npj | Computational Materials

www.nature.com/npjcompumats

Corrected: Publisher Correction

ARTICLE OPEN

Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm

Stephen Wu[1,2], Yukiko Kondo[3], Masa-aki Kakimoto[3], Bin Yang[4], Hironao Yamada[1], Isao Kuwajima[3], Guillaume Lambard[3], Kenta Hongo[3,5,6], Yibin Xu[3], Junichiro Shiomi[3,7], Christoph Schick[4,8], Junko Morikawa[3,9] and Ryo Yoshida[1,2,3]

The use of machine learning in computational molecular design has great potential to accelerate the discovery of innovative materials. However, its practical benefits still remain unproven in real-world applications, particularly in polymer science. We demonstrate the successful discovery of new polymers with high thermal conductivity, inspired by machine-learning-assisted polymer chemistry. This discovery was made by the interplay between machine intelligence trained on a substantially limited amount of polymeric properties data, expertise from laboratory synthesis and advanced technologies for thermophysical property measurements. Using a molecular design algorithm trained to recognize quantitative structure—property relationships with respect to thermal conductivity and other targeted polymeric properties, we identified thousands of promising hypothetical polymers. From these candidates, three were selected for monomer synthesis and polymerization because of their synthetic accessibility and their potential for ease of processing in further applications. The synthesized polymers reached thermal conductivities of 0.18–0.41 W/mK, which are comparable to those of state-of-the-art polymers in non-composite thermo-plastics.

npj Computational Materials (2019)5:66 ; https://doi.org/10.1038/s41524-019-0203-2

INTRODUCTION

The ability of machine intelligence trained on massive amounts of data to match or even outperform humans has been demonstrated in intellectually demanding tasks across various fields.[1–3] As such, there is growing interest in the use of machine learning (ML) to reap substantial time and cost savings in the development of new materials.[4,5] In particular, remarkable advances have recently been made in ML for de novo molecular design.[6–10] The goal of computational molecular design is the identification of new promising molecules whose physicochemical properties meet arbitrary given requirements. Despite the growing potential of ML in materials science, its practical impacts have not been fully verified. To the best of our knowledge, the emphasis of recent studies has largely been on algorithmic developments, whereas much less work has been done on the experimental verification of computationally designed materials (except for a few works[11,12]). In the particular case of polymers, it is unprecedented that designed polymers were synthesized and experimentally confirmed. Major challenges in polymer informatics, for example, arise from the lack of data on polymeric properties and from the structural complexity/diversity of polymers.[13–15] In this study, we demonstrate the successful discovery of new polymers with high thermal conductivity that were designed by our ML algorithm, referred to as Bayesian molecular design.[16] This proof-of-concept study intended to highlight a promising new example of polymer informatics and to raise several issues that should be addressed to enable the widespread use of ML.

This study focused on the design of a chemical structure in the repeat unit of a polymer. The objective of molecular design is to generate promising hypothetical chemical structures that exhibit a set of desired properties. The chemical space of small organic molecules is known to consist of as many as 10^{60} potential candidates,[17] whereas the total number of currently known compounds is at most 10^8.[18] The emergence of ML algorithms, which can exhaustively search this very large space, can contribute significantly to expanding the frontier of the vast chemical universe. In the history of chemical informatics, there have been extensive studies into computational molecular design. Their origin dates back to the pioneering work by Venkatasubramanian et al.[19] Most such studies have focused on the use of a limited number of chemical fragments and their stochastic recombination to sequentially transform starting compounds into desired targets.[20,21] However, this approach significantly narrows the design space. To broaden the search space, more advanced ML techniques using probabilistic language models have appeared in recent years.[7,10,22,23] The Bayesian method developed in our previous work has also contributed to technological advancement in this stream.[16]

Despite remarkable methodological innovations in computational molecular design, there are still barriers to achieving a successful proof of concept. Such barriers arise mainly from the substantially limited amount of polymeric properties data, in addition to the synthetic difficulty of designed candidates, disagreements between expert knowledge and machine-acquired

[1]The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tachikawa, Tokyo 190-8562, Japan; [2]The Graduate University for Advanced Studies, Tachikawa, Tokyo 190-8562, Japan; [3]Center for Materials research by Information Integration (CMI^2), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Ibaraki 305-0047, Japan; [4]Institute of Physics and Competence Centre CALOR, University of Rostock, 18059 Rostock, Germany; [5]Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan; [6]PRESTO, JST, Kawaguchi, Saitama 332-0012, Japan; [7]The University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan; [8]Tokyo Tech World Research Hub Initiative (WRHI), Tokyo Institute of Technology, Tokyo 226-8503, Japan and [9]Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan
Correspondence: Junko Morikawa (morikawa.j.aa@m.titech.ac.jp) or Ryo Yoshida (yoshidar@ism.ac.jp)

Received: 21 October 2018 Accepted: 28 May 2019
Published online: 21 June 2019

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

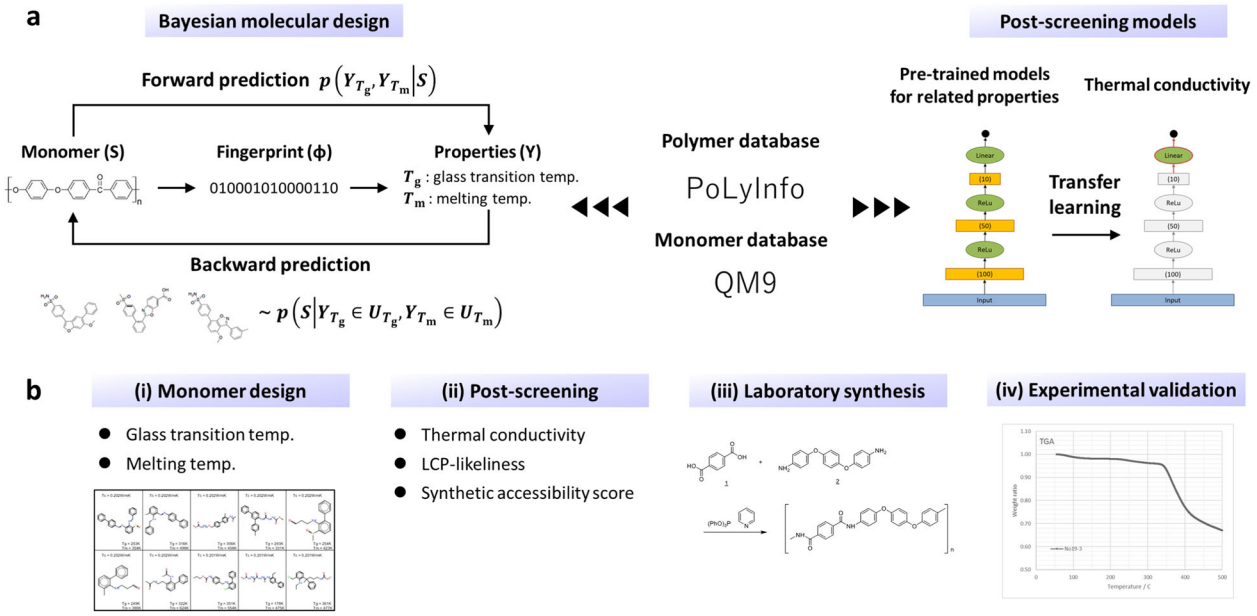

Fig. 1 Machine learning (ML)-assisted de novo design and experimental validation of new polymers. **a** The objective of forward prediction is to derive a model that describes polymeric properties (e.g., glass transition temperature (T_g) and melting temperature (T_m)) as a function of chemical structures in the constitutional repeat units. The forward model trained on the data set from PoLyInfo was inverted to obtain a backward model, which was conditioned by desired property regions (U_{T_g} and U_{T_m}). The backward model produced a library of hypothetical chemical structures that exhibit the desired properties. In addition, we developed a prediction model of thermal conductivity, which was utilized in the post-screening of the produced library. Here, an ML framework called transfer learning was used to overcome the issue of limited data on thermal conductivity: prediction models of proxy properties were pre-trained on given large data sets from PoLyInfo and QM9, and then the pre-trained models were fine-tuned using the limited data on the target property. We did not use the transferred models directly for the molecular design calculation because their generalization capability would likely be restricted on the design space spanned by the few training polymers. **b** Analytic workflow consisting of four internal steps towards materials discovery

intelligence and the difficulty of meeting stringent requirements in practical applications. Indeed, the experimental data set on thermal conductivity that we used was limited in size, as it consisted of only 28 training instances. The limited amount of training data rendered ordinal ML methods impractical for prediction, as demonstrated. In addition, as a second-rank tensor, thermal conductivity can vary substantially across polymer processing operations, such as laminating films and spinning fibres, where anisotropic molecular orientation is introduced. Most of these variations have not been recorded in the current database. Therefore, we failed to derive practically useful prediction models directly from the given data.

Our ML workflow was designed to overcome the issue of limited data. A solution to mitigate this barrier was to exploit proxy properties related to thermal conductivity as alternative design targets. In the Bayesian molecular design process that generated a library of virtual chemical structures, we specified a higher region of glass transition temperatures and melting temperatures as alternative design targets, for which sufficient data were given to obtain reliable prediction models. We know empirically that polymers with higher glass transition temperatures tend to be achieved by rigid structures, which result in higher thermal conductivity. In addition, taking into account the ease of processing of polymers, we selected designed candidates by eliminating those with exceedingly high glass transition temperatures. Furthermore, an ML framework referred to as "transfer learning" was introduced to obtain a thermal conductivity model with the given small data set. For the given target property to be predicted from the limited supply of data, models on physically related proxy properties were pre-trained using an adequate amount of data, which captured common features relevant to the target task of predicting thermal conductivity. Repurposing such machine-acquired features for the target task produced an outstanding achievement in the prediction accuracy even with the exceedingly small data set. We used the transferred thermal conductivity model to screen promising candidates over the virtual library that was produced by targeting the glass transition and melting temperatures, and then proceeded with laboratory synthesis and experimental characterization of the thermophysical properties. Figure 1 outlines the analytic workflow of this study. R codes to reproduce key results are available at https://github.com/stewu5/HighTCond_Polymer_iqspr.

Finally, three chemical structures were selected from a list of 1000 designed candidates on the basis of criteria involving synthetic accessibility (SA) and ease of processing, which are required for the practical use of enhanced newly designed polymers with high thermal conductivity. Then, the monomers of these candidates were synthesized and polymerized using retrosynthetic routes designed by synthetic chemists. The synthesized polymers exhibited a glassy state, and two of them were crystallized by annealing. We also observed the change in the crystal system resulting from additional chemical reaction during annealing. Their thermal conductivities reached 0.18–0.41 W/mK within non-composite thermo-plastics in amorphous and semi-crystalline states.

RESULTS

Data

PoLyInfo[24] has recorded approximately one hundred kinds of polymeric properties of chemical structures in terms of the constitutional repeat units. Narrowing the focus to 14,423 unique homopolymers in the database, we generated ML models that describe a set of properties as a function of the chemical structures. We extracted a total of 38,310 structure–property relationships with respect to thermal conductivity (λ), glass transition temperature (T_g), melting temperature (T_m) and density

Table 1. Summary of the structure–property relationship data sets from PoLyInfo and QM9 and their classification by use

Use	Database	Property	Number of structures	Number of samples	Max σ of within-polymer fluctuation	Range of temperature
CMD, TL_λ	PoLyInfo	T_g	5917	17,001	30 °C	N/A
CMD, TL_λ	PoLyInfo	T_m	3234	12,374	30 °C	N/A
TL_λ	PoLyInfo	ρ	1516	8613	0.50 g/cm^3	10–35 °C
TL_λ	QM9	C_V	133,805	133,885	0.97 cal/molK	25 °C
Post-screening	PoLyInfo	λ	28	322	0.10 W/mK	10–35 °C

For the PoLyInfo data sets, only homopolymers that have linearly connected structures with no additives or fillers were selected: CMD, used for forward modelling in the molecular design calculation; post-screening, used for transfer learning to obtain a screening model of λ; TL_λ, used to obtain pre-trained source models for transfer learning; σ, standard deviation; T_g glass transition temperature, T_m melting temperature

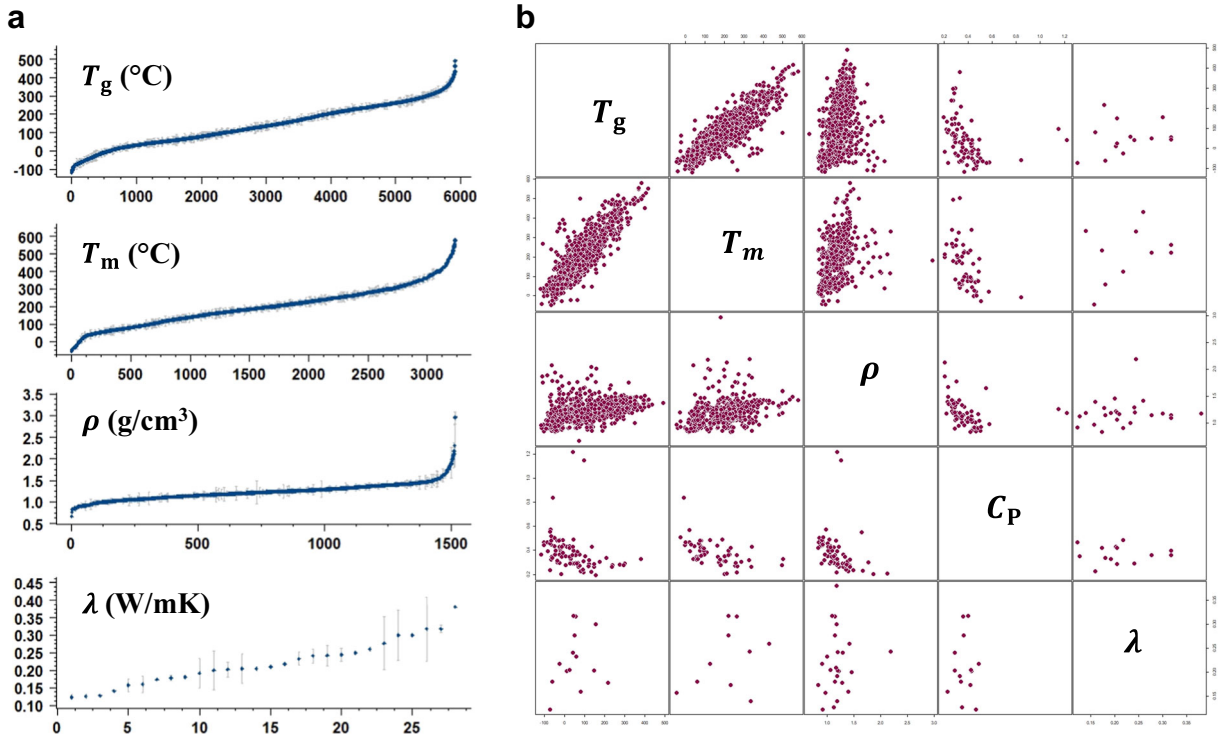

Fig. 2 Summary of PoLyInfo data. **a** Average properties of recorded polymers are plotted in ascending order with error bars indicating ±1σ (σ: standard deviation). **b** Scatterplot matrix that summarizes the joint distribution of the five polymeric properties. C_P denotes specific heat capacity at constant pressure

(ρ), as summarized in Table 1. When multiple property values were recorded for a polymer under the same experimental condition, they were reduced to the mean value.

The volume of data varies significantly across different properties. For example, PoLyInfo recorded multiple values of T_g and T_m for 5917 and 3234 unique homopolymers, respectively. In contrast, there were 322 observations for only 28 homopolymers with respect to λ around room temperature (10–35 °C). Moreover, λ varied considerably even within the same polymer, as shown in Fig. 2a (unreliable data were removed by curation). Such within-polymer fluctuations could arise from differences in processing operations, higher-order molecular structures or any other measurement conditions that varied in different studies. Unfortunately, such information was mostly not recorded in the database. Consequently, supervised learning directly using the given data on λ failed to reach desirable levels of prediction accuracy (Fig. 3d).

The lack of data in terms of both quantity and quality prompted us to pursue a strategic solution based on the use of T_g and T_m as proxy target properties in the de novo design calculation, as described later. In addition, we applied transfer learning to obtain a prediction model on λ, which was used in the post-screening process. In the construction of pre-trained models for transfer learning, we utilized the four data sets from PoLyInfo and the QM9 data set[25,26] that records the computational data of specific heat capacity at constant volume (C_V) for 133,805 small organic molecules, which were calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry.

Overview of Bayesian molecular design

The objective of the de novo design calculation is to algorithmically create a chemical structure S in a polymer repeat unit, that is, monomer, for which n polymeric properties $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ lie in a desired region U. The chemical structure S that represents a configuration of atoms and chemical bonding is encoded as a sequence of SMILES symbols (simplified molecular-input line-entry system[27]) in which $S = s_1 s_2 \ldots s_g$ forms a variable-length string, here consisting of g letters. For example, a SMILES string representing

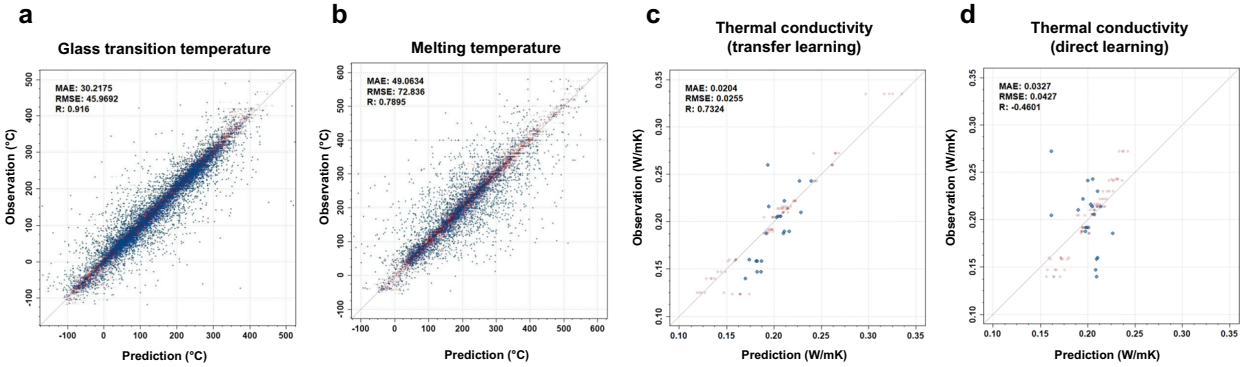

Fig. 3 Performance of forward prediction models. **a**, **b** Five-fold cross-validation of trained linear models for glass transition temperature (T_g) and melting temperature (T_m). All predicted values in the five validation sets are plotted against observed values, denoted by blue dots (red for the training). The mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R) are shown in each plot. **c**, **d** Validation results for the prediction model on λ that exhibited the best transferability (MAE = 0.0204 W/mK) out of 1000 pre-trained models on T_m. The prediction results of the best transferred model and a random forest model trained directly using the 28 data points for λ (MAE = 0.0327 W/mK) are shown in **c**, **d** respectively

phenol (C6H6O) is C1=CC=C(C=C1)O, where C and O indicate the aliphatic carbon and oxygen atoms, and = indicates the double bond. The start and terminal of a ring closure are designated by a common digit, 1 in this case, and the side chain is enclosed in parentheses, "(" and ")".

The Bayesian molecular design framework relies on the statement of Bayes' law:

$$p(S|\mathbf{Y} \in U) \propto p(\mathbf{Y} \in U|S)p(S), \quad (1)$$

where $p(A|B)$ denotes the conditional probability distribution of A given B. ML models on n properties were trained with structure–property relationship data sets that define the forward model $p(\mathbf{Y}|S) = \prod_{i=1}^{n} p(Y_i|S)$. Imposing the desired region U on $\boldsymbol{Y}$ provides $p(\boldsymbol{Y} \in U|S)$ on the right-hand side of Eq. (1). This probability evaluates the goodness of fit of S with respect to the property requirements. The prior distribution $p(S)$ serves to reduce the occurrence of chemically unfavourable or unrealistic structures in designed molecules as it assigns zero or lower probability masses to invalid or unrealistic chemical structures. For a given $p(S)$, Bayes' law inverts the forward model ($S \to \boldsymbol{Y}$) to obtain the backward model $p(S|\boldsymbol{Y} \in U)$ ($\boldsymbol{Y} \to S$). We then draw a random sample of the SMILES string (S) from high-probability regions of the backward model using a sequential Monte Carlo (SMC) method[28] to identify promising monomers that exhibit the desired U. The R language library iqspr 1.0[16] (the latest version is 2.4) that we developed was used to pipeline the forward and backward calculations.

The SMC method shares a common algorithmic structure with genetic algorithms. The prior $p(S)$ constitutes the most important factor that influences the structural features of the produced sample. In the implementation of iqspr, the prior is modelled by a probabilistic language model that we call the extended n-gram, which takes the form $p(S) = p(s_1)\prod_{i=2}^{g} p(s_i|s_{i-1}, \ldots, s_1)$. The occurrence probability of the ith letter, s_i, depends on the preceding s_{i-1}, …, s_1. The conditional probability $p(s_i|s_{i-1}, \ldots, s_1)$ is estimated by the frequencies of substring patterns in a training set of existing chemical structures. The trained language model is anticipated to successfully learn structural patterns of the existing compounds or implied contexts of "chemically favourable or realistic" structures. For a given randomly chosen substring s_{i-1}, …, s_1, the trained probabilistic model is used to modify the rest of the components by recursively adding subsequent letters according to the conditional probabilities, which encode the acquired chemical reality. In this way, a currently given set $\{S_1, \ldots, S_M\}$ of M chemical structures could be consecutively updated to a new population. The fitness scores of the updated structures are assessed based on the forward model. Structures with better fitness have a better chance to survive in the next generation. This process is iterated many times, and at the end, samples from the targeted posterior are produced. The algorithmic details are shown in Ikebata et al.[16]

As mentioned in the beginning, molecular design techniques using probabilistic language models have appeared rapidly since 2017. The present method has some distinctive methodological features, which are briefly noted here. One of the distinctive features of our method is that it relies on the Bayesian framework, which provides a natural way to pipeline the workflow between the forward and backward prediction processes. In addition, the Bayesian approach benefits from the principle-based handling of "uncertainty" in the prediction models. A chemical structure S is designed based on $P(S|Y \in U)$, the probability that for a given S, its property Y lies in a desired region U in the presence of prediction uncertainty in the trained forward model $S \to Y$. The design results depend strongly on whether or not the uncertainty is considered.

Another feature lies in the architecture of probabilistic language models. One major difficulty of constructing a SMILES generator is associated with the rules of grammar regarding the expression of rings and branching components. To be specific, unclosed ring and branch indicators must be prohibited. For instance, any strings extended rightward from a given $s_{1:6}$ = CC(C(C should eventually contain two closing letters, ")". In addition, the issue of "long-term dependency" must be addressed: neighbours in a string are not always adjacent in the original molecular graph. For example, the occurrence probability of the last carbon in a structure expressed by CCCCC(CCCCC)C should be affected more by the letters in the main chain, that is, the first five "C" in the string, than by the adjacent letters because the substring in the parentheses constitutes a branch from the main chain. It is quite difficult for ordinal language models to capture such intrinsic patterns in SMILES representations without any special operations. Most recent works have relied on deep neural networks (DNNs) as molecular generators, such as Recurrent Neural Networks[7] or Variational Autoencoder.[8] In general, massive numbers of training instances are needed for such DNNs to learn the underlying high-level contexts of chemical rules and the grammatical rules of SMILES in fully data-driven analysis without any prior knowledge. However, the extended n-gram that we developed is a highly engineered model specifically developed for the ML of the SMILES language. It required significantly less data to train than the DNNs.

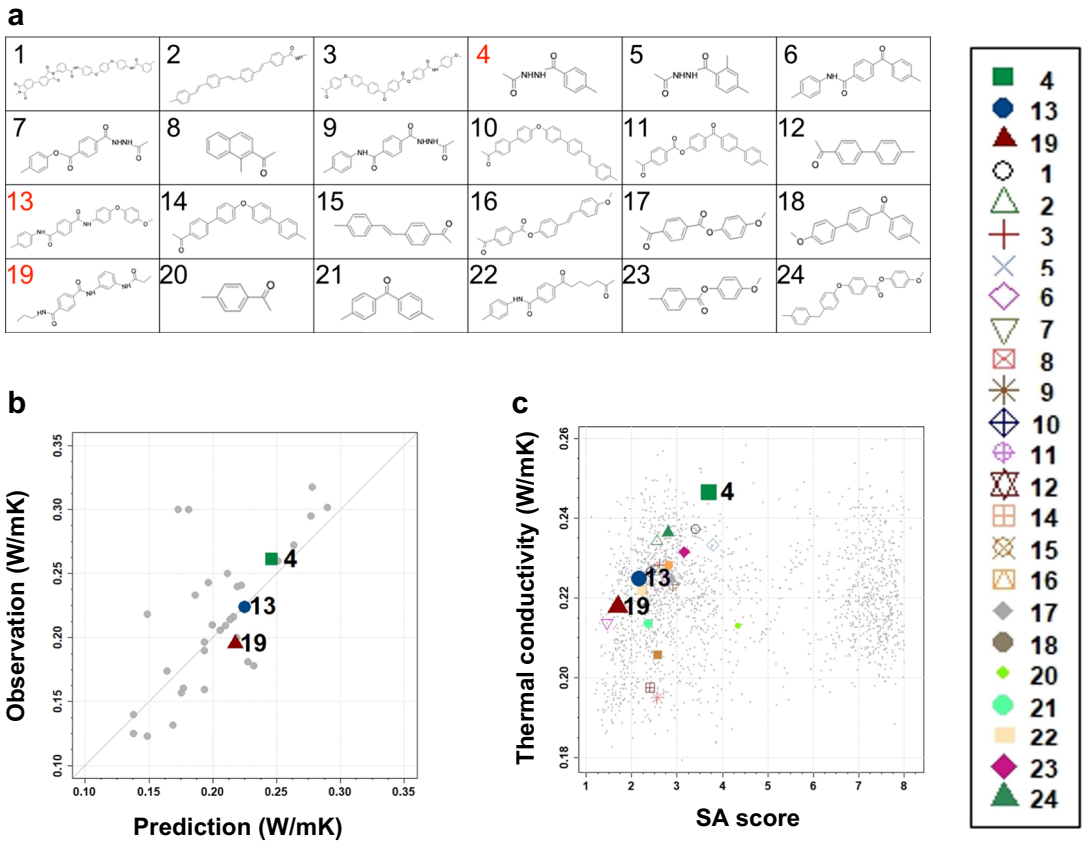

Fig. 4 Summary of screening results. **a** Repeat units of 24 screened polymers. The synthesized polymers are numbered in red. A zoomed version is available in SI (Fig. S3) **b** The predicted and observed values of λ for the 28 existing polymers recorded in PoLyInfo (grey) and the three synthesized polyamides (coloured and numbered). **c** Predicted properties are shown on λ vs. SA scores. Grey dots denote the 1000 designed candidates, and the 24 screened candidates are colour coded as described in the legends on the right-hand side. The numbers are assigned to the newly synthesized polyamides

Forward prediction on T_g and T_m

Forward models on T_g and T_m were used as the proxy targets in the Bayesian design calculation. The chemical structure of a monomer was encoded into a descriptor vector of binary digits comprised of multiple molecular fingerprints, such as the extended connectivity fingerprint.[29] For T_g or T_m, a linear regression model, which described the polymeric property as a function of molecular fingerprints, was trained on a random selection of 80% of the instances of the given data in PoLyInfo. Figure 3a, b show the prediction performance of these models on the validation data set.

Forward prediction on λ

For the post-screening, we developed neural network models for λ using a transfer learning technique to break the barrier of the exceedingly limited data. First, we generated 1000 pre-trained neural networks for T_g, T_m and ρ using the data from PoLyInfo, as well as 1000 models for C_V with the QM9 data set. Each neural network consisted of a fully connected pyramid structure in which the size of layers and the number of neurons were randomly chosen. For a given pre-trained model, we refined the weight parameters using the small data set on λ, for which the initial values of parameters were taken from the pre-trained neural networks of the related tasks. Among the 1000 pre-trained models of each property, we identified the best transferable model of predicting λ that exhibited the highest generalization capability on the five validation sets, each randomly constructed from 20% of the given data. Figures 3c and 4b show the performance of two models on λ that were transferred from T_m and C_V, respectively. The model that performed best in predicting λ was transferred from a pre-trained model of the monomer-level C_V. The prediction accuracy of the transferred model reached 0.0204 W/mK of the mean absolute error (MAE), as the MAE was reduced by 40% compared with that of a random forest model trained directly using the 28 data points (MAE = 0.0327 W/mK) (see Fig. 3c, d). Further details are described in the Supplementary Information (SI), for example, a successful transfer from T_g to λ.

Design targets

Transfer learning has substantially improved the accuracy of predicting λ. Nevertheless, we could not dispel uncertainty in the generalization capability because the given model was validated only on an input subspace spanned by the 28 training polymers, which was rather small with respect to the entire materials space. The use of such a unreliable forward model, in turn, could lead to significant inaccuracy or bias in designed molecules. Thus, instead of directly targeting λ in the design calculation, we decided to use the relatively reliable models on T_g and T_m as intermediate targets, and the transferred model on λ was used in the post-screening step. Though the connection between λ and these surrogate properties has not yet been fully understood, there is some evidence to support our strategy.

It is widely known that increasing the rigidity of polymer chains can increase the values of T_g and T_m, consequently leading to high values of λ. For example, it has been reported that the maximum value of λ in a glassy phase depends on the level of T_g.[30,31] Theoretically, lattice heat conduction in crystals can be conceived in terms of the kinetics of propagating phonons, where thermal

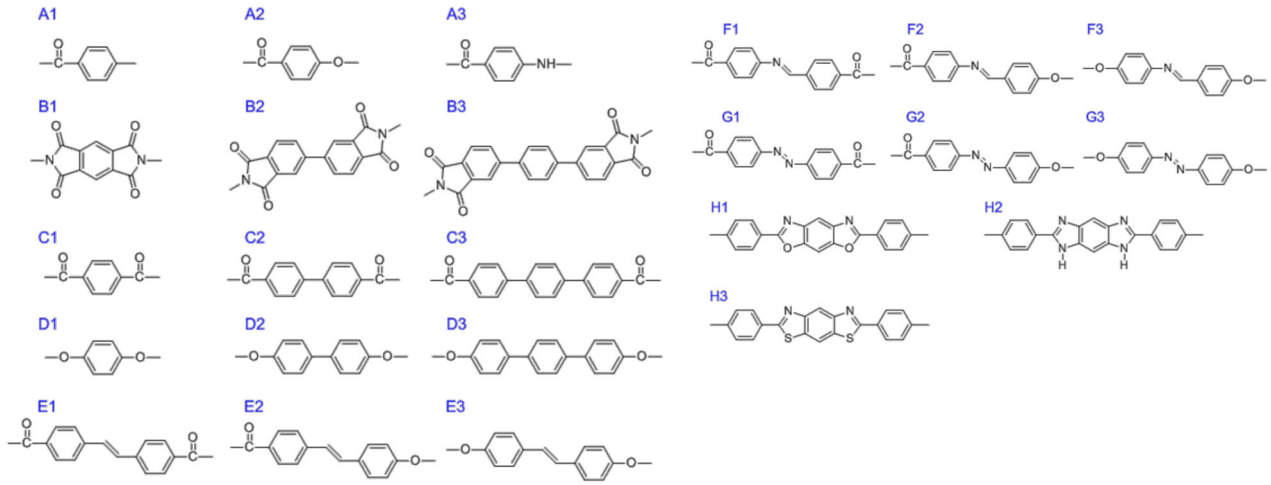

Fig. 5 List of fragments compiled on the liquid-crystalline polymer (LCP)-likeliness filter that were used in the de novo design and post-screening process

conductivity is determined by the heat capacity, group velocity and mean free path (or velocity times lifetime) of phonons. Here, velocity can be related to harmonic interatomic/intermolecular force constants (IFCs) and the lifetime of anharmonic IFCs. Of course, polymers are often disorderly in structure, which reduces the mean free path so that phonons no longer propagate, and their thermal conductivity can be expressed by the heat capacity and mode diffusivity obtainable by harmonic IFCs.[32] However, this does not mean that disorder terminates the propagation of all phonons. Even in amorphous polymers, some phonons can still propagate depending on the frequency. In both scenarios (harmonic and anharmonic), the strength of intermolecular forces affects thermal conductivity. Therefore, we expect to see some correlation, either directly or indirectly, between thermal conductivity and T_g and T_m, which are also strongly affected by the strength of intermolecular forces, as transition fundamentally involves the breaking of bonds or a cooperative mode change, where harmonic and anharmonic forces correspond to small and large intermolecular displacements.

The observed data also showed weak positive correlations between T_g, T_m and λ, as shown in Fig. 2b. Indeed, the success of the model transfer from T_g or T_m to λ constitutes evidence in favour of using T_g and T_m as proxy design targets (Fig. 3c, d and SI). We have chosen a target design range of 200–500 °C and 300–600 °C for T_g and T_m, respectively.

High λ is produced not only by rigid polymer main chains with high T_g or T_m but also by the highly oriented molecular chain that is often observed in ultra-drawn fibres, axially oriented thin films and injection-moulded pieces.[33] In addition, processing ease is indispensable for the practical use of polymeric materials to shape them as films, fibres, moulding and so on. From the perspective of further developments and industrial applications, we targeted liquid-crystalline polymers (LCPs) in both the de novo design calculation and the post-screening process. We chose this particular target because of its practical importance in effective thermal management applications, heat exchangers and energy storage. In general, polymers have quite low thermal conductivity, typically 0.1–0.2 W/mK, because of their semi-crystalline, electrically insulating structures. The side chains or main chains of LCPs make up a family of thermoplastics that exhibit high heat resistance and tolerance, high electrical resistance and high chemical resistance.[34–36] The ordered stacked orientation along one direction of LCPs significantly increases their thermal conductivity in the direction of the molecular orientation. In this study, LCP likeliness was set as a design objective because of the intrinsic processability and rigidity of LCPs to enhance thermal conductivity in further applications. We compiled a list of LCP-like substructures (Fig. 5) based on expert knowledge. During the de novo design calculation, sequentially generated structures were scored higher if they contained one or more fragments in the list so as to create a library of LCP-like structures. Thus, the forward model in Eq. (1) takes the form $p(Y \in U|S) \propto p(Y_{T_g} \in U_{T_g}|S)p(Y_{T_m} \in U_m|S)\theta^{1(Y_f(S)\cap U_f \neq \phi)}$, where $1(\cdot)$ denotes the indicator function, which takes the value one if the argument is true and zero otherwise. In addition to the probabilities that T_g and T_m of S lie in the desired regions, U_{T_g} and U_{T_m}, the additional score $\theta > 1$ is assigned to S if its substructures $Y_f(S)$ coincide with at least one fragment listed in the LCP-likeliness filter U_f. Furthermore, in the post-screening step, we once again screened out LCP-like candidates that contained one or more fragments while assessing the predicted values of λ and SA.

Backward prediction: generation of candidates

The iqspr package consists of two main modules: (1) ML algorithms to train the forward prediction models and the prior distribution and (2) the Monte Carlo generation of de novo molecules from the backward model. The preparation of the forward model has already been described. The prior distribution $p(S)$ takes the form of a probabilistic language model. We then trained the model on the SMILES strings of the 14,423 unique homopolymers recorded in PoLyInfo. The trained prior implicitly encoded frequently appearing atomic configuration and chemical bonding in the existing polymers with the given instances of the SMILES character sequences. Monte Carlo samples drawn from this prior are anticipated to recognize implied contexts in the chemical language such as exclusion rules of invalid chemical bonding, SA and chemical stability.

With the prior and the forward model to form the backward model, the SMC calculation was executed to successively refine SMILES strings of seed molecules such that their resulting properties lay in the desired property region. The iqspr script that we used is provided at the GitHub repository, https://github.com/stewu5/HighTCond_Polymer_iqspr, along with the models trained on T_g and T_m, and the chemical language model. We generated 1000 promising synthetic targets with predicted polymeric properties lying in the prescribed ranges of T_g and T_m. Examples of the generated chemical structures are depicted in Fig. 4a. Supplementary Movie S1 shows the process of transforming chemical structures and refining the target properties.

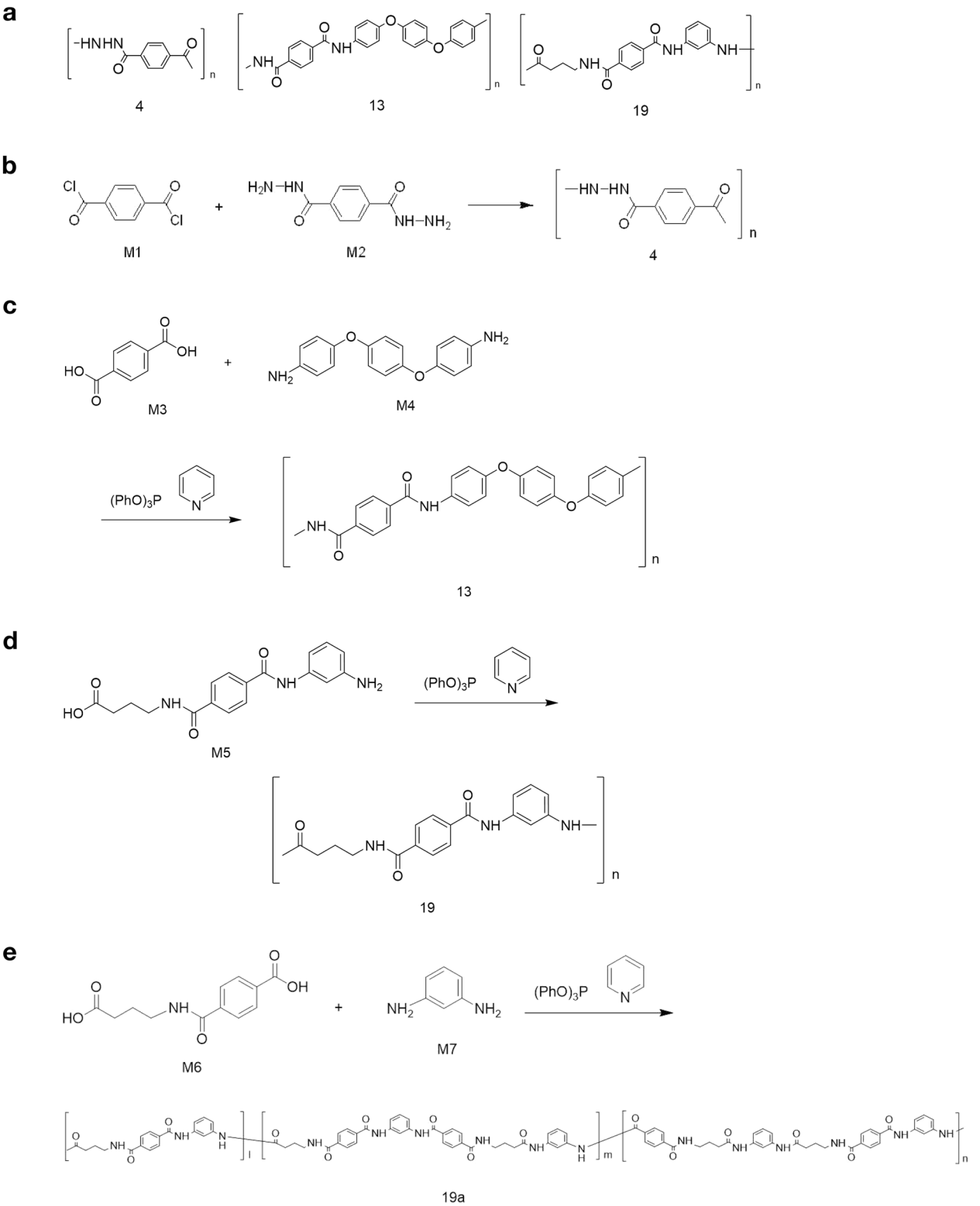

Fig. 6 Details of the three synthesized polymers. **a** Chemical structures of three synthesized polymers and **b**–**e** designed synthetic routes to the targets (see SI for further details)

Selection of synthetic targets

To assist in the selection of synthetic targets, we imposed screening steps on the 1000 designed candidates. First, to identify LCP-like structures, candidates that exhibited one or more components on the list in Fig. 5a were moved forward. Next, we evaluated their synthesizability using Schuffenhauer's SA scores.[37] Finally, considering the ease of processing required in industry, we prioritized candidates with $T_g \leq 300$. As a result, 24 candidates were identified for the further investigation of potential routes of chemical synthesis (Fig. 4a). Eventually, the synthetic routes of

Table 2. Experimental properties of the three newly synthesized polymers compared with predictions from ML models

Polymer	4 (pre)	4 (obs)	4 (anneal)	13 (pre)	13 (obs)	13 (anneal)	19 (pre)	19a (obs)
T_g (°C) (DSC)	286	N/A[a]	–	228	N/A[a]	–	121	194
T_g (°C) (FSC)	286	221	–	228	226	–	121	191
T_m (°C) (FSC)	404	513	–	426	494	–	321	303
λ (W/mK)	0.246	0.261	0.408[b]	0.225	0.224	0.387	0.218	0.195
Xc	–	0.16	–	–	0.30	0.30	–	0.09[c]

Compressed film-shaped samples were used in all cases except the X-ray diffraction of polymer 19a. We report values from prediction (pre), observation (obs) and observation after annealing (anneal)
DSC differential scanning calorimetry, FSC fast scanning calorimetry, T_g glass transition temperature, T_m melting temperature
[a]T_g values, and instead, FSC was introduced to determine T_g and T_m
[b]Thermal conductivity of annealed polymer 4 was obtained using the heat capacity and density measured for non-heat-treated samples
[c]Crystallinity (Xc) of polymer 19a was measured in powder form

three kinds of polyamides could be identified (Fig. 6a) and successfully synthesized (see SI for more details), namely, polyamides 4, 13 and 19, a wholly aromatic polyamide, an aromatic polyhydrazide, and an aliphatic–aromatic polyamide, respectively. Figure 4c shows the predicted values of λ with the SA scores of the three polyamides. In the decision-making process, we placed particular importance on the SA and the ease of processing for the created polymers. As a consequence, the predicted values of λ for the three selected polyamides were not particularly high.

Experimental validation

As shown in Fig. 6b–e, polymers 4 and 13 were prepared by the reaction between dicarboxylic acids (dicarboxyl chloride) and diamines, whereas polymer 19 was prepared starting from a self-condensation AB type monomer. In addition, an analogous polyamide to 19, denoted as 19a, was prepared from asymmetric dicarboxylic acid monomer M6 and *m*-phenylenediamine M7. Polyamide 19a has three different sequences, as shown in Fig. 6e. The monomers for 19 and 19a were newly synthesized, and the preparative procedure is described in SI.

Among the three synthesized polyamides (4, wholly aromatic polyamide; 13, aromatic polyhydrazide; 19 or 19a, aliphatic–aromatic polyamide), 19 is a completely new substance. Chemical analysis was carried out by elemental analysis, nuclear magnetic resonance (^{1}H NMR) and infrared (Fourier-transform infrared) spectroscopy. The thermophysical properties of inherent viscosity, thermal diffusivity, specific heat capacity at constant pressure (C_P), ρ, T_g and T_m were measured using an Ostwald viscometer, the temperature wave method (TWA),[38,39] differential scanning calorimetry (DSC), Archimedes' method and a fast scanning calorimeter (FSC).[40] Thermogravimetric analysis and thermomechanical analysis suggested that the weight loss of polymers 4 and 13 was as low as 5 and 20%, respectively, even at 500 °C, and heat resistance was high. By utilizing the FSC technique, the T_g and T_m of all three polymers were observed; the values were not detectable by conventional DSC except for T_g of polymer 19 at 500 °C or less. We confirmed the crystallinity of the polymers by X-ray diffraction measurements. For thermal conductivity near room temperature, compressed polymers 4 and 13 reached 0.26 and 0.22 W/mK, respectively. Polymer 19 was soluble in organic solvent; thus, film formation is possible. Polymer 19 could be categorized as an amorphous polymer with T_g 194 °C clearly observed by conventional DSC; its thermal conductivity, 0.195 W/mK, is notably high for an amorphous polymer. Polymers 13 and 4 reached thermal conductivities of 0.39 and 0.41 W/mK after annealing at 370 or 420 °C, respectively. These values were comparable to those of state-of-the-art polymers in non-composite thermoplastics. As summarized in Table 2 and Fig. 4c, the experimentally confirmed T_g, T_m and λ were highly consistent with their predicted values for polymers 4, 13 and 19a. A full summary of all the material properties tested is available in SI (Table S2).

DISCUSSION

The high-level agreement between the predicted and experimental thermal conductivities validates the ML protocols as the first stage of molecular design in this study. The absolute prediction errors in 4, 13 and 19a were 0.015, 0.001 and 0.017 W/mK for λ and 65, 2 and 70 °C for T_g, respectively.

In addition, to evaluate the thermophysical properties of the limited amount of synthesized new polymers, recent measurement techniques have been introduced. Thermal diffusivity was measured by the micro-scale temperature wave analysis (TWA) originally developed for the small-scale measurement of polymers (TWA,[33] Fig. S13 in SI). Thermal conductivity was calculated from the measured thermal diffusivity along with the measured density and specific heat capacity (Table S2). The ultra-fast scanning nano-scale calorimetric technique (FSC,[36] Fig. S9 in SI) has been applied for the measurement of T_g and T_m of aromatic polyamides for the first time, as these temperatures have not been observed because of thermal degradation when measured by conventional DSC. By using the scan rate of 30,000 K/s, we could experimentally observe T_g, T_m, and in the case of polymer 13, cold crystallization phenomena.

The thermal conductivity of new and existing polymers is compared in Table 3. The new polymers, three kinds of polyamide containing mesogen groups, as depicted in Fig. 4a, were compared with typical polyimide films utilized in electronic applications. The typical polyimides, such as Kapton and Upilex, in the amorphous state exhibited thermal conductivity values of approximately 0.17–0.22 W/mK, whereas the thermal conductivity of the new polymers was 18–80% higher, in the range of 0.20–0.41 W/mK. The post-screening by LCP filter successfully produced a liquid-crystalline-like polymer with the not-so-high targeted T_g (<300 °C) based on the consideration of other important factors, such as SA and the ease of processing required in industry. A film-shaped polymer was realized for the synthesized polymer 19a, which is soluble in organic solvent.

To conclude, we have demonstrated the discovery of new thermally conductive polymers by the use of a series of ML methods in combination with a comprehensive database of polymer properties, expertise from organic synthesis and advanced measurement technologies for thermal properties. In particular, the experimentally confirmed properties of the computationally designed polymers are highly consistent with the predicted values from ML. We discovered a retrosynthesis route to designed monomers, which have actually been synthesized and polymerized. Some of the resulting polymers exhibited

Table 3. Comparison of the thermal conductivity of new and existing polymers at approximately 300 K, as reported in the literature and as measured by temperature wave analysis in this study

No.	Film grad	Manufacturer	Chemical structures	*d* (μm)	λ (W/mK) in thickness	Ref.
	Kapton	Toray	PMDA/ODA	7.3	0.198	38
		Toray	PMDA/ODA	12.7	0.194	38
		Toray	PMDA/ODA	25	0.194	38
		Toray	PMDA/ODA	50	0.186	38
		Toray	PMDA/ODA	76.4	0.189	38
		Toray	PMDA/ODA	124.6	0.189	38
		Toray	PMDA/ODA	175	0.191	38
	UPILEX-S	Ube	BPDA/*p*-PDA	7.5	0.168	48
		Ube	BPDA/*p*-PDA	12.6	0.211	48
		Ube	BPDA/*p*-PDA	20.5	0.216	48
	UPILEX-R	Ube	BPDA/ODA	7.5	0.183	48
		Ube	BPDA/ODA	12.2	0.186	48
		Ube	BPDA/ODA	20.5	0.194	48
4	Predicted	this study	Fig. 4a	–	0.246	This study
4	Observed	this study	Fig. 4a	97	0.261	This study
4	Annealed	this study	Fig. 4a	–	0.408	This study
13	Predicted	this study	Fig. 4a	–	0.225	This study
13	Observed	this study	Fig. 4a	112	0.224	This study
13	Annealed	this study	Fig. 4a	–	0.387	This study
19	Predicted	this study	Fig. 4a	–	0.218	This study
19	Observed	this study	Fig. 4a	103	0.195	This study

PMDA/ODA pyromellitic dianhydride and 4,4 -oxydianiline, *BPDA/p-PDA* 3,3, 4,4 -biphenyltetracarboxylic dianhydride and *p*-phenylenediamine, *BPDA/ODA* 3,3, 4,4 -biphenyltetracarboxylic dianhydride and 4,4 -oxydianiline, *d* thickness of the plate/film-shaped specimens

crystallinity, glassy states and promising thermal properties. Their potential processability and ability to act as casting films provide the basis for revealing further optimized properties.

To fully enjoy the great potential of ML-driven polymer chemistry, there are still some hurdles to be overcome. A wide variety of databases have been developed in various fields of materials science, which provide the starting point for data-intensive and ML-centric workflows (Materials Project,[41] AtomWorks,[42] OQMD[43] and so on). However, very little such work has been done for polymers; there are no comprehensive databases of polymeric properties other than PoLyInfo and Polymer Genome,[44] at least in the public domain. In addition, where polymers are concerned, high-throughput, automated computations such as molecular dynamics simulations are currently difficult to execute. In this study, the available data on thermal conductivity were too sparse to obtain models generally applicable to a diverse set of input materials. Even for the indeterminate target property T_g, the available data set would be more or less uncertain, as it consists of several thousand polymers spanning only a tiny fraction of the vast polymer landscape. Therefore, our workflow was constructed on the premise that predicted properties have a certain level of discrepancy from reality, and computationally designed candidates were used as a guideline for chemists' decision-making. Furthermore, this study focused only on considerably simplified models that ignored any key covariates other than the chemical structures of repeat units. The inability of the current models to account for observed within-polymer fluctuations in polymeric properties might be largely due to the lack of data on processing parameters, higher-order molecular structures and so on. This lack of data is one of the most fundamental issues in polymer informatics.

Another issue concerns the lack of ML methods to facilitate chemical synthesis. In this study, synthesized polymers were selected by emphasizing synthetic accessibly over the novelty of designed structures and thermal properties. In recent years, several researchers have begun to develop ML methods for chemical synthesis.[45,46] Unfortunately, many chemists are still unconvinced of the utility of such strategies, as well as of de novo design methods, because their practical impacts remain unexplored in real-world applications. In future work, ML methods for design and synthesis should be pipelined and practised.[5] We hope that this proof-of-concept study could contribute to the widespread use of such ML platforms, opening up new opportunities in the next generation of polymer chemistry.

METHODS

Polymer design using iqspr

The 1000 candidates were generated using iqspr 1.0. The script available at https://github.com/stewu5/HighTCond_Polymer_iqspr can be used to reproduce the results of this study. To summarize, first, for the prior $p(S)$, we used the extended *n*-gram of order $n = 10$ as the chemical language model for SMILES strings; this approach was developed in our previous study.[16] The language model was trained on 14,423 homopolymers recorded in PoLyInfo. The forward models consisted of two Bayesian linear models trained on 5917 and 3234 instances of T_g and T_m, respectively. The training was performed with default hyperparameters. The descriptor was calculated by combining seven different fingerprints implemented in iqspr: *standard*, *extended*, *hybridization*, *maccs*, *circular* and *pubchem* (see https://cran.r-project.org/web/packages/rcdk/rcdk.pdf for descriptions of these fingerprints). One hundred structures randomly selected from the 14,423 existing polymers were sequentially modified over 500 iterations; molecules created in the burn-in period (first 100 iterations) were discarded. In the SMC run, annealing was scheduled to lower the

temperature linearly from $T = 30$ to $T = 1$ at every step during the burn-in period and maintain $T = 1$ after the burn-in. As described in the Results section, we applied the LCP-likeliness filter, $\theta^{1(Y_f(S) \cap U_f \neq \phi)}$, in every step of the SMC run. The score was set to $\theta = 10$. Note that iqspr 1.0 does not permit the use of such additional filters. Therefore, we customized the original execution command lines by simple scripting. Finally, we selected 1000 candidates with the highest values of $p(Y_{T_g}|S)p(Y_{T_m}|S)$ among all the generated structures.

Transfer learning

We used the MXNet package[47] to train the pre-trained neural networks models for predicting T_g, T_m, ρ of polymers and C_V of monomers. Then, a pre-trained model was re-trained by fine-tuning it to the limited available data on λ.

We started to build a "shotgun pre-trained model library" for T_g, T_m, ρ and C_V. For each property, we generated and trained 1000 neural networks with randomly constructed different network structures. Each network formed a fully connected pyramid in which the number of hidden layers was randomly chosen from {3, 4}. The size of the input layer consisted of a randomly selected subset of 400–600 of the descriptors composed entirely of all the fingerprints. Then, the number of neurons was randomly reduced by 20–80% in each of the following layers, and the number of neurons in the last hidden layer was bounded by 10–30 (pre-determined randomly). Neurons in all hidden layers were activated by ReLU (Rectified Linear Unit), and a linear activation function was configured on the output layer. The details of the predictive performance of the best transferred model for λ among the 1000 fine-tuning trials are shown in SI.

Monomer and polymer synthesis

Details on the synthesis of monomers and polymers are provided in SI.

Measurement of thermophysical properties

Detailed procedures for the measurement of the polymer properties are provided in SI. In particular, recent measurement techniques were introduced to evaluate the limited number of new polymers with high T_g and T_m. Thermal diffusivity was measured by micro-scale TWA[38] (see Fig. S13 in SI). Ultra-fast scanning nano-scale calorimetry (FSC,[40] see Fig. S9 in SI) was introduced to execute a 30,000 K/s temperature scan to observe T_g, T_m and cold crystallization λ, which is unique among the semi-crystalline polymers.

DATA AVAILABILITY

The digital data in PoLyInfo were manually extracted because acquisition using an application programming interface is not currently supported. The QM9 data are publicly available at http://quantum-machine.org/datasets/. The trained models, constructed using R, and other data are available upon request.

ACKNOWLEDGEMENTS

This work was supported in part by the "Materials Research by Information Integration" Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST) and a Grant-in-Aid for Scientific Research (B) 15H02672 from the Japan Society for the Promotion of Science (JSPS). S.W. gratefully acknowledges financial support from JSPS KAKENHI Grant Number JP18K18017. K.H. gratefully acknowledges financial support from JSPS KAKENHI Grant Number JP17K17762, a Grant-in-Aid for Scientific Research on Innovative Areas (16H06439) and PRESTO (JPMJPR16NA). C.S. gratefully acknowledges financial support from the Ministry of Education and Science of the Russian Federation (Grant 14.Y26.31.0019), and J.M. acknowledges partial financial support by JSPS KAKENHI Grant Number JP16K06768.

AUTHOR CONTRIBUTIONS

S.W., M.-a.K., J.M. and R.Y. planned the study. Computational design calculations were performed by S.W., H.Y., I.K., G.L., K.H., Y.X., J.S. and R.Y. Polymer synthesis and property measurements were performed by Y.K., M.-a.K., B.Y., C.S. and J.M. All authors discussed the results and were involved in the development and writing of the manuscript, as well as taking the accountability for all aspects of the work in this manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-019-0203-2).

Competing interests: The authors declare one potential patent application to be submitted in the near future. Patent applicant (whether author or institution): NIMS. Name of inventor(s): R.Y., S.W., M.K., J.M. and Y.X. Application number: not yet available. Status of application: in preparation. Specific aspect of manuscript covered in patent application: not specified.

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

REFERENCES

1. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 770–778 (IEEE, Las Vegas, NV, USA, 2016).
2. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
3. Brown, N. & Sandholm, T. Superhuman AI for heads-up no-limit poker: libratus beats top professionals. *Science* **359**, 418–424 (2017).
4. Green, M. L. et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.* **4**, 011105 (2017).
5. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
6. Yoshikawa, N. et al. Population-based de novo molecule generation, using grammatical evolution. *Chem. Lett.* **47**, 1431–1434 (2018).
7. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
8. Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform.* **10**, 31 (2018).
9. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. Drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **14**, 3098–3104 (2017).
10. Yang, X., Zhang, J., Yoshizoe, K., Terayama, K. & Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **18**, 972–976 (2017).
11. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
12. Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2017).
13. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
14. Audus, D. J. & de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).
15. Peerless, J. S., Milliken, N. J. B., Oweida, T. J., Manning, M. D. & Yingling, Y. G. Soft matter informatics: current progress and challenges. *Adv. Theory Simul.* **2**, 1800129 (2019).
16. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des.* **31**, 379–391 (2017).
17. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
18. Kim, S. et al. Pubchem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
19. Venkatasubramanian, V., Chan, K. & Caruthers, J. M. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **18**, 833–844 (1994).
20. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
21. Venkatraman, V. & Alsberg, B. Designing high-refractive index polymers using materials informatics. *Polymers* **10**, 103 (2018).
22. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
23. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **3610**, 360–365 (2018).

24. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. Polyinfo: polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*. 22–29 (Tirana, Albania, 2011).
25. Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
26. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
27. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
28. Del Moral, P., Doucet, A. & Jasra, A. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B* **68**, 411–436 (2006).
29. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
30. Morikawa, J., Tan, J. & Hashimoto, T. Study of change in thermal diffusivity of amorphous polymers during glass transition. *Polymers* **36**, 4439–4443 (1995).
31. Morikawa, J. & Hashimoto, T. Study on thermal diffusivity of poly(ethylene terephthalate) and poly(ethylene naphthalate). *Polymers* **38**, 5397–5400 (1997).
32. Allen, P. B. & Feldman, J. L. Thermal conductivity of disordered harmonic solids. *Phys. Rev. B* **480**, 12581–12588 (1993).
33. Shen, S., Henry, A., Tong, J., Zheng, R. & Chen, G. Polyethylene nanofibres with very high thermal conductivities. *Nat. Nanotechnol.* **5**, 251–255 (2010).
34. Sugimoto, A., Yoshioka, Y., Kang, S. & Tokita, M. Thermal diffusivity of side-chain-polymer smectic liquid crystals. *Polymers* **106**, 35–42 (2016).
35. Shin, J. et al. Thermally functional liquid crystal networks by magnetic field driven molecular orientation. *ACS Macro Lett.* **5**, 955–960 (2016).
36. Wang, M. et al. Homeotropically-aligned main-chain and side-on liquid crystalline elastomer films with high anisotropic thermal conductivities. *Chem. Commun.* **52**, 4313–4316 (2016).
37. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
38. Morikawa, J. & Hashimoto, T. Thermal diffusivity of aromatic polyimide thin films by temperature wave analysis. *J. Appl. Phys.* **105**, 113506 (2009).
39. Tawade, B. V., Valsange, N. G. & Wadgaonkar, P. P. Synthesis and characterization of polyhydrazides and poly(1,3,4-oxadiazole)s containing multiple arylene ether linkages and pendent pentadecyl chains. *High. Perform. Polym.* **29**, 836–848 (2017).
40. Gao, Y. L. et al. Calorimetric measurements of undercooling in single micron sized snagcu particles in a wide range of cooling rates. *Thermochim. Acta* **482**, 1–7 (2009).
41. Jain, A. et al. The Materials Project: materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 15010 (2013).
42. Xu, Y., Yamazaki, M. & Villars, P. Inorganic materials database for exploring the nature of material. *Jpn. J. Appl. Phys.* **50**, 11RH02 (2011).
43. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
44. Huan, T. D. et al. A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
45. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
46. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
47. Chen, T. et al. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv* https://arxiv.org/abs/1512.01274 (2015).
48. Choy, C. L., Leung, W. P. & Ng, Y. K. Thermal diffusivity of polymer films by the flash radiometry method. *J. Polym. Sci. Part B* **25**, 1779–1799 (1987).