like film thickness and roughness.[62] Spectral phonon descriptors such as dominant phonon frequencies, group velocities and relaxation times allow ML models to utilize classical phonon transport relations while maintaining interpretability under data-scarce conditions.[11,63] At the atomic level, descriptors like Smooth Overlap of Atomic Positions (SOAP) and radial distribution functions (RDF) capture local atomic environments and spatial correlations crucial for phonon transport modeling. Integrating

these physics-guided features introduces an inductive bias, substantially improving model extrapolation to unseen material pairs.

With recent advancements in deep learning, there is a shift toward automatically learned descriptors, which dynamically extract representations directly from data. Two prevalent methods are graph-based or language-based approaches. Graph-based representations exploit molecular topology by modeling molecules as graphs, with atoms as nodes and bonds as edges. GNNs[64] leverage graph structures to learn hierarchical and spatially aware features, implicitly capturing complex interactions critical for predicting thermal transport properties and generating novel structures. Language-based embeddings, inspired by natural language processing, treat SMILES strings as input sequences. Models such as PE[51] and transformer-based architectures directly encode chemical semantics and structural context from data, eliminating the need for manually defined descriptors. Recent models like polyBERT[65] and TransPolymer[66] demonstrate comparable or superior accuracy compared to classical descriptors (e.g., ECFP, Polymer Genome[67]) in polymer property predictions. Furthermore, Guo et al.[68] introduced GraSeq, a joint graph and sequence representation learning model which outperformed single-modality models across multiple molecular property prediction benchmarks, showing the complementary nature of information extracted from these two representation strategies. Nonetheless, handcrafted descriptors remain competitive, particularly when datasets are small or computational resources limited,[69] and the relative advantage of automatically learned descriptors depends on the specific dataset and task.

Feature selection is also crucial for eliminating irrelevant or redundant features to reduce complexity and dimensionality in structure-property relationship modeling.[70] This process can be automated using various strategies. For instance, the Least Absolute Shrinkage and Selection Operator (LASSO)[71] automatically eliminates irrelevant functional group signals through ℓ_1-regularization or symbolic regression, reducing overfitting in high-dimensional fingerprint scenarios. The Sure Independence Screening and Sparsifying Operator (SISSO)[72] constructs sparse, physically meaningful descriptors from large candidate pools. Moreover, encoder architectures such as multilayer perceptron (MLP) and autoencoders can automatically capture

salient chemical and structural patterns, performing dimensionality reduction or feature selection by encoding inputs into a low-dimensional latent vectors. These techniques help identify and retain relevant features, enhancing the efficiency and effectiveness of ML models.

4 Applications of ML in Nanoscale Thermal Transport

4.1 ML for Thermal Property Prediction

4.1.1 Polymer properties prediction

Early efforts applied ML to individual polymer chains to understand intrinsic thermal transport properties. For example, Zhu et al.[73] applied kernel ridge regression (KRR), a feed-forward ANN, and a CNN to predict the lattice TC of diverse single polymer chains, which shows that CNN yielded the highest prediction accuracy with a MAE of 5.20 W/mK, RMSE of 6.83 W/mK. In a recent high-throughput study, Huang et al.[11] developed an interpretable ML framework for exploring high TC polymer chains. They selected twenty optimized descriptors from 320 physics-based descriptors (e.g., bonding, rigidity metrics) and enabled tree-based models and an MLP to achieve $R^2 > 0.80$ in TC prediction. The model revealed that π-conjugation and stiff backbones lead to higher single-chain TC due to high phonon group velocities. These findings agree well with previous MD simulations of aligned polymer chains,[74] showing that ML can correctly reveal certain structure-property relationships.

In nanoscale thermal transport, T_g is an essential polymer properties when studying bulk homopolymer. Luo and co-workers[52] compared different polymer representations for T_g prediction across different ML algorithms and found that the learned continuous "molecular embeddings" gave the best T_g correlation ($R^2 \approx 0.865$), outperforming traditional Morgan fingerprints and molecular graph. Meanwhile, deep learning has shown promise in polymer thermal property modeling. For example, Miccio and Schwartz[77] trained a CNN on images of polymer repeat units to predict T_g, achieving 6% average error. Likewise, Park et al.[78] introduced a graph convolutional network (GCN) model to predict multiple thermal properties, including T_g and melting temperature (T_m). Their GCN attained high accuracy for T_g on a broad dataset, slightly surpassing the performance of conventional circular fingerprints. Tao et al.[79] proposed comprehensive benchmarks through

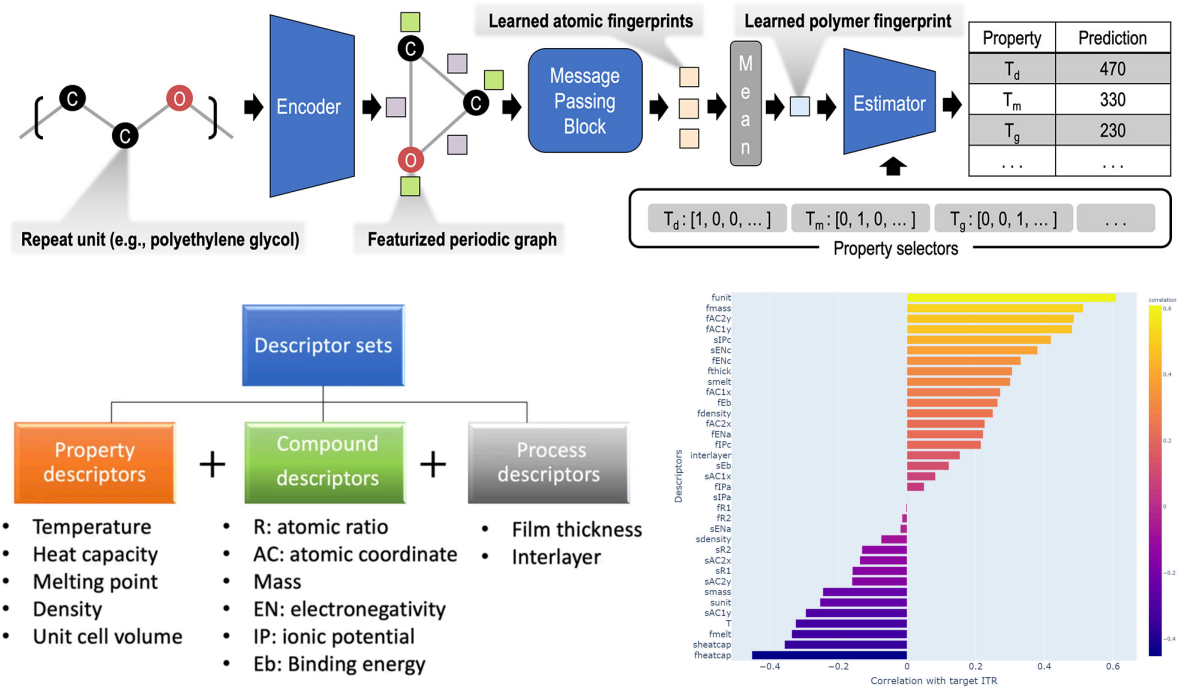

Fig. 4 (a) PolyGNN architecture. The Encoder converts polymer SMILES into periodic graphs, generating atomic (green) and bond (purple) fingerprints. The Message Passing Block further refines atomic fingerprints (yellow), which are averaged into polymer-level fingerprints (blue) and used by the Estimator to predict thermal properties (e.g., T_d, T_m, T_g). (Reprinted from Gurnani et al.[75] under CC-BY 4.0.) (b) List of the three descriptor sets: property descriptors, compound descriptors, and process descriptors (Reproduced from Wu et al.[14] under CC BY 4.0.) (c) Pearson correlation coefficient map between all 35 descriptors and the target ITR in a training dataset. (Reproduced from Chen et al.[76] under CC BY 4.0.)

evaluating 79 models (varying representations and algorithms) for T_g prediction. Their study confirmed that both advanced deep learning, such as CNNs and DNNs, and traditional models (RFs, SVMs, etc.) can achieve strong performance when appropriately tuned.

To further boost generalization, multitask learning has been applied. Gurnani et al.[75] introduced multi-task GNNs for predicting 34 polymer properties, as shown in Figure 4a. Their results showed that when data for a given property are limited, this PolyGNN model exceeded the accuracy of single-property models. Another notable trend in thermal property prediction is combining simulations with ML. Ma et al.[80] computed the TC of 365 amorphous polymers based on high-throughput MD, then trained an ML regression model to quantify the structural-TC relation. They identified 122 polymers with predicted TC $>$ 0.3 W/mK, demonstrating how ML can guide discovery of thermally conductive polymer.

Copolymers introduce additional complexity in thermal property prediction due to their various types, such as alternating, random, and block copolymers. Recent work has extended polymer informatics beyond

homopolymers to tackle this challenge. Kuenneth et al.[81] reported a multi-task DNN for copolymer informatics that incorporated meta-learning techniques. They curated over 18,000 data points of T_g, T_m, and thermal degradation temperature (T_d) for polymers composed of one or two monomers and trained the model which could accurately predict all three thermal properties for both homopolymers and binary random copolymers. This approach achieved screening of copolymer to identify optimal T_g/T_m combinations. Besides, Tao et al.[82] explored multiple ML strategies for copolymer property prediction. They formulated four model architectures - a standard feed-forward neural network (FFNN), a CNN, an recurrent neural networks (RNN), and a combined FFNN/RNN (fusion) mode, and evaluated them across four different datasets. The RNN architecture, which allows the sequence distribution to be processed both forward and backward, is found to be the best-suited model for copolymers with good generalization ability. A more recent study by Queen et al.[83] introduced PolymerGNN for polyesters, a GNN-based multitask model capable of handling copolymer compositions and branching. Each polymer (linear or branched, homopolymer or co-polyester) is represented by its constituent monomer graphs, and the network jointly learns multiple targets like T_g and inherent viscosity.

4.1.2 Thermal boundary conductance modeling

TBC is critical in thermal transport across heterogeneous interfaces in nanostructured materials, composite systems, and electronic devices. However, accurately modeling TBC is difficult due to its dependence on a complex interplay of atomic structure, interfacial chemistry, and phonon mismatch. ML offers a promising approach to address these challenges by enabling data-driven prediction and optimization of interfacial thermal resistance (ITR) and TBC across diverse material systems. Recent studies have integrated experimental data, MD simulations, and first-principles calculations with ML models to accelerate interface screening and provide interpretable insights into thermal transport mechanisms.

Wu et al.[14] developed ML models to predict ITR based on experimental data using three physical, chemical, and material descriptors. A list of the three descriptor sets is shown in Figure 4b. Their models achieved a predictive performance of 96%, using inputs from over 80,000 material systems composed of 293 materials. Among the top 100 high-ITR predictions, 25 material systems were consistently identified

by at least two algorithms, including Bi/Si, which showed ultra-low TC. Building on this, Chen et al.[76] employed ML and deep learning algorithms to predict ITR by addressing its high-dimensional nature. After exploratory data analysis, they used XGBoost to identify the most significant descriptors and built concise predictive models with XGBoost, KRR, and DNNs. An ensemble learning approach combined these models to predict high melting points and high-ITR material systems relevant for spacecraft, automotive, and building insulation applications. Pearson correlation coefficient map between all 35 descriptors and the target ITR in a training dataset is shown in Figure 4c. Complementing these data-driven methods, Rustam et al.[84] presented a scalable Bayesian optimization framework to enhance TBC by dynamically running parallel MD simulations via Message Passing Interface (MPI). They optimized heat transfer at the silicon/aluminum (Si/Al) interface, achieving up to a 50% increase in TBC through a two-layer intermixed region with a higher silicon concentration. The study highlighted that the randomness of intermixing and the stochastic nature of MD contribute to the variance in TBC. Jin et al.[85] combined MD simulations with ML to optimize interfacial thermal transport in Si/Ge heterostructures through interfacial nanostructuring. They introduced three structural parameters to describe nanostructures and demonstrated that ITR exhibits nonmonotonic dependencies on nanostructure height, density, and angle. The study revealed that optimal heat dissipation occurs at specific structural configurations, with minima gradually disappearing as nanostructure angles increase.

In a broader scope, Foss et al.[86] used first-principles phonon dispersion calculations combined with a 2D-3D Boltzmann transport model to compute the TBC of 156 unique 2D/3D interface pairs. They developed ML models, including neural networks and Gaussian processes, achieving high predictive accuracy with RMSE < 5 MW/m^{2}K and $R^2 > 0.99$. Sensitivity analysis identified key descriptors impacting TBC, and decision-tree models demonstrated good transferability to unseen materials. Anandakrishnan et al.[87] proposed a data-driven method to model TBR at nanoscale solid-liquid interfaces using macroscopic observables. Using heuristic algorithms, they correlated TBR with thermodynamic state variables, material properties, and geometric parameters, deriving generalized predictive relationships. Their analysis showed that interfacial liquid layering strongly correlates with TBR, while work on adhesion and system geometry also becomes

significant under phonon size effects and extreme conditions. Together, these studies demonstrate how ML enhances interfacial thermal transport's prediction, understanding, and optimization.

4.2 ML for Materials Design and Optimization

4.2.1 Design of Polymers using ML

Forward screening and inverse design are two of the most prominent supervised learning tasks for polymer design. Early efforts in polymer design relied on high-throughput virtual screening (HTVS), which explores large, predefined libraries using ML predictors or high-throughput simulations (e.g., DFT or MD). These screenings are guided by expert intuition, with iterative feedback between theory and experiment refining the candidate pool. For example, Li and co-workers[88] trained a DNN model to screen one million hypothetical polymers, identifying over 65,000 candidates with T_g greater than 200 °C, showing the efficiency of in expanding high T_g polymer discovery. They also successfully identified multi-functional polyimides with high T_g and tensile strength through high-throughput screening combined with explainable ML models.[89] Huang et al.[90] proposed a hybrid workflow integrating physical feature engineering with automated MD simulations to explore amorphous polymers with high TC. Based on trained ML models using 25 optimized descriptors, 104 high TC candidates were identified out of over 790,000 polymers, over half exceeding 0.35 W/mK. The study suggested that the conjugated molecular structure, chain stiffness, and maximum relative atomic mass in polymers have a strong influence on TC. These mostly agree with physics-based MD simulations.[91]

Unlike forward screening, inverse polymer design enables the direct, on-demand generation of candidate structures tailored to target properties or performance requirements. Typically, candidate structures are iteratively generated by applying sequence perturbations or interpolations within learned latent spaces (generator), then evaluated by property predictors, until structures meet the desired properties (Figure 5a). Crucially, additional filtering steps, such as synthetic accessibility (SA) scores or retrosynthesis rankings, can ensure chemical validity of generated polymers. Recent advances in generative models, such as RNNs, variational autoencoders (VAEs), RL, and GANs, have enabled inverse design of polymers with target properties. For instance, Batra et al.[93] coupled the unsupervised syntax-directed VAE with the supervised GPR method to identify polymers with high T_g and large band gap. They first encode known polymers meeting

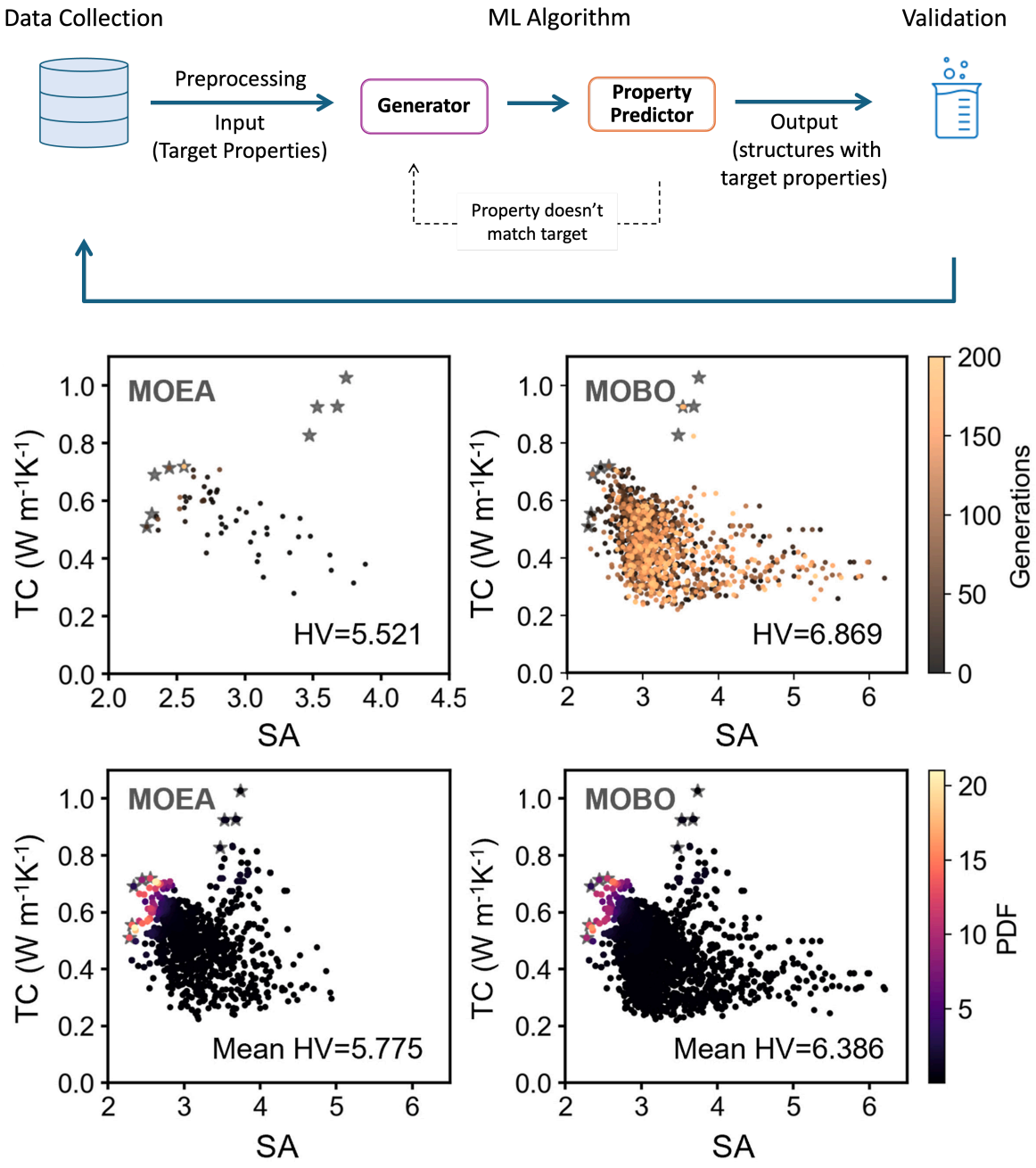

Fig. 5 (a) General ML workflow for inverse design. (b, c) Optimization trajectories from single runs of multi-objective evolutionary algorithm (MOEA) and multi-objective Bayesian optimization (MOBO). (d, e) Probability density maps in objective space based on 20 independent runs of MOEA and MOBO, respectively. (Reproduced with permission from Huang et al.[92] Copyright 2024 Elsevier)

design criteria to identify promising regions in the latent space, then interpolate within these regions using GPR predictions to select latent points, from which the decoder generates new polymer SMILES. Hundreds of polymers achieving the target properties were successfully generated. Another good example is that Ma

et al.[13] employed RL framework based on RNN and MD-labeled TC data to discover amorphous polymers with high predicted TC above 0.4 W/mK, with the top candidate reaching 0.69 W/mK after MD validation. Liu et al.[94] employed an invertible graph generative model to generate hypothetical polymers with promising properties, particularly focusing on high-temperature polymer dielectrics. Moreover, a recent benchmark compared six deep generative models on real and hypothetical polymer datasets, identifying CharRNN, REINVENT and GraphINVENT as the most effective for producing valid, diverse structures.[95] These three models were successfully further trained on real polymers using RL methods, targeting the generation of hypothetical high-temperature polymers for extreme environments.

4.2.2 Search-based Optimization

Beyond predictive analytics, ML plays a key role in structural optimization for thermal transport, using methods like simulated annealing, particle swarm optimization, genetic algorithms (GAs), Bayesian optimization (BO), and Monte Carlo tree search (MCTS) to identify structures with desired thermal properties. Global Optimization like GAs have been widely applied in polymer design as a template-based generative approach. GAs iteratively evolve polymer structures to optimize desired properties by mimicking natural selection processes. For example, Pilania et al.[96] integrated the ML surrogate model with a GA, efficiently identifying multicomponent polymer compositions with a prespecified T_g. Kim et al.[97] employed a GA guided by ML property predictors to design polymers that simultaneously exhibit ultra-high T_g (>500 K) and wide bandgaps (>6 eV), demonstrating GA's ability to find polymers beyond known property limits. Similarly, Zhou et al.[98] combined MD with a GA to optimize the specific sequence of polyethylene-polypropylene (PE-PP) copolymers for maximum TC, showing that the monomer sequence has a crucial effect on thermal energy transport of the copolymers. Beyond GAs, other global search algorithms have also emerged. For instance, Nagoya et al.[99] applied a MCTS with rapid Green-Kubo MD evaluations to autonomously explore new amorphous polyimides. After screening 1,000 evaluations, their MCTS approach found an optimal polymer sequence with TC of 0.25 W/mK.

BO has also been applied to both single- and multi-objective optimizations for discovering desirable polymers. In BO, each iteration fits an updated ML model, such as GPR, to predict the target property

and its uncertainty, then selects the next candidate by maximizing an acquisition function (e.g., expected improvement). This closed-loop "AL" approach is well-suited for expensive evaluations. In thermal transport material design, optimization is often needed to balance multiple competing objectives simultaneously. Multi-objective EAs (MOEAs) and multi-objective BO (MOBO) techniques have been applied to generate Pareto-optimal sets of polymer candidates that balance such trade-offs. Huang et al.[92] developed a framework combining a deep-learning surrogate model with two advanced multi-objective optimizers to design sequence-defined polymers with high intrinsic TC and easy synthesizability. In their workflow, a unified NSGA-III (U-NSGA-III) genetic algorithm and a qNEHVI (q-noisy expected hypervolume improvement) Bayesian optimization were used in parallel to explore polymer sequence space (Figure 5b-c). They found that the GA-based U-NSGA-III quickly generated candidate solutions but sometimes converged to local optima, whereas the BO-based qNEHVI method can identify a diverse set of Pareto-optimal polymers (Figure 5d-e).

4.3 ML Interatomic Potentials for Thermal Transport

4.3.1 High-fidelity ML interatomic potentials

Machine learning interatomic potentials (MLIPs) have emerged as powerful surrogates for quantum mechanical simulations, delivering near-DFT accuracy at a fraction of the computational cost.[104–107] ML potentials have significantly improved atomistic simulations by offering a flexible, data-driven alternative to empirical interatomic potentials. Traditional models, such as Lennard-Jones or EAM potentials, rely on fixed analytical expressions that often fail to capture the full complexity of many-body interactions. In contrast, MLIPs learn these interactions directly from high-fidelity data, typically obtained from DFT calculations, leading to notable improvements in accuracy and transferability.

Among the foundational MLIP models, Behler-Parrinello Neural Network Potentials (BPNNPs)[108] utilize high-dimensional descriptors and deep learning architectures to capture complex many-body effects. Deep Potential Molecular Dynamics (DPMD) by Zhang et al.[109] extends NNPs with hierarchical representations to improve accuracy across diverse chemical environments. Kernel-based models, such as the Gaussian Approximation Potential (GAP)[110] and Moment Tensor Potentials (MTPs)[111], leverage symmetry-aware descriptors to enhance accuracy and transferability. For instance, Deringer et al.[112] demonstrated that ML

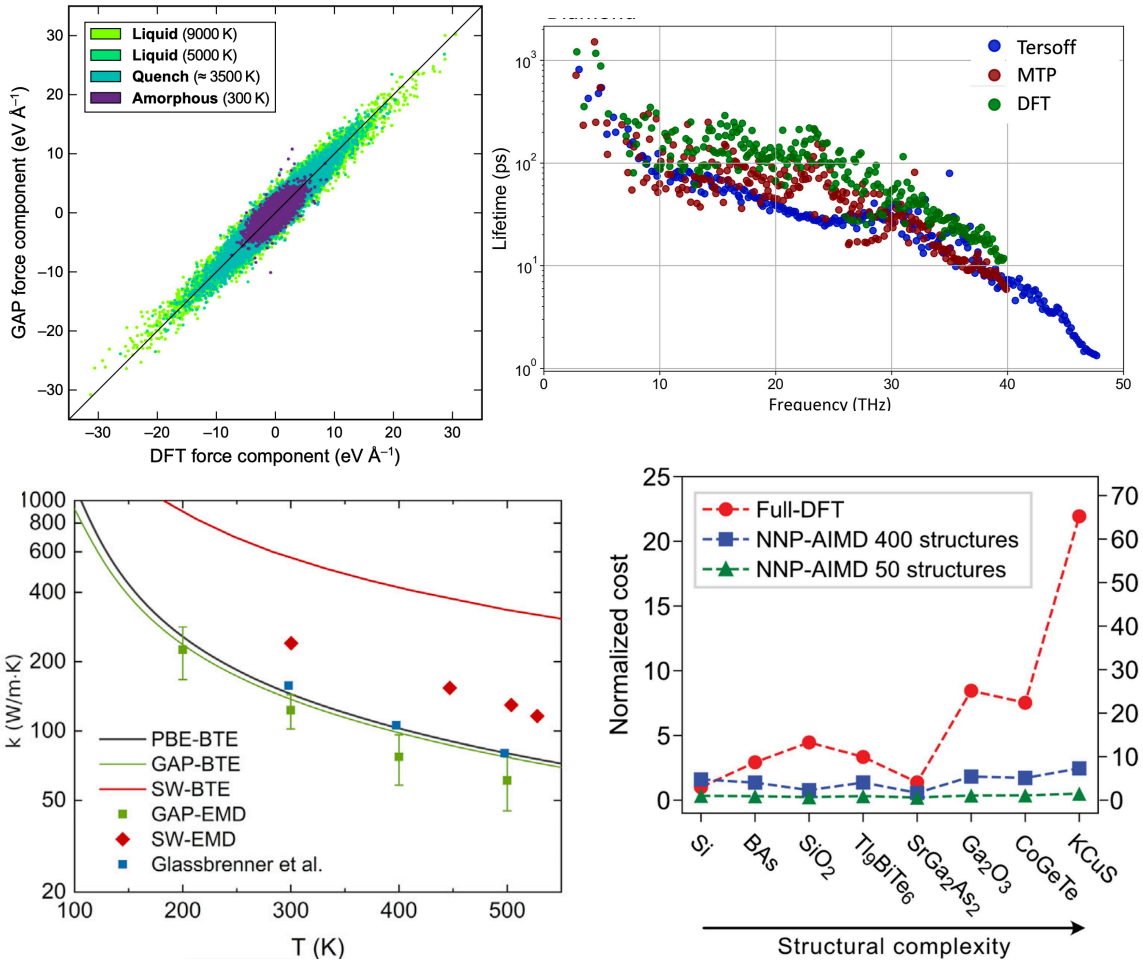

Fig. 6 (a) DFT-computed versus GAP-predicted force components in a set of 125-atom snapshots of liquid and amorphous carbon, emphasizing the GAP's overall magnitude of forces to learn at various parts of the melt-quench trajectories. (Reprinted with permission from Deringer et al.[100] Copyright 2017 by the American Physical Society) (b) Phonon lifetime calculated with MTP and Tersoff potential and compared with DFT calculations for bulk diamond calculated at 300 K. (Reprinted with permission from Rajabpour et al.[101] Copyright 2025 Elsevier Masson SAS) (c) TC of crystalline silicon derived from the GAP model by EMD and BTE, compared with other works. (Reprinted with permission from Qian et al.[102] Copyright 2019 Elsevier Ltd) (d) The computational cost of calculating κ_l by DFT and NNPs trained by 400 or 50 structures. The computational cost is normalized with respect to Si computed within the fully DFT approach (left scale) or the NNP trained with 50 structures (right scale). (Reprinted with permission from Choi et al.[103] Copyright 2022 Elsevier Ltd)

potentials can achieve orders of magnitude speed increases compared to conventional DFT while accurately modeling phase transitions, defect dynamics, and catalytic behavior. Despite these advancements, the predictive performance of MLIPs is fundamentally limited by the quality and diversity of the training data, which can hinder their ability to generalize to new chemical environments.

In the study of thermal transport, MLIPs facilitate large-scale and long-timescale simulations that would be computationally prohibitive with first-principles methods. When trained on high-fidelity datasets, MLIPs

can accurately reproduce atomic interactions and phonon dynamics, which are critical for modeling heat transfer in both crystalline and disordered materials.[113] A pioneering application of GAP is the work by Deringer et al.,[100] who applied it to simulate liquid and amorphous elemental carbon. This model accurately captured energetic and structural properties across a wide range of densities using a hierarchical set of two-, three-, and many-body structural descriptors, outperforming state-of-the-art empirical potentials. Applications to diamondlike amorphous carbon surfaces further demonstrated the model's ability to simulate surface energies and high-temperature graphitization processes. As shown in Figure 6a, comparing DFT-computed and GAP-predicted force components for 125-atom liquid and amorphous carbon snapshots illustrates the range and magnitude of forces the model learns across melt-quench trajectories. Similarly, Rybin et al.[114] developed MTPs for atomistic simulations of the α and β phases of Ga_2O_3, achieving excellent agreement with DFT-calculated phonon dispersion and lattice thermal conductivity. Their study also highlighted the importance of active learning in generating robust and accurate potentials with a moderate training dataset.

More recent developments have expanded the MLIP landscape to include Neuro-evolution Machine Learning Potentials (NEPs),[115] which employ evolutionary algorithms to optimize both structural descriptors and neural network architectures. This approach achieves high accuracy while maintaining computational efficiency and has been successfully applied to directly calculate per-atom heat currents, enabling efficient thermal transport modeling. In parallel, universal pre-trained MLIPs such as CHGNet[116] and MACE[117] have been trained on chemically diverse datasets, allowing fine-tuning for specific materials with minimal additional data. These models reduce the cost of developing new potentials and are particularly useful in high-throughput screening for thermal property discovery. Furthermore, equivariant message-passing architectures like NequIP[118] encode rotational and translational symmetries, improving accuracy in anisotropic systems, while active-learning frameworks such as DP-GEN[119] iteratively expand the training set to cover unexplored yet physically relevant regions of the configuration space. Together, these advances strengthen the applicability of MLIPs in capturing complex atomic-scale processes critical for accurate thermal transport predictions.

To provide a broader perspective, Wang et al.[120] reviewed the current state of MLIPs, highlighting their role in bridging the gap between computationally expensive DFT and less accurate classical MD. They discussed four essential stages of MLIP development: data generation, structural descriptors, machine learning algorithms, and available software tools. Applications of MLIPs were examined across areas such as phase-change memory materials, structure searching, property prediction, and pre-trained universal models. The review also addressed future perspectives, emphasizing the importance of standard datasets, transferability, generalization, and balancing accuracy with model complexity.

4.3.2 Applications to nanoscale systems

MLIPs are particularly valuable for modeling thermal transport in nanoscale systems, where atomic-level accuracy is essential and traditional empirical potentials often fail to capture complex interfacial phenomena. Mortazavi et al.[121] demonstrated that MLIPs trained on short *ab initio* MD trajectories can enable first-principles multiscale modeling of lattice TC. Using graphene/borophene heterostructures as a case study, they bridged DFT, classical MD, and finite element method (FEM) simulations to predict thermal properties at both atomic and continuum scales efficiently. Their MLIP-based approach accurately captured the lattice TC of pristine phases and thermal conductance across interfaces. Expanding on interfacial modeling, Wyant et al.[122] developed MLIPs for modeling interfacial heat transport in Ge/GaAs systems, addressing limitations of traditional analytic potentials and mixing rules. They assessed *ab initio* harmonic force constants (IFC2s) near interfaces, showing convergence to bulk-like values within 1 nm while highlighting complex interface behavior. Two MLIPs based on the linear spectral neighborhood analysis potential (SNAP) were constructed: a standard SNAP fit to total forces and a hybrid SNAP combined with a harmonic Taylor expansion. Each potential was evaluated for bulk thermal properties, interface behavior, and stability, guiding future modeling of interfacial thermal transport.

Further demonstrating the advantages of MLIPs in interface modeling, Rajabpour et al.[101] calculated the TBC between silicon and diamond using an MLIP trained on DFT data. Non-equilibrium MD (NEMD) simulations showed that the ML potential produced TBC values in much closer agreement with experiments than traditional semi-empirical potentials like Tersoff[123] and Brenner[124]. The ML model accurately captured

phonon dispersion relations and lifetimes, providing insights into the frequency-dependent heat transfer spectrum. Phonon lifetimes calculated using the MTP and Tersoff potential, compared against DFT results for bulk diamond, are shown in Figure 6b, highlighting the improved agreement achieved with ML-based models.

4.3.3 Accelerating thermal transport calculations

High-fidelity simulation of thermal transport properties, especially lattice TC, traditionally relies on DFT combined with phonon transport calculations. However, these methods are computationally expensive, especially for large or disordered systems. MLIPs provide a promising alternative, offering significant speedups while retaining near-DFT accuracy. Qian et al.[102] demonstrated this potential by developing MLIPs trained on DFT data to model the TC of both crystalline and amorphous silicon. By stochastically sampling the potential energy surface, they trained potentials that enabled equilibrium MD simulations at scales beyond those achievable by first-principles methods. The predicted thermal conductivities agreed with experimental measurements for both material forms. Compared with other studies, the TC of crystalline silicon derived from the GAP model using EMD and the BTE is shown in Figure 6c.

Expanding the applicability across a broader material spectrum, Choi et al.[103] employed neural network interatomic potentials to compute LTCs at 300 K for 25 materials with diverse structural symmetries and thermal conductivities spanning three orders of magnitude (10^{-1} to 10^{3} W/mK). They found that training on MD trajectories between 50-700 K consistently achieved near DFT-level accuracy while providing a uniform computational cost and a speed gain of 2-10 times compared to pure DFT methods and using a reduced training set further increased efficiency by up to 50 times without significant loss of accuracy. The computational cost of calculating κ_l using DFT and NNPs trained on 400 and 50 structures is illustrated in Figure 6d.

Further broadening the application of MLIPs to both bulk and interfacial systems, Li et al.[125–127] developed deep learning-based interatomic potentials trained on *ab initio* data to model thermal transport across silicon (in crystalline, liquid, and amorphous phases), β-Ga_2O_3 (with anisotropic phonon transport), and GaN/SiC interfaces with AlN transition layers. Their MLIPs accurately captured phonon dispersions, phase

transitions, and TCs in good agreement with experiments. Notably, their GaN/SiC interface study revealed that AlN transition layers enhance TBC via phonon bridging and improved crystalline quality. Collectively, these works demonstrate the versatility of MLIPs in capturing complex nanoscale thermal transport phenomena across diverse materials and interfaces. These studies demonstrate how MLIPs are accelerating thermal transport modeling by enabling accurate, efficient, and scalable simulations across crystalline, disordered, and interfacial systems. As ML techniques mature, they are poised to become indispensable tools for materials discovery and design in thermal management and energy technologies.[120]

4.4 ML for Thermal Data Analysis

4.4.1 Data interpretability

While ML models for thermal transport prediction have achieved remarkable accuracy, their widespread adoption in materials science is often hindered by many algorithms' "black-box" nature. This opacity challenges scientific understanding and rational materials design, underscoring the need for interpretable approaches. Explainable ML (XML) techniques aim to identify which structural, chemical, or physical features most strongly influence predictions, thereby enhancing trust, guiding design strategies, and revealing new physical insights. Several methods have been adopted to address this challenge. Techniques such as feature importance analysis, Shapley Additive exPlanations (SHAP), and symbolic regression are increasingly applied in thermal transport research to bridge the gap between predictive performance and interpretability.[128] Attention mechanisms in neural networks enhance model transparency by highlighting influential features within input structures. Feature importance analysis, including SHAP, quantifies the contribution of individual input features (e.g., bond types, functional groups, or atomic configurations) to the model's prediction. This enables researchers to identify which physical or chemical attributes are key drivers of thermal transport behavior, facilitating rational design of materials with targeted properties.[129] Symbolic regression, on the other hand, can derive human-readable mathematical expressions that describe complex structure-property relationships, allowing domain experts to uncover underlying physical principles from data-driven models.[130] Attention mechanisms in GNNs further help pinpoint substructures or motifs within polymer chains or crystalline lattices that dominate transport behavior.[131]

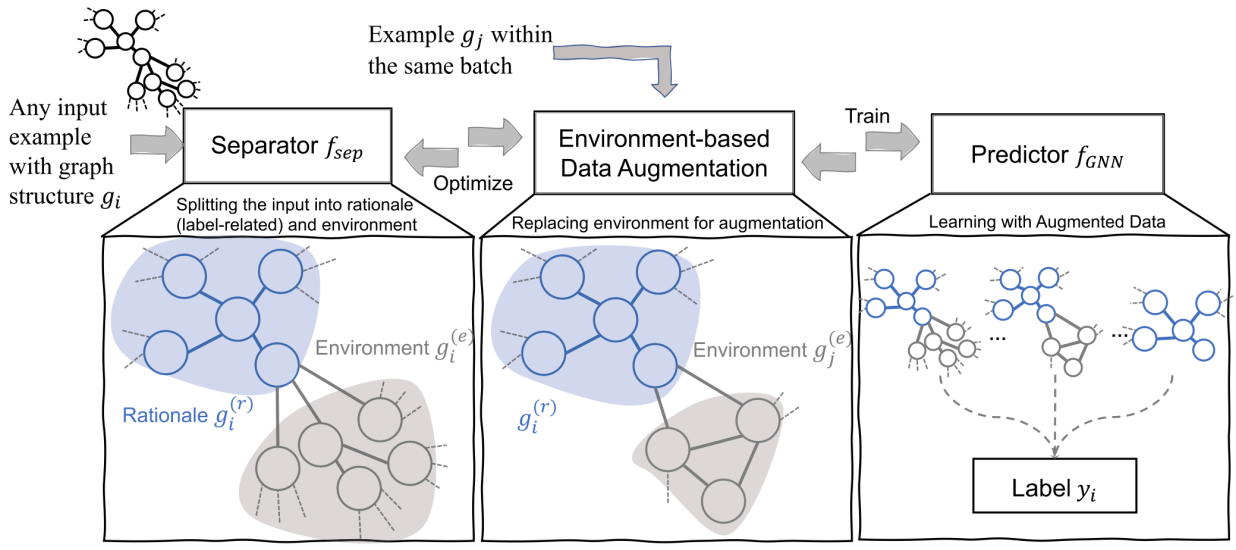

Fig. 7 An overview of the RGDA (Rationalize GNN predictions with Data Augmentation) framework. (Reprinted with permission from Liu et al.[132] Copyright 2024 by the Association for Computing Machinery)

For example, Liu et al.[132,133] developed a unified framework (as shown in Figure 7), RGDA (Rationalize GNN predictions with Data Augmentation), which rationalizes GNN predictions for both node-level and graph-level tasks by identifying key environment-based subgraph structures. Their method incorporates environment-based data augmentation to enhance rationale extraction and bypasses costly graph decoding by learning representations in latent space. This allows researchers to trace predictions back to chemically meaningful substructures, thereby improving trust and model transparency. This strategy enhances graph-based models' interpretability and generalization capabilities for materials science applications.

To integrate interpretability and deep learning, Zeng et al.[134] introduced an interpretable deep learning framework for rapid and accurate prediction of lattice TC, addressing the trade-off between model accuracy and interpretability. Using this framework, they identified and validated four promising thermal conductors and insulators through DFT and MD simulations. The interpretable model helped in screening candidate materials by revealing which descriptors (e.g., atomic mass, bonding environment, or symmetry factors) significantly influenced thermal conductivity predictions. Sensitivity analysis and DFT calculations provided novel insights into phonon thermal transport mechanisms.

Other studies have applied SHAP-based feature importance analysis on experimental and hybrid datasets to uncover dominant physicochemical drivers behind TC, demonstrating the utility of explainable ML in

diverse material systems.[135,136] For instance, SHAP analysis revealed that atomic density, chain flexibility, and bonding motifs were the most influential features across different polymer families, directly informing design strategies for high- or low-conductivity materials.

4.4.2 Physics-informed ML

Recent advancements in PINNs can also enhance thermal data analysis, particularly through the integration of physical principles into predictive models. As PINNs embed governing equations directly into the learning process, they can be used to improve predictions from datasets, especially when dealing with sparse and noisy datasets. For instance, Gokhale et al.[137] proposed PINN-based models that integrate measured data with underlying physics to perform thermal modeling for buildings. The models demonstrate more data-efficient and accurate predictions for control-oriented thermal modeling. Apart from solving forward problems, PINNs also demonstrate the advantage in handling inverse problems when data are noisy and sparse. Li et al.[138] integrated Bayesian PINNs with nonparametric variational inference to model phonon BTE with uncertainty quantification. The model can not only recover temperature and heat flux distributions from noisy data, but infer parameters such as the Knudsen number. Liao et al.[139] developed a hybrid thermal modeling approach for additive manufacturing processes, combining partial temperature observations with physical laws to predict full-field temperature histories and identify unknown material and process parameters. Similarly, Sripada et al.[140] utilized PINNs to robustly extract thermal properties from noisy data obtained by the laser-based Angstrom method, showing the method's ability in performing inverse parameter fitting under experimental uncertainties. These studies collectively underscore the versatility and robustness of PINNs in thermal data analysis.

4.4.3 ML for analyzing experimental and simulation thermal data

Traditional thermal metrology techniques, such as time- or frequency-domain thermoreflectance (TDTR/FDTR)[141,142] and 3ω method[143,144], produce complex, multi-variable datasets that encode properties like TC, heat capacity, and TBC. Traditional nonlinear fitting methods for property extraction can be laborious and sensitive to initial parameter guesses and noise, especially when multiple parameters are involved. ML offers a great alternative for thermal transport data analysis, by learning direct mappings from

experimental data to material properties, thereby improving the robustness of analysis. For example, one study[145] proposed to apply deep learning approach to nanosecond transient thermoreflectance technique for high-throughput experimental data processing. Compared to the conventional non-linear fitting method (such as Global Optimization), the computation time of the new model is 1,000 times lower. Additionally, ML models can be designed to output multiple target parameters at once, allowing simultaneous determination of TC, TBC, and volumetric heat capacity (ρC_p) from a single experiment, something that is difficult for standard fitting due to parameter correlations. Xiang et al.[146] used KRR-based reconstruction method for extracting depth-dependent TC directly from pump-probe phase signals, without requiring pre-knowledge about the functional form of the profile. Moreover, hybrid simulation-experiment modeling has emerged as an effective strategy for thermal data analysis. Shen et al.[147] proposed a novel data analysis method to build DL-FDTR, which can predict TC, ρC_p and TBC with mean errors below 5% for bulk samples coated with Au and supply an initial guess for subsequent least-square fitting. Similarly, Pang et al.[148] generated over 10,000 synthetic TDTR curves to train an ANN. Without further retraining, their model can directly extract TC and TBC from experimental TDTR data collected from various samples under different modulation frequencies and laser spot sizes.

5 Perspectives

5.1 Enlarging databases via automated data generation

Developing accurate ML models for nanoscale thermal transport requires much larger and more diverse datasets than are currently available. One promising approach is sustainable data acquisition from the literature. Advanced NLP methods and large language models (LLMs), such as BERT-based models and ChemDataExtractor[149], can be leveraged to mine published papers for thermal properties of polymers, nanocomposites, and other materials. Such tools can automatically extract values of TC, interfacial conductance, heat capacity, etc., but they face challenges like polymer name normalization (many synonyms for the same polymer) and the lack of standardized dictionaries for thermal terms. Domain-specific solutions are needed, for example, polymer-focused NER taggers and ontologies , and future systems must move beyond

text, extracting data from figures and tables when critical information (like molecular structures or data points) is only present in graphical form.[150]

Another strategy is data augmentation, which can multiply the effective size of training sets. For polymer-based thermal predictions, graph-based augmentations are especially useful. Examples include node reordering (to augment graph representations without altering the polymer),[151] repeat-unit expansion (simulating longer chain segments or varying chain lengths),[75] and substructure recombination (mixing key "rationale" motifs of a polymer with different chemical environments)[132] can generate new virtual samples from existing ones. By judiciously applying such augmentations - tailored to preserve relevant thermal transport physics - researchers can combat data scarcity and improve model robustness without requiring entirely new experiments.

5.2 Integrating Physics and ML

The integration of physics and advanced ML techniques is rapidly reshaping the landscape of thermal transport research, particularly at micro- and nanoscale regimes where classical diffusion models fail. By embedding physical laws (e.g., the phonon BTE) into ML architectures, researchers can overcome data scarcity, improve prediction accuracy, and ensure physically consistent solutions. Unlike traditional black-box ML modeling, integrating physics with ML can increase model interpretability.

Recent developments in neural operators[152,153] and conditioning neural fields[154,155] provide powerful frameworks for learning mappings in function spaces of PDE-governed systems. They can be used for high-resolution predictions from sparse and noisy experimental measurements. And the conditioning allows models to generalize well across geometries and boundary conditions. Furthermore, the fast developments of foundation models and LLMs[156,157] present new opportunities for incorporating prior physical knowledge, symbolic reasoning, and multimodal data fusion into thermal modeling. As the architecture of ML models continues to evolve, the coupling with domain-specific knowledge offers a promising direction towards data-efficient, interpretable, and robust modeling of non-diffusive thermal effects, paving the way for breakthroughs in thermal metrology, material design, and microelectronic thermal management.

5.3 Generative and foundation models for thermal transport

Generative machine-learning models, ranging from graph- and sequence-based generators to large pre-trained "foundation" models (e.g., transformer LLMs), offer new routes to propose novel nanoscale materials with tailored thermal properties. By exploring vast chemical spaces in an inverse-design framework, they could accelerate polymer discovery for targeted thermal behavior.[158] LLMS have been used to generate small molecules, peptides and even polymers in property-driven design tasks. However, direct applications to thermal transport remain scarce.

Recent efforts illustrate the emerging promise. Transformer-based models have shown high accuracy in polymer generation. For instance, PolyTAO (a Transformer-Assisted model)[159] produced valid polymers with 99.3% chemical validity and strong agreement ($R^2 \approx 0.96$) between generated and target properties. Multimodal LLMs like Llamole (which interleaves graph-based molecular generation with text-based retrosynthetic planning)[160] demonstrate how future models might integrate synthesis and domain knowledge. Such examples suggest that generative/foundation models could possibly support (a) property-conditioned design: Models can be fine-tuned or conditioned to generate polymers optimized for desired thermal metrics (e.g., high conductivity or insulation); (b) Synthesis-aware generation: Generative networks can incorporate synthetic-accessibility or retrosynthetic planning (as in Llamole) to filter candidates for realistic synthesis; (c) Multi-modal integration: Foundation models may combine structural (graph) inputs with text or simulation data (e.g., literature, thermophysical predictions) to enrich design; early multimodal systems illustrate this potential.

Generally, generative/foundation models conceptually promise to expand thermal materials discovery by automating property-driven material design and by linking generation with practical constraints. However, their use in nanoscale thermal transport is still in early stages. Careful, application-specific development and benchmarking will be needed, and current claims should be viewed as proof-of-principle rather than fully realized solutions.

5.4 Transfer learning and cross-domain application

One of the most promising frontiers in ML for materials science and thermal transport is the growing application of TL and cross-domain frameworks. These strategies allow models trained on large, readily available datasets to generalize across different tasks, including those with limited or high-cost data. By transferring learned representations across properties, material classes, or simulation fidelities, TL significantly reduces data requirements, improves model accuracy, and accelerates materials discovery workflows.[161] This capability is especially valuable in thermal transport research, where high-fidelity data (e.g., experimental or DFT-calculated) is often scarce and expensive to generate. Instead of training models in isolation, TL allows knowledge reuse across properties, material systems, or simulation fidelities. For example, Gupta et al.[162] demonstrated that deep TL models trained on elemental fractions outperform baseline models across numerous material properties, including experimental datasets. Such cross-property learning is well-suited for TC, a property often lacking large, labeled datasets.

This layered approach, pretraining on computational data and fine-tuning on high-fidelity targets, also encourages the integration of multi-fidelity datasets. While not an example of TL, the study by Carrete et al.[163] illustrates how high-throughput *ab-initio* modeling, combined with simple ML rules, can accelerate the discovery of high-performance thermoelectric materials. Such efforts highlight the potential of physics-informed data pipelines to support downstream ML applications, including TL. Looking ahead, the real opportunity lies in designing TL workflows aware of domain shifts across materials classes, property types, or fidelity levels. As reviewed by Chen et al.[161], the challenges include how to transfer knowledge and when and what to transfer. Representational mismatch, model adaptability, and physical interpretability remain open yet are increasingly tractable with emerging techniques in domain adaptation, meta-learning, and uncertainty quantification. In thermal transport, where scarce labels and complex mechanisms often limit models, TL could be a unifying framework, connecting simulations, experiments, and surrogate models under a shared learning paradigm. Rather than discarding low-fidelity or auxiliary data, the next generation of thermal models may treat them as stepping stones toward high-fidelity insight.

5.5 AI agents

AI agents in scientific research are autonomous, goal-directed systems that perceive information and execute multi-step plans. In AI, an agent is defined as "an autonomous intelligent entity capable of performing appropriate and contextually relevant actions" to achieve goals.[164] Such agents can iteratively plan and execute complex workflows rather than returning a single prediction. For example, agents can automate literature reviews by querying databases, retrieving relevant papers, and synthesizing insights.[165] Agents can also integrate predictive models to screen material candidates: for instance, an agent might propose polymer structures or nanocomposite designs, use ML predictors for TC to rank them, and iterate until promising candidates emerge. In experimental planning, an agent could recommend measurement strategies (e.g., temperature sweeps) based on prior results, much like an automated active-learning approach. Similarly, in computational workflows, an agent could orchestrate simulation pipelines: e.g., scheduling multiscale MD or phonon calculations and analyzing outputs to inform the next steps. Crucially, these capabilities are achievable with current technology. LLMs can serve as the reasoning core, and frameworks like LangChain allow them to invoke external tools in a chain of actions. Retrieval-augmented generation can supply up-to-date data from literature or repositories, grounding the agent's reasoning in current knowledge. In essence, an AI agent for thermal transport would continuously reason, plan and act - autonomously curating knowledge, suggesting candidates and experiments, and tying together modeling tasks, in contrast to static ML models that only output a single inference.[165]

5.6 Autonomous experimental labs

Autonomous experimental laboratories are redefining how materials discovery is conducted by coupling ML, robotics, and AL into fully integrated platforms. These systems shift experimentation from human-led trial-and-error to closed-loop, data-driven processes capable of designing, executing, and interpreting experiments with minimal intervention. As scientific discovery increasingly depends on navigating high-dimensional chemical and processing spaces, autonomous labs offer a scalable solution to accelerate synthesis, characterization, and optimization.[166]

Recent advances have begun to demonstrate the practical potential of these systems. Szymanski et al.[167] introduced A-Lab, an autonomous platform for solid-state synthesis that combined computational modeling, historical data, and AL. Over 17 days, it synthesized 41 novel compounds and adapted its strategies based on success and failure, showcasing the effectiveness of autonomous exploration in complex chemical spaces. Beaucage et al.[168] described the emergence of self-driving labs as a fundamental shift in experimental science. They emphasized their role in studying rare phenomena, improving reproducibility, and redefining the role of human researchers. The authors also highlighted the need for interdisciplinary collaboration, infrastructure investment, and open data and hardware integration standards. Tom et al.[166] provided a broad review of self-driving laboratories (SDLs) across scientific domains. They detailed enabling technologies, real-world implementations, and key challenges such as system interoperability and data representation. The review emphasized SDLs' potential to transform research and industrial applications.

Autonomous platforms could be pivotal in systematically exploring process-structure-property relationships in thermal transport and materials design. Combined with predictive simulations and AL, they may enable closed-loop workflows where hypotheses are tested, refined, and re-evaluated autonomously. As with TL, integrating experimental autonomy does not replace human intuition. Instead, it expands the scope of scientific exploration by allowing researchers to focus on reasoning and interpretation while delegating routine experimentation to machines.

5.7 Optimizing nanocomposites and interfaces

High thermal performance in polymer nanocomposites depends on complex interactions between fillers, matrices, and interfaces. Moving forward, the integration of ML with experiments and simulations is expected to play a central role in predicting effective TC, estimating ITR, and guiding the rational design of composite architectures. Recent works have shown that combining data-driven models with multiscale insights can streamline the discovery and optimization of multifunctional nanocomposites.[169]

Efforts in this direction have begun to converge across different modeling strategies and material systems. For instance, Champa-Bujaico et al.[170] emphasized the role of ML in designing polymer nanocomposites by identifying critical features, enabling optimization, and supporting uncertainty quantification. Their review

positioned ML as a predictive tool and a means to complement and extend traditional experimental and computational approaches. Lu et al.[171] proposed a framework that couples high-throughput simulations with ML and minimal experimental input to estimate ITR in particulate-filled composites. Their approach captures both polymer-filler, and filler-filler interactions, yielding physically grounded parameters that are otherwise difficult to access experimentally. A similar philosophy is evident in the work of Fathidoost et al.,[172] who linked microstructural image data with finite element simulations and ML to predict effective TC in ultrahigh-temperature ceramic nanocomposites. Their method enabled inverse estimation of interfacial resistance and strongly agreed with experimental data, supporting the development of robust surrogate models. These studies reflect a broader shift toward integrated, data-informed frameworks that unify structure, processing, and property relationships.

6 Conclusion

The integration of ML into nanoscale thermal transport research marks a significant advancement, overcoming traditional limitations of computational intensity and experimental complexity. By effectively bridging physics-based insights and data-driven predictions, ML methods have demonstrated considerable potential across polymers, interfaces, and nanostructured materials. ML techniques, particularly when coupled with traditional methods such as DFT and MD, provide robust solutions for predicting and optimizing material properties, accelerating the discovery of materials with tailored thermal characteristics.

Several emerging ML strategies have shown remarkable promise. TL and multi-fidelity models effectively leverage limited high-quality data by incorporating extensive lower-fidelity information, significantly improving predictive accuracy. AL algorithms streamline data acquisition by strategically identifying the most informative experiments or simulations, substantially reducing the time and resources required for thermal property exploration. Furthermore, the recent development of PINNs addresses challenges in sparse-data regimes by explicitly embedding governing physical equations into ML architectures, ensuring physically consistent and interpretable outcomes.

Despite these advances, challenges remain, notably in data quality, interpretability, and ensuring reliable extrapolation beyond training datasets. Addressing these issues necessitates continued development of standardized, high-quality datasets, along with transparent models that can elucidate the underlying physics driving thermal transport phenomena. Future directions include deeper integration of ML methods into autonomous experimental setups, further development of generative models for inverse design, and greater utilization of AI-driven optimization in designing advanced polymer composites and interfaces.

The ongoing integration of computational simulations, experimental methods, and ML will drive significant advancements in materials science. This interdisciplinary approach accelerates the identification and deeper understanding of novel nanoscale thermal phenomena, paving the way toward robust, efficient, and sustainable solutions for next-generation thermal management technologies.

Acknowledgements

This work was supported in part by the Lucy Family Institute for Data & Society, Postdoctoral Fellowship and National Science Foundation grants (2332270 and 2102592), and DARPA grant (HR00112390112).

References

[1] Luo, T., Chen, G.: Nanoscale heat transfer–from computation to experiment. Physical Chemistry Chemical Physics **15**(10), 3389–3412 (2013)

[2] Cahill, D.G., Braun, P.V., Chen, G., Clarke, D.R., Fan, S., Goodson, K.E., Keblinski, P., King, W.P., Mahan, G.D., Majumdar, A., Maris, H.J., Phillpot, S.R., Pop, E., Shi, L.: Nanoscale thermal transport. ii. 2003–2012. Applied Physics Reviews **1**(1), 011305 (2014)

[3] Hopkins, P.E.: Thermal transport across solid interfaces with nanoscale imperfections: Effects of roughness, disorder, dislocations, and bonding on thermal boundary conductance. International Scholarly Research Notices **2013**(1), 682586 (2013)

[4] Pop, E.: Energy dissipation and transport in nanoscale devices. Nano Research **3**, 147–169 (2010)

[5] Snyder, G.J., Toberer, E.S.: Complex thermoelectric materials. Nature materials **7**(2), 105–114 (2008)

[6] Wei, X., Wang, Z., Tian, Z., Luo, T.: Thermal transport in polymers: a review. Journal of Heat Transfer **143**(7), 072101 (2021)

[7] Kim, J., Liu, Y., Luo, T., Tian, Z.: Molecular dynamics simulations in nanoscale heat transfer: A mini review. ASME Journal of Heat and Mass Transfer **147**(3) (2025)

[8] Muraleedharan, M.G., Gordiz, K., Rohskopf, A., Wyant, S.T., Cheng, Z., Graham, S., Henry, A.: Understanding phonon transport properties using classical molecular dynamics simulations. arXiv preprint arXiv:2011.01070 (2020)

[9] Ju, S., Shiomi, J.: Materials informatics for heat transfer: recent progresses and perspectives. Nanoscale and Microscale Thermophysical Engineering **23**(2), 157–172 (2019)

[10] Guo, R., Cao, B.-Y., Luo, T., McGaughey, A.J.: Machine learning for thermal transport. Journal of Applied Physics **136**(16) (2024)

[11] Huang, X., Ma, S., Zhao, C., Wang, H., Ju, S.: Exploring high thermal conductivity polymers via interpretable machine learning with physical descriptors. npj Computational Materials **9**(1), 191 (2023)

[12] Babaei, H., Guo, R., Hashemi, A., Lee, S.: Machine-learning-based interatomic potential for phonon transport in perfect crystalline si and crystalline si with vacancies. Physical Review Materials **3**(7), 074603 (2019)

[13] Ma, R., Zhang, H., Luo, T.: Exploring high thermal conductivity amorphous polymers using reinforcement learning. ACS Applied Materials & Interfaces **14**(13), 15587–15598 (2022)

[14] Wu, Y.-J., Fang, L., Xu, Y.: Predicting interfacial thermal resistance by machine learning. npj Computational Materials **5**(1), 56 (2019)

[15] Dou, Y., Shimizu, K., Carrete, J., Fujioka, H., Watanabe, S.: Machine-learning potential for phonon transport in aln with defects in multiple charge states. Physical Review Materials **9**(3), 034601 (2025)

[16] Hu, R., Iwamoto, S., Feng, L., Ju, S., Hu, S., Ohnishi, M., Nagai, N., Hirakawa, K., Shiomi, J.: Machine-learning-optimized aperiodic superlattice minimizes coherent phonon heat conduction. Physical Review X **10**(2), 021050 (2020)

[17] Le, T., Epa, V.C., Burden, F.R., Winkler, D.A.: Quantitative structure–property relationship modeling of diverse materials properties. Chemical reviews **112**(5), 2889–2919 (2012)

[18] Liu, Z., Jiang, M., Luo, T.: Leveraging low-fidelity data to improve machine learning of sparse high-fidelity thermal conductivity data via transfer learning. Materials Today Physics **28**, 100868 (2022)

[19] Ghiringhelli, L.M., Vybiral, J., Levchenko, S.V., Draxl, C., Scheffler, M.: Big data of materials science: critical role of the descriptor. Physical review letters **114**(10), 105503 (2015)

[20] Zhu, C., Bamidele, E.A., Shen, X., Zhu, G., Li, B.: Machine learning aided design and optimization of thermal metamaterials. Chemical Reviews **124**(7), 4258–4331 (2024)

[21] Chauhan, N.K., Singh, K.: A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 347–352 (2018). IEEE

[22] Lei, J.: Cross-validation with confidence. Journal of the American Statistical Association **115**(532), 1978–1997 (2020)

[23] Wei, H., Bao, H., Ruan, X.: Perspective: Predicting and optimizing thermal transport properties with machine learning methods. Energy and AI **8**, 100153 (2022)

[24] Panigrahi, S., Nanda, A., Swarnkar, T.: A survey on transfer learning. In: Mishra, D., Buyya, R., Mohapatra, P., Patnaik, S. (eds.) Intelligent and Cloud Computing, pp. 781–789. Springer, Singapore (2021)

[25] Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N.D., Karniadakis, G.E.: Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **473**(2198), 20160751 (2017)

[26] Wu, S., Kondo, Y., Kakimoto, M.-a., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., *et al.*: Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. Npj Computational Materials **5**(1), 66 (2019)

[27] Ju, S., Yoshida, R., Liu, C., Wu, S., Hongo, K., Tadano, T., Shiomi, J.: Exploring diamondlike lattice thermal conductivity crystals via feature-based transfer learning. Physical Review Materials **5**(5), 053801 (2021)

[28] Hu, Y., Wang, Q., Ma, H.: Machine-learning-assisted searching for thermally conductive polymers: A mini review. Journal of Applied Physics **135**(12) (2024)

[29] Lookman, T., Balachandran, P.V., Xue, D., Yuan, R.: Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Computational Materials **5**(1), 21 (2019)

[30] Kim, C., Chandrasekaran, A., Jha, A., Ramprasad, R.: Active-learning and materials design: the example of high glass transition temperature polymers. Mrs Communications **9**(3), 860–866 (2019)

[31] Xu, J., Luo, T.: Unlocking enhanced thermal conductivity in polymer blends through active learning. npj Computational Materials **10**(1), 74 (2024)

[32] Zhang, R., Xu, J., Zhang, H., Xu, G., Luo, T.: Active learning-guided exploration of thermally conductive polymers under strain. Digital Discovery **4**(3), 812–823 (2025)

[33] Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. arXiv preprint arXiv:1711.10561 (2017)

[34] Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal

of Computational physics **378**, 686–707 (2019)

[35] McClenny, L.D., Braga-Neto, U.M.: Self-adaptive physics-informed neural networks. Journal of Computational Physics **474**, 111722 (2023)

[36] Wang, S., Teng, Y., Perdikaris, P.: Understanding and mitigating gradient flow pathologies in physics-informed neural networks. SIAM Journal on Scientific Computing **43**(5), 3055–3081 (2021)

[37] Jagtap, A.D., Kawaguchi, K., Karniadakis, G.E.: Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. Journal of Computational Physics **404**, 109136 (2020)

[38] Jagtap, A.D., Kharazmi, E., Karniadakis, G.E.: Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. Computer Methods in Applied Mechanics and Engineering **365**, 113028 (2020)

[39] Jagtap, A.D., Karniadakis, G.E.: Extended physics-informed neural networks (xpinns): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations. Communications in Computational Physics **28**(5) (2020)

[40] Lu, L., Meng, X., Mao, Z., Karniadakis, G.E.: Deepxde: A deep learning library for solving differential equations. SIAM review **63**(1), 208–228 (2021)

[41] Li, R., Lee, E., Luo, T.: Physics-informed neural networks for solving multiscale mode-resolved phonon boltzmann transport equation. Materials Today Physics **19**, 100429 (2021)

[42] Li, R., Wang, J.-X., Lee, E., Luo, T.: Physics-informed deep learning for solving phonon boltzmann transport equation with large temperature non-equilibrium. npj Computational Materials **8**(1), 29 (2022)

[43] Zhou, J., Li, R., Luo, T.: Chapter five - physics-informed neural networks for modeling mesoscale heat transfer using the boltzmann transport equation. Advances in Heat Transfer, vol. 55, pp. 211–238. Elsevier (2023)

[44] Li, R., Lee, E., Luo, T.: Physics-informed deep learning for solving coupled electron and phonon boltzmann transport equations. Physical Review Applied **19**(6), 064049 (2023)

[45] Na, G.S., Chang, H.: A public database of thermoelectric materials and system-identified material representation for data-driven discovery. npj Computational Materials **8**(1), 214 (2022)

[46] 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**(31), 17575–17585 (2018)

[47] 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, pp. 22–29 (2011). IEEE

[48] Mondejar, M.E., Regidor, M., Krafczyk, J., Ihmels, C., Schmid, B., Kontogeorgis, G.M., Haglind, F.: An open-access database of the thermophysical properties of nanofluids. Journal of Molecular Liquids **333**, 115140 (2021)

[49] Lee, Y.-L., Lee, H., Jang, S., Shin, J., Kim, T., Byun, S., Chung, I., Im, J., Chang, H.: Texplorer.org: Thermoelectric material properties data platform for experimental and first-principles calculation results. APL Materials **11**(4) (2023)

[50] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner,

D., Ceder, G., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)

[51] Ma, R., Luo, T.: Pi1m: a benchmark database for polymer informatics. Journal of Chemical Information and Modeling **60**(10), 4684–4690 (2020)

[52] Ma, R., Liu, Z., Zhang, Q., Liu, Z., Luo, T.: Evaluating polymer representations via quantifying structure–property relationships. Journal of chemical information and modeling **59**(7), 3110–3119 (2019)

[53] Xu, J., Liu, G., Guo, R., Jiang, M., Luo, T.: Point^2: A polymer informatics training and testing database. arXiv preprint arXiv:2503.23491 (2025)

[54] Shen, J., Griesemer, S.D., Gopakumar, A., Baldassarri, B., Saal, J.E., Aykol, M., Hegde, V.I., Wolverton, C.: Reflections on one million compounds in the open quantum materials database (oqmd). Journal of Physics: Materials **5**(3), 031001 (2022)

[55] Schleder, G.R., Padilha, A.C., Acosta, C.M., Costa, M., Fazzio, A.: From dft to machine learning: recent approaches to materials science–a review. Journal of Physics: Materials **2**(3), 032001 (2019)

[56] Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences **28**(1), 31–36 (1988)

[57] Krenn, M., Häse, F., Nigam, A., Friederich, P., Aspuru-Guzik, A.: Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Machine Learning: Science and Technology **1**(4), 045024 (2020)

[58] Lin, T.-S., Coley, C.W., Mochigase, H., Beech, H.K., Wang, W., Wang, Z., Woods, E., Craig, S.L., Johnson, J.A., Kalow, J.A., *et al.*: Bigsmiles: a structurally-based line notation for describing macromolecules. ACS central science **5**(9), 1523–1531 (2019)

[59] Bicerano, J.: Prediction of Polymer Properties, 3rd edn. CRC Press, Boca Raton, FL (2002)

[60] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling **50**(5), 742–754 (2010)

[61] Nilakantan, R., Bauman, N., Dixon, J.S., Venkataraghavan, R.: Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. Journal of Chemical Information and Computer Sciences **27**(2), 82–85 (1987)

[62] Wu, Y.-J., Zhan, T., Hou, Z., Fang, L., Xu, Y.: Physical and chemical descriptors for predicting interfacial thermal resistance. Scientific Data **7**(1), 36 (2020)

[63] Guo, Z., Roy Chowdhury, P., Han, Z., Sun, Y., Feng, D., Lin, G., Ruan, X.: Fast and accurate machine learning prediction of phonon scattering rates and lattice thermal conductivity. npj Computational Materials **9**(1), 95 (2023)

[64] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)

[65] Kuenneth, C., Ramprasad, R.: polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Nature communications **14**(1), 4099 (2023)

[66] Xu, C., Wang, Y., Barati Farimani, A.: Transpolymer: a transformer-based language model for polymer property predictions. npj Computational Materials **9**(1), 64 (2023)

[67] Doan Tran, H., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J.P., Gurnani, R., Shetty, P., et al.: Machine-learning predictions of polymer properties with polymer genome. Journal of Applied Physics **128**(17) (2020)

[68] Guo, Z., Yu, W., Zhang, C., Jiang, M., Chawla, N.V.: Graseq: graph and sequence fusion learning for molecular property prediction. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 435–443 (2020)

[69] Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., Hou, T.: Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. Journal of cheminformatics **13**(1), 12 (2021)

[70] Bellman, R.E., Dreyfus, S.E.: Applied Dynamic Programming. Princeton Legacy Library. Princeton University Press, Princeton, NJ (2015). Originally published 1962

[71] Janet, J.P., Kulik, H.J.: Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships. The Journal of Physical Chemistry A **121**(46), 8939–8954 (2017)

[72] Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., Ghiringhelli, L.M.: Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Physical Review Materials **2**(8), 083802 (2018)

[73] Zhu, M.-X., Song, H.-G., Yu, Q.-C., Chen, J.-M., Zhang, H.-Y.: Machine-learning-driven discovery of polymers molecular structures with high thermal conductivity. International Journal of Heat and Mass Transfer **162**, 120381 (2020)

[74] Zhang, T., Wu, X., Luo, T.: Polymer nanofibers with outstanding thermal conductivity and thermal stability: fundamental linkage between molecular characteristics and macroscopic thermal properties. The Journal of Physical Chemistry C **118**(36), 21148–21159 (2014)

[75] Gurnani, R., Kuenneth, C., Toland, A., Ramprasad, R.: Polymer informatics at scale with multitask graph neural networks. Chemistry of Materials **35**(4), 1560–1567 (2023)

[76] Chen, M., Li, J., Tian, B., Al-Hadeethi, Y.M., Arkook, B., Tian, X., Zhang, X.: Predicting interfacial thermal resistance by ensemble learning. Computation **9**(8), 87 (2021)

[77] Miccio, L.A., Schwartz, G.A.: From chemical structure to quantitative polymer properties prediction through convolutional neural networks. Polymer **193**, 122341 (2020)

[78] Park, J., Shim, Y., Lee, F., Rammohan, A., Goyal, S., Shim, M., Jeong, C., Kim, D.S.: Prediction and interpretation of polymer properties using the graph convolutional network. ACS Polymers Au **2**(4), 213–222 (2022)

[79] Tao, L., Varshney, V., Li, Y.: Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. Journal of Chemical Information and Modeling **61**(11), 5395–5413 (2021)

[80] Ma, R., Zhang, H., Xu, J., Sun, L., Hayashi, Y., Yoshida, R., Shiomi, J., Wang, J.-x., Luo, T.: Machine learning-assisted exploration of thermally conductive polymers based on high-throughput molecular dynamics simulations. Materials Today Physics **28**, 100850 (2022)

[81] Kuenneth, C., Schertzer, W., Ramprasad, R.: Copolymer informatics with multitask deep neural networks. Macromolecules **54**(13), 5957–5961 (2021)

[82] Tao, L., Byrnes, J., Varshney, V., Li, Y.: Machine learning strategies for the structure-property

relationship of copolymers. Iscience **25**(7) (2022)

[83] Queen, O., McCarver, G.A., Thatigotla, S., Abolins, B.P., Brown, C.L., Maroulas, V., Vogiatzis, K.D.: Polymer graph neural networks for multitask property learning. npj Computational Materials **9**(1), 90 (2023)

[84] Rustam, S., Schram, M., Lu, Z., Chaka, A.M., Rosenthal, W.S., Pfaendtner, J.: Optimization of thermal conductance at interfaces using machine learning algorithms. ACS Applied Materials & Interfaces **14**(28), 32590–32597 (2022)

[85] Jin, S., Zhang, Z., Guo, Y., Chen, J., Nomura, M., Volz, S.: Optimization of interfacial thermal transport in si/ge heterostructure driven by machine learning. International Journal of Heat and Mass Transfer **182**, 122014 (2022)

[86] Foss, C., Aksamija, Z.: Machine learning enables robust prediction of thermal boundary conductance of 2d substrate interfaces. Applied Physics Letters **122**(6) (2023)

[87] Anandakrishnan, A., Sathian, S.P.: A data driven approach to model thermal boundary resistance from molecular dynamics simulations. Physical Chemistry Chemical Physics **25**(4), 3258–3269 (2023)

[88] Tao, L., Chen, G., Li, Y.: Machine learning discovery of high-temperature polymers. Patterns **2**(4) (2021)

[89] Tao, L., He, J., Munyaneza, N.E., Varshney, V., Chen, W., Liu, G., Li, Y.: Discovery of multifunctional polyimides through high-throughput screening using explainable machine learning. Chemical Engineering Journal **465**, 142949 (2023)

[90] Huang, X., Ma, S., Wu, Y., Wan, C., Zhao, C., Wang, H., Ju, S.: High-throughput screening of amorphous polymers with high intrinsic thermal conductivity via automated physical feature engineering. Journal of Materials Chemistry A **11**(38), 20539–20548 (2023)

[91] Zhang, T., Luo, T.: Role of chain morphology and stiffness in thermal conductivity of amorphous polymers. The Journal of Physical Chemistry B **120**(4), 803–812 (2016)

[92] Huang, X., Zhao, C., Wang, H., Ju, S.: Ai-assisted inverse design of sequence-ordered high intrinsic thermal conductivity polymers. Materials Today Physics **44**, 101438 (2024)

[93] Batra, R., Dai, H., Huan, T.D., Chen, L., Kim, C., Gutekunst, W.R., Song, L., Ramprasad, R.: Polymers for extreme conditions designed using syntax-directed variational autoencoders. Chemistry of Materials **32**(24), 10489–10500 (2020)

[94] Liu, D.-F., Zhang, Y.-X., Dong, W.-Z., Feng, Q.-K., Zhong, S.-L., Dang, Z.-M.: High-temperature polymer dielectrics designed using an invertible molecular graph generative model. Journal of Chemical Information and Modeling **63**(24), 7669–7675 (2023)

[95] Yue, T., Tao, L., Varshney, V., Li, Y.: Benchmarking study of deep generative models for inverse polymer design. Digital Discovery (2025)

[96] Pilania, G., Iverson, C.N., Lookman, T., Marrone, B.L.: Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. Journal of Chemical Information and Modeling **59**(12), 5013–5025 (2019)

[97] Kim, C., Batra, R., Chen, L., Tran, H., Ramprasad, R.: Polymer design using genetic algorithm and machine learning. Computational Materials Science **186**, 110067 (2021)

[98] Zhou, T., Wu, Z., Chilukoti, H.K., Muller-Plathe, F.: Sequence-engineering polyethylene–polypropylene

copolymers with high thermal conductivity using a molecular-dynamics-based genetic algorithm. Journal of Chemical Theory and Computation **17**(6), 3772–3782 (2021)

[99] Nagoya, A., Kikkawa, N., Ohba, N., Baba, T., Kajita, S., Yanai, K., Takeno, T.: Autonomous search for polymers with high thermal conductivity using a rapid green–kubo estimation. Macromolecules **55**(9), 3384–3395 (2022)

[100] Deringer, V.L., Csányi, G.: Machine learning based interatomic potential for amorphous carbon. Physical Review B **95**(9), 094203 (2017)

[101] Rajabpour, A., Mortazavi, B., Mirchi, P., El Hajj, J., Guo, Y., Zhuang, X., Merabia, S.: Accurate estimation of interfacial thermal conductance between silicon and diamond enabled by a machine learning interatomic potential. International Journal of Thermal Sciences **214**, 109876 (2025)

[102] Qian, X., Peng, S., Li, X., Wei, Y., Yang, R.: Thermal conductivity modeling using machine learning potentials: application to crystalline and amorphous silicon. Materials Today Physics **10**, 100140 (2019)

[103] Choi, J.M., Lee, K., Kim, S., Moon, M., Jeong, W., Han, S.: Accelerated computation of lattice thermal conductivity using neural network interatomic potentials. Computational Materials Science **211**, 111472 (2022)

[104] Behler, J.: Perspective: Machine learning potentials for atomistic simulations. The Journal of chemical physics **145**(17) (2016)

[105] Khot, K., Xiao, B., Han, Z., Guo, Z., Xiong, Z., Ruan, X.: Phonon local non-equilibrium at al/si interface from machine learning molecular dynamics. Journal of Applied Physics **137**(11) (2025)

[106] Guo, Z., Han, Z., Feng, D., Lin, G., Ruan, X.: Sampling-accelerated prediction of phonon scattering rates for converged thermal conductivity and radiative properties. npj Computational Materials **10**(1), 31 (2024)

[107] Guo, Z., Han, Z., Alkandari, A., Khot, K., Ruan, X.: First-principles prediction of thermal conductivity of bulk hexagonal boron nitride. Applied Physics Letters **124**(16) (2024)

[108] Behler, J., Parrinello, M.: Generalized neural-network representation of high-dimensional potential-energy surfaces. Physical review letters **98**(14), 146401 (2007)

[109] Zhang, L., Han, J., Wang, H., Car, R., E, W.: Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. Physical review letters **120**(14), 143001 (2018)

[110] Bartók, A.P., Payne, M.C., Kondor, R., Csányi, G.: Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters **104**(13), 136403 (2010)

[111] Shapeev, A.V.: Moment tensor potentials: A class of systematically improvable interatomic potentials. Multiscale Modeling & Simulation **14**(3), 1153–1173 (2016)

[112] Deringer, V.L., Caro, M.A., Csányi, G.: Machine learning interatomic potentials as emerging tools for materials science. Advanced Materials **31**(46), 1902765 (2019)

[113] Korotaev, P., Novoselov, I., Yanilkin, A., Shapeev, A.: Accessing thermal conductivity of complex compounds by machine learning interatomic potentials. Physical Review B **100**(14), 144308 (2019)

[114] Rybin, N., Shapeev, A.: A moment tensor potential for lattice thermal conductivity calculations of α and β phases of ga2o3. Journal of Applied Physics **135**(20) (2024)

[115] Fan, Z., Zeng, Z., Zhang, C., Wang, Y., Song, K., Dong, H., Chen, Y., Ala-Nissila, T.: Neuroevolution machine learning potentials: Combining high accuracy and low cost in atomistic simulations and application to heat transport. Physical Review B **104**(10), 104309 (2021)

[116] Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C.J., Ceder, G.: Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence **5**(9), 1031–1041 (2023)

[117] Batatia, I., Kovacs, D.P., Simm, G., Ortner, C., Csányi, G.: Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in neural information processing systems **35**, 11423–11436 (2022)

[118] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.: E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature communications **13**(1), 2453 (2022)

[119] Zhang, Y., Wang, H., Chen, W., Zeng, J., Zhang, L., Wang, H., *et al.*: Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models. Computer Physics Communications **253**, 107206 (2020)

[120] Wang, G., Wang, C., Zhang, X., Li, Z., Zhou, J., Sun, Z.: Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. Iscience (2024)

[121] Mortazavi, B., Podryabinkin, E.V., Roche, S., Rabczuk, T., Zhuang, X., Shapeev, A.V.: Machine-learning interatomic potentials enable first-principles multiscale modeling of lattice thermal conductivity in graphene/borophene heterostructures. Materials Horizons **7**(9), 2359–2367 (2020)

[122] Wyant, S., Rohskopf, A., Henry, A.: Machine learned interatomic potentials for modeling interfacial heat transport in ge/gaas. Computational Materials Science **200**, 110836 (2021)

[123] Tersoff, J.: Empirical interatomic potential for carbon, with applications to amorphous carbon. Physical Review Letters **61**(25), 2879 (1988)

[124] Brenner, D.W.: Empirical potential for hydrocarbons for use in simulating the chemical vapor deposition of diamond films. Physical review B **42**(15), 9458 (1990)

[125] Li, R., Hussain, K., Liao, M.E., Huynh, K., Hoque, M.S.B., Wyant, S., Koh, Y.R., Xu, Z., Wang, Y., Luccioni, D.P., *et al.*: Enhanced thermal boundary conductance across gan/sic interfaces with aln transition layers. ACS applied materials & interfaces **16**(6), 8109–8118 (2024)

[126] Li, R., Lee, E., Luo, T.: A unified deep neural network potential capable of predicting thermal conductivity of silicon in different phases. Materials Today Physics **12**, 100181 (2020)

[127] Li, R., Liu, Z., Rohskopf, A., Gordiz, K., Henry, A., Lee, E., Luo, T.: A deep neural network interatomic potential for studying thermal conductivity of β-ga2o3. Applied Physics Letters **117**(15) (2020)

[128] Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., Han, T.Y.-J.: Explainable machine learning in materials science. npj computational materials **8**(1), 204 (2022)

[129] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[130] Udrescu, S.-M., Tegmark, M.: Ai feynman: A physics-inspired method for symbolic regression. Science advances **6**(16), 2631 (2020)

[131] Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable

prediction of material properties. Physical review letters **120**(14), 145301 (2018)

[132] Liu, G., Inae, E., Luo, T., Jiang, M.: Rationalizing graph neural networks with data augmentation. ACM Transactions on Knowledge Discovery from Data **18**(4), 1–23 (2024)

[133] 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, pp. 1069–1078 (2022)

[134] Zeng, Y., Cao, W., Zuo, Y., Peng, T., Hou, Y., Miao, L., Wang, Z.: Exploring lattice thermal conductivity models via interpretable deep learning to accelerate the discovery of novel materials. arXiv preprint arXiv:2412.05948 (2024)

[135] Kanti, P.K., Paramasivam, P., Wanatasanappan, V.V., Dhanasekaran, S., Sharma, P.: Experimental and explainable machine learning approach on thermal conductivity and viscosity of water based graphene oxide based mono and hybrid nanofluids. Scientific Reports **14**(1), 30967 (2024)

[136] Barua, N.K., Hall, E., Cheng, Y., Oliynyk, A.O., Kleinke, H.: Interpretable machine learning model on thermal conductivity using publicly available datasets and our internal lab dataset. Chemistry of Materials **36**(14), 7089–7100 (2024)

[137] Gokhale, G., Claessens, B., Develder, C.: Physics informed neural networks for control oriented thermal modeling of buildings. Applied Energy **314**, 118852 (2022)

[138] Li, R., Zhou, J., Wang, J.-X., Luo, T.: Physics-informed bayesian neural networks for solving phonon boltzmann transport equation in forward and inverse problems with sparse and noisy data. ASME Journal of Heat and Mass Transfer **147**(3) (2025)

[139] Liao, S., Xue, T., Jeong, J., Webster, S., Ehmann, K., Cao, J.: Hybrid thermal modeling of additive manufacturing processes using physics-informed neural networks for temperature prediction and parameter identification. Computational Mechanics **72**(3), 499–512 (2023)

[140] Sripada, S., Gaitonde, A.U., Weibel, J.A., Marconnet, A.M.: Robust inverse parameter fitting of thermal properties from the laser-based ångstrom method in the presence of measurement noise using physics-informed neural networks (pinns). Journal of Applied Physics **135**(22) (2024)

[141] Huxtable, S., Cahill, D.G., Fauconnier, V., White, J.O., Zhao, J.-C.: Thermal conductivity imaging at micrometre-scale resolution for combinatorial studies of materials. Nature Materials **3**(5), 298–301 (2004)

[142] Yang, J., Maragliano, C., Schmidt, A.J.: Thermal property microscopy with frequency domain thermoreflectance. Review of Scientific Instruments **84**(10) (2013)

[143] Cahill, D.G.: Thermal conductivity measurement from 30 to 750 k: the 3ω method. Review of scientific instruments **61**(2), 802–808 (1990)

[144] Birge, N.O.: Specific-heat spectroscopy of glycerol and propylene glycol near the glass transition. Physical Review B **34**(3), 1631 (1986)

[145] Mao, Y., Zhou, S., Tang, W., Wu, M., Zhang, H., Sun, H., Yuan, C.: Deep learning-based data processing method for transient thermoreflectance measurements. Journal of Applied Physics **135**(9) (2024)

[146] Xiang, Z., Pang, Y., Qian, X., Yang, R.: Machine learning reconstruction of depth-dependent thermal conductivity profile from pump–probe thermoreflectance signals. Applied Physics Letters **122**(14) (2023)

[147] Shen, W., Vaca, D., Kumar, S.: Reconsidering uncertainty from frequency domain thermoreflectance measurement and novel data analysis by deep learning. Nanoscale and microscale thermophysical engineering **24**(3-4), 138–149 (2020)

[148] Pang, Y., Jiang, P., Yang, R.: Machine learning-based data processing technique for time-domain thermoreflectance (tdtr) measurements. Journal of Applied Physics **130**(8) (2021)

[149] Sierepeklis, O., Cole, J.M.: A thermoelectric materials database auto-generated from the scientific literature using chemdataextractor. Scientific Data **9**(1), 648 (2022)

[150] Shetty, P., Rajan, A.C., Kuenneth, C., Gupta, S., Panchumarti, L.P., Holm, L., Zhang, C., Ramprasad, R.: A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. npj Computational Materials **9**(1), 52 (2023)

[151] Hu, J., Li, Z., Lin, J., Zhang, L.: Prediction and interpretability of glass transition temperature of homopolymers by data-augmented graph convolutional neural networks. ACS Applied Materials & Interfaces **15**(46), 54006–54017 (2023)

[152] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895 (2020)

[153] Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A.: Neural operator: Learning maps between function spaces with applications to pdes. Journal of Machine Learning Research **24**(89), 1–97 (2023)

[154] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[155] Pan, S., Brunton, S.L., Kutz, J.N.: Neural implicit flow: a mesh-agnostic dimensionality reduction paradigm of spatio-temporal data. Journal of Machine Learning Research **24**(41), 1–60 (2023)

[156] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[157] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186 (2019)

[158] Pyzer-Knapp, E.O., Manica, M., Staar, P., Morin, L., Ruch, P., Laino, T., Smith, J.R., Curioni, A.: Foundation models for materials discovery–current state and future directions. npj Computational Materials **11**(1), 61 (2025)

[159] Qiu, H., Sun, Z.-Y.: On-demand reverse design of polymers with polytao. npj Computational Materials **10**(1), 273 (2024)

[160] Liu, G., Sun, M., Matusik, W., Jiang, M., Chen, J.: Multimodal large language models for inverse molecular design with retrosynthetic planning. arXiv preprint arXiv:2410.04223 (2024)

[161] Chen, A., Wang, Z., Vidaurre, K.L.L., Han, Y., Ye, S., Tao, K., Wang, S., Gao, J., Li, J.: Knowledge-reused transfer learning for molecular and materials science. Journal of Energy Chemistry **98**, 149–168 (2024)

[162] Gupta, V., Choudhary, K., Tavazza, F., Campbell, C., Liao, W.-k., Choudhary, A., Agrawal, A.: Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data.

Nature communications **12**(1), 6595 (2021)

[163] Carrete, J., Mingo, N., Wang, S., Curtarolo, S.: Nanograined half-heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study. Advanced Functional Materials **24**(47), 7427–7432 (2014)

[164] Gridach, M., Nanavati, J., Abidine, K.Z.E., Mendes, L., Mack, C.: Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. arXiv preprint arXiv:2503.08979 (2025)

[165] Ferrag, M.A., Tihanyi, N., Debbah, M.: From llm reasoning to autonomous ai agents: A comprehensive review. arXiv preprint arXiv:2504.19678 (2025)

[166] Tom, G., Schmid, S.P., Baird, S.G., Cao, Y., Darvish, K., Hao, H., Lo, S., Pablo-García, S., Rajaonson, E.M., Skreta, M., *et al.*: Self-driving laboratories for chemistry and materials science. Chemical Reviews **124**(16), 9633–9732 (2024)

[167] Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T., Milsted, D., McDermott, M.J., Gallant, M., Cubuk, E.D., Merchant, A., *et al.*: An autonomous laboratory for the accelerated synthesis of novel materials. Nature **624**(7990), 86–91 (2023)

[168] Beaucage, P.A., Sutherland, D.R., Martin, T.B.: Automation and machine learning for accelerated polymer characterization and development: Past, potential, and a path forward. Macromolecules **57**(18), 8661–8670 (2024)

[169] Liu, B., Vu-Bac, N., Rabczuk, T.: A stochastic multiscale method for the prediction of the thermal conductivity of polymer nanocomposites through hybrid machine learning algorithms. Composite Structures **273**, 114269 (2021)

[170] Champa-Bujaico, E., García-Díaz, P., Díez-Pascual, A.M.: Machine learning for property prediction and optimization of polymeric nanocomposites: a state-of-the-art. International Journal of Molecular Sciences **23**(18), 10712 (2022)

[171] Lu, X., Huang, J., Cheng, N., Cheng, Z., Ren, L., Xu, J., Sun, R., Lu, J.: Data-driven framework toward accurate prediction of interfacial thermal resistance in particulate-filled composites. ACS Applied Materials & Interfaces **15**(36), 43169–43182 (2023)

[172] Fathidoost, M., Yang, Y., Thor, N., Bernauer, J., Pundt, A., Riedel, R., Xu, B.-X.: Thermal conductivity analysis of polymer-derived nanocomposite via image-based structure reconstruction, computational homogenization, and machine learning. Advanced Engineering Materials **26**(17), 2302021 (2024)