

<https://doi.org/10.1038/s41524-025-01606-5>

# Scaling Law of Sim2Real transfer learning in expanding computational materials databases for real-world predictions

Check for updates

Shunya Minami<sup>1,5</sup>, Yoshihiro Hayashi<sup>1,2,5</sup>, Stephen Wu<sup>1,2</sup>, Kenji Fukumizu<sup>1,2</sup>, Hiroki Sugisawa<sup>3</sup>, Masashi Ishii<sup>4</sup>, Isao Kuwajima<sup>4</sup>, Kazuya Shiratori<sup>3</sup> & Ryo Yoshida<sup>1,2,4</sup>✉

To address the challenge of limited experimental materials data, extensive physical property databases are being developed based on high-throughput computational experiments, such as molecular dynamics simulations. Previous studies have shown that fine-tuning a predictor pretrained on a computational database to a real system can result in models with outstanding generalization capabilities compared to learning from scratch. This study demonstrates the scaling law of simulation-to-real (Sim2Real) transfer learning for several machine learning tasks in materials science. Case studies of three prediction tasks for polymers and inorganic materials reveal that the prediction error on real systems decreases according to a power-law as the size of the computational data increases. Observing the scaling behavior offers various insights for database development, such as determining the sample size necessary to achieve a desired performance, identifying equivalent sample sizes for physical and computational experiments, and guiding the design of data production protocols for downstream real-world tasks.

Machine learning holds great potential for revolutionizing the methodology of materials science. Recent studies have demonstrated that models trained using materials data can accurately predict various physicochemical properties for diverse material systems<sup>1,2</sup>. Conventionally, a model defines the mapping from compositional or structural features of a given material to its thermal, electrical, mechanical, and energetic properties, as well as higher-order structural features. Assessing a large library of candidate materials using such models has led to the discovery of various materials, such as polymers<sup>3</sup>, inorganic crystalline compounds<sup>4,5</sup>, high-entropy alloys<sup>6</sup>, catalysts<sup>7,8</sup>, and quasiperiodic materials<sup>9–11</sup>. The success of such data-driven research depends on the quantity and quality of the data, and researchers often face the critical issue of data scarcity. Generating experimental data requires time-consuming, multi-stage workflows involving synthesis, sample preparation, property measurements, phase identification, and other laborious trial-and-error processes. More critically, researchers lack the incentive to disclose their laboratory data to open communities due to concerns regarding information confidentiality<sup>12</sup>, which hampers the creation of an open data foundation.

Large-scale databases based on computer experiments such as first-principles calculations and molecular dynamics (MD) simulations are being

developed to overcome the barriers posed by limited experimental data. For inorganic compounds, extensive first-principles property databases, including tens of thousands or more of crystal structures, have been developed, such as Materials Project<sup>13</sup>, AFLOWLIB<sup>14</sup>, NOMAD<sup>15</sup>, OQMD<sup>16,17</sup>, and GNoME<sup>4</sup>. The QM9 database<sup>18</sup> comprises over 130,000 small organic molecules, providing molecular structures and their properties obtained from quantum mechanical calculations, which serves as a dataset for machine-learning-based property prediction tasks<sup>1,3</sup>. Although there is currently no comprehensive computational database for polymeric materials, RadonPy<sup>19</sup> is being developed as a Python library for fully automated all-atom classical MD simulations to generate data resources for machine learning.

The methodology of transfer learning, particularly simulation-to-real (Sim2Real) transfer, enables the integration of extensive simulation data with limited quantitative experimental data<sup>20–23</sup>. Transfer learning is beneficial when training a model from scratch on the target task is impractical due to data scarcity; it leverages data or pretrained models from a source domain to enhance machine learning tasks in a target domain. This becomes increasingly advantageous as the relevance between the source and target domains increases. For example, in computer vision, Sim2Real transfer is

<sup>1</sup>The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Japan. <sup>2</sup>The Graduate Institute for Advanced Studies, SOKENDAI, Tachikawa, Japan. <sup>3</sup>Science & Innovation Center, Mitsubishi Chemical Corporation, Yokohama, Japan. <sup>4</sup>Research and Service Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Japan. <sup>5</sup>These authors contributed equally: Shunya Minami, Yoshihiro Hayashi.

✉ e-mail: [yoshidar@ism.ac.jp](mailto:yoshidar@ism.ac.jp)

crucial for adapting vision models trained in simulation environments to real-world applications, such as autonomous vehicles, by leveraging insights gained from simulated environments. Sim2Real transfer is also widely used in materials research. For instance Wu et al.<sup>3</sup> developed a predictive model for the thermal conductivity of polymeric materials using experimentally observed data for 28 amorphous polymers. Leveraging a large dataset of specific heat capacity generated through quantum chemistry calculations as the source task, they successfully derived the Sim2Real-transferred model in the target domain. Similarly Aoki et al.<sup>2</sup> employed a machine learning framework called multitask learning to integrate a quantum chemistry dataset with biased and quantitatively limited experimental data, successfully building a predictive model for polymer–solvent miscibility for a wide range of chemical spaces. Ju et al.<sup>24</sup> employed transfer learning to build a model predicting lattice thermal conductivity of inorganic crystalline materials. With only 45 samples for the target property, ordinary supervised learning failed to meet accuracy requirements. To address this, they utilized the first-principles calculations of scattering phase space as the source task and applied transfer learning to achieve sufficient accuracy.

Given the inherent domain gap between computer experiments and real-world systems, it is uncertain whether increasing the volume of simulation data enhances the generalization performance of Sim2Real-transferred models. To clarify this Mikami et al.<sup>25</sup>, provided theoretical and experimental evidence showing that the generalization of Sim2Real transfer learning improves according to a power-law relationship with the expansion of simulation data. Specifically, experimental validation of the scaling law was demonstrated for the Sim2Real scenario in computer vision tasks. Observing the scaling behavior of Sim2Real transfer, and estimating its convergence rate and asymptotic behavior offer valuable insights for advancing database development.

Here, we present a statistical measure for quantitatively evaluating the transferability and scalability of a growing computational materials database. Our work reveals the existence of a scaling law in transfer learning across diverse prediction tasks in materials research involving polymers and inorganic material systems. Specifically, it encompasses three scenarios: (1) Sim2Real prediction of polymer properties by re-purposing neural networks pretrained with all-atom classical MD simulations; (2) multitask machine learning integrating expansive data from quantum chemistry calculations and a limited experimental dataset to predict the miscibility of polymer–solvent binary mixtures; and (3) validation of the Wiedemann–Franz (WF) law between thermal and electrical conductivities of inorganic material systems through transfer learning. Notably, since both the source and target datasets in the third case were obtained from real experiments, the concept shown here can extend beyond Sim2Real scenarios. By experimentally observing the scaling behavior of transferred predictors, we can estimate their expected generalization performance upon further increasing the volume of simulation data, serving as an indicator of the database’s potential value. Moreover, multidimensional scaling, considering both physical and computer experiments, provides a statistical estimate for the equivalent sample size of experimental and computational data. This aids in decision-making for the design of data production protocols. Additionally, by observing the scaling behavior of individual materials, we can individualize database design guidelines and gain insights into the existence of material groups that share physical mechanisms across different material systems.

## Results

### Outline

Sim2Real transfer learning involves adapting a predictive model that is pretrained in a virtual environment to real-world scenarios. In materials research, the predictor defines a mathematical mapping from a descriptor representing the composition or structural features of a given material to its physicochemical properties. In the source task, the model is trained using a dataset of size  $n$  generated from computer experiments, such as first-principles ab initio calculations or MD simulations. In the target task, this pretrained model is repurposed and transferred to predict experimentally observed properties, utilizing an experimental dataset of size  $m$ , where  $m$  is

typically much smaller than  $n$ . Mikami et al.<sup>25</sup> presented a general theory, under certain assumptions, stating that in the fine-tuning of neural networks, the generalization error  $\mathbb{E}[L(f_{n,m})]$  with the squared loss  $L(f)$  of a transferred model  $f_{n,m}$  for the real-world system is bounded from above by a function  $R(n, m)$ :

$$R(n, m) := (An^{-\alpha} + B)m^{-\beta} + \epsilon, \quad (1)$$

where  $A, B, \alpha, \beta, \epsilon \geq 0$  are constants independent of  $n, m$ . In particular, considering the case of a fixed number of experimental samples at  $m$ , the upper bound for the generalization error is expressed as follows:

$$\mathbb{E}[L(f_{n,m})] \leq R(n) := Dn^{-\alpha} + C, \quad (2)$$

where  $D := Am^{-\beta}$  and  $C := Bm^{-\beta} + \epsilon$ . According to this law, as  $n$  increases, the generalization error of predicting experimentally observed properties for the transferred network converges to a reachable limit  $C \geq 0$ , called the transfer gap, with a decay rate  $\alpha \geq 0$ . Mikami et al.<sup>25</sup> demonstrated that these power-law relations hold empirically in Sim2Real transfer in computer vision tasks.

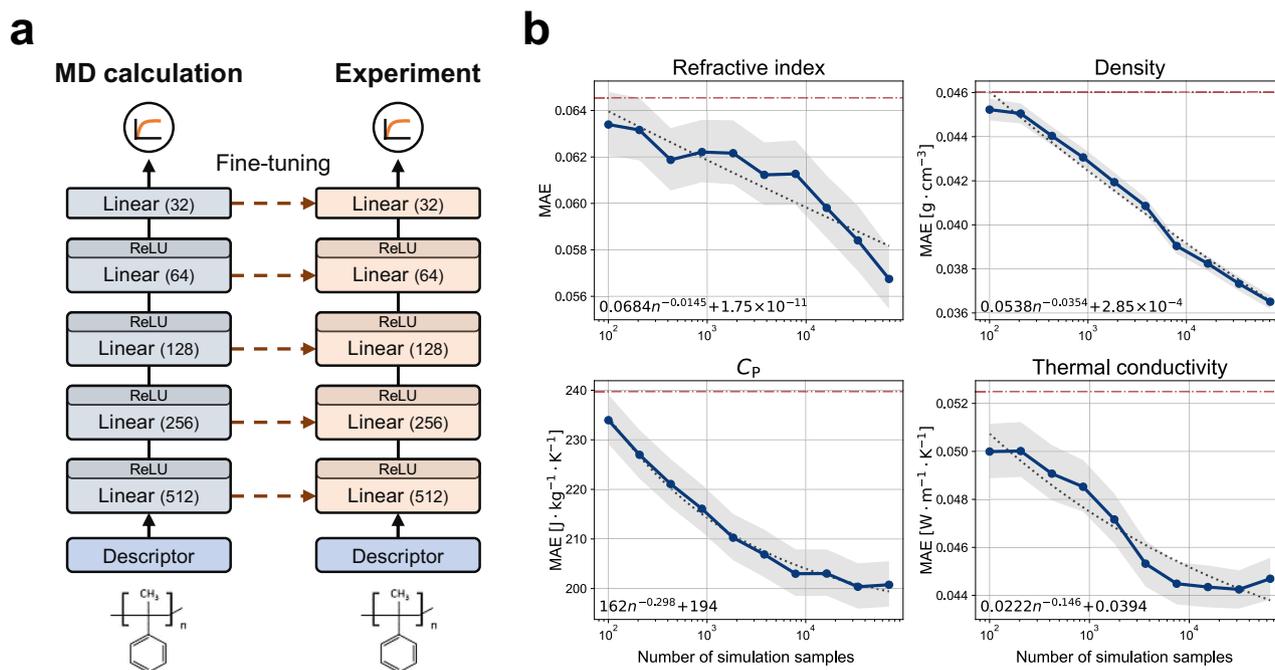
While increasing the data size for pretraining, the reduction in the generalization error of the transferred model is measured experimentally, which can be used to evaluate whether the transferred model can attain the desired prediction performance in the target task or to estimate the required sample size based on the estimated  $(D, \alpha, C)$ . Additionally, the observed scaling behaviors provide guidelines for designing the source database. For a predefined set of downstream tasks leveraging the database, the simulation environment can be tailored to accelerate scaling to real systems, such as selecting empirical interatomic potentials or polymerization degrees in MD simulations. For instance Mikami et al.<sup>25</sup>, applied Sim2Real transfer for computer vision tasks and showed that intentionally increasing the diversity of the appearance, luminosity, and background in a synthetic image set leads to an increase in the scaling factor  $\alpha$  and a partial improvement in  $C$ . Creating a data-generation scheme that results in a negligible  $C$  is the ultimate objective in developing foundational source data. Intuitively, the consistency of simulations to real-world scenarios and the methodology employed in transfer learning mainly affect the magnitude of  $C$ .

In the following sections, we describe the benefits and utility of analyzing scaling laws in transfer learning, based on three case studies of different material systems and their databases.

### Polymer property predictions with Sim2Real transfer learning using MD simulations

We demonstrate that the scaling law of Sim2Real transfer learning holds in polymer property prediction using all-atom classical MD simulations. The target properties to be predicted are the refractive index, density, specific heat capacity at constant pressure ( $C_p$ ), and thermal conductivity. Using RadonPy<sup>19</sup>, an open-source Python library developed to fully automate MD simulations of polymeric materials using LAMMPS (large-scale atomistic/molecular massively parallel simulator)<sup>26</sup>, we generated a source dataset comprising the four physical properties of approximately  $7 \times 10^4$  amorphous polymers (see Table S1 for the number of samples). Details of the MD calculations are provided in the Methods section. We randomly selected  $n$  samples from this dataset for the pretraining of neural networks, where  $n$  was varied across 10 equally spaced points on a logarithmic scale between 100 and the maximum number of samples.

The property predictor used a 190-dimensional descriptor vector that represents the compositional and structural features of the chemical structure of a polymer repeating unit. This vectorized polymer was mapped to each property using a conventional fully connected multi-layer neural network (see Fig. 1a and the Method section). With experimental data, we fine-tuned each pretrained neural network to a predictor of the experimental properties. The experimental datasets were extracted from the PoLyInfo database<sup>27–29</sup>. The number of polymers in each property dataset was 234 for refractive index, 607 for density, 104 for  $C_p$ , and 39 for thermal



**Fig. 1 | Transfer learning of polymer property predictions using all-atom classical MD simulations.** **a** Neural network architecture. **b** Scaling behavior of Sim2Real transfer for four different properties, namely refractive index, density, specific heat capacity ( $C_p$ ), and thermal conductivity. The horizontal axis represents the simulation data size, and the vertical axis shows the MAE averaged over 500 independent

trials with 90% confidence interval calculated by performing bootstrapping sampling. The dashed line is the estimated power-law with the estimated equation given at the bottom left, and the horizontal red line indicates the mean MAE for direct learning with no pretraining.

conductivity. To transfer a pretrained model to each target domain, we randomly selected 80% of the experimental datasets and evaluated the model's predictive performance on the remaining samples. This process was repeated 500 times independently for each  $n$ , observing scaling behaviors with the average of the mean absolute errors (MAEs) with their 90% confidence interval calculated by performing bootstrapping sampling.

As shown in Fig. 1b, the empirical generalization error for the experimental refractive index decays almost linearly on a logarithmic scale across the observed range of  $n$ . The parameters for power-law scaling were estimated as  $D = 0.0684$ ,  $\alpha = 0.0145$ , and  $C = 1.75 \times 10^{-11}$ . For the density, the prediction error also linearly decreases, and as  $n$  grows infinitely large, the MAE is expected to approach zero. The prediction error for  $C_p$  linearly decreases until around  $n = 10^4$ , after which the decay begins to slow. Regarding the thermal conductivity, the generalization performance rapidly improves until  $n < 10^4$ , followed by a plateau as  $n$  further increases. In summary, all tasks are notably scaled as the volume of MD-calculated data increases. Moreover, as shown in Fig. 1b, the generalization performance of transfer learning notably surpasses that of direct learning without transfer. The potential cross-domain transferability becomes more evident when contrasting direct and transfer learning based on scaling behavior rather than at a fixed  $n$ .

For the refractive index and density, the MD-calculated values exhibit remarkably high consistency with the experimental observations from our previous study<sup>19</sup>. Therefore, the observed strong scaling is verified because increasing the amount of simulation data directly improves the generalization performance for real-world scenarios. The  $C_p$  calculations with the classical MD simulations (neglecting quantum effects) introduced systematic biases compared to the corresponding experimental values, resulting in significant overestimation of the  $C_p$ <sup>19</sup>. Furthermore, the effect of random sampling of the initial structures in the MD simulations is more pronounced for the MD-calculated  $C_p$  than for the refractive index and density. Similarly, there are slight systematic biases and inherently large fluctuations in the MD-calculated thermal conductivity values. Hence, these findings suggest that simulation uncertainty due to the randomness of initial

structures is one of the critical factors influencing the scaling strength. This aspect will be discussed in more detail later.

Furthermore, we investigated the relationship between the similarity of polymers in the source and target datasets and the scaling behavior. By artificially creating multiple source datasets with varying degrees of task similarity (Fig. S1 in the Supplementary Information), we observed the scaling behavior. The results showed no significant differences (Fig. S2 in the Supplementary Information). This suggests that a greater structural similarity between the datasets does not necessarily lead to stronger scaling.

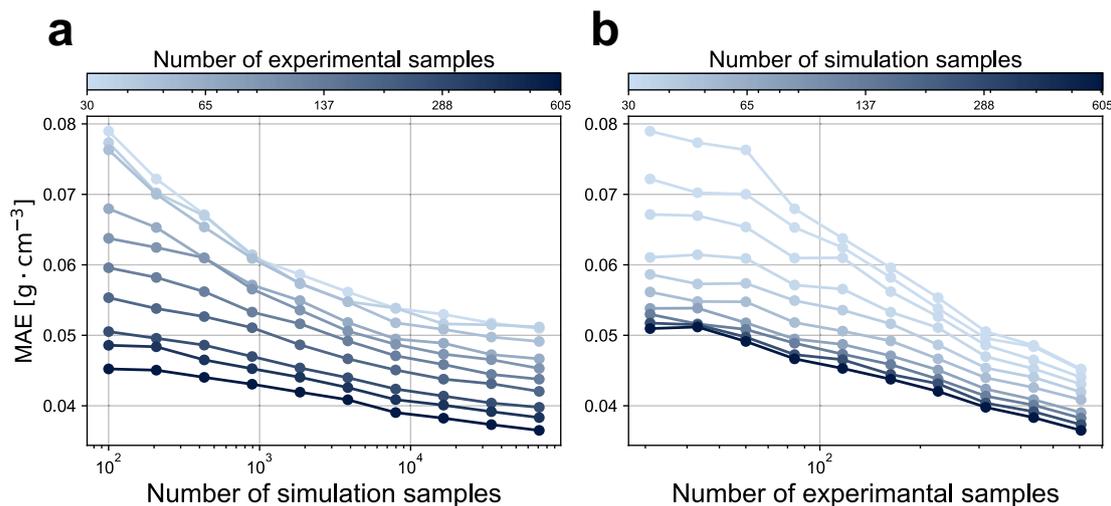
Here, we discuss the multidimensional scaling of simulated and experimental data. We examined the scaling behavior of Sim2Real predictions for the density while simultaneously varying the sizes of simulation and experimental datasets, as shown in Fig. 2. The empirical generalization errors for both types of data show a monotonic decreasing trend. In particular, the increase in the size of the experimental dataset results in a significantly larger gain than the increase in simulation dataset size. The power-law curve in Eq. (1) was estimated as follows:

$$R(n, m) = (0.0192 + 0.338n^{-0.265})m^{-0.239} + 0.0535 \quad (3)$$

As predicted theoretically, the scaling effect on the simulation data weakens as the experimental dataset size increases. Likewise, it was confirmed that the scaling effect on experimental data also decreases as the amount of simulation data increases.

Furthermore, by applying the concept of a marginal rate of substitution from microeconomics<sup>30</sup> to this estimated surface, we estimated the number of simulation samples equivalent to one experimental sample. For example, at the current maximum sample sizes of  $n = 71,068$  and  $m = 601$ , the partial derivatives of the estimated error function, analogous to marginal utility in economic theory, are given as follows:

$$\left. \frac{\partial R}{\partial n} \right|_{n=71068, m=601} = -1.41 \times 10^{-8}, \quad \left. \frac{\partial R}{\partial m} \right|_{n=71068, m=601} = -3.13 \times 10^{-6}.$$



**Fig. 2 | Multidimensional scaling of Sim2Real transfer learning, illustrated by the density prediction of amorphous polymers. a** Scaling to increase the amount of simulation data across various experimental dataset sizes, and **b** scaling to increase

the amount of experimental data for different sizes of simulation datasets. Each line represents the MAE averaged over 500 independent trials.

Taking the ratio of these coefficients provides an estimate of the marginal rate of substitution between experiments and simulations. Specifically, on the set of  $(m, n)$  pairs that maintain the same level of generalization error  $R(n, m) = r$  (referred to as indifference curves in microeconomics), the marginal rate of substitution  $dm/dn$  is given by  $\frac{dm}{dn} = -\frac{\partial R}{\partial n} / \frac{\partial R}{\partial m}$  (see the Method section). In this case, 221 simulation samples are equivalent to one experimental sample.

### Sim2Real multitask learning for polymer–solvent miscibility

While the theoretical implications presented by Mikami et al.<sup>25</sup> were derived under the assumption of neural fine-tuning, here we explored the scaling behavior for Sim2Real multitask learning scenarios. The task is to predict the Flory–Huggins  $\chi$  parameter between any given polymer and solvent, which is a critical dimensionless quantity governing the miscibility of polymer–solvent binary mixtures. The dataset comprises  $\chi$  parameters for 9575 polymer–solvent pairs calculated via COSMO-RS simulations based on density functional calculations<sup>31</sup>, which were generated in our previous work<sup>2</sup>, and 1,190 experimentally observed  $\chi$  parameters for 766 unique polymer–solvent pairs compiled from Orwoll and Arnold<sup>32</sup>. Aoki et al.<sup>2</sup> demonstrated that integrating both simulated and experimental  $\chi$  parameters into multitask learning significantly enhanced the generalization capability of the resulting predictors for experimental  $\chi$  parameters. In particular, this strategy effectively addressed limitations of molecular diversity, data size constraints, and inherent distributional biases in the experimental dataset.

We slightly modified the model structure developed by Aoki et al.<sup>2</sup> as the multitasking network architecture inspired by domain knowledge, known as the Hansen solubility parameter, as shown in Fig. 3. The 325-dimensional descriptor encodes the chemical structure of a given polymer repeating unit or solvent. Specifically, it comprises a 190-dimensional kernel mean force field descriptor<sup>33</sup> and a 135-dimensional RDKit descriptor<sup>34</sup>, where irrelevant features with zero variance within the given dataset were removed from the 207-dimensional RDKit descriptor. Additionally, we included a binary flag representing polymer (1) or solvent (0), temperature  $T$ , and its inverse  $1/T$  as additional inputs. For a given polymer or solvent, the input descriptor was passed through three hidden layers to map it to a 32-dimensional latent space. The distance between the polymer and solvent in this latent space was calculated, and two separate head networks were employed to output the experimental and simulated  $\chi$  parameters. See the Methods section for further details on the model structure.

To assess the generalization performances, 20% of the experimental data was randomly allocated as a test set, while the remaining samples, along

with  $n$  randomly selected simulation data points, were used for model training. This procedure was repeated 100 times independently with different data splits. The  $n$  ranged across 10 evenly spaced points on a logarithmic scale within the interval [100, 9575].

Figure 4 a shows the observed scaling curves, which exhibit strong linear decay in the generalization error on a logarithmic scale. The estimated  $C$  was  $4.52 \times 10^{-10}$ , suggesting that expanding COSMO-RS simulations can yield high-performance predictors for real systems. This observation suggests that scaling behavior occurs even in multitask learning, although the Sim2Real scaling law was theoretically derived for fine-tuning scenarios in Mikami et al.<sup>25</sup>.

Here, we discuss the multidimensional scaling of multitask learning. Figure 4b, c. describe the observed scaling behaviors when simultaneously varying the sizes of simulation and experimental datasets. Interestingly, unlike the fine-tuning results shown in the previous section, in multitask learning, as the experimental dataset size increases, the absolute gradient of the scaling curve becomes steeper. Similarly, by increasing the simulation dataset size, the improvement in generalization performance per experimental sample also increases. In other words, the experimental results suggest a mechanism where simulations and experiments mutually enhance their impact on improving generalization performance through synergistic effects.

This observation differs from the theoretical implication of Eq. (1). This is thought to be due to the difference in the choice of fine-tuning and multitask learning. Consequently, the parameter estimation for two-dimensional scaling with Eq. (1) is invalid. Instead, the equivalent sample size was calculated by approximating the gradient based on the observed increment of the MAE for increasing simulation and experimental dataset sizes. The estimated scaling curves and the gradients at the current dataset sizes ( $n = 9129$  and  $m = 612$ ) were computed respectively as follows:

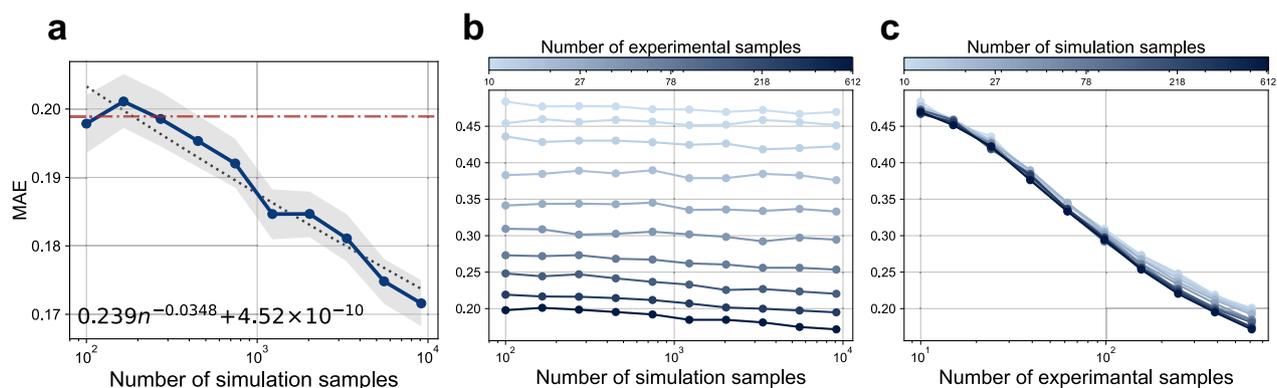
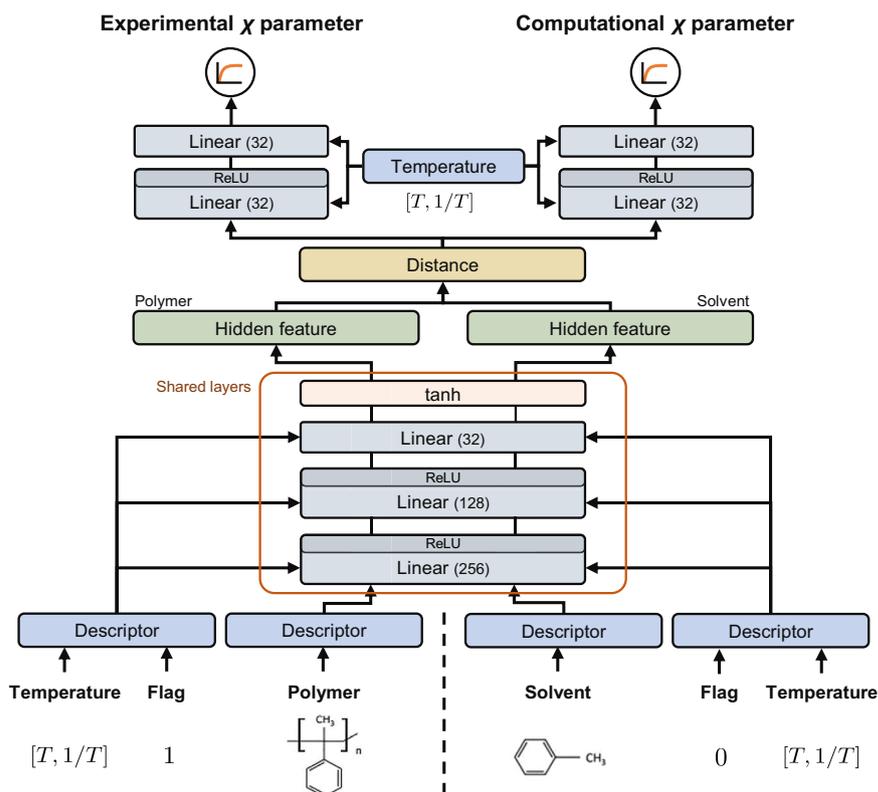
$$R(n, 612) = 0.239n^{-0.0348} + 4.52 \times 10^{-10} \left. \frac{\partial R(n, 612)}{\partial n} \right|_{n=9129} = -6.63 \times 10^{-7},$$

$$R(9129, m) = 0.863m^{-0.240} + 6.36 \times 10^{-16} \left. \frac{\partial R(9129, m)}{\partial m} \right|_{m=612} = -7.25 \times 10^{-5}.$$

By taking the ratio of these gradients, the marginal rate of substitution was estimated to be 109, indicating that for the  $\chi$  parameter prediction task, one real-world experiment is worth 109 COSMO-RS simulations.

Although we investigated whether the overall generalization performance scales across various materials as a whole, it is important to verify whether individual materials scale or not. Figure 5a summarizes the

**Fig. 3 | Model architecture of Sim2Real multitask learning used for predicting the Flory–Huggins  $\chi$  parameter.** The multitasking network architecture was inspired by domain knowledge, known as the Hansen solubility parameter, by calculating the distance between the polymer and solvent in the 32-dimensional latent space.

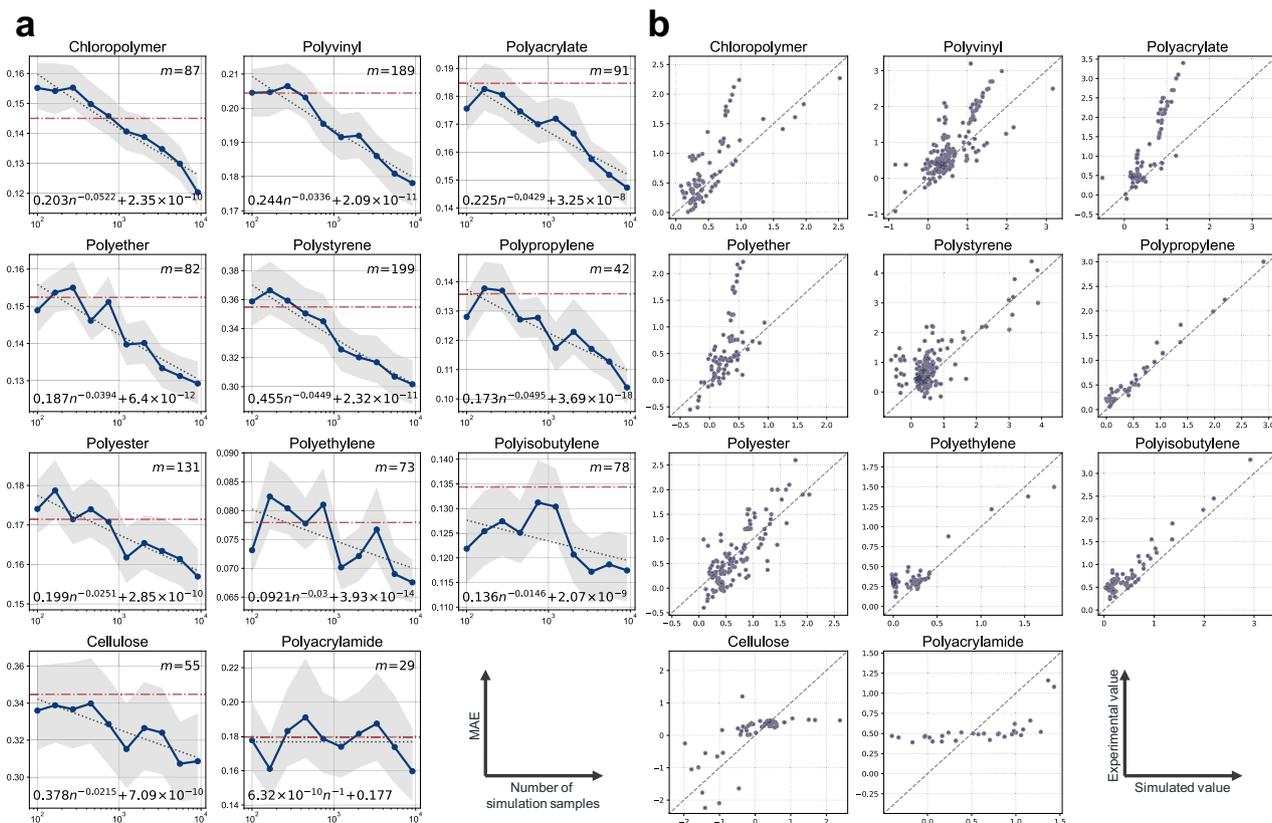


**Fig. 4 | Scaling law observed in the Flory–Huggins  $\chi$  parameter prediction task.** **a** Scaling behavior when increasing the simulation dataset size. The horizontal axis represents the number of polymer–solvent pairs used as the simulation dataset, and the vertical axis shows the average MAE of 100 independent trials with 90% confidence interval calculated via bootstrapping. The dashed line is the estimated

power-law with the estimated equation given at the bottom left, and the horizontal red line indicates the average MAE for direct learning without pretraining. **b** Scaling behaviors across different sizes of experimental data, and **c** scaling to increase the experimental dataset for different simulation dataset sizes. Each line shows the average MAE over 100 trials.

observed scaling behaviors to increasing  $n$  for different polymer classes, where test instances of polymer–solvent pairs were classified into 11 classes based on structural features of the polymers. The generalization performances of chloropolymers, polyvinyls, polyacrylates, polyethers, and polystyrenes were strongly scaled, while, other materials showed almost no improvement, e.g., polyacrylamides. Observing the scalability of each material class provides valuable insights for planning data generation. In the development of a simulation database, limited computational resources should be allocated primarily to scalable polymer classes. For non-scalable polymer classes, some modifications are needed in the data production protocol to improve scalability. To devise strategies, it is important to identify the governing factor that determines the scalability. Fig. 5b shows

parity plots of  $\chi$  parameters obtained from COSMO-RS simulations and experimental values for each polymer class. In comparison with the scaling behaviors shown in Fig. 5a, it is evident that the observed scalability of polymer species can be largely explained by the predictive capability of COSMO-RS simulations. However, in some polymer classes such as polystyrenes and polyesters, where the predictive ability of COSMO-RS simulations is weak, the generalization performance of transferred models scales strongly, reaching levels far beyond those of the simulations. This indicates the essence of Sim2Real transfer. Additionally, it is important to note that many of the non-scalable polymer classes had extremely limited experimental data (see Fig. 5a). For example, the  $m$  values for celluloses and polyacrylamides are 55 and 29, respectively. In such cases, model training



**Fig. 5 | Observation of Sim2Real scaling behaviors and predictive capability of COSMO-RS simulations for different polymer classes in the  $\chi$  parameter prediction task. **a** Observation of Sim2Real scaling behaviors for different polymer classes in the  $\chi$  parameter prediction task. Test instances of polymer–solvent pairs**

were classified into 11 classes based on structural features. The  $m$  value is denoted in the upper-right corner of each panel. **b** Predictive capability of COSMO-RS simulations (horizontal axis) against experimental values (vertical axis) for each of the 11 polymer classes in the  $\chi$  parameter predictions.

becomes extremely challenging. Furthermore, empirical generalization errors approximated with the small sample sets may significantly deviate from true generalization errors. Even if a polymer class appears to be non-scalable, it cannot be conclusively determined that there is no transferability or scalability.

### Transfer learning for thermal and electrical conductivity of inorganic materials

By definition, the scaling laws of transfer learning hold not only for Sim2Real scenarios but also for real-to-real (Real2Real) transfer scenarios. Here, we highlight an interesting aspect of the scaling analysis by showing the Real2Real scaling behavior in transfer learning from thermal conductivity to electrical conductivity for inorganic compounds.

We compiled a dataset from Starrydata<sup>35</sup>, comprising 5910 inorganic compounds with experimentally observed thermal conductivities and 3640 compounds with experimental electrical conductivities, all derived at 300 K. Starrydata is a comprehensive experimental database of thermoelectric materials that was collected from published papers. Figure 6b illustrates the dependency and discrepancy between the two physical properties across 1757 materials, where both thermal and electrical conductivity measurements were obtainable. According to the WF law<sup>36</sup>, the ratio of thermal conductivity ( $\kappa$ ) to electrical conductivity ( $\sigma$ ) of a metal is proportional to temperature ( $T$ ), expressed as:

$$\frac{\kappa}{\sigma} = LT,$$

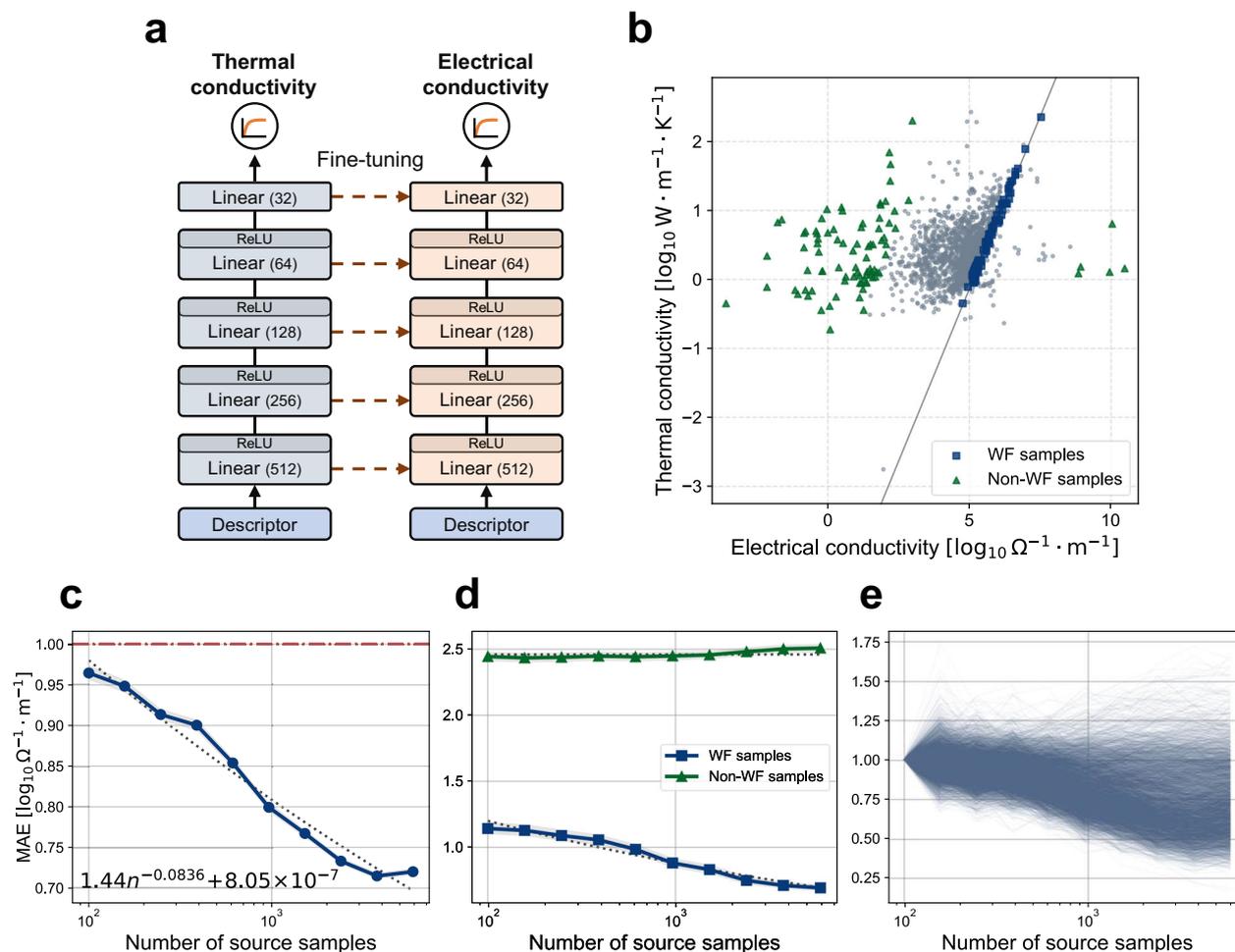
where  $L = 2.44 \times 10^{-8} W \Omega K^2$  is the Lorentz number. The gray line in Fig. 6b depicts the WF law on the joint distribution of thermal and electrical conductivities at 300 K. While the WF law holds for metallic materials, where

the free electrons are mainly responsible for both of these properties, it does not necessarily hold for non-metallic materials. Since the data includes both metallic and non-metallic materials, some of the samples deviated from this line.

For model building, we encoded the compositional features of an input compound into a 580-dimensional kernel mean descriptor<sup>33</sup>. Subsequently, a fully connected neural network was pretrained to learn the mapping from the vectorized composition to thermal conductivity. The network architecture is illustrated in Fig. 6a. The  $n$  used to train the thermal conductivity predictor was increased logarithmically in 10 steps from 100 to 5910. During the transfer learning phase, 80% of the electrical conductivity data were randomly selected for fine-tuning, and the remaining 20% served as the test set to evaluate the performance of the transferred model. This procedure was repeated 500 times with different randomly selected sample sets.

Figure 6c shows the observed scaling behaviors. The predictive performance improved linearly on the logarithmic scale, with the estimated power-law function  $1.44n^{-0.0836} + 8.05 \times 10^{-7}$ . Since the WF law holds for metallic materials, the transfer is expected to be more successful for metallic materials than for non-metallic ones. To investigate the difference in the transferability for metallic and non-metallic materials, we extracted samples that followed the WF law (blue square dots in Fig. 6b) and those that deviate from it (green triangular dots in Fig. 6b). Fig. 6c shows the scaling behaviors separately for each of the two sample sets. As expected, strong scaling was observed for materials for which the WF law holds, but for non-metallic materials that deviate from this law, increasing the amount of thermal conductivity data did not improve predictive performance.

Furthermore, observing transferability individually for different materials allows us to infer the presence or absence of common physical mechanisms between different physical systems. Fig. 6e illustrates the scaling behaviors of all test cases, which clearly distinguishes material species



**Fig. 6 | Real2Real transfer learning from thermal to electrical conductivities.** **a** Model architecture. **b** Parity plot showing values of electrical conductivity (horizontal axis) and thermal conductivity (vertical axis) at 300 K on a logarithmic scale. The dashed line represents the WF law. Blue square dots represent the top 10% of compounds with the smallest deviation from the WF rule (WF samples), while green triangular dots correspond to the top 10% of compounds with the largest deviation (non-WF samples). **c** Scaling behavior. The horizontal axis represents the source data size of the electrical conductivity prediction task, while the vertical axis shows

the average MAE for the thermal conductivity prediction task over 500 independent trials (solid line) with the standard deviation and 90% confidence interval calculated using bootstrapping. The black dashed line represents the estimated power-law (equation provided at the bottom left), and the red dashed line indicates the MAE for direct learning. **d** Scaling behavior for the two extracted datasets. The line color corresponds to the color of the dots in the parity plot in (b). **e** Scaling behavior for each of the 3640 compounds in the dataset.

where transfer does or does not scale. The presence or absence of scaling laws and the observed scaling strength could be used to characterize individual materials. Moreover, observing individual transferability provides valuable insights for planning data generation. The overall average performance transitioned into a plateau around  $n = 4 \times 10^3$  (Fig. 6c). However, there were several material groups where predictive performance continued to improve logarithmically; for example, metallic materials. (Fig. 6d, e). Intuitively, it would be efficient to halt the production of source data for material groups where improvement has plateaued and reallocate resources to groups more likely to scale. Analyzing only the overall average generalization performance overlooks the existence of material groups with the potential to scale even further.

## Discussion

This study discussed the significance and utility of analyzing the scalability in Sim2Real and Real2Real transfer learning in materials science. Across diverse case studies encompassing polymers and inorganic materials, it was consistently observed that as the size of the computational pretraining data set increases, the prediction error relative to the experimental data improves according to a power-law relationship. These findings highlight the

importance of synergistic effects between computational and experimental approaches. By observing the scaling law for Sim2Real transfer, we can estimate the required size of computational datasets to achieve the desired predictive performance in downstream real-world tasks. Additionally, we provide a microeconomic framework for determining the optimal allocation of computational and experimental resources during the creation of data platforms by analyzing multi-dimensional scaling behaviors. This approach guides decisions related to the allocation of resources for data collection efforts for maximum impact.

The scaling laws of transfer learning provide guiding principles for designing computational databases. It is desirable to create transferable computational databases that scale the generalization performance of downstream tasks for specified target tasks in the real-world domain. Alternatively, it is important to discover real-world tasks and analytical workflows that can be transferred scalably from computational databases. While various computational material-property databases have been developed to date, there are no reported cases of the values being quantified from the perspective of scaling laws. Strong scalability of transfer to diverse real-world tasks serves as a measure of the usefulness of the computational database.

It is important to see that discrepancies always exist between simulated and experimental properties. Additionally, experimental data are subject to biases and fluctuations due to unobserved factors related to the experimental conditions, sample fabrication, noise in measurement systems, and selection bias of the researchers. Therefore, transfer learning plays a key role in bridging the gap between complex and uncertain real-world scenarios and imperfect computational models. To this end, it is crucial to explicitly demonstrate the transferability and benefits of expanding datasets to downstream tasks. Finding a scheme with the scalability of Sim2Real transfer is a goal of developing materials databases using simulated data.

## Methods

### PoLyInfo polymer property datasets

Experimental property datasets for refractive index, density,  $C_p$ , and thermal conductivity were extracted from the polymer property database PoLyInfo<sup>27–29</sup>. The data for density, specific heat capacity, and thermal conductivity were restricted to measurements in amorphous states near room temperature (273–323K). The number of polymers in each property dataset was 39 for thermal conductivity, 104 for  $C_p$ , 607 for density, and 234 for refractive index.

### RadonPy polymer property datasets

To construct the simulation datasets, all-atom classical MD simulations were conducted using RadonPy, a Python library that automates polymer property calculations through high-throughput MD simulations<sup>19</sup>. Input parameters include the chemical structure of polymer repeating units represented by a simplified molecular input line entry system (SMILES)<sup>37</sup>, polymerization degree, number of polymer chains forming a simulation cell, temperature, and pressure. The automated calculation workflow consists of the following steps: (1) conformation search for a monomer with the given repeating unit, (2) calculation of electronic properties such as atomic charges using the density functional theory (DFT) method, (3) search for initial configuration of polymer chains using the self-avoiding random walk algorithm, (4) assignment of force field parameters using the general Amber force field version 2 (GAFF2), (5) generation of isotropic amorphous cells, (6) MD simulations for equilibration, (7) determination of whether to reach equilibrium, (8) non-equilibrium MD (NEMD) simulations for thermal conductivity calculation, and (9) property calculation in the post-processing stage. The DFT calculations and the MD simulations were executed using Psi4<sup>38</sup> and LAMMPS, respectively, within the RadonPy interface. An amorphous cell was created with 10 polymer chains comprising approximately 10,000 atoms. Following the initial configuration of polymer chains using the self-avoiding random walk and a 1 ns NVT simulation, the simulation cell was packed isotropically to achieve a density of 0.8 g cm<sup>-3</sup> at 700 K. The amorphous cell was equilibrated following Larsen's 21-step compression/decompression equilibration protocol<sup>39</sup>, undergoing temperature cycles between 300 and 600 K, repeating the ascent and descent for stabilization. After completing the 21-step equilibration process, NpT simulations were conducted for over 5 ns at 300 K and 1 atm until reaching equilibrium. The property calculation methods for the density, specific heat capacity at constant pressure, refractive index, and thermal conductivity were described previously<sup>19</sup> and detailed in the Supplementary Information.

In collaboration with an academia–industry consortium, we generated the property datasets of approximately  $7 \times 10^4$  linear polymers in amorphous states using RadonPy (see Table S1). The virtual polymers were generated using an N-gram-based polymer structure generator<sup>40</sup> for each of the 20 polymer classes, such as polyimides, polyesters, and polystyrenes, following the classification rule established by PolyInfo. The chemical structure  $X$  of an existing compound used in the training dataset is described by the SMILES representation, where  $X$  is represented by a string of length  $p$  as  $X = x_1x_2 \dots x_p$ . By using the string set of synthesized polymers belonging to each polymer class, an N-gram language model was trained to obtain a structure generator that mimicked the patterns, such as frequent fragments and appropriate chemical bonding rules, observed for the existing polymers. The 20 class-specific SMILES generators were used to create the virtual

library. A list of the 20 polymer classes with their dataset sizes is provided in Table S1.

### Equivalent sample size for experimental and simulation data

Differentiating the generalization error  $R(n, m)$  in Eq. (1) with respect to  $n$  and  $m$ , we obtain the following expression:

$$dR(n, m) = \frac{\partial R}{\partial n} dn + \frac{\partial R}{\partial m} dm. \quad (4)$$

On the set of equivalent samples  $(n, m)$  that maintain the same level of  $R(n, m) = r$ ,  $R(n, m)$  remains constant at  $r$ , thus satisfying  $dR(n, m) = 0$ . Therefore,  $\frac{dm}{dn} = -\frac{\partial R}{\partial n} / \frac{\partial R}{\partial m}$  holds.

### Polymer–solvent solubility datasets

Aoki et al.<sup>2</sup> used the experimental values of the  $\chi$  parameter for 1190 polymer–solvent pairs, consisting of 46 different polymers and 140 different solvent molecules, to train the model. The data were compiled from a supplementary table of Orwoll and Arnold<sup>32</sup>. The dataset also included measurements of the  $\chi$  parameter for different temperatures and polymer–solvent compositions. The molecular species of the polymers/solvents in the dataset were distributed over a limited region of the entire chemical space. In addition, in certain experimental systems, it is difficult to measure the  $\chi$  parameters of the polymer–solvent system in an immiscible state, resulting in a significant bias in the distribution of the data. Therefore, models trained using only this dataset generally have narrow predictive applicability.

To tackle this issue, we utilized the COSMO-RS simulation<sup>41–44</sup> to generate a dataset of  $\chi$  parameters for 9129 pairs of polymers and solvents at the BP-SVP-AM1 level<sup>2</sup>. The calculations were performed using the TURBOMOLE<sup>45</sup> and COSMOtherm<sup>46</sup> software packages for creating COSMO files by density functional calculations and the calculations of  $\chi$  parameters from the COSMO files, respectively. For polymers, a structure comprising three repeating units was created, in which the two endpoints were replaced by methyl groups. After creating the COSMO files, the COSMOmeso function was executed to calculate the  $\chi$  parameters using the activity coefficients obtained from the COSMO files.

### Data preprocessing

In all experiments, variable transformations were applied to the model inputs and outputs to enhance the efficiency of machine-learning model training. The methods included logarithmic transformation, normalization, Yeo–Johnson transformation<sup>47</sup>, and min–max transformation (i.e., scaling each feature to a range of [0, 1]). These methods were tailored for the input and output variables in each of the three applications, as summarized in Table 1.

**Table 1 | Data preprocessing methods for input and output variables in three different applications: polymer property prediction using RadonPy (RadonPy), multitask learning of polymer–solvent miscibility ( $\chi$  parameter), and transfer learning between thermal and electrical conductivities using Starraydata (Starrydata)**

	RadonPy	$\chi$ parameter	Starrydata
Input	log → norm → YJ	FFKM <sup>1</sup> : log → norm → YJ → 0-1	log → norm → YJ
		RDKit + temp + inv-temp <sup>2</sup> : 0-1	
Output	norm → YJ	norm → YJ	norm → YJ → 0-1

A combination of four methods — logarithmic transformation (log), normalization (norm), Yeo–Johnson transformation (YJ), and min–max transformation (0-1) — was applied in the order described below.

<sup>1</sup> Force-field kernel mean (FFKM) descriptor<sup>33</sup>.

<sup>2</sup> Concatenation of RDKit descriptor<sup>34</sup> (RDKit), thermodynamic temperature (temp), and inverse temperature (inv-temp).

**Table 2 | Hyperparameter settings for model training in the three applications: RadonPy,  $\chi$  parameter, and Starrdata**

	RadonPy	$\chi$ parameter	Starrydata
Optimizer	Adam <sup>48</sup>	Adam <sup>48</sup>	Adam <sup>48</sup>
Learning rate	0.001	0.001	0.001 (source task) 0.0001 (target task)
Batch size	32	16	32
Early stopping patience	5	10	10

In the fine-tuning experiments (RadonPy and  $\chi$  parameter), the same hyperparameters were used for the source and target tasks, except for the learning rate in Starrdata.

### Model fine-tuning

In Sim2Real transfer learning using RadonPy and transfer learning between thermal and electrical conductivities, neural fine-tuning was employed. Specifically, the weights of the neural networks pretrained on the source data (MD-calculated property data or experimental data for thermal conductivity) were used as initial values and updated with the target data (experimental observation of polymeric properties or electrical conductivity data). The hyperparameters, such as learning rate and batch size, are listed in Table 2.

### Multitask learning

For the  $\chi$  parameter prediction task, we employed multitask learning with empirical risk minimization as follows:

$$\min_{\lambda} \frac{\lambda}{|\mathcal{D}_{\text{sim}}|} \sum_{(\chi, T, p, s) \in \mathcal{D}_{\text{sim}}} \{\chi - f(T; p, s)\}^2 + \frac{1-\lambda}{|\mathcal{D}_{\text{exp}}|} \sum_{(\chi, T, p, s) \in \mathcal{D}_{\text{exp}}} \{\chi - f(T; p, s)\}^2,$$

where  $\mathcal{D}_{\text{sim}}$  and  $\mathcal{D}_{\text{exp}}$  denote the dataset of  $\chi$  parameters obtained by the COSMO-RS simulation and the experimental dataset, respectively. The neural network model  $f(T, p, s)$  is a function of temperature  $T$ , polymer  $p$ , and solvent  $s$ . The first term fits the simulated dataset, while the second term fits the experimentally observed dataset. The hyperparameter  $\lambda$  controls the relative importance between these two terms. In this study, we set  $\lambda = 0.5$ , which is consistent with the value employed in Aoki et al.<sup>2</sup>, resulting in learning from both simulation and real systems with equal importance. Other hyperparameters are listed in Table 2.

### Data availability

The data supporting the findings of this study will be made available upon reasonable request to the corresponding author. The datasets of the experimental and computational  $\chi$  parameters can be accessed via Figshare [https://github.com/yoshida-lab/MTL\\_ChiParameter](https://github.com/yoshida-lab/MTL_ChiParameter). The datasets of thermal and electrical conductivity are accessible through Starrdata [https://figshare.com/projects/Starrydata\\_datasets/155129](https://figshare.com/projects/Starrydata_datasets/155129).

### Code availability

The code for multitask learning for the  $\chi$  parameter prediction task is available at the Github [https://github.com/yoshida-lab/MTL\\_ChiParameter](https://github.com/yoshida-lab/MTL_ChiParameter). Other codes are available upon reasonable request to the corresponding author.

Received: 30 August 2024; Accepted: 12 March 2025;

Published online: 24 May 2025

### References

- Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
- Aoki, Y. et al. Multitask machine learning to predict polymer–solvent miscibility using Flory–Huggins interaction parameters. *Macromolecules* **56**, 5446–5456 (2023).
- Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Comput. Mater.* **5**, 66 (2019).
- Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
- Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
- Rao, Z. et al. Machine learning-enabled high-entropy alloy discovery. *Science* **378**, 78–85 (2022).
- Zhong, M. et al. Accelerated discovery of CO<sub>2</sub> electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
- Kim, M. et al. Artificial intelligence to accelerate the discovery of N<sub>2</sub> electroreduction catalysts. *Chem. Mater.* **32**, 709–720 (2019).
- Liu, C. et al. Machine learning to predict quasicrystals from chemical compositions. *Adv. Mater.* **33**, 2102507 (2021).
- Liu, C. et al. Quasicrystals predicted and discovered by machine learning. *Phys. Rev. Mater.* **7**, 093805 (2023).
- Uryu, H. et al. Deep learning enables rapid identification of a new quasicrystal from multiphase powder diffraction patterns. *Adv. Sci.* **11**, 2304546 (2024).
- Martin, T. B. & Audus, D. J. Emerging trends in machine learning: a polymer perspective. *ACS Polym. Au* **3**, 239–258 (2023).
- Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
- Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
- Saal, J. E., Kirklín, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom* **65**, 1501–1509 (2013).
- Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 1–15 (2015).
- Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1–7 (2014).
- Hayashi, Y., Shiomi, J., Morikawa, J. & Yoshida, R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput. Mater.* **8**, 222 (2022).
- Su, H., Qi, C. R., Li, Y. & Guibas, L. J. Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proc. IEEE international conference on computer vision*, 2686–2694 (IEEE, 2015).
- Movshovitz-Attias, Y., Kanade, T. & Sheikh, Y. How useful is photo-realistic rendering for visual learning? In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III 14*, 202–217 (Springer, 2016).
- Georgakis, G., Mousavian, A., Berg, A. C. & Kosecka, J. Synthesizing training data for object detection in indoor scenes. *Rob. Sci. Syst.* **043** (2017).
- Tremblay, J. et al. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 969–977 (IEEE, 2018).
- Ju, S. et al. Exploring diamondlike lattice thermal conductivity crystals via feature-based transfer learning. *Phys. Rev. Mater.* **5**, 053801 (2021).
- Mikami, H. et al. A scaling law for syn2real transfer: How much is your pre-training effective? In *Machine Learning and Knowledge Discovery in Databases*, 477–492 (Springer Nature Switzerland, 2023).

26. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
27. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. PoLyInfo: polymer database for polymeric materials design. *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22–29 (IEEE, 2011).
28. Ishii, M., Ito, T., Sado, H. & Kuwajima, I. NIMS polymer database PoLyInfo (I): an overarching view of half a million data points. *Sci. Technol. Adv. Mater. Methods* **0**, 2354649 (2024).
29. PoLyInfo. <https://polymer.nims.go.jp/>.
30. Varian, H. R. *Intermediate Microeconomics with Calculus: A Modern Approach*. (WW Norton & Company, New York, NY, 2014).
31. Loschen, C. & Klamt, A. Prediction of solubilities and partition coefficients in polymers using COSMO-RS. *Ind. Eng. Chem. Res.* **53**, 11478–11487 (2014).
32. Orwoll, R. A. & Arnold, P. A. *Polymer–Solvent Interaction Parameter  $\chi$*  (pp. 233–257. Springer New York, New York, NY, 2007).
33. Kusaba, M., Hayashi, Y., Liu, C., Wakiuchi, A. & Yoshida, R. Representation of materials by kernel mean embedding. *Phys. Rev. B* **108**, 134107 (2023).
34. RDKit: Open-source cheminformatics. <https://www.rdkit.org/>.
35. Katsura, Y. et al. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. *Sci. Technol. Adv. Mater.* **20**, 511–520 (2019).
36. Jones, W. & March, N. H. *Theoretical solid state physics*, vol. 35 (Courier Corporation, 1985).
37. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Computer Sci.* **28**, 31–36 (1988).
38. Smith, D. G. et al. PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108 (2020).
39. Larsen, G. S., Lin, P., Hart, K. E. & Colina, C. M. Molecular simulations of PIM-1-like polymers of intrinsic microporosity. *Macromolecules* **44**, 6944–6951 (2011).
40. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des.* **31**, 379–391 (2017).
41. Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).
42. Klamt, A., Jonas, V., Bürger, T. & Lohrenz, J. C. W. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **102**, 5074–5085 (1998).
43. Eckert, F. & Klamt, A. Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J.* **48**, 369–385 (2002).
44. Klamt, A. COSMO-RS: From quantum chemistry to fluid phase thermodynamics and drug design; Elsevier Science: Amsterdam (2005).
45. TURBOMOLE V7.5.1 2021, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007. <https://www.turbomole.org>.
46. BIOVIA COSMOtherm. Release 2022; Dassault Systèmes. <http://www.3ds.com>.
47. Yeo, I.-K. & Johnson, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954–959 (2000).
48. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (San Diego, 2015).

## Acknowledgements

We express our sincere gratitude to all members of ISM-MCC Frontier Materials Design Laboratory, a joint laboratory of Mitsubishi Chemical Corporation (MCC) and the Institute of Statistical Mathematics (ISM), for their valuable contributions to the discussion of this study. This research received support from MEXT as “Program for Promoting Researches on the Super-computer Fugaku” (project ID: hp210264), JST CREST (Grant Numbers JPMJCR19I3, JPMJCR22O3, JPMJCR2332), MEXT/JSPS KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (19H05820), Grant-in-Aid for Scientific Research (A) (19H01132), Grant-in-Aid for Research Activity Start-up (23K19980), and Grant-in-Aid for Scientific Research (C) (22K11949). Computational resources were provided by Fugaku at the RIKEN Center for Computational Science, Kobe, Japan (hp210264) and the supercomputer at the Research Center for Computational Science, Okazaki, Japan (project: 23-IMS-C113, 24-IMS-C107).

## Author contributions

R.Y. and S.M. devised the project, main conceptual ideas, and outline proof. S.M. and Y.H. implemented the machine-learning algorithms and conducted the experiments with the support of R.Y., S.W., H.S., and K.S. Y.H. performed the MD simulations using RadonPy to generate the polymer property data. K.S. generated the  $\chi$  parameter dataset using the COSMO-RS simulations. K.F. performed a theoretical analysis of the Sim2Real transfer learning. H.S. examined the results from a physicochemical point of view. M.I. and I.K. extracted and structured data from PoLyInfo. S.M. and R.Y. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01606-5>.

**Correspondence** and requests for materials should be addressed to Ryo Yoshida.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025