npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics

Yoshihiro Hayashi[1,2]✉, Junichiro Shiomi[1,2,3], Junko Morikawa[1,4] and Ryo Yoshida[1,5,6]✉

The spread of data-driven materials research has increased the need for systematically designed materials property databases. However, the development of polymer databases has lagged far behind other material systems. We present RadonPy, an open-source library that can automate the complete process of all-atom classical molecular dynamics (MD) simulations applicable to a wide variety of polymeric materials. Herein, 15 different properties were calculated for more than 1000 amorphous polymers. The MD-calculated properties were systematically compared with experimental data to validate the calculation conditions; the bias and variance in the MD-calculated properties were successfully calibrated by a machine learning technique. During the high-throughput data production, we identified eight amorphous polymers with extremely high thermal conductivity (>0.4 $W \cdot m^{-1} \cdot K^{-1}$) and their underlying mechanisms. Similar to the advancement of materials informatics since the advent of computational property databases for inorganic crystals, database construction using RadonPy will promote the development of polymer informatics.

npj Computational Materials (2022)8:222 ; https://doi.org/10.1038/s41524-022-00906-4

INTRODUCTION

Materials informatics (MI) is a growing interdisciplinary field of materials science, attracting significant attention in recent years. MI utilizes machine learning to model, predict, and optimize the properties of new materials[1,2]. Naturally, the most essential resource in MI is data. Hence, significant efforts have been made to develop open databases for inorganic materials and light-weight organic molecules, such as the Materials Project[3] (~140,000 inorganic compounds), the Automatic-Flow[4] (AFLOW: ~3,000,000 inorganic compounds), the Open Quantum Materials Database[5] (OQMD: ~1,000,000 inorganic compounds), and QM9[6] (~134,000 organic molecules). In particular, the huge databases of computational properties built using high-throughput first-principles calculations have brought remarkable progress in MI and their widespread use in science and technology. However, for polymeric materials, despite their industrial usefulness and unique characteristics, such as lightness, high tenacity, elasticity, and ease of processing, the development of open databases has considerably lagged behind other material systems[7]. This is due to the following reasons: (1) high costs of data production, (2) the difficulty in creating common data due to the diversity of polymeric materials in terms of structures and processing conditions, and (3) cultural barriers to avoiding information leakage to competitors[2]. In addition, the computational difficulty in performing high-throughput calculations and their high computational costs have hindered the development of computational property databases for polymeric materials.

PoLyInfo[8] is the current largest database of polymer properties, built from manually collected literature data. Currently, it contains ~100 properties of more than 18,000 polymers. However, the overall data in PoLyInfo are rather sparse as there are few cases where more than one property is simultaneously recorded for one polymer. Polymer Genome[9–12] is a database containing seven different electronic and optical properties of crystalline and single chain states of polymers from first-principles calculations and several experimental properties of amorphous polymers. The computational properties include the crystal bandgap (562 polymers), polymer chain bandgap (3881 polymers), static dielectric constant of polymer crystals (383 polymers), and refractive index of polymer crystals (383 polymers). A common feature of these databases is that they do not provide application programming interfaces (APIs) and therefore do not allow automatic batch downloading of the data. Therefore, the creation of data resources conducive to data-driven research is vital for advancing polymer informatics.

Large-scale data of computational properties have proven to be an essential resource for machine learning applications in MI. For example, such big data have been used as source data for transfer learning when dealing with limited data in materials research. Transfer learning represents a statistical methodology for reusing knowledge, data, or models acquired in one domain (source domain) to another (target domain)[13,14]. Suppose that directly establishing a machine learning model from scratch is difficult due to the lack of sufficient amount of experimental data, in such cases, a model is trained on a large amount of computational property data, and the pretrained model is fine-tuned using a small amount of experimental data, to build a highly accurate prediction model in the target domain. Successful examples of cross-domain transfer between computational and experimental data have been reported for various material systems[15–20],

[1]The Institute of Statistical Mathematics (ISM), Research Organization of Information and Systems, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. [2]Department of Mechanical Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan. [3]Institute of Engineering Innovation, The University of Tokyo, 2-11 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan. [4]Department of Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan. [5]Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan. [6]Department of Statistical Science, School of Multidisciplinary Science, The Graduate University of Advanced Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. ✉email: yhayashi@ism.ac.jp; yoshidar@ism.ac.jp

including our previous work on the prediction and synthesis of thermally conductive amorphous polymers using neural networks transferred from computational properties in which only 28 samples were available in the target domain[21].

For polymer properties, even computational data are rather scarce. Polymer Genome[9–12] is the only existing database, which is constructed using first-principles electronic structure calculations of polymers in crystalline states. However, currently, the number of samples is small, and the calculation is limited to seven electrical and optical properties. For the all-atom classical molecular dynamics (MD) simulations, which are powerful techniques for computing the equilibrium and non-equilibrium properties of the condensed-phase systems of polymeric materials, there are only a few reported works that have constructed large datasets with high-throughput calculations[22–24]. Afzal et al. created a dataset of 315 polymers using high-throughput MD simulations; however, the target properties were limited to the glass transition temperature (T_g) and thermal expansion coefficient[24]. To build a computational polymer property database, a workflow of high-throughput MD simulations should be established, which is considered technically challenging. The entire workflow of an MD simulation comprises several sub-modules, such as the specification of an empirical potential, the initialization of polymer chains, equilibrium and nonequilibrium MD simulations, and the calculation of the properties from simulated molecular trajectories, which complicate the job control and error handling when fully automating the workflow. While this workflow can be partially streamlined using the pysimm Python package[25], there is still no open-source software that facilitates the building of the entire workflow. In addition, various types of conditional parameters, such as the degree of polymerization, number of polymer chains, and annealing schedules, need to be determined appropriately. Furthermore, a unified platform is required to create various polymeric states such as amorphous structures, oriented structures, and polymer blends. It also requires vast computational resources. For example, an equilibrium MD simulation of a conventional amorphous polymer requires an average run time of more than 30–50 h based on our experiments conducted on a workstation with a dual CPU (Intel Xeon Gold 6148; 2.4 GHz) having 40 cores.

Herein, we present RadonPy (https://github.com/RadonPy/RadonPy), which is an open-source Python library for fully automated calculation, for a comprehensive set of polymer properties, using all-atom classical MD simulations. For a given polymer repeating unit with its chemical structure, the entire process of the MD simulation can be performed fully automatically, including molecular modeling, equilibrium and nonequilibrium MD simulations, automatic determination of the completion of equilibration, scheduling of restarts in case of failure to converge, and property calculations in the post-process step. In this first release, the library comprises the calculation of 15 properties, such as the thermal conductivity, density, specific heat capacity, thermal expansion coefficient, and refractive index, in the amorphous state. In this study, we calculated 15 properties for more than 1000 unique amorphous polymers. These calculated properties were systematically validated with respect to experimental values obtained from PoLyInfo. In particular, the focus here is on the thermal conductivity of polymers, which will be an important performance metric for designing polymeric materials used as insulating resins, molding resins, adhesives, and coating agents for mobile devices, given the increase in heat generation brought on by miniaturization and performance improvement of mobile devices. During the high-throughput data production, we computationally identified eight amorphous polymers with extremely high thermal conductivities (>0.4 $W \cdot m^{-1} \cdot K^{-1}$), including six polymers with unreported thermal conductivities. These polymers exhibited a high density of hydrogen bonding units or rigid, linear backbones. In addition, a decomposition analysis of the heat conduction, which is implemented in RadonPy, revealed the underlying mechanisms that yield such a high thermal conductivity: heat transfer via hydrogen bonds and dipole–dipole interactions between polymer chains having hydrogen bonding units or via covalent bonds of polymer backbones with high rigidity and linearity.

RESULTS AND DISCUSSION

Software overview

RadonPy is compatible with Python 3.7 to 3.9. RadonPy is designed to be used jointly with the chemoinformatics Python library RDkit[26], with high compatibility between the input/output systems of each module in RadonPy and those of RDKit. The input parameter set for RadonPy comprises a simplified molecular input line entry system (SMILES)[27] string with two asterisks representing the connecting points of a repeating unit, the polymerization degree, the number of polymer chains in a simulation cell, and temperature. Subsequently, the following processes are fully automated (Fig. 1): the conformation search for the repeating unit, calculation of the electronic properties, such as the atomic charge and dipole polarizability, based on the density functional theory (DFT), generation of initial configurations of polymer chains based on the self-avoiding random walk, assignment of the force field parameters, creation of a simulation cell such as an isotropic amorphous cell, MD simulation to equilibrate the system, determination of whether to reach equilibrium, execution of nonequilibrium MD simulation (NEMD) for thermal conductivity calculation, and calculation of various physical property values. RadonPy is mainly designed to run on a supercomputer; multiple polymers are calculated independently in parallel using many computation nodes in a supercomputer. The DFT and MD calculations were performed using the Psi4[28] and Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[29], respectively, through the RadonPy interface. Each step will be detailed in the Methods section (see also Supplementary Notes in the Supplementary Information for a "Getting started with RadonPy" guide).

With the current release, the following properties are calculated from the equilibrium calculations: density, radius of gyration (Rg), specific heat capacities at constant pressure (C_P) and at constant volume (C_V), isothermal/isentropic compressibility, isothermal/isentropic bulk modulus, volume expansion coefficient, linear expansion coefficient, self-diffusion coefficient, refractive index, static dielectric constant, and nematic order parameter. Thermal conductivity and thermal diffusivity are calculated from the NEMD.

RadonPy outputs and stores trajectory data, including atomic coordinates and velocities, and thermodynamic data in the text-based dump files of LAMMPS. Calculated physical properties are stored in CSV format. In addition, the final system state, including atomic coordinates and velocities, in the equilibration and the NEMD are saved as Python pickle files, allowing the final system state to be restored to restart further MD calculations.

Dataset

The PoLyInfo database contains 15,335 homopolymers, which have only organic 10 element species, H, C, N, O, F, P, S, Cl, Br, and I. Among these, we selected 1138 unique homopolymers as the calculation target in this study, for which as many experimental properties as possible were recorded. The selected polymer set was composed of a wide variety of polymer backbones, such as polystyrenes, polyvinyl, polyacrylates, polyamides, polycarbonates, polyurethanes, and polyimides. The validation data of the density, thermal conductivity, refractive index, specific heat capacity C_P, linear expansion coefficient, and volume expansion coefficient were collected from PoLyInfo. The data used were limited to homopolymers and those meeting the following conditions: their

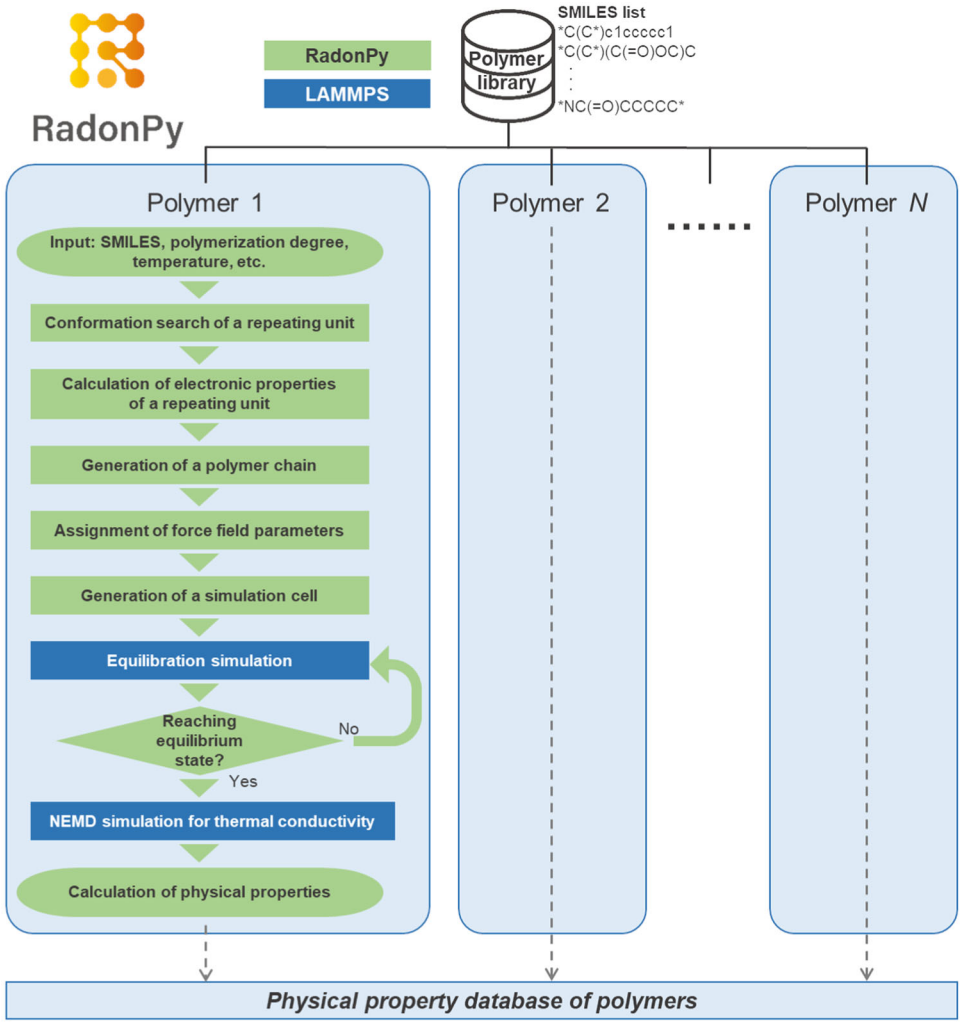


Fig. 1 Flowchart of the automated MD calculation workflow for polymer properties using RadonPy. RadonPy can automate each process to perform all-atom classical molecular dynamics simulations. Multiple polymers are calculated independently in parallel using many computation nodes in a supercomputer.

material type was labeled as one of neat resin, samples contained no additives, fillers, and dopants, the measured temperature was in the range of 273–323 K, the postforming state was amorphous or unidentified, and the topology of the polymers was linear or unidentified.

Distribution of calculated polymers in chemical space

The automated MD calculations were conducted for the 1138 homopolymers selected from the PoLyInfo database. Of the five independent calculations, the automatic calculations succeeded at least once for 1070 polymers, more than thrice for 1001 polymers, and in all the five cases for 759 polymers. The failed calculations were classified into four cases: the structural optimization of the DFT calculation did not converge, the MD simulation did not reach equilibrium, the system was partially oriented (nematic order parameter > 0.1), and the temperature gradient in the NEMD calculation did not become linear.

To investigate the distribution of the backbones of the 1070 calculated polymers over the 15,335 polymers in PoLyInfo, their chemical structures were visualized onto a 2D space using the uniform manifold approximation and projection (UMAP)[30]. The chemical structure of each polymer was transformed into a 2048 bit vector with an extended connectivity fingerprint with a radius of three atoms (ECFP6)[31]. To consider the repeating structure of polymers, the ECFP6 descriptor was constructed after generating the macrocyclic oligomer with 10-mer of the repeating unit. The UMAP with the Hamming distance was used to create the 2D representation of the 15,335 fingerprinted polymers, as shown in Fig. 2a, in which its subset corresponding to the 1070 polymers successfully calculated at least once is shown in Fig. 2b. The plot colors indicate the 21 classes of the polymer backbones according to the PoLyInfo database. The two distributions exhibited similar patterns in the UMAP plot, confirming no significant selection bias in the calculated polymers. In addition, the calculated polymers were selected to cover the 20 classes except for the class of others.

Validation of the calculated physical properties

To evaluate the performance of the automated MD pipeline, the calculated properties were systematically compared with the experimental values from PoLyInfo (Fig. 3). In the validation process, we used the 1001 polymers for which the automatic calculation was successfully completed at least thrice out of the five independent trials. We also examined the effect of the simulation box size on the calculated properties (see "Examination of box size effects" in the Supplementary Discussion in the

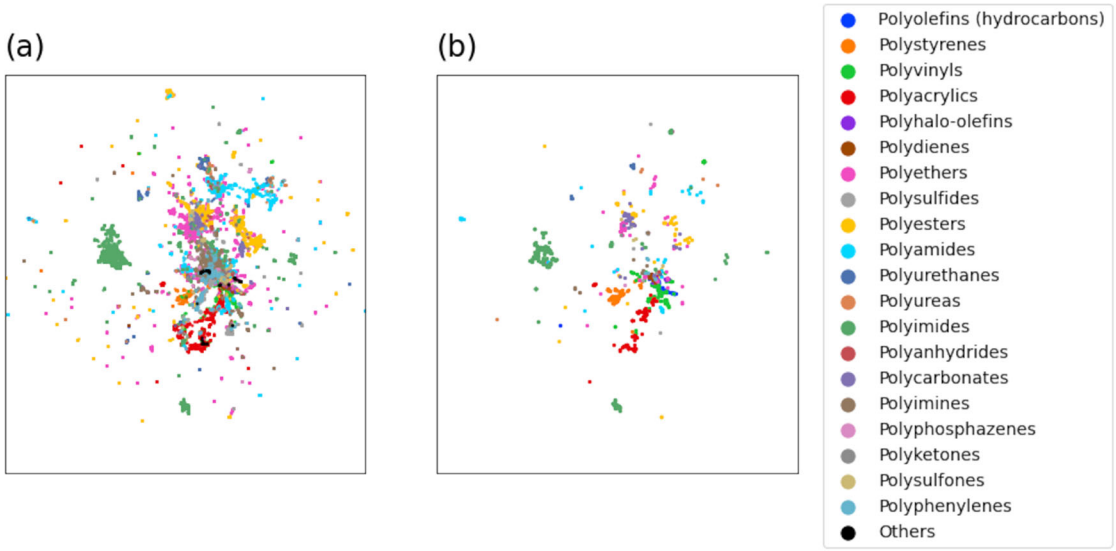


Fig. 2 UMAP plot visualizing the distribution of the polymer backbones. The UMAP plots show the distribution of (**a**) 15,335 homopolymers in PoLyInfo and (**b**) 1070 homopolymers calculated in this study. The 21 classes of the polymer backbones are color-coded according to the definition of PoLyInfo.

Supplementary Information). For the validation test, 28 different polymers were chosen by taking structural variations into account (Supplementary Fig. 11). The number of polymer chains varied from 10 to 50 with the number of atoms in each of a polymer chain set to 1000–2000. According to the experimental results, the box size had no significant effect on the calculated properties in these conditions (Supplementary Figs. 6–10).

The calculated density well reproduced the experimental values ($R^2 = 0.890$), albeit with a slight underestimation, as the slope of the fitted straight line in the parity plot was equal to 0.805 in Fig. 3a. The standard deviation (SD) of the calculated values in the five independent trials was low. The slight underestimation can be explained as follows: since the polymerization degree in the present MD simulations is smaller than that in the experimental conditions in PoLyInfo, the mobility of the simulated polymer chains becomes larger than that observed in the real systems, resulting in an overestimation of the free volume. In addition, unobservable partial crystallization behind experimental data could result in higher experimental density than in the amorphous states. Note also that, as reported in a previous study, the MD calculation of the density of the organic molecule liquids using the GAFF2 force field is often poorly performed in high-density regions[32]. On the other hand, in our calculation, such a discrepancy never occurred in the high-density regions. This is because of the use of the modified force field parameters developed by Träg and Zahn[33] for fluorocarbon polymers (see "Assignment of force field parameters" in the Methods section).

The calculated thermal conductivities also showed good agreement with the experimental values in PoLyInfo ($R^2 = 0.490$), as shown in Fig. 3b. However, the correlation was not so high. This could be because the experimental values of the thermal conductivity involve fluctuations due to differences in the measurement methods and temperature dependence. Moreover, there is a gap between the real and model systems, owing to the differences in various factors, including the degree of polymerization and its distribution, degree of orientation, crystallinity, impurities, and polymer chain entanglement. In addition, as the level of the thermal conductivity increases, the fluctuation in the calculated values within the independent trials increases significantly (Supplementary Fig. 1). Thus, for polymers with potentially high thermal conductivity, the number of independent trials of MD calculations should be increased to improve the accuracy.

The calculated refractive index well reproduced the PoLyInfo dataset ($R^2 = 0.809$) with a trivial underestimation where the slope of the fitted straight line in the parity plot was equal to 0.839 (Fig. 3c). The slight underestimation would arise from the reported underestimation of the density because the refractive index is defined to be the increasing function of the density. The variation in the MD simulations was quite small. It can be concluded that a sufficiently high prediction accuracy was obtained for the refractive index.

Figure 3d shows the correlation of C_P between the calculated and experimental values ($R^2 = 0.602$). The calculated C_P showed an evident overestimation as the fitted slope in the parity plot was 1.430. This observation is inevitable in the classical MD because classical MD calculations do not include quantum effects: the vibrational energy in a classical harmonic oscillator is significantly higher than that in a quantum harmonic oscillator at the same frequency. The quantum-corrected-to-classical C_P ratio decreases monotonically with increasing frequency of vibration. Thus, the ratio of the C_P in PoLyInfo ($C_P^{PoLyInfo}$) to the MD-calculated value (C_P^{MD}) should decrease with the increasing mean of the bond-stretching and -bending force constants. The theoretical consideration and experimental observations are described in the Supplementary Discussion and Supplementary Fig. 12 in the Supplementary Information, respectively. Fortunately, since the observed correlation is relatively clear, it would not be difficult to correct the systematic bias by applying e.g., transfer learning or multi-fidelity learning.

The linear and volume expansion coefficients showed weak correlations ($R^2 = 0.178$ and 0.217) between the calculated and experimental values (Fig. 3e, f). The variations in the linear and volume expansion coefficients within the same polymer were significant both experimentally and computationally. Previous studies also reported that MD-calculated values of the volume expansion coefficients in molecular organic liquids are highly variable and less reproducible with respect to the experimental values[32]. Possibly, the timescale and simulation cell size of the MD simulations should be sufficiently large for an accurate simulation.

Data distribution

The marginal distributions of the six properties for the calculated 1070 unique amorphous polymers are presented in the diagonal panels of Fig. 4, and their statistics are summarized in Table 1. The

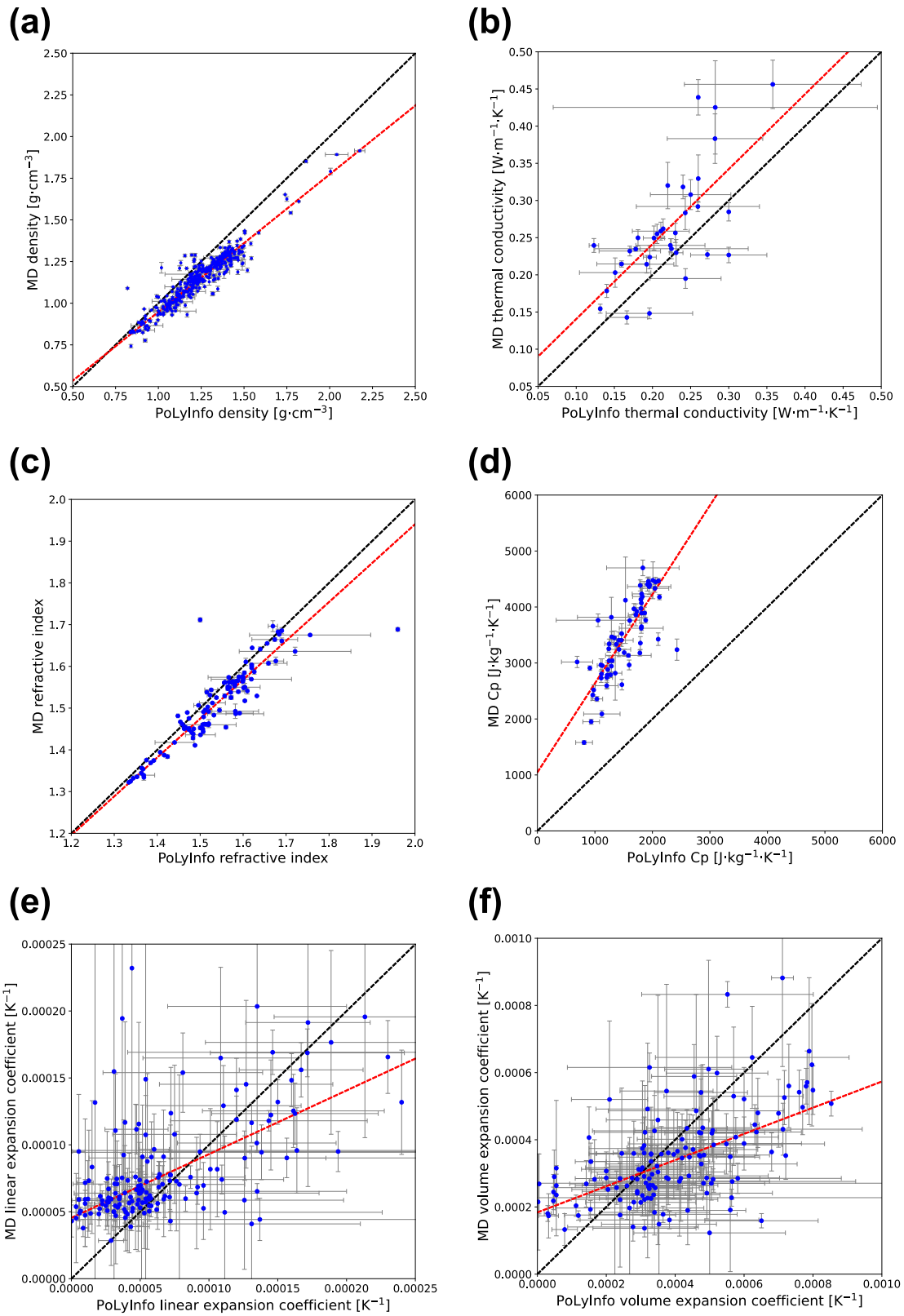


Fig. 3 Comparison between the MD-calculated properties of various amorphous polymers (vertical axis) and their experimental values in PoLyInfo (horizontal axis). The six panels show parity plots of (**a**) density ($N = 382$), (**b**) thermal conductivity ($N = 34$), (**c**) refractive index ($N = 107$), (**d**) specific heat capacity ($N = 66$), (**e**) linear expansion coefficient ($N = 165$), and (**f**) volume expansion coefficient ($N = 144$). The error bar indicates the standard deviation of the calculated or measured properties within the same polymer. The dashed black line indicates the $y = x$ line. The red line is the regression line fitted to the calculated and experimental values.

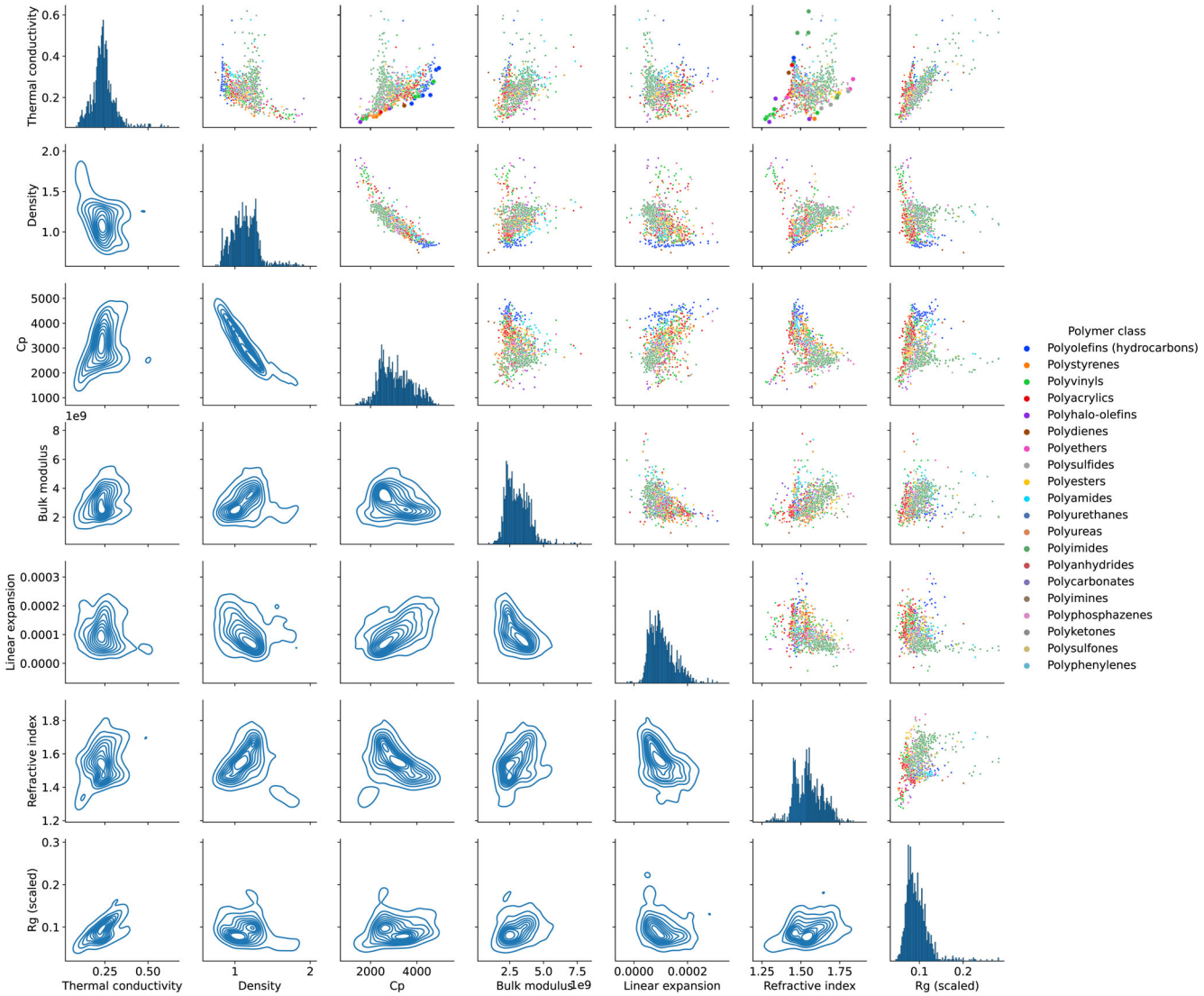


Fig. 4 Joint distribution of the six properties calculated from the automatic MD simulation, including the thermal conductivity ($W \cdot m^{-1} \cdot K^{-1}$), density ($g \cdot cm^{-3}$), specific heat capacity C_P ($J \cdot kg^{-1} \cdot K^{-1}$), bulk modulus (Pa), linear expansion coefficient (K^{-1}), refractive index, and radius of gyration Rg (scaled). The diagonal panels represent the histograms of the individual property values. In the upper off-diagonal panels, a scatter plot of each pair of properties is shown with its Pareto front set displayed as large dots that indicates higher and lower bounds of specific heat capacity and thermal conductivity, refractive index and thermal conductivity, and thermal conductivity and refractive index. The lower off-diagonal panels represent the kernel density estimation of the bivariate joint distributions, which is displayed with contours.

calculated thermal conductivities were distributed between 0.082 and 0.619 $W \cdot m^{-1} \cdot K^{-1}$, with their mean being 0.240 $W \cdot m^{-1} \cdot K^{-1}$. The thermal conductivity of the unoriented polymers in the amorphous states is known to be typically less than 0.3 $W \cdot m^{-1} \cdot K^{-1}$. On the other hand, few of the calculated polymers exhibited exceedingly high thermal conductivities. However, as mentioned above, in the high-thermal-conductivity regions, the fluctuations in the MD-calculated properties became significant. Thus, we narrowed down to eight highly reliable polymers, as shown in Fig. 5, with small variation in the repeated calculations (SD < 0.05 $W \cdot m^{-1} \cdot K^{-1}$). For polyethylene (**PI1**) and poly(vinyl alcohol) (**PI241**), the experimental thermal conductivities were recorded in the PoLyInfo. The calculated thermal conductivity of **PI1** was 0.456 $W \cdot m^{-1} \cdot K^{-1}$, which is consistent with the reported values (0.39–0.53 $W \cdot m^{-1} \cdot K^{-1}$) of the polyethylene neat resin in PoLyInfo. On the other hand, the calculated thermal conductivity of **PI241** was 0.439 $W \cdot m^{-1} \cdot K^{-1}$, which is overestimated compared to the reported value (0.31 $W \cdot m^{-1} \cdot K^{-1}$) of the poly(vinyl alcohol) neat resin in PoLyInfo. The experimental thermal conductivities of the other six polymers were unrecorded in the PoLyInfo. Apart from polyethylene (**PI1**), the structural features of these polymers fall into three types: (1) polymers with a high density of hydrogen bonding units (**PI241** and **PI305**), (2) aromatic polyamides with rigid, linear backbones (**PI626**), and aromatic polyimides (**PI687**, **PI711**, **PI715**, and **PI1093**).

In addition, the calculated values of the density and refractive index sufficiently correlated with the experimental values; therefore, we investigated the distributions of these properties from a quantitative viewpoint. The calculated density values were distributed between 0.742 and 1.914 $g \cdot cm^{-3}$ with their mean being 1.133 $g \cdot cm^{-3}$. Twelve amorphous polymers were identified as having a high-density state: >1.75 $g \cdot cm^{-3}$. These polymers were found to contain rich halogen atoms (Supplementary Fig. 2). The calculated values of the refractive index ranged from 1.274 to 1.839 with their mean equal to 1.550. Nine polymers were identified as high-refractive-index polymers in

Table 1. Summary statistics of the calculated properties, including the mean, standard deviation (SD), minimum, and maximum.

Properties	Number of polymers	Mean	SD	Minimum	Maximum
Thermal conductivity ($W{\cdot}m^{-1}{\cdot}K^{-1}$)	1070	0.240	6.562×10^{-2}	8.220×10^{-2}	0.619
Thermal diffusivity ($m^2{\cdot}s^{-1}$)	1069	7.100×10^{-8}	2.014×10^{-8}	2.957×10^{-8}	2.273×10^{-7}
Density ($g{\cdot}cm^{-3}$)	1077	1.133	0.180	0.742	1.914
Radius of gyration (Å)	1077	20.59	8.149	10.37	85.68
Self-diffusion coefficient ($m^2{\cdot}s^{-1}$)	1076	6.747×10^{-13}	8.693×10^{-13}	8.939×10^{-15}	1.098×10^{-11}
C_P ($J{\cdot}kg^{-1}{\cdot}K^{-1}$)	1076	3086	691.6	1345	4955
C_V ($J{\cdot}kg^{-1}{\cdot}K^{-1}$)	1076	2993	644.3	1331	4579
Compressibility (GPa^{-1})	1076	0.360	0.107	0.129	1.299
Isentropic compressibility (GPa^{-1})	1076	0.349	0.102	0.128	1.255
Bulk modulus (GPa)	1076	3.062	0.835	0.921	7.766
Isentropic bulk modulus (GPa)	1076	3.144	0.842	0.935	7.862
Linear expansion coefficient (K^{-1})	1076	1.048×10^{-4}	4.611×10^{-5}	-2.598×10^{-5}	3.115×10^{-4}
Volume expansion coefficient (K^{-1})	1076	3.144×10^{-4}	1.383×10^{-4}	-7.794×10^{-5}	9.345×10^{-4}
Static dielectric constant	1075	4.866	10.535	1.674	130.8
Refractive index	1075	1.550	8.857×10^{-2}	1.274	1.839
Properties of a repeating unit					
HOMO (eV)	1077	−9.205	0.918	−7.458	−12.647
LUMO (eV)	1077	1.016	1.416	−5.997	3.188
Dipole moment (Debye)	1077	2.426	1.948	6.175×10^{-7}	12.14
Dipole polarizability ($Å^3$)	1077	35.45	25.73	3.842	139.6

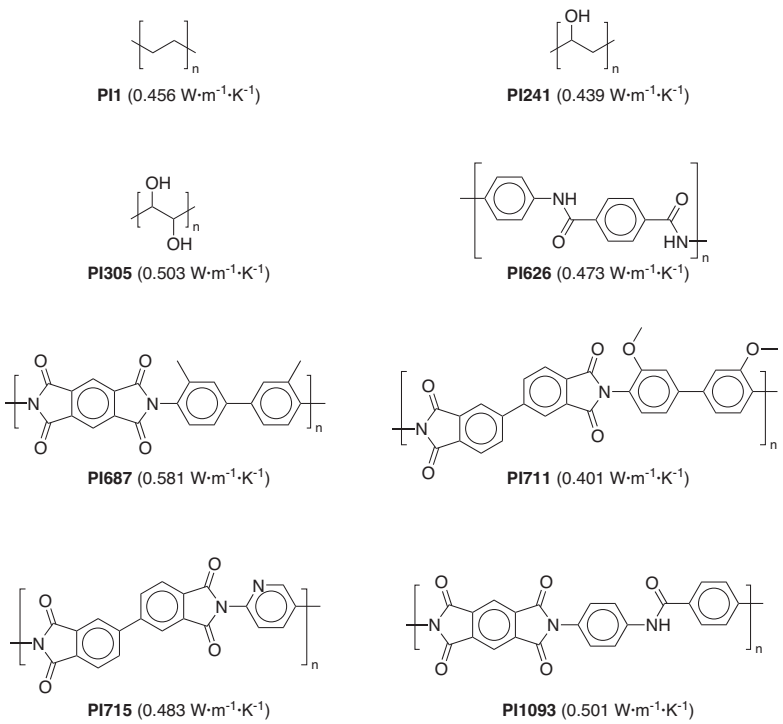


Fig. 5 Repeating units of identified polymers exhibiting a high thermal conductivity in amorphous states. The compound identifier corresponds to the polymer ID in the calculated dataset.

amorphous states, with their refractive index being greater than 1.75. These polymers had large π-conjugated backbones (Supplementary Fig. 3), indicating that the calculated high refractive index originated from the high polarizability of the large π-conjugation.

An observation of the joint distribution of the multiple properties, as shown in the off-diagonal panels of Fig. 4, provides hypothetical insights into the hidden dependency of the multiple properties, and the existence and location of the Pareto frontiers with the chemical features of the constituent polymers. The

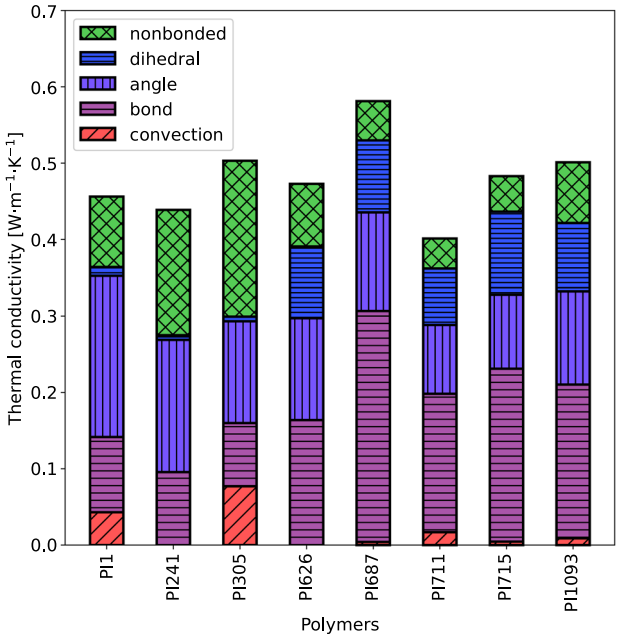


Fig. 6 Contributions of convection and different types of interactions to the calculated high thermal conductivities of the eight polymers. The colors in the bar chart mean convection (red), bond stretching (purple), bond angle bending (violet), dihedral (blue), and nonbonded (green) terms.

observed Pareto frontier of the specific heat capacity and thermal conductivity suggests the difficultly of achieving both high specific heat capacity and low thermal conductivity in amorphous polymers. Polymers distributed around the Pareto frontier included mainly polystyrenes, polyacrylates, and hydrocarbon polymers. On the other hand, no Pareto frontier was observed in the region of higher thermal conductivity. The joint distribution of the thermal conductivity and refractive index shows that there are still unexplored regions of amorphous polymers reaching lower thermal conductivity with higher refractive index and higher thermal conductivity with a lower refractive index. The thermal conductivity was approximately proportional to the scaled Rg. The scaled Rg was defined as Rg scaled by $1/M^{0.6}$ to remove molecular weight (M) dependency based on the following scaling rule[34].

$$\mathrm{Rg} \propto M^{0.6} \tag{1}$$

Another study confirmed computationally that thermal conductivity is positively correlated with Rg for amorphous polyethylene[35]. Our study demonstrated that this dependency holds for a wide variety of amorphous polymers. The specific heat capacity was inversely proportional to the density. This observation can be explained by the Dulong–Petit law[36]. The specific heat capacity is inversely proportional to the mean atomic weights (Supplementary Fig. 4) because the heat capacity of a mole is typically almost a constant in materials. On the other hand, the density is proportional to the mean atomic weights (Supplementary Fig. 5). In the joint distribution of the density and refractive index, their correlation was unclear. According to the Lorentz–Lorenz equation (Eq. 17 in the Methods section), the refractive index is described as a function of the density and polarizability. The observed distribution implies that, for polymers in amorphous states, the polarizability is dominant in determining the refractive index.

Decomposition analysis of thermal conductivity

The decomposition analysis was performed to understand the mechanism of the eight polymers (Fig. 5) that exhibited a high thermal conductivity (see "NEMD simulation for thermal conductivity calculation" in the Methods section). As shown in Fig. 6, for each calculated thermal conductivity, the decomposition analysis quantified the contribution of the six components corresponding to convection, bond, angle, dihedral, improper, and nonbonded, where the nonbonded contribution represents the sum of the pairwise and K-space contributions described in Eq. 4 in Methods section. Since the contribution of the improper term was negligible, it is shown as a dihedral term in Fig. 6. Notably, the AMBER-type force field describes the dihedral potential as the sum of the dihedral term and nonbonded 1–4 interactions; thus, a part of the nonbonded contribution is essentially attributed to the dihedral contribution[37].

The calculated thermal conductivity of **PI1** (polyethylene) was $0.456\,\mathrm{W\cdot m^{-1}\cdot K^{-1}}$. The high thermal conductivity of **PI1** was due to the significant contribution of bond bending.

The high thermal conductivities of **PI241** (polyvinyl alcohol) and **PI305** (poly(vinylene) carbonate) were largely due to the contributions of nonbonded interactions. The polymer chains of **PI241** and **PI305** contain highly condensed hydroxyl groups. This indicates that a high density of hydrogen bonding units provides large intermolecular interactions via the creation of hydrogen bonds and dipole–dipole interactions, resulting in a significant contribution of nonbonded interactions. Thus, the thermal conductivities of **PI241** and **PI305** are enhanced by the heat transfer via hydrogen bonds and dipole–dipole interactions.

In the aromatic polyamide **PI626** (poly-*p*-phenyleneterephthalamide a.k.a. Kevlar), the bond, angle, dihedral, and nonbonded interactions showed moderately large contributions. The results can be explained as follows: the backbone of **PI626** is relatively rigid, resulting in a significant contribution to the thermal conductivity through covalent bonds, and **PI626** can create the interaction of hydrogen bonds and dipole–dipole interactions with its amide groups, resulting in moderately high contributions through nonbonded interactions.

Thermally conductive behaviors in the aromatic polyimides **PI687**, **PI711**, **PI715**, and **PI1093** were largely due to the contributions of bond stretching. The **PI687** had a significantly large contribution of bond stretching. Aromatic polyimides have rigid backbones, particularly **PI687**, which has high rigidity and linearity. The results show that the rigid and linear characteristics of a polymer backbone can help enhance the thermal conductivity through the contribution of bond stretching. The **PI1093** is an aromatic polyimide containing an amide group. The contribution of nonbonded interactions of **PI1093** was the largest in the four identified aromatic polyimides. This suggests that polymers containing hydrogen bonding units and having rigid and linear backbones can help further increase the thermal conductivity in amorphous states.

Figure 7 shows the joint distribution of the total thermal conductivity with each quantified contribution. The correlations with the total thermal conductivity can be clearly observed in the bond, angle, dihedral, and nonbonded terms. On the other hand, the convection term did not correlate significantly with the thermal conductivity. In summary, thermally conductive amorphous polymers can be designed, in principle, by increasing the contributions of the bond, angle, dihedral, and nonbonded terms.

Transfer learning from MD values to experimental values

As described above, several properties showed significant discrepancies, including systematic bias, between the experimental and the MD-calculated values. The dependence of the MD simulations on initial conditions resulted in large fluctuations in the calculated properties, especially in the linear expansion and volume expansion coefficients. The experimental values of these two properties also fluctuated considerably, making them insufficiently reliable as a validation set. We believe that the

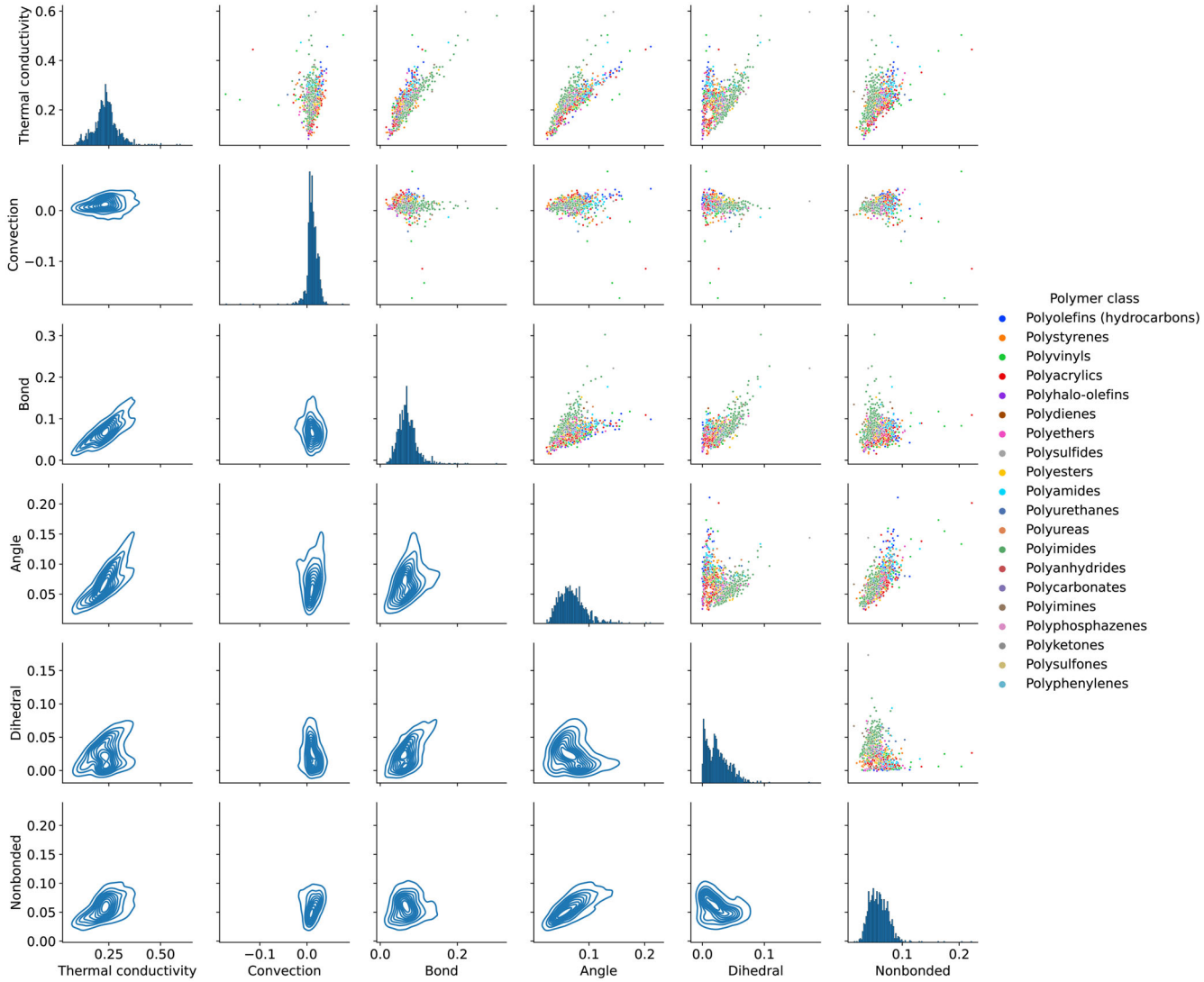


Fig. 7 Distribution of the thermal conductivity ($W \cdot m^{-1} \cdot K^{-1}$) and categorization in terms of its contributions from convection, bond, angle, dihedral, and nonbonded terms. Diagonal panels represent the histograms of individual quantities. In the upper off-diagonal panels, the scatter plots of the six quantities are displayed. The lower off-diagonal panels represent their kernel density estimation, which is displayed with contours.

application of machine learning can contribute to the reduction of these biases and variances. Hereafter, we demonstrate an example of calibrating the discrepancy of MD simulations by using transfer learning.

The target properties to be predicted were the specific heat capacity, linear expansion coefficient, and volume expansion coefficient. As discussed previously, the specific heat capacity exhibited a large bias between the experimental and MD-calculated values, which would originate from the presence or absence of quantum effects, and the latter two had significantly large variations even within the same polymer in both experimental and calculated properties. For each property, the source task of the transfer learning was to predict the MD-calculated properties, and the target task was to predict the experimental properties in PoLyInfo. A predictive model defines a mapping from a fingerprinted chemical structure of a given polymer repeating unit to the experimental or MD-calculated property. The workflow of the shotgun transfer learning[18] is outlined as follows (see the Supplementary Methods in the Supplementary Information for more details):

1. All samples in the MD properties dataset with their polymers included in the PoLyInfo experimental dataset were removed to obtain the dataset for the source task.
2. Using the source dataset, we trained 100 neural networks with randomly generated network structures.
3. We randomly selected 80%, 10%, and 10% of the experimental dataset for the training, validation, and test datasets, respectively, for the target task.
4. Each pretrained neural network was fine-tuned using the training set of the target task, in order to obtain a transferred calibration model.
5. The root mean squared error (RMSE) of each transferred model with respect to the validation set was calculated, and the prediction performance on the test set was examined using the model exhibiting the best transferability that achieved the smallest validation RMSE.

As shown in Fig. 8, for all the three properties, the transferred models showed significant improvements in predicting the experimental data, compared to the direct predictions from the MD calculations. The systematic bias in the specific heat capacity almost disappeared. Interestingly, for the linear expansion

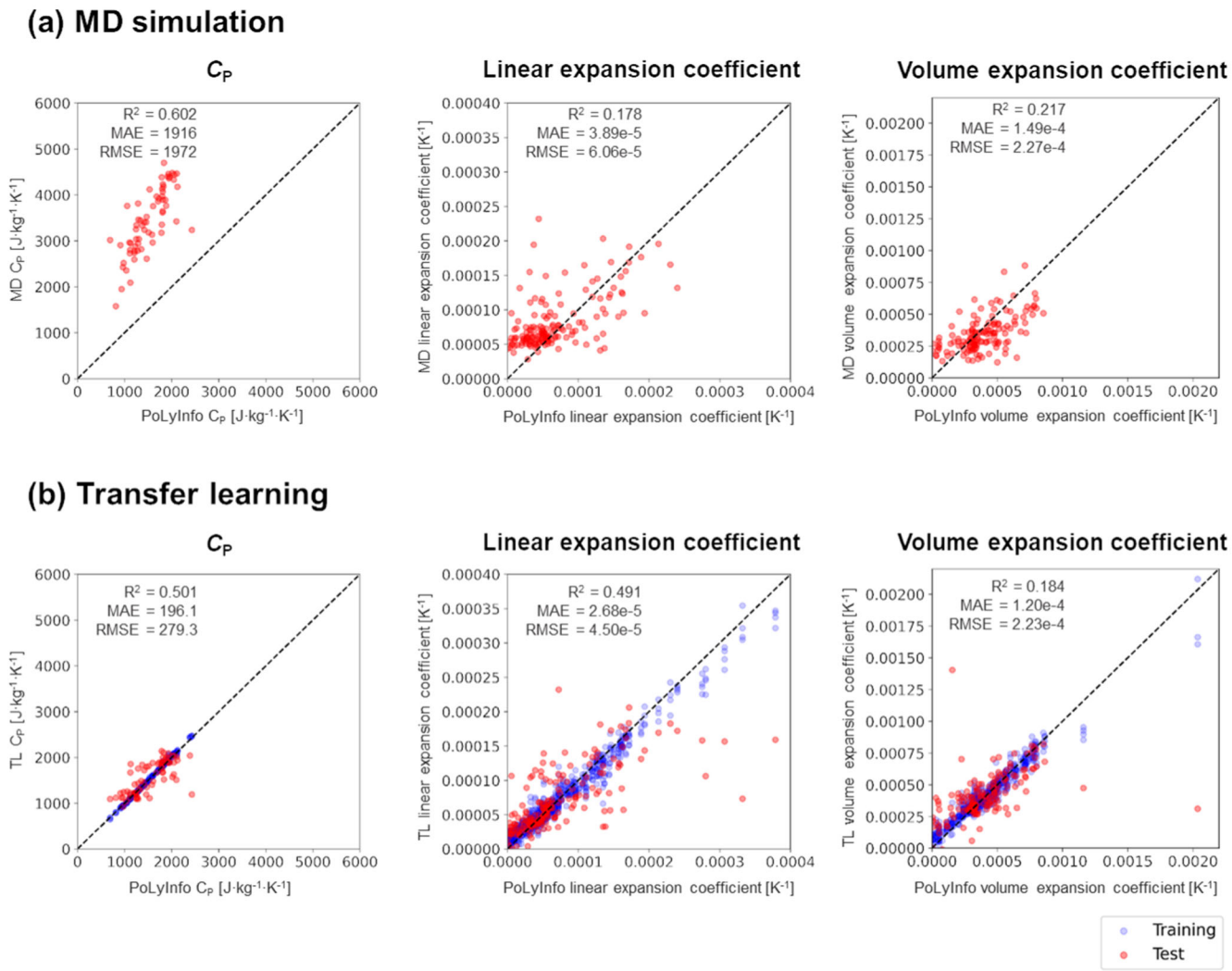


Fig. 8 Comparison of the predictive performance of MD simulation and transfer learning. The parity plots of (**a**) MD simulation and (**b**) transfer learning for experimental values of specific heat capacity (C_P), linear expansion coefficient, and volume expansion coefficient. The experimental and predicted values are shown on the horizontal and vertical axes, color-coded by red (blue: fits to the training data for transfer learning).

coefficient and volume expansion coefficient, the transferred model not only corrected for the systematic bias of the MD properties, but also significantly reduced the variability of the experimental values. The improvement in MAE for these property predictions reached 69% and 87%, respectively (Table 2), compared to that in the MD-based predictions.

It is inevitable that any dataset mass-produced from a fully automated MD simulation will be subject to various kinds of biases and variances, because there are no calculation conditions universally applicable to a wide variety of polymer systems. The results shown here imply that machine learning techniques have the great potential to bridge the gap between real systems and inherently incomplete computational models.

Table 2. Comparison of prediction performance on the PoLyInfo experimental dataset between the MD simulations and the machine learning model based on transfer learning (TL).

Properties	Metrics	MD	TL
C_P [J·kg^{-1}·K^{-1}]	R^2	**0.602**	0.501
	MAE	1916	**196.1**
	RMSE	1972	**279.3**
Linear expansion coefficient [K^{-1}]	R^2	0.178	**0.491**
	MAE	3.89×10^{-5}	$\mathbf{2.68 \times 10^{-5}}$
	RMSE	6.06×10^{-5}	$\mathbf{4.50 \times 10^{-5}}$
Volume expansion coefficient [K^{-1}]	R^2	**0.217**	0.184
	MAE	1.49×10^{-4}	$\mathbf{1.30 \times 10^{-4}}$
	RMSE	2.27×10^{-4}	$\mathbf{2.23 \times 10^{-4}}$

The bold values indicate the best result for prediction performance.

Summary and outlook

We presented RadonPy, which is the first open-source Python library to fully automate polymer property calculations using all-atom classical MD simulations. The high-throughput calculation using RadonPy was successfully performed for more than 1000 unique amorphous polymers with a wide variety of thermo-physical properties, such as the thermal conductivity, refractive index, density, and specific heat capacity C_P. For systems other than amorphous homopolymers, such as copolymers, blend polymers, and uniaxially oriented systems, as well as for other properties, automated calculation capabilities have already been

implemented; however, no calculation protocols based on experimental data have been established. In RadonPy, automatic calculation protocols for various polymer properties can be implemented as an add-on feature. We will continue to promote the development of RadonPy.

In this study, the agreement between a total of six properties obtained from the high-throughput MD calculation and experimental values was comprehensively verified. As a result, the refractive index, density, and thermal conductivity successfully reproduced the experimental values quantitatively. The calculated values of the specific heat capacity were also highly correlated with the experimental values, although the classical MD calculation had a systematic bias due to its inability to represent quantum effects. For the linear and volume expansion coefficients, the correlation between the calculated and experimental values was weak due to large variations and uncertainties in both the calculations and experiments. There has been no previous work on a comprehensive validation of high-throughput MD simulations of polymer properties on such a scale. More rigorous comprehensive validation with experimental values, including other properties not discussed in this study, should be conducted to determine appropriate calculation conditions and protocols.

This study also revealed various issues related to the creation of a polymer properties database using high-throughput MD calculations. Properties, such as glass transition temperature, dielectric loss tangent, and cohesive energy density, are expected to be predictable under the same or nearly the same setting as the current calculation conditions. On the other hand, mechanical and viscoelastic properties, which are largely affected by polymer chain entanglement, would be difficult to predict with the current settings for molecular weight, timescale, generation of initial structure, high-order structure, etc. It is necessary to determine appropriate conditions for automated calculations according to individual properties. It is also important to produce temperature-dependent and molecular weight-dependent physical property profiles. We will then be faced with the problem of lacking a comprehensive set of experimental data necessary to determine appropriate calculation conditions. In this study, PoLyInfo was employed as the benchmark dataset, but as indicated by the observed large fluctuations of linear expansion coefficients and volume expansion coefficients, the quality and reliability of the current data are far from satisfactory. Data cleansing, which involves an enormous amount of work to trace back to the original paper for each record, will need to be performed. Alternatively, we will eventually be faced with the need to construct an experimental dataset for benchmarking, acquired in a controlled environment.

In addition, several issues related to data storage need to be further considered when building a large database. Currently, RadonPy stores and outputs all intermediate trajectory data, including atomic coordinates and velocities, in LAMMPS dump files. However, in the future database development, intermediate trajectory files may be discarded, except for the final states and the last several nanoseconds in the equilibration and the NEMD because of the enormous data size (~20 GB per polymer on average). The issue of data storage when building a large database is unavoidable. In addition, the issue of data formatting will be an obstacle. In the first stage, we focused only on linear polymers, so their representation could be handled with the SMILES notation. However, in the future, block or alternating copolymers and branched polymers will also be included in the automated calculation pipeline. Then, it will be necessary to introduce an advanced notation for polymers such as BigSMILES[38].

Compared with other material systems, polymer research has lagged in terms of constructing open databases available for data-driven research. The primary objective in the development of RadonPy was to use it to create a systematically designed polymer property database. In the early days of MI in inorganic chemistry, the development of an open database was strategically promoted. In particular, huge computational property databases constructed using high-throughput first-principles calculations drove the evolution and widespread applications of MI. Large-scale computational property data have historically proven to be an important resource in MI, and RadonPy was designed for the rapid production of large amounts of polymer property data using highly parallel computers such as supercomputers. In this study, more than 1000 unique amorphous polymers were computed in ~2 months mainly using the supercomputer, Fugaku. In the future, our growing data will significantly facilitate the evolution of polymer informatics, just like the first-principles computational database for inorganic crystals.

METHODS

Conformation search of a repeating unit

For a given SMILES string of a polymer repeating unit, 3D atomic coordinates of up to 1000 different molecular conformations were generated using the ETKDG version 2 method[39–41] implemented in the Python library RDKit[26]. The SMILES string has two asterisk symbols for representing two attachment points of the repeating unit. These symbols were capped with hydrogen atoms. The potential energy of each conformation of a repeating unit was evaluated using the molecular mechanics calculation with the general Amber force field version 2 (GAFF2)[37,42] after the geometry optimization. Subsequently, the optimized conformers were clustered by performing the Butina clustering[43] based on the torsion fingerprint deviation[44]. The most stable four conformations were further optimized by performing DFT calculations with the ωB97M-D3BJ functional[45,46] combined with the 6–31 G(d,p)[47,48] basis set. The most stable conformation was determined based on the DFT total energies.

Calculation of electronic property of a repeating unit

The atomic charges of a repeating unit were calculated using the restrained electrostatic potential (RESP) charge model[49] with a single-point calculation of the Hartree–Fock method[50] combined with the 6–31 G(d) basis set on the optimized geometry of the most stable conformation. The total energy, the highest occupied molecular orbital (HOMO) energy level, the lowest unoccupied molecular orbital (LUMO) level, and the dipole moment were calculated with the single-point calculation using the ωB97M-D3BJ functional combined with the 6–311 G(d,p) basis set[45,46,51–53] for H, C, N, O, F, P, S, Cl, and Br atoms and with the LanL2DZ basis set[54] for I atom. In addition, the dipole polarizability tensor was obtained by applying the finite field method under an electric field of 1.0×10^{-4} a.u. using the ωB97M-D3BJ functional combined with the 6–311 + G(2d,p) basis set[45,46,51–53,55,56] for H, C, N, O, F, P, S, and Cl atoms, with the 6–311 G(d,p) for Br atom, and with the LanL2DZ basis set for I atom. The reason for using the 6–311 + G(2d,p) basis set is that a basis set, including double polarization and diffuse functions, is required for appropriate polarizability calculations[57]. The isotropic dipole polarizability was defined as the mean of the diagonal values of the dipole polarizability tensor.

Generation of polymer chains

A polymer chain was constructed by connecting a repeating unit with the self-avoiding random walk algorithm. To prevent unintended chiral inversions and cis/trans conversions due to a large strain structure in the polymer chain growth, the bond between the head and capped atoms in a growing polymer chain and the bond of the tail and capped atoms in the next repeating unit were arranged to be coaxial and anti-parallel, the two capped atoms were deleted, and a new bond between the head and tail

atoms was created. The length of the new bond was 1.5 Å, and the dihedral angle around the new bond was randomized in the range of −180° to +180° during the self-avoiding step. Charge neutrality was ensured by summing the charges of capped H atoms to the atoms to which they are bonded. In this study, polymer chains were created to include ~1000 atoms; thus, the degree of polymerization varies across polymers. By taking the number of atoms at the same level for different polymers, the molecular weights were controlled to be almost the same. Thus, all calculated properties were obtained under conditions where the molecular weights were set to be approximately the same. In addition, we investigated the sensitivity of the calculated properties to the change in the number of atoms. As a result, we confirmed that the number of atoms in the simulation cell has a trivial effect on the calculated properties, which is detailed in "Validation of the calculated physical properties" in the Results and Discussion section and in the Supplementary Discussion in the Supplementary Information. The tacticity of a polymer chain could also be controlled in this process using RadonPy. In this study, all the polymers were generated as atactic polymers.

Assignment of force field parameters

The GAFF2 force field is expressed as follows[37]:

$$E_{\mathrm{MM}} = \sum_{\mathrm{bonds}} K_{\mathrm{b}}(r-r_0)^2 + \sum_{\mathrm{angles}} K_{\mathrm{a}}(\theta-\theta_0)^2 + \sum_{\mathrm{dihedrals}} K_{\mathrm{d}}[1+\cos(n_{\mathrm{d}}\varphi-\delta)] + \sum_{\mathrm{impropers}} K_{\mathrm{i}}(\chi-\chi_0)^2 + \sum_{\mathrm{i,j}} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{\mathrm{i,j}} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad (2)$$

where r, θ, φ, χ, and r_{ij} are the bond length, bond angle, dihedral angle, improper angle, and distance between atoms i and j, respectively; K_b, K_a, K_d, and K_i denote the force constants of the bond, bond angle, dihedral angle, and improper angle, respectively; r_0, θ_0, and χ_0 are the equilibration structural parameters of the bond, bond angle, and improper angle, respectively; n_d is the multiplicity, and δ is the phase angle for the torsional angle parameters; and q_i and q_j are the atomic charges of atoms i and j, and ε_0 is the dielectric constant of vacuum; ε_{ij} and σ_{ij} are the Lennard–Jones parameters determining the depth of the energy potential and equilibrium distance, respectively. Compared to those by GAFF, GAFF2 has improved the parameter values of K_b, r_0, K_a, and θ_0 to reproduce molecular geometries, vibrational spectra, and potential energy surfaces from higher level quantum mechanics calculations and improved the non-bonded parameters to better reproduce ab initio interaction energies and experimental neat liquid properties[42]. The parameter set was suitable for thermal conductivity calculations because the reproducibility of the vibrational properties was considered. The modified parameters for fluorocarbon developed by Träg and Zahn[33] were used for fluorocarbon polymers. The GAFF2 parameters were automatically assigned to each polymer chain in RadonPy. If the pre-defined parameter set lacked the bond angle parameters of K_a and θ_0 for a certain atom group, these parameter values were empirically estimated in the same manner as GAFF2.

Generation of a simulation cell

A simulation cell containing amorphous polymers was constructed by randomly arranging and rotating 10 polymer chains such that they did not overlap with each other, resulting in an amorphous cell having ~10,000 atoms. Initially, the density of the amorphous cell was set to 0.05 g·cm^{-3} and was then increased by conducting a packing simulation as described below.

Packing simulation

The initial structure of the generated amorphous cell had a very low density. A packing simulation was performed to increase the density of the amorphous polymers to an appropriate value for subsequent calculations. A 1 ns *NVT* simulation with a Nosé−Hoover thermostat was performed while the temperature was increased from 300 K to 700 K; in the next 1 ns *NVT* simulation, the calculation cell was isotropically reduced to a density of 0.8 g·cm^{-3} at 700 K. In this packing simulation, to prevent the self-aggregation of a polymer chain by intramolecular interactions leading to a globule-like structure, the Coulomb interaction was turned off, and the cutoff of the Lennard–Jones potential was set to 3.0 Å. Under this condition, the polymer chains remain random coil structures and could not pass through each other. Thus, the polymer chains were entangled in the final structure of the packing simulation. The time step was set to 1 fs, the periodic boundary condition (PBC) was applied, and all the bonds and angles, including those of the hydrogen atoms, were constrained by the SHAKE algorithm[58] in this packing simulation.

Equilibration simulation

The amorphous polymers after the packing simulation were equilibrated by the 21-steps compression/decompression equilibration protocol[59] proposed by Larsen and co-workers. In this protocol, a temperature rise to 600 K and a drop to 300 K were repeated for ~1.5 ns while the system was compressed to 50,000 atm and then decompressed to 1 atm by combining the *NVT* and *NpT* simulations with a Nosé−Hoover thermostat and a barostat. After the 21-steps equilibration, *NpT* simulations were run for more than 5 ns at 300 K and 1 atm until equilibrium was achieved. The achievement of the equilibrium was checked each 5 ns after the 21-steps equilibration. In this study, the equilibrium state was defined as being reached when the following conditions were met: the relative standard deviations (RSD) of the total, kinetic, bonding, bond angle, dihedral, van der Waals (vdW), and long-range coulomb energy fluctuations were less than 0.05, 0.05, 0.1, 0.1, 0.2, 0.2, and 0.1%, respectively, and the RSDs of the density and radius of gyration fluctuations were less than 0.1 and 1%, respectively. In this study, calculations that did not achieve equilibrium after 50 ns of equilibration calculations were treated as failures. The time step was set to 1 fs, the PBC was applied, and the SHAKE constraint[58] was applied to all the bonds and angles, including those of the hydrogen atoms in this equilibration simulation. The twin-range cutoff method[60] was used for nonbonded interactions with a short cutoff of 8 Å and a long cutoff of 12 Å. The long-range Coulomb interaction was treated using the particle-particle particle-mesh (PPPM) method[61]. When the nematic order parameter decreased below 0.1, it was judged that the amorphous structure was appropriately generated; otherwise, it was treated as a failed calculation and removed from the data.

NEMD simulation for thermal conductivity calculation

To calculate the thermal conductivity, we performed the reverse NEMD simulation[62] proposed by Müller-Plathe. The simulation box of the reverse NEMD was constructed by triplication of an equilibrated amorphous cell in the *x*-axis direction under the PBC. The reverse NEMD simulation involved dividing the simulation box into *N* slabs along the direction of the heat flux, which was generated in the system with temperature gradients induced by exchanging the velocity between the coldest atom in slab *N*/2 and the hottest atom in slab 0, as shown in Fig. 9. As a result, slab *N*/2 becomes the hottest in the cell, and the temperature gradually decreases towards slab 0 and slab *N* because of using the PBC. To prevent the occurrence of temperature shifts due to cell replication, the preheating step with *NVT* ensemble was run for 2 ps at 300 K. Subsequently, the reverse NEMD with *NVE* ensemble was run for 1 ns. The number of slabs was set to 20, and the frequency of velocity swapping was set to 200 fs.

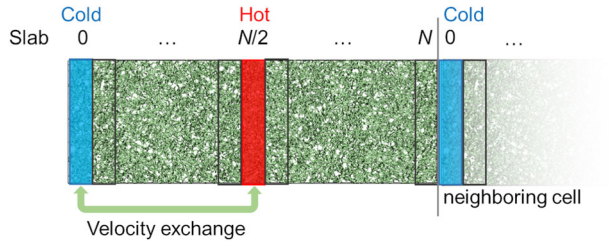


Fig. 9 Schematic representation of the simulation box for reverse nonequilibrium molecular dynamics. The red and blue slabs are the hottest and coldest region, respectively, in the simulation box.

The time step was set to 0.2 fs, and the SHAKE constraint was not applied in the reverse NEMD simulation. The twin-range cutoff method was used for nonbonded interactions with a short cutoff of 8 Å and a long cutoff of 12 Å. The long-range Coulomb interaction was treated using the PPPM method. As a validation of the adequacy of the reverse NEMD calculation, RadonPy confirmed a linearity in the temperature gradient. A calculation result with a poor linearity in the temperature gradient ($R^2 < 0.95$) was treated as a failure and removed from the data.

A thermal conductivity decomposition analysis was performed for 100 ps. For the Irving-Kirkwood equation[63] modified by Torii and co-workers[64], the energy flux can be expressed as follows:

$$J_u = \frac{1}{V}\left\{\sum_{i\in V} e_i v_{i,u} + \sum_{i\in V} (\mathbf{S}_i \mathbf{v}_i)_u\right\} \tag{3}$$

where J_u is the energy flux along the direction of unit vectors u, V is the volume, e_i is the per-atom potential and kinetic energy, $v_{i,u}$ is the velocity of the atom, $\mathbf{S}_i$ is the per-atom stress tensor, and i is the index of atoms. The first and second terms represent the contribution to the energy flux via convection and interatomic interactions, respectively. The second term can be further divided into each component of the interactions. The component (a, b) of the stress tensor can be written as[65–67]

$$\begin{aligned}\mathbf{S}_{ab} = &\sum_{n=1}^{N_p} \mathbf{r}_{i0,a}\mathbf{F}_{i,b} + \sum_{n=1}^{N_b} \mathbf{r}_{i0,a}\mathbf{F}_{i,b} + \sum_{n=1}^{N_a} \mathbf{r}_{i0,a}\mathbf{F}_{i,b} + \sum_{n=1}^{N_d} \mathbf{r}_{i0,a}\mathbf{F}_{i,b} \\ &+ \sum_{n=1}^{N_i} \mathbf{r}_{i0,a}\mathbf{F}_{i,b} + \mathrm{Kspace}(\mathbf{r}_{i,a}, \mathbf{F}_{i,b})\end{aligned} \tag{4}$$

The first to fifth terms denote the pairwise, bond, angle, dihedral, and improper contributions, respectively, where $\mathbf{F}_i$ denotes the force acting on atom i due to the interaction, $\mathbf{r}_{i0}$ denotes the relative position of atom i to the geometric center of its interacting atoms, and N_p, N_b, N_a, N_d, and N_i are the numbers of non-bonding atom pairs, bonds, bond angles, dihedral angles, and improper angles, respectively. The sixth term is the K-space contribution from the long-range Coulombic interactions. The partial thermal conductivity $\lambda_{partial}$ is given by

$$\lambda_{partial} = \frac{J_{partial}}{J_{total}}\lambda_{total} \tag{5}$$

where J_{total} is the total heat flux calculated by Eq. 3, $J_{partial}$ is the partial heat flux subdivided by each term of Eqs. 3 and 4, and λ_{total} is the total thermal conductivity calculated by the reverse NEMD.

Calculation of physical properties

The density in a *NpT* simulation was computed using the mass m and volume V of the system as follows:

$$\rho = \frac{m}{\langle V\rangle} \tag{6}$$

where the angular brackets $\langle\cdot\rangle$ represent time averaging.

The radius of gyration Rg was calculated using the following equation:

$$\mathrm{Rg} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\mathbf{r}_k - \mathbf{r}_{mean})^2} \tag{7}$$

where $\boldsymbol{r}_k$ is the position of a repeating unit k, and $\boldsymbol{r}_{mean}$ denotes the mean position of the repeating units in a polymer chain.

The specific heat capacity at constant pressure C_p was calculated from the fluctuations in the enthalpy H[68]:

$$C_P = \frac{\langle \delta H^2\rangle}{k_B T^2 m} \tag{8}$$

where k_B is the Boltzmann constant, and T is the temperature. The enthalpy was calculated using the constant pressure of 1 atm because the calculated pressure value in the *NpT* simulations has a significant fluctuation, leading to inaccurate C_P calculation.

The isothermal compressibility β_T and isothermal bulk modulus K_T were calculated from the fluctuations in the volume V[68]:

$$\beta_T = \frac{\langle \delta V^2\rangle}{k_B T\langle V\rangle} \tag{9}$$

$$K_T = \frac{1}{\beta_T} \tag{10}$$

The volume expansion coefficient α_P was calculated from the covariance of the volume V and enthalpy H[68]:

$$\alpha_P = \frac{\langle \delta V \delta H\rangle}{k_B T^2\langle V\rangle} \tag{11}$$

Here, the enthalpy was calculated at a constant pressure of 1 atm. The linear expansion coefficient $\alpha_{P,l}$ in the isotropic systems was calculated using the following equation[68]:

$$\alpha_{P,l} = \frac{\alpha_P}{3} \tag{12}$$

The specific heat capacity at constant volume C_V was calculated from the following equation, associated with C_P, α_P, and β_T[68]:

$$C_V = C_P - \frac{\alpha_P^2 T\langle V\rangle}{\beta_T m} \tag{13}$$

The isentropic compressibility β_S and isentropic bulk modulus K_S were calculated using the following equations:

$$\beta_S = \beta_T \frac{C_V}{C_P} \tag{14}$$

$$K_S = \frac{1}{\beta_S} \tag{15}$$

The self-diffusion coefficient was calculated using the Einstein equation[68]:

$$D = \lim_{t\to\infty}\frac{1}{6t}\left\langle |\mathbf{r}(t+t_0) - \mathbf{r}(t_0)|^2\right\rangle \tag{16}$$

where t is the time, and $\boldsymbol{r}$ denotes the atomic position at the time.

The refractive index n was obtained from the Lorentz–Lorenz equation:

$$\frac{n^2-1}{n^2+2} = \frac{4\pi}{3}\frac{\rho}{M}\alpha_{polar} \tag{17}$$

where α_{polar} is the isotropic dipole polarizability of a repeating unit computed from the DFT calculation, and M is the molecular weight of a repeating unit.

The static dielectric constant $\varepsilon(0)$ was calculated using the equation[69]:

$$\varepsilon(0) = \frac{\langle \boldsymbol{\mu}^2 \rangle - \langle \boldsymbol{\mu} \rangle^2}{3\varepsilon_0 k_B T \langle V \rangle} + \varepsilon_{el} \tag{18}$$

where $\boldsymbol{\mu}$ is the dipole moment of the system, ε_0 is the dielectric constant of vacuum, and ε_{el} is the contribution of the electronic polarization in the dielectric constant, which is evaluated as the square of the refractive index n^2.

The nematic order parameter was calculated as the highest eigenvalue of the second rank ordering tensor[68] $\mathbf{Q}_{\alpha\beta}$, following equation:

$$\mathbf{Q}_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(3\mathbf{u}_{i\alpha}\mathbf{u}_{i\beta} - \boldsymbol{\delta}_{\alpha\beta}\right) \tag{19}$$

where $\mathbf{u}_{i\alpha}$ and $\mathbf{u}_{i\beta}$ (α, β = x, y, or z) are the unit vectors of the molecular axis of a repeating unit i, $\boldsymbol{\delta}_{\alpha\beta}$ is the Kronecker delta, and N is the number of repeating units. The molecular axis of each repeating unit is defined as the long axis found from the inertia tensor. The nematic order parameter takes on a value between 0 for an isotropic structure and 1 for a completely ordered structure.

The thermal conductivity λ was calculated according to Fourier's law:

$$\lambda = \frac{J_Q}{(\partial T/\partial x)} = \frac{\Delta E}{2A\Delta t(\partial T/\partial x)} \tag{20}$$

where J_Q is the heat flux, and $\partial T/\partial x$ is the temperature gradient of the NEMD simulation. The heat flux J_Q can be calculated from the exchanged energy obtained using the Müller-Plathe algorithm ΔE, the cross-sectional area in the heat flux direction A, and the simulation time Δt. The thermal diffusivity κ was obtained from the calculated thermal conductivity λ, density ρ, and heat capacity C_P:

$$\kappa = \frac{\lambda}{\rho C_P} \tag{21}$$

DATA AVAILABILITY

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

CODE AVAILABILITY

The source code of RadonPy is available from the GitHub site (https://github.com/RadonPy/RadonPy).

Received: 5 May 2022; Accepted: 13 October 2022;
Published online: 08 November 2022

REFERENCES

1. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).
2. Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
3. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
4. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
5. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
6. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
7. Sha, W. et al. Machine learning in polymer informatics. *InfoMat* **3**, 353–361 (2021).
8. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. In *2011 International Conference on Emerging Intelligent Data and Web Technologies* 22–29 (IEEE, 2011).
9. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer Genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C.* **122**, 17575–17585 (2018).
10. Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).
11. Doan Tran, H. et al. Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **128**, 171104 (2020).
12. Chen, L. et al. Polymer informatics: current status and critical next steps. *Mater. Sci. Eng. R. Rep.* **144**, 100595 (2021).
13. Jiang, J., Shu, Y., Wang, J. & Long, M. Transferability in deep learning: a survey. Preprints at https://arxiv.org/abs/2201.05867 (2022).
14. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **4**, 3320–3328 (2014).
15. Kong, S., Guevarra, D., Gomes, C. P. & Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **8**, 021409 (2021).
16. Lee, J. & Asahi, R. Transfer learning for materials informatics using crystal graph convolutional neural network. *Comput. Mater. Sci.* **190**, 110314 (2021).
17. Li, X. et al. A transfer learning approach for microstructure reconstruction and structure-property predictions. *Sci. Rep.* **8**, 13461 (2018).
18. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
19. Torres, P. et al. Descriptors of intrinsic hydrodynamic thermal transport: screening a phonon database in a machine learning approach. *J. Phys. Condens. Matter* **34**, 135702 (2022).
20. Ju, S. et al. Exploring diamondlike lattice thermal conductivity crystals via feature-based transfer learning. *Phys. Rev. Mater.* **5**, 053801 (2021).
21. 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).
22. Hruska, E., Gale, A., Huang, X. & Liu, F. AutoSolvate: a toolkit for automating quantum chemistry design and discovery of solvated molecules. *J. Chem. Phys.* **156**, 124801 (2022).
23. Ma, R., Zhang, H. & Luo, T. Exploring high thermal conductivity amorphous polymers using reinforcement learning. *ACS Appl. Mater. Interfaces* **14**, 15587–15598 (2022).
24. Afzal, M. A. F. et al. High-throughput molecular dynamics simulations and validation of thermophysical properties of polymers for various applications. *ACS Appl. Polym. Mater.* **3**, 620–630 (2021).
25. Demidov, A. G., Perera, B. L. A., Fortunato, M. E., Lin, S. & Colina, C. M. Update 1.1 to "pysimm: a python package for simulation of molecular systems". *SoftwareX* **15**, 100749 (2021).
26. Landrum, G. *RDKit: Open-Source Cheminformatics Software.* https://www.rdkit.org/ (2020).
27. Weininger, D. SMILES, a chemical language and information system. 1. *Introduction Methodol. encoding rules. J. Chem. Inf. Model.* **28**, 31–36 (1988).
28. Smith, D. G. A. et al. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108 (2020).
29. 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).
30. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprints at https://arxiv.org/abs/1802.03426 (2020).
31. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
32. Caleman, C. et al. Force field benchmark of organic liquids: density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *J. Chem. Theory Comput.* **8**, 61–74 (2012).
33. Träg, J. & Zahn, D. Improved GAFF2 parameters for fluorinated alkanes and mixed hydro- and fluorocarbons. *J. Mol. Model.* **25**, 39 (2019).
34. Teraoka, I. Models of Polymer Chains. *Polymer Solutions: An Introduction to Physical Properties*. (John Wiley & Sons, Inc., 2002).
35. Zhang, T. & Luo, T. Role of chain morphology and stiffness in thermal conductivity of amorphous polymers. *J. Phys. Chem. B* **120**, 803–812 (2016).
36. Petit, A. T. & Dulong, P. L. Study on the measurement of specific heat of solids. *Ann. Chim. Phys.* **10**, 395 (1819).
37. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
38. Lin, T. S. et al. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019).
39. Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
40. Hawkins, P. C. D. Conformation generation: the state of the art. *J. Chem. Inf. Model.* **57**, 1747–1756 (2017).

41. Wang, S., Witek, J., Landrum, G. A. & Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **60**, 2044–2058 (2020).
42. Case, D. A. et al. *Amber 21*. http://ambermd.org/index.php (2021).
43. Butina, D. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
44. Schulz-Gasch, T., Schärfer, C., Guba, W. & Rarey, M. TFD: Torsion fingerprints as a new measure to compare small molecule conformations. *J. Chem. Inf. Model.* **52**, 1499–1512 (2012).
45. Mardirossian, N. & Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110 (2016).
46. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
47. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **54**, 724–728 (1971).
48. Francl, M. M. et al. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **77**, 3654–3665 (1982).
49. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).
50. Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69–89 (1951).
51. Krishnan, R., Binkley, J. S., Seeger, R. & Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **72**, 650–654 (1980).
52. McLean, A. D. & Chandler, G. S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z = 11–18. *J. Chem. Phys.* **72**, 5639–5648 (1980).
53. Binning, R. C. & Curtiss, L. A. Compact contracted basis sets for third-row atoms: Ga-Kr. *J. Comput. Chem.* **11**, 1206–1216 (1990).
54. Wadt, W. R. & Hay, P. J. Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *J. Chem. Phys.* **82**, 284–298 (1985).
55. Clark, T., Chandrasekhar, J., Spitznagel, G. W. & Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F. *J. Comput. Chem.* **4**, 294–301 (1983).
56. Frisch, M. J., Pople, J. A. & Binkley, J. S. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *J. Chem. Phys.* **80**, 3265–3269 (1984).
57. Ando, S. Efficient hybrid functional and basis set functions for DFT calculation of refractive indices and Abbe numbers of organic compounds. *Chem. Lett.* **47**, 1494–1497 (2018).
58. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
59. 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).
60. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
61. Hockney, R. & Eastwood, J. *Computer Simulation Using Particles*. *Computer Simulation Using Particles* (CRC Press, 1988).
62. Müller-Plathe, F. A simple nonequilibrium molecular dynamics method for calculating the thermal conductivity. *J. Chem. Phys.* **106**, 6082–6085 (1997).
63. Irving, J. H. & Kirkwood, J. G. The statistical mechanical theory of transport processes. IV. The equations of hydrodynamics. *J. Chem. Phys.* **18**, 817–829 (1950).
64. Torii, D., Nakano, T. & Ohara, T. Contribution of inter- and intramolecular energy transfers to heat conduction in liquids. *J. Chem. Phys.* **128**, 044504 (2008).
65. Surblys, D., Matsubara, H., Kikugawa, G. & Ohara, T. Application of atomic stress to compute heat flux via molecular dynamics for systems with many-body interactions. *Phys. Rev. E* **99**, 051301 (2019).
66. Surblys, D., Matsubara, H., Kikugawa, G. & Ohara, T. Methodology and meaning of computing heat flux via atomic stress in systems with constraint dynamics. *J. Appl. Phys.* **130**, 215104 (2021).
67. Boone, P., Babaei, H. & Wilmer, C. E. Heat flux for many-body interactions: corrections to LAMMPS. *J. Chem. Theory Comput.* **15**, 5579–5587 (2019).
68. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids* 2nd edn (Oxford University Press, 2017).
69. Farahvash, A., Leontyev, I. & Stuchebrukhov, A. Dynamic and electronic polarization corrections to the dielectric constant of water. *J. Phys. Chem. A* **122**, 9243–9250 (2018).

ACKNOWLEDGEMENTS

The numerical calculations were conducted on the five supercomputer systems, (1) Fugaku at the RIKEN Center for Computational Science, Kobe, Japan, (2) the supercomputer at the Research Center for Computational Science, Okazaki, Japan (Project: 21-IMS-C126, 22-IMS-C125), (3) the supercomputer Ohtaka at the Supercomputer Center, the Institute for Solid State Physics, the University of Tokyo, Tokyo, Japan, (4) the supercomputer TSUBAME3.0 at the Tokyo Institute of Technology, Tokyo, Japan, and (5) the supercomputer ABCI at the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. This work was supported by the following five grants: (1) a JST CREST (Grant Number JPMJCR19I3 to J.M. and R.Y.), (2) the MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Project ID: hp210264 to R.Y.), (3) the Grant-in-Aid for Scientific Research (A) from the Japan Society for the Promotion of Science (19H01132 to R.Y.), (4) the Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (22K11949 to Y.H.), and (5) the HPCI System Research Project (Project ID: hp210213 to Y.H.). The authors are grateful to Hiroki Sugisawa from Mitsubishi Chemical Corporation and the RadonPy consortium members for helpful discussions about molecular dynamics simulations. The authors are also grateful to the PoLyInfo development team at National Institute for Materials Science for providing benchmark data.

AUTHOR CONTRIBUTIONS

Y.H. and R.Y. conceived the study. Y.H. developed the Python library RadonPy, performed all MD calculations, analyzed the data, and drafted the manuscript. J.S., J.M., and R.Y. reviewed 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/s41524-022-00906-4.

Correspondence and requests for materials should be addressed to Yoshihiro Hayashi or Ryo Yoshida.

Reprints and permission 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.

Supplementary Information

RadonPy: Automated Physical Property Calculation using All-atom Classical Molecular Dynamics Simulations for Polymer Informatics

Yoshihiro Hayashi,[1,2]* Junichiro Shiomi,[1,2,3] Junko Morikawa,[1,4] Ryo Yoshida[1,5,6]*

[1] The Institute of Statistical Mathematics (ISM), Research Organization of Information and Systems, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

[2] Department of Mechanical Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan

[3] Institute of Engineering Innovation, The University of Tokyo, 2-11 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

[4] Department of Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

[5] Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

[6] Department of Statistical Science, School of Multidisciplinary Science, The Graduate University of Advanced Studies (SOKENDAI), 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

E-mail: (Y.H.) yhayashi@ism.ac.jp, (R.Y.) yoshidar@ism.ac.jp

Table of Contents

Supplementary Figures

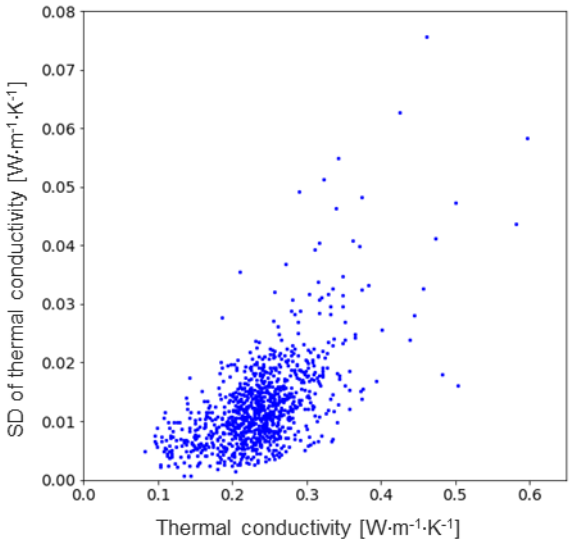
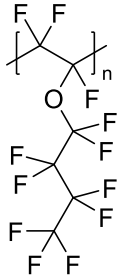
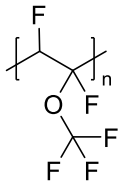


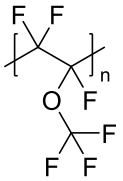
Figure 1. Relationship between thermal conductivity and its standard deviation (SD).



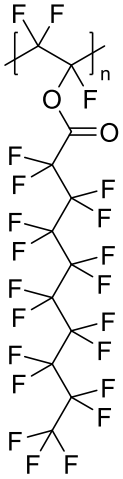
PI310 (1.780)



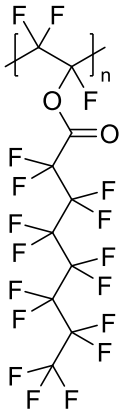
PI314 (1.750)



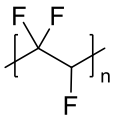
PI315 (1.816)



PI349 (1.803)



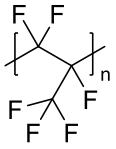
PI351 (1.766)



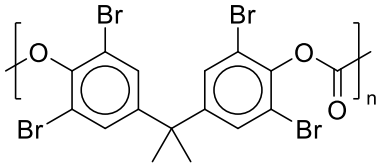
PI512 (1.799)

PI513 (1.914)

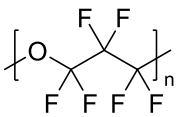
PI514 (1.892)



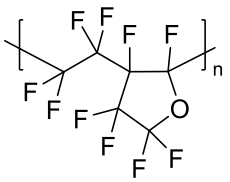
PI516 (1.791)



PI797 (1.816)

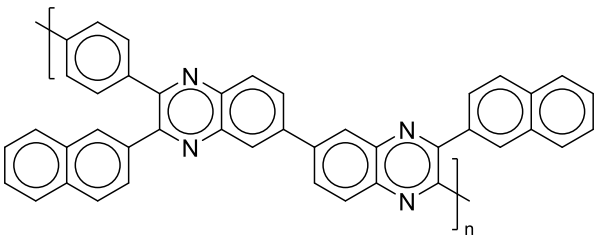


PI875 (1.851)

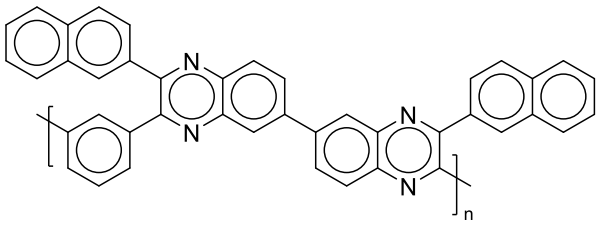


PI897 (1.822)

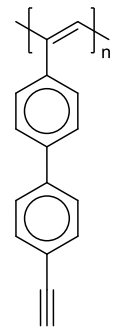
Figure 2. Chemical structure of high-density (>1.75 g·cm^{-3}) polymers and their calculated density (g cm^{-3}).



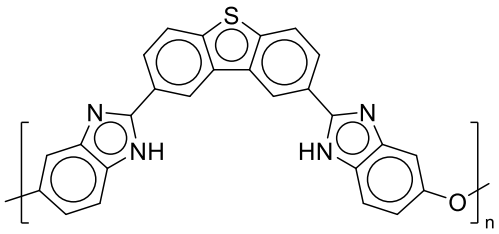
PI821 (1.776)



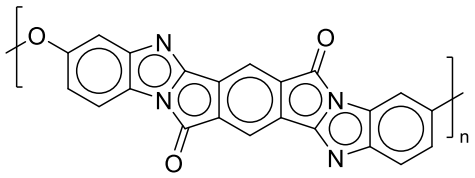
PI822 (1.799)



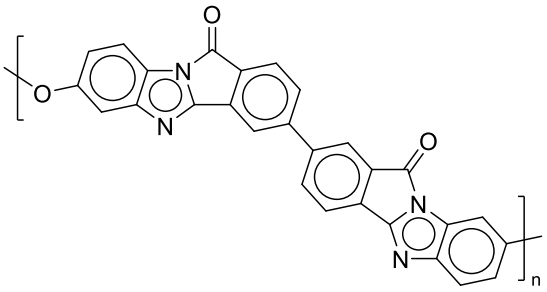
PI844 (1.765)



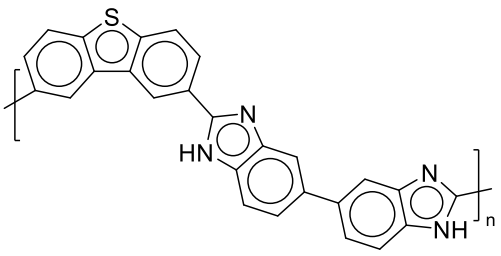
PI874 (1.820)



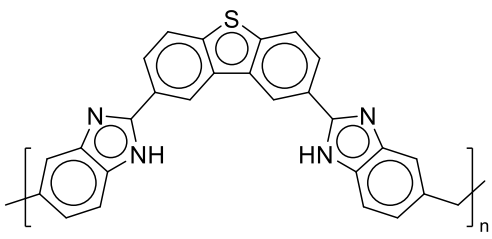
PI910 (1.804)



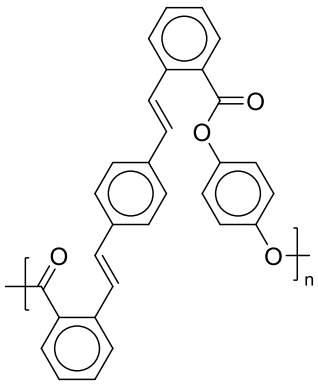
PI911 (1.839)



PI917 (1.805)



PI918 (1.808)



PI941 (1.764)

Figure 3. Chemical structure of high-refractive-index (>1.75) polymers and their calculated refractive index.

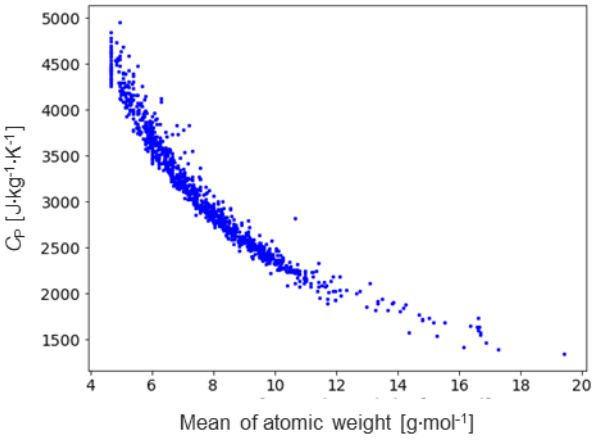


Figure 4. Correlation of mean atomic weights with specific heat capacity.

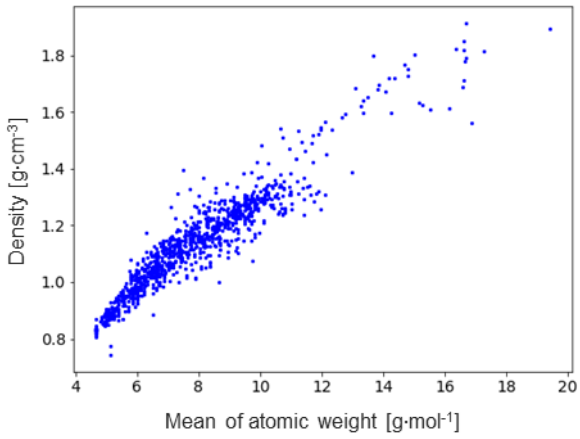


Figure 5. Correlation between mean atomic weight and density.

Supplementary Discussion

1. Examination of box size effects

We tested the effect of box size on several properties. For the evaluation, 28 different polymers were selected by taking structural variations into account. Supplementary Figures 6–10 below show the change of five properties (density, C_P, refractive index, linear expansion coefficient, and volume expansion coefficient) of the 28 polymers, of which chemical structures are shown in Supplementary Figure 11, when the number of polymer chains was varied from 10 to 50 with the number of atoms in each of a polymer chain set to 1,000, and when the number of atoms in each of a polymer chain was increased to 2,000 and the number of polymer chains was set to 25. According to the experimental results, the box size had no significant effect on the calculated properties. Therefore, we decided to perform the MD calculations with 10 polymer chains with approximately 1,000 atoms in each of a polymer chain.

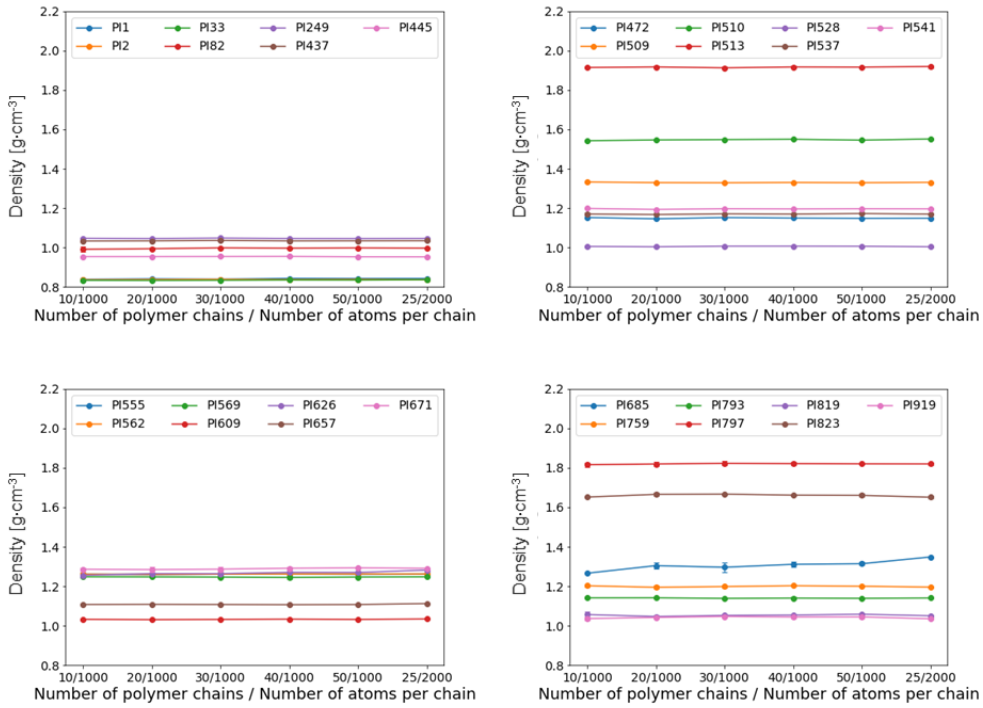


Figure 6. Box size effect on density.

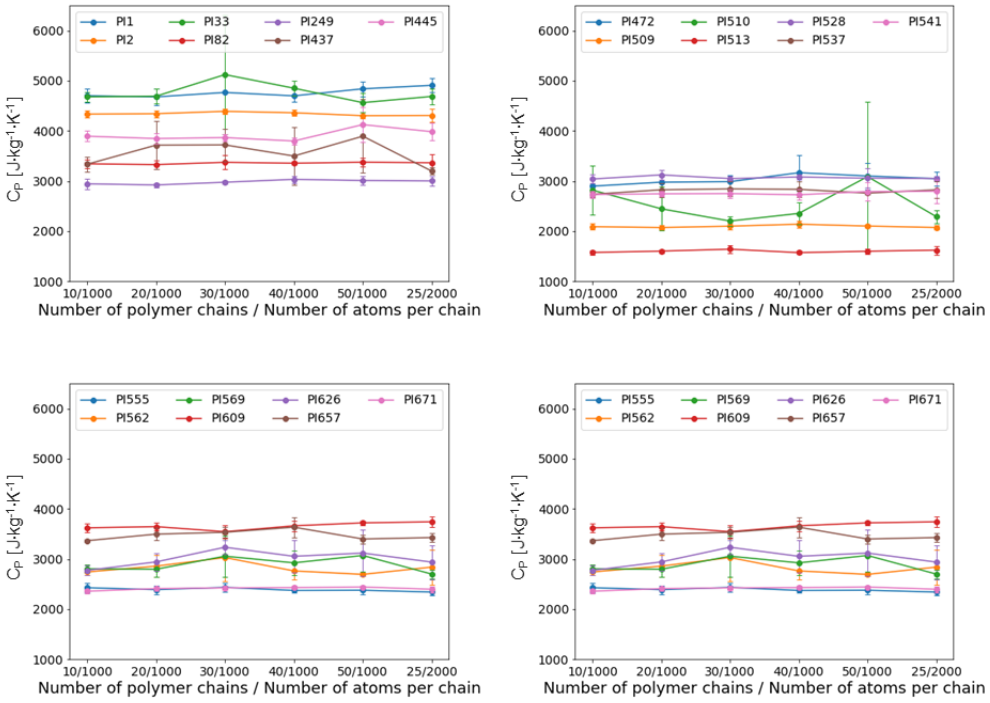


Figure 7. Box size effect on specific heat capacity (C_P).

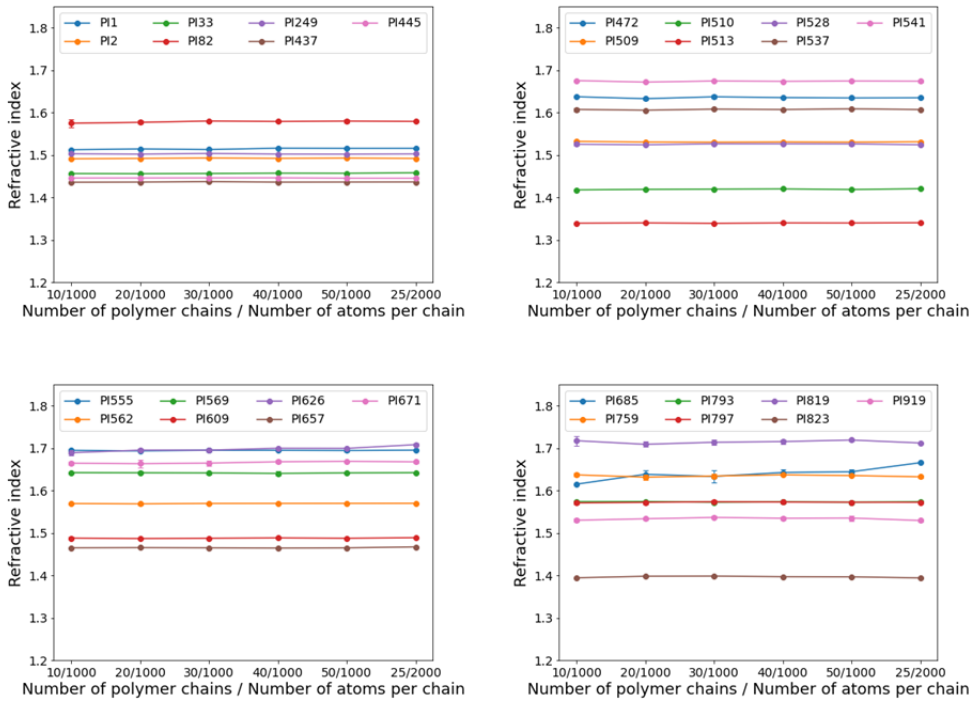


Figure 8. Box size effect on refractive index.

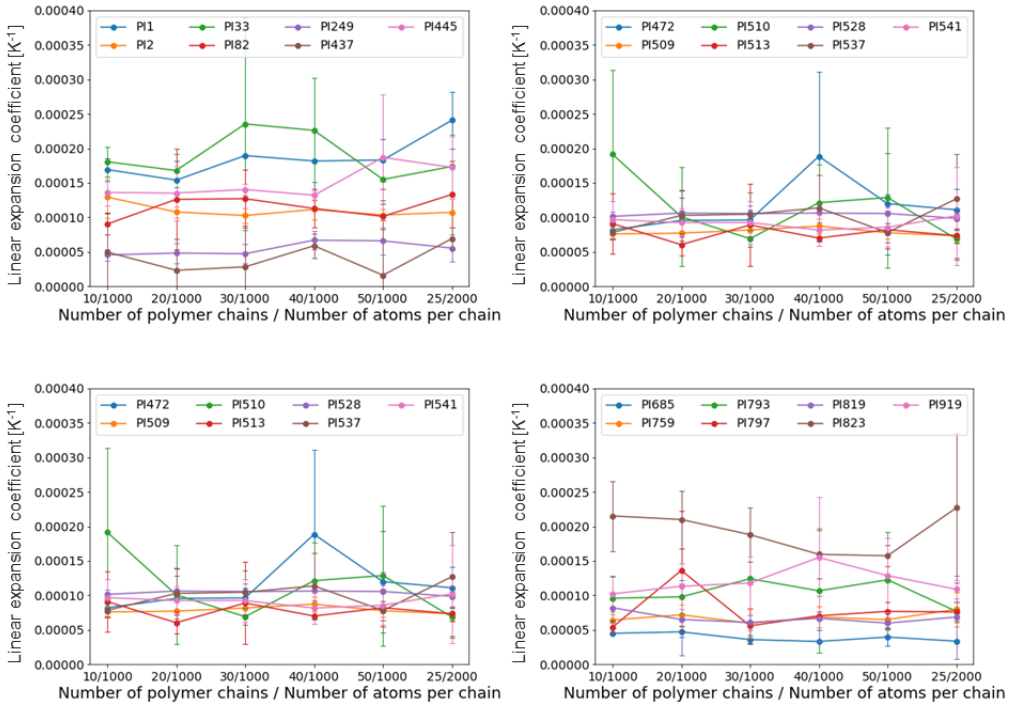


Figure 9. Box size effect on linear expansion coefficient.

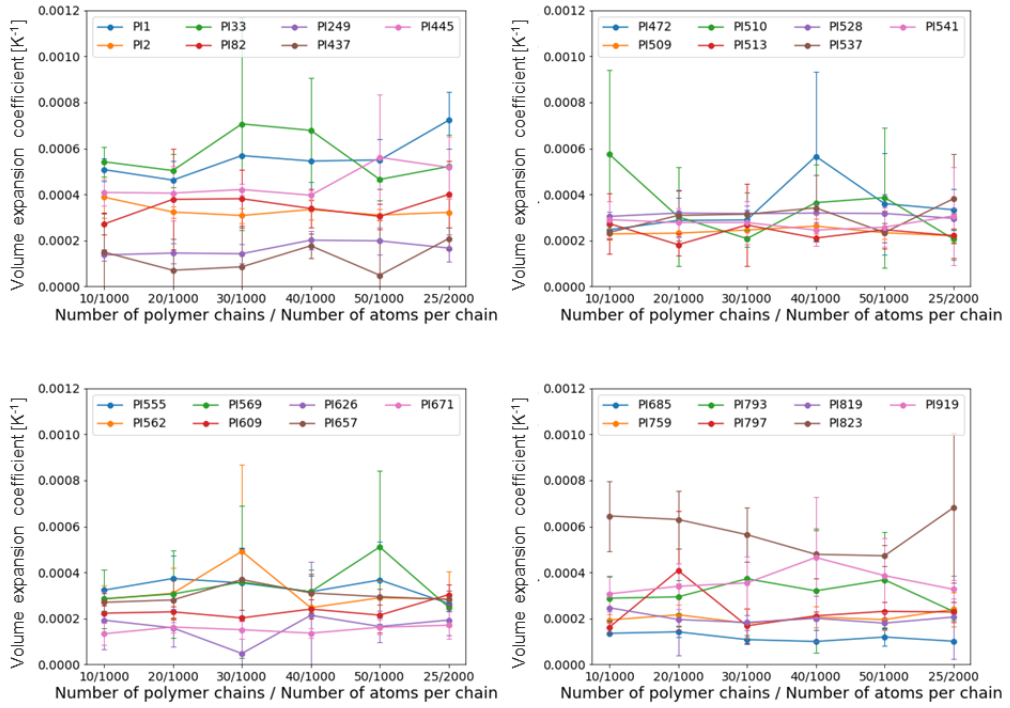


Figure 10. Box size effect on volume expansion coefficient.

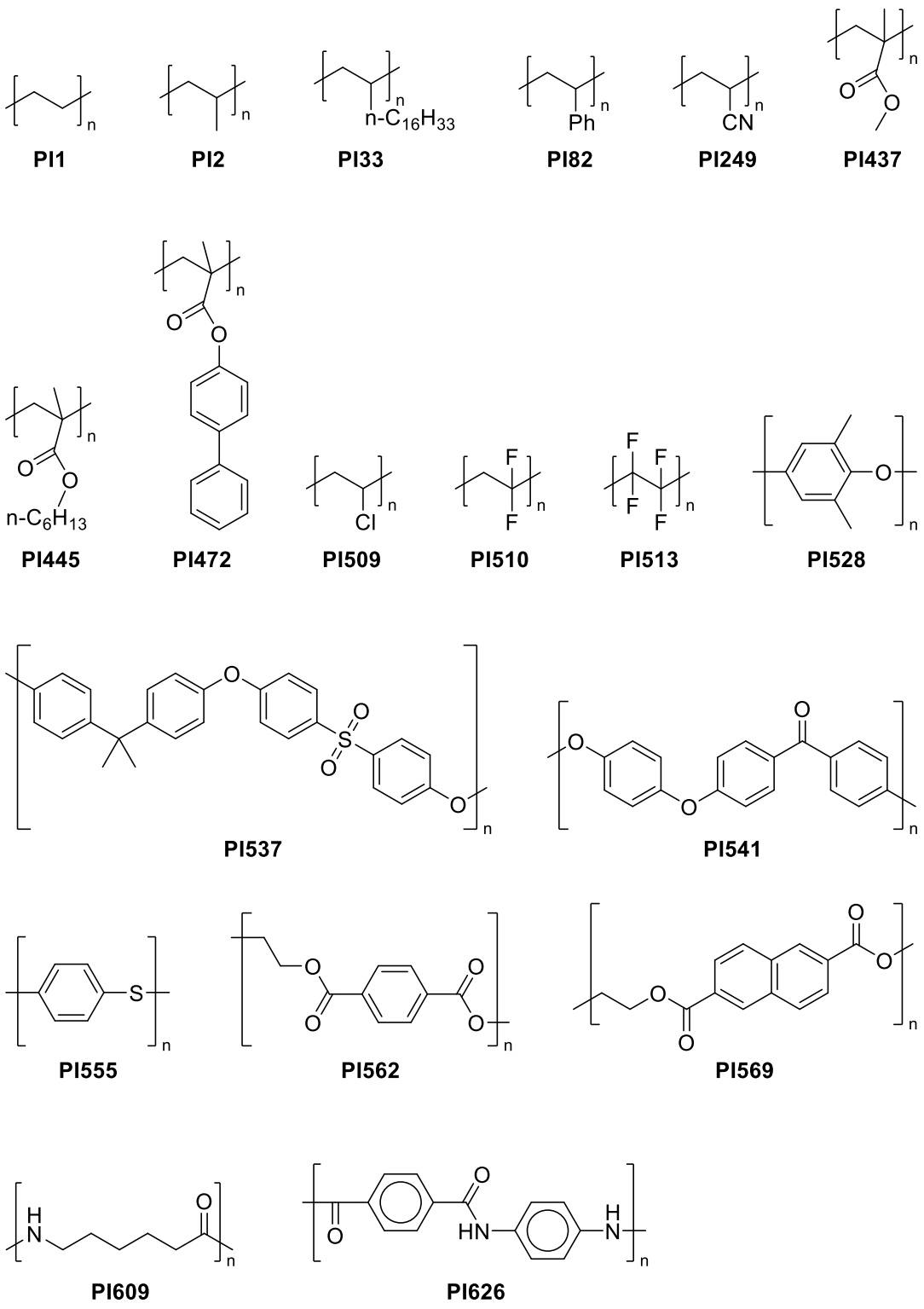


Figure 11. Chemical structure of 28 polymers in examination of box size effects.

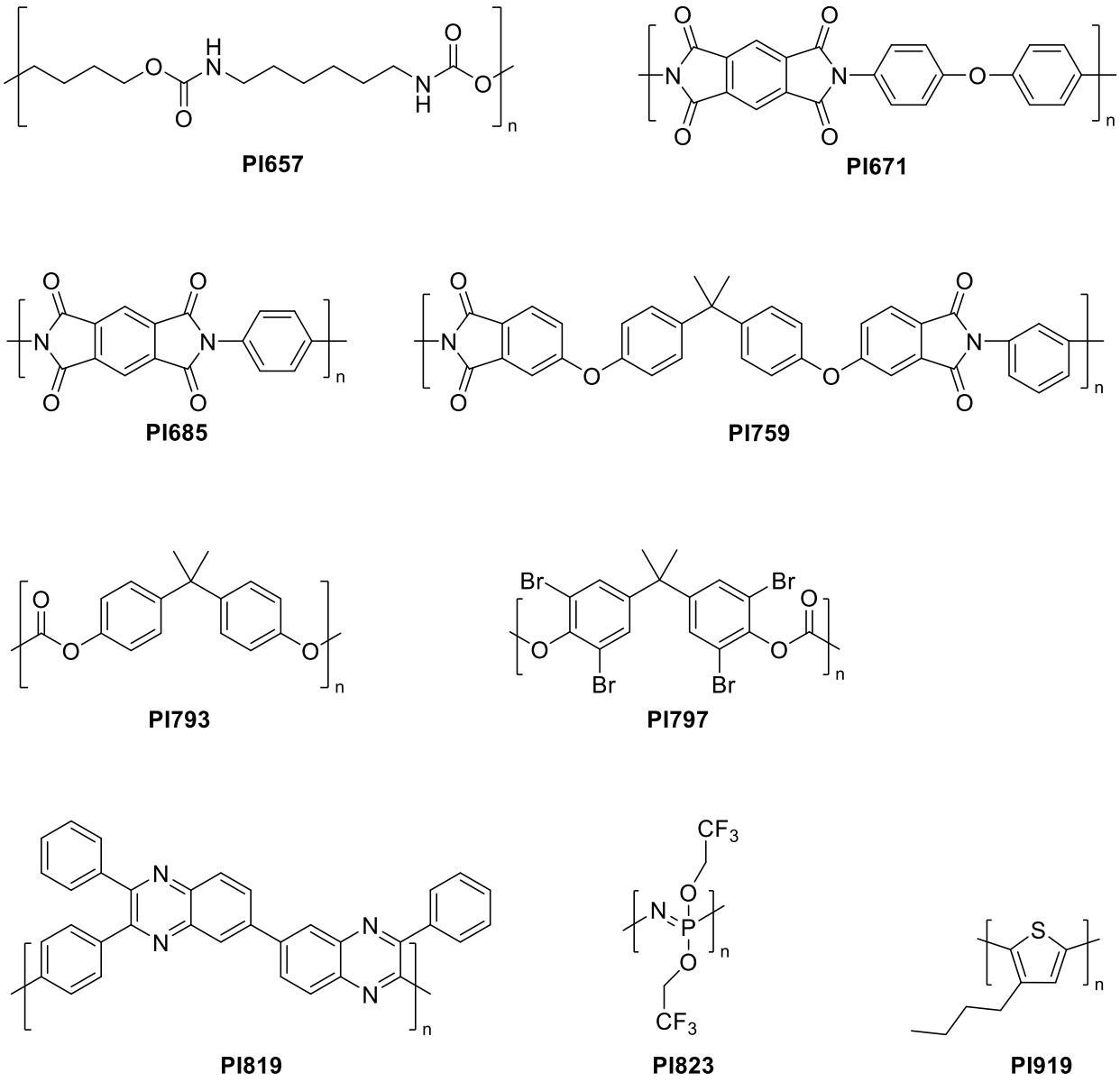


Figure 11. Continued.

2. Correlation between the $C_P^{PoLyInfo}/C_P^{MD}$ ratio and bond-stretching and -bending force constants

Einstein's formulas for the quantum-corrected ($C^{quantum}$) and classical ($C^{classical}$) heat capacities are respectively expressed as

$$C^{\mathrm{quantum}} = 3Nk_{\mathrm{B}}\left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)^2 \frac{\exp\left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)}{\left[\exp\left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)-1\right]^2} \quad (1)$$

$$C^{\mathrm{classical}} = 3Nk_{\mathrm{B}} \quad (2)$$

where N is the number of atoms, k_{B} is the Boltzmann constant, $\hbar$ is the Planck constant, ω is the frequency, and T is the temperature. Hence, the $C^{quantum}/C^{classical}$ ratio is defined as

$$\frac{C^{\mathrm{quantum}}}{C^{\mathrm{classical}}} = \left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)^2 \frac{\exp\left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)}{\left[\exp\left(\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)-1\right]^2} \quad (3)$$

This ratio decreases monotonically with increasing frequency ω. The MD value of C_P (C_P^{MD}) was computed using Eq 8 in the main text, the value of which is the classical C_P without quantum effects. On the other hand, the experimental value of C_P in PoLyInfo includes quantum effects. The frequency of the bond stretching and bending increases with increasing force constants. Therefore, the observed C_P in PoLyInfo ($C_P^{PoLyInfo}$) to MD-calculated C_P (C_P^{MD}) ratio should decrease with the increasing mean of the bond-stretching and -bending force constants. Such correlation can be seen in Supplementary Figure 12.

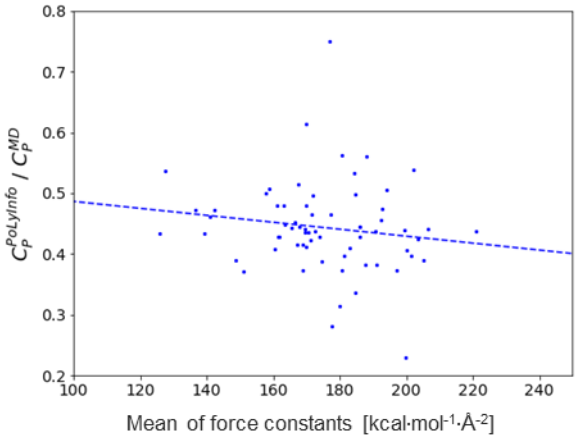


Figure 12. Correlation between the experimental C_P in PoLyInfo ($C_P^{PoLyInfo}$) to MD-calculated C_P (C_P^{MD}) ratio and the mean of the bond-stretching and -bending force constants. The GAFF2 force field parameters K_b and K_a of Eq 2 in the main text were used as the values of the force constants.

Supplementary Notes

1. Getting started with RadonPy

RadonPy is distributed at the GitHub (https://github.com/RadonPy/RadonPy/tree/main). Documentation, including installation guides, system requirements, and a detailed description of the module set and user interface, can be accessed through the tutorial on the GitHub site (https://github.com/RadonPy/RadonPy/blob/main/docs/RadonPy_tutorial_20220331.pdf).

Supplementary Figure 13 shows sample code to perform equilibrated and non-equilibrated MD calculations for a polystyrene (SMILES string: *C(C*)c1ccccc1). See the online tutorial for details on the code and function and options defined in each module.

```python
from radonpy.core import utils, poly
from radonpy.ff.gaff2_mod import GAFF2_mod
from radonpy.sim import qm
from radonpy.sim.preset import eq, tc

smiles = '*C(C*)c1ccccc1'
ter_smiles = '*C'
temp = 300
press = 1.0
omp_psi4 = 10
mpi = 10
omp = 1
gpu = 0
mem = 10000
work_dir = './work_dir'
ff = GAFF2_mod()

if __name__ == '__main__':
    # Conformation search
    mol = utils.mol_from_smiles(smiles)
    mol, energy = qm.conformation_search(mol, ff=ff, work_dir=work_dir,
                                         psi4_omp=omp_psi4, mpi=mpi, omp=omp, memory=mem, log_name='monomer1')

    # Electronic propety calculation
    qm.assign_charges(mol, charge='RESP', opt=False, work_dir=work_dir, omp=omp_psi4, memory=mem, log_name='monomer1')
    qm_data = qm.sp_prop(mol, opt=False, work_dir=work_dir, omp=omp_psi4, memory=mem, log_name='monomer1')
    polar_data = qm.polarizability(mol, opt=False, work_dir=work_dir, omp=omp_psi4, memory=mem, log_name='monomer1')

    # RESP charge calculation of a termination unit
    ter = utils.mol_from_smiles(ter_smiles)
    qm.assign_charges(ter, charge='RESP', opt=True, work_dir=work_dir, omp=omp_psi4, memory=mem, log_name='ter1')

    # Generate polymer chain
    dp = poly.calc_n_from_num_atoms(mol, 1000, terminal1=ter)
    homopoly = poly.polymerize_rw(mol, dp, tacticity='atactic')
    homopoly = poly.terminate_rw(homopoly, ter)
```

Figure 13. Sample code of performing equilibrated and non-equilibrated MD calculations for polystyrene using RadonPy.

```
# Force field assignment
result = ff.ff_assign(homopoly)
if not result:
    print('[ERROR: Can not assign force field parameters.]')

# Generate simulation cell
ac = poly.amorphous_cell(homopoly, 10, density=0.05)

# Equilibration MD
eqmd = eq.EQ21step(ac, work_dir=work_dir)
ac = eqmd.exec(temp=temp, press=1.0, mpi=mpi, omp=omp, gpu=gpu)

analy = eqmd.analyze()
prop_data = analy.get_all_prop(temp=temp, press=1.0, save=True)
result = analy.check_eq()

# Additional equilibration MD
for i in range(4):
    if result: break
    eqmd = eq.Additional(ac, work_dir=work_dir)
    ac = eqmd.exec(temp=temp, press=press, mpi=mpi, omp=omp, gpu=gpu)
    analy = eqmd.analyze()
    prop_data = analy.get_all_prop(temp=temp, press=press, save=True)
    result = analy.check_eq()

if not result:
    print('[ERROR: Did not reach an equilibrium state.]')

# Non-equilibrium MD for thermal condultivity
else:
    nemd = tc.NEMD_MP(ac, work_dir=work_dir)
    ac = nemd.exec(decomp=True, temp=temp, mpi=mpi, omp=omp, gpu=gpu)

    nemd_analy = nemd.analyze()
    TC = nemd_analy.calc_tc(decomp=True, save=True)
    if not nemd_analy.Tgrad_data['Tgrad_check']:
        print('[ERROR: Low linearity of temperature gradient.]')

    print('Thermal conductivity: %f' % TC)
```

Figure 13. Continued.

The specified SMILES string is converted into a Mol object by the "utils.mol_from_smiles" function. After converting the specified SMILES string into a Mol object, the quantum chemistry calculation module ("qm") is used to perform a conformation search ("qm.conformation_search") and charge assignment ("qm.assign_charges") for the input repeating unit. The sample code specifies the RESP charge model as an option. Other electronic properties such as HOMO and LUMO energies, dipole moments, and polarizability tensors are also calculated using this module. The "poly.calc_n_from_num_atoms", "poly.polymerize_rw", and "poly.terminate_rw" functions generate a polymer chain for a specified number of atoms (e.g., 1,000) with a self-avoiding random walk

algorithm. Although not shown in this paper, the current version also provides modules to generate alternating, random, and block copolymers, as described in the online tutorial. The "ff.ff_assign" method assigns the parameters of the GAFF2 force field. The "poly.amorhous_cell" function generates a simulation cell for amorphous polymers. In the sample code, 10 polymer chains are randomly arranged so that they do not overlap each other and have a density of 0.05 g·cm^{-3}. The user can also build simulation cells for polymer-polymer or polymer-solvent mixture systems. The user then specifies the temperature ("temp") and pressure ("press"), and performs an equilibrium MD calculation (the "exec" method of the "eq.EQ21step" class). The convergence of energy and density is monitored every 5 ns to determine if the system has reached equilibrium ("analy.check_eq"). If the equilibration is not reached, the equilibrium calculation is scheduled to restart from the saved previous state (the "exec" method of the "eq.Additional" class). If the equilibrium is successfully achieved, the NEMD simulation (the "exec" method of the "tc.NEMD_MP" class) is performed. When running a hybrid parallel computing with MPI (message passing interface), OpenMP (open multi-processing), and GPU (graphics processing unit), the number of MPI processes ("mpi"), OpenMP threads ("omp"), and GPUs ("gpu") can be specified according to the available computer architecture.

Supplementary Methods

1. Shotgun transfer learning

The objective is to predict experimental properties from the chemical structure of any given polymer repeating unit. For the model input, we used a subset of the 1006-dimensional descriptor vector that concatenated eight different fingerprints implemented in the RDKit package, including the extended-connectivity fingerprint (ECFP),[1] the functional-class fingerprint (FCFP),[1] topological torsion fingerprint,[2] atom pair fingerprint,[2] MACCS key, RDKit fingerprint,[3] pattern fingerprint,[3] and layered fingerprint.[3] For transfer learning, the source task was to predict the MD-calculated properties, and the target task was to predict the experimental properties in PoLyInfo.

From the MD dataset for the source task, all polymers in the experimental dataset to be predicted were excluded. Using the MD dataset, we trained 100 neural networks having randomly constructed different network structures. Each model was designed to form a fully connected pyramid-shaped layers in which the number of layers was randomly selected from three and four, and the number of neurons monotonically decreased from the input layer to the output one. All neurons in the hidden layers were activated by a rectified linear unit (ReLU),[4] and a linear transformation function form the output layer. For each model, we used a randomly chosen subset of the 400-dimensional descriptor.

We randomly selected 80%, 10%, and 10% of the experimental dataset for the training, validation, and testing for the target task, respectively. Using the training set, each pretrained source model was fine-tuned into a calibration model for the target task. Here, the pretrained model was used as a starting point for the model training. The given weights on the last output layers of the pretrained model were randomly initialized, while the learned parameters of the remaining layers were used as the initial values of training. All those parameters were then fine-tuned at a small learning rate.

The root mean square error (RMSE) of each transferred model with respect to the validation set was calculated, and the prediction performance on the test dataset was examined using the model showing the best transferability that achieved the lowest validation RMSE.

Supplementary References

1. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
2. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85 (1987).
3. Landrum, G. RDKit: open-source cheminformatics software. https://www.rdkit.org/ (2020).
4. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. *Proc. Fourteenth Int. Conf. Artif. Intell. Stat.* **15**, 315–323 (2011).