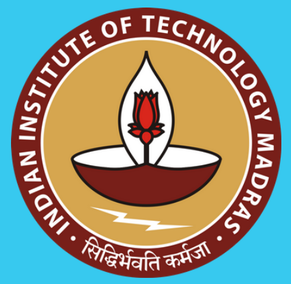
ML assisted discovery of polymers with high thermal conductivity using a molecular design algorithm

-Stephen Wu, Yukiko Kondo et al.



CH5650: Molecular Data Science & Informatics

Raj Jain | CH17B066

AGENDA

1. Introduction
2. The Framework
3. Dataset
4. Bayesian Molecular Design Algorithm
5. Transfer Learning
6. Results
7. Why it works?
8. Synthesis

INTRODUCTION

The authors demonstrate the successful discovery of new polymers with high thermal conductivity, inspired by machine-learning-assisted polymer chemistry.

CHALLENGES

- The objective of this molecular design in this study is to generate promising hypothetical chemical structures that exhibit a set of desired properties.

- Barriers arise mainly from-
  - substantially limited amount of polymeric properties data
  - synthetic difficulty of designed candidates
  - disagreements between expert knowledge and machine-acquired intelligence.

- The experimental data set on thermal conductivity that they 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.

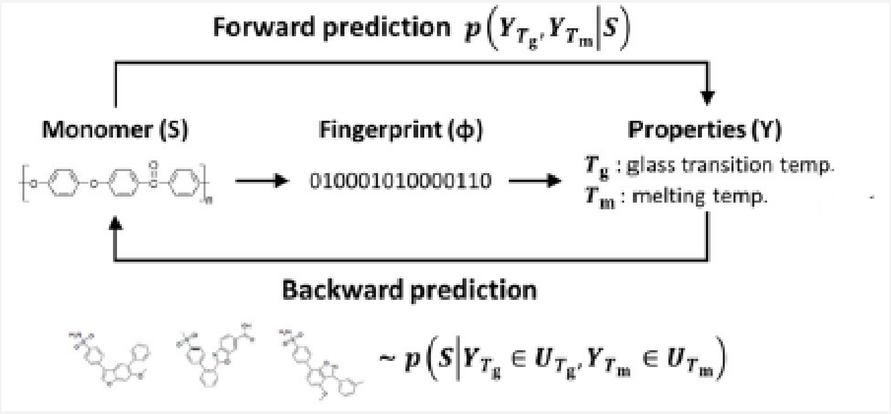
- A solution to mitigate this barrier was to exploit proxy properties related to thermal conductivity as alternative design targets and then use transfer learning to tune the screened model on thermal conductivity targets.

THE FRAMEWORK

Bayesian Molecular Design with Transfer Learning

Bayesian Molecular Design for Alternate Design Targets

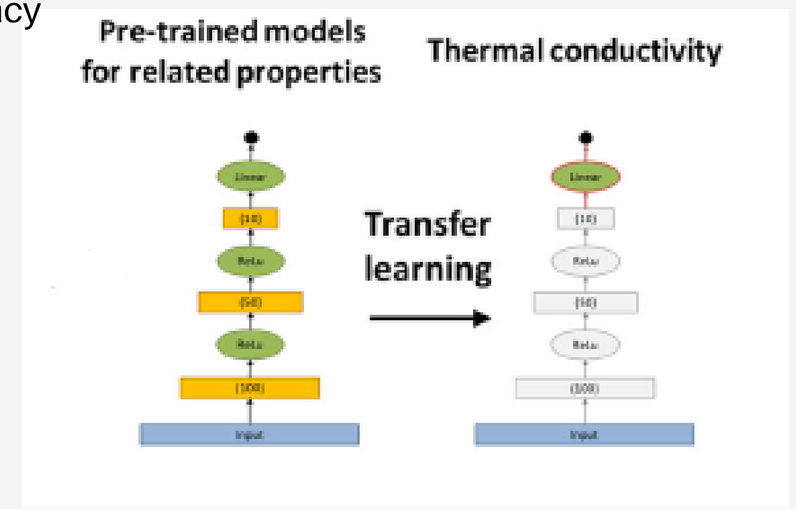
- A solution to mitigate above barriers was to exploit proxy properties related to thermal conductivity as alternative design targets.
- In the Bayesian molecular design process that, they specified a higher region of glass transition temperatures and melting temperatures as alternative design targets, 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.



Screening

Post Screening Models

- 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



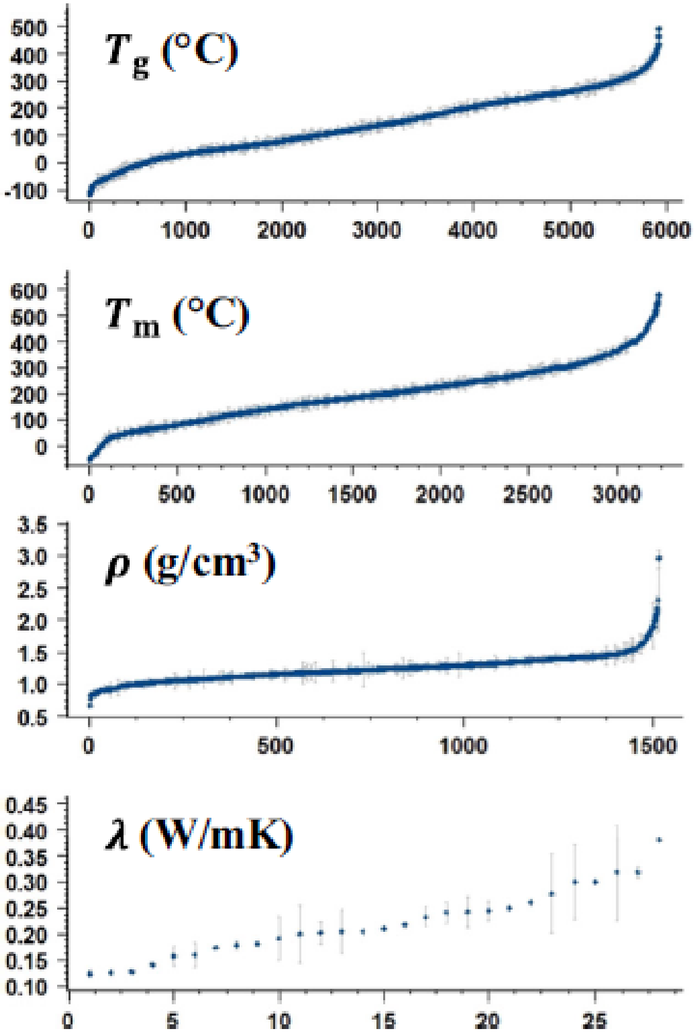
DATASET

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

- **PoLyInfo** 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.
- Extracted a total of 38,310 structure–property relationships with respect to thermal conductivity (λ), glass transition temperature (Tg), melting temperature (Tm) and density (ρ)
- PoLyInfo recorded multiple values of Tg and Tm 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.
- In the construction of pre-trained models for transfer learning, utilized the four data sets from PoLyInfo and the QM9 data set that records the computational data of specific heat capacity at constant volume (Cv)

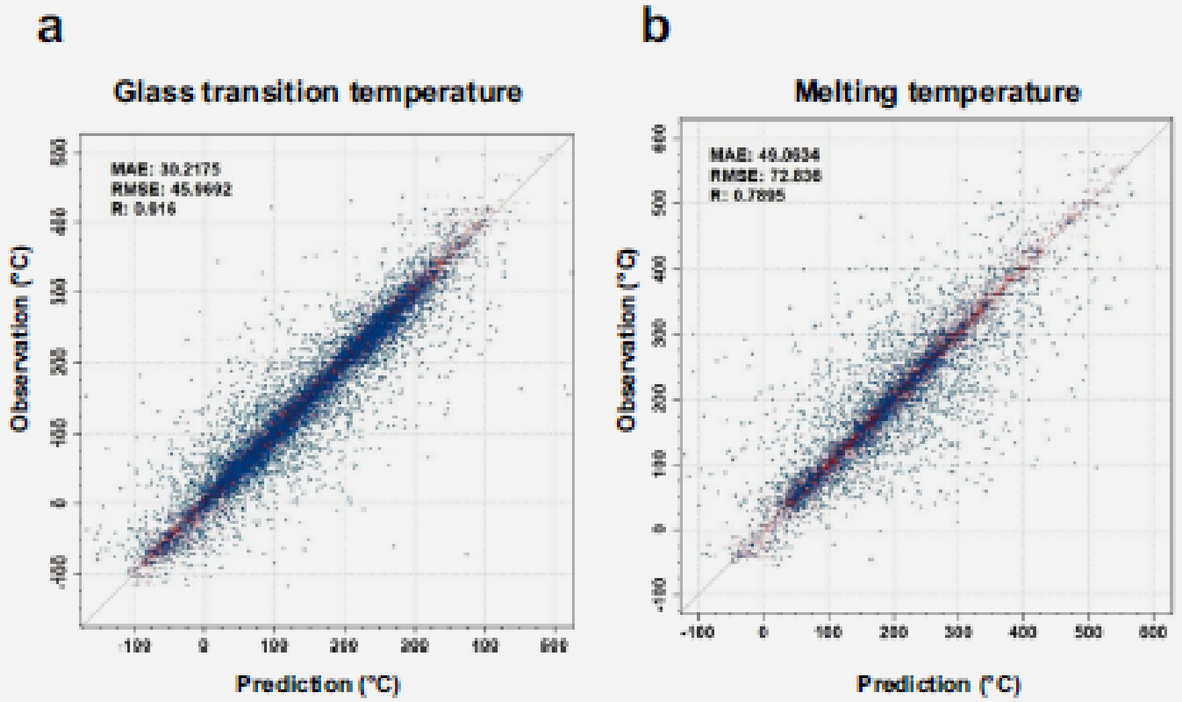


BAYESIAN MOLECULAR DESIGN ALGORITHM

The Bayesian molecular design framework relies on the statement of Bayes' law: $p(S|\mathbf{Y} \in U) \propto p(\mathbf{Y} \in U|S)p(S)$

Forward prediction on Tg and Tm: p(Y∈U|S)

- Forward models on Tg and Tm 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.
- For Tg or Tm, 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.



BAYESIAN MOLECULAR DESIGN ALGORITHM

Prior distribution p(S)

- The molecular structures are encoded as sequence of SMILES symbols (simplified molecular-input line-entry system) [e.g. Phenol: C1=CC=C(C=C1)O]

- The prior is modelled by a probabilistic language model that we call the extended n-gram, which takes the form - $p(S) = p(s_1) \prod_{i=2}^{q} p(s_i|s_{i-1}, \ldots, s_1)$

- The occurrence probability of the ith letter, Si, depends on the preceding S(i−1), …, S1. The conditional probability p(Si|S(i−1), …, S1) 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.

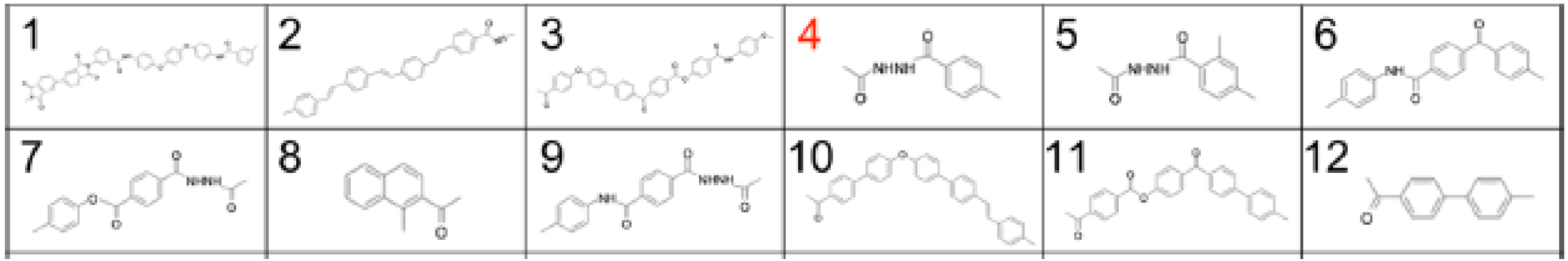
- Unclosed ring and branch indicators must be prohibited. For instance, any strings extended rightward from a given s1:6 = CC(C(C should eventually contain two closing letters, “)”.

- Issue of “long-term dependency” must be addressed: neighbours in a string are not always adjacent in the original molecular graph. e.g., 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 than the side chain

BAYESIAN MOLECULAR DESIGN ALGORITHM

Backward Prediction p(S|Y∈U) & Generation of Candidates

- The preparation of the forward model has already been described. The prior distribution p(S) takes the form of a probabilistic language model.
- 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, and chemical stability.
- Generated 1000 promising synthetic targets with predicted polymeric properties lying in the prescribed ranges of Tg and Tm.



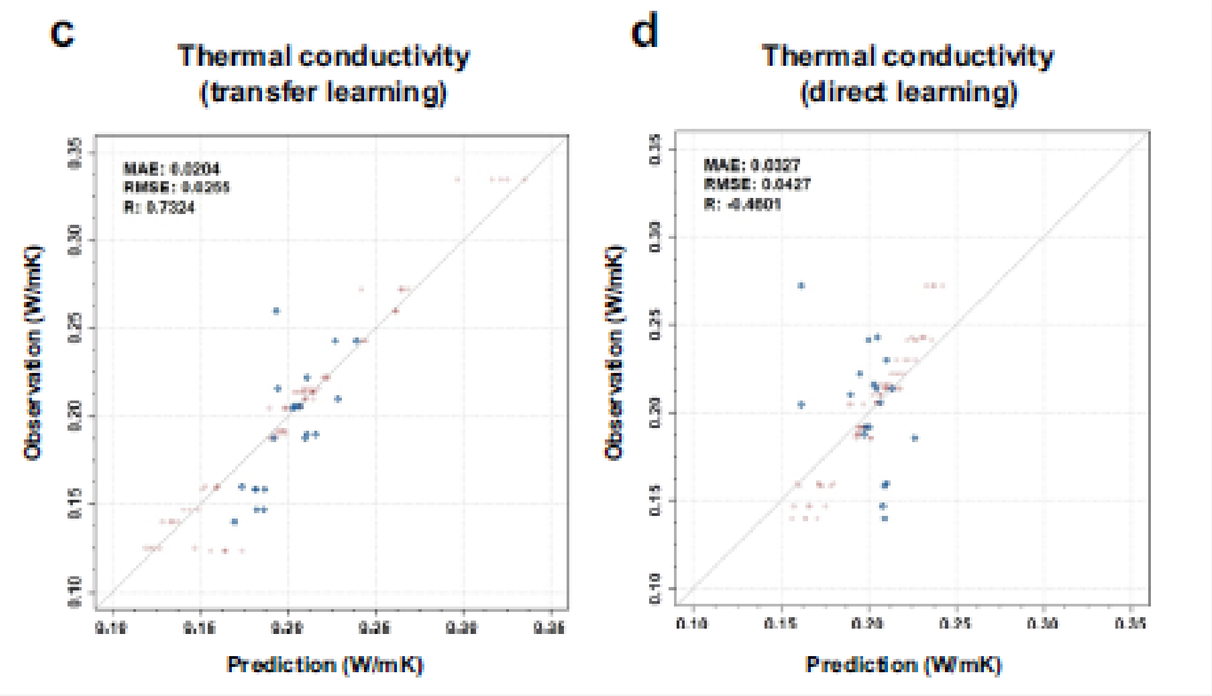
TRANSFER LEARNING

Thermal Conductivity Model

- First, we generated 1000 pre-trained neural networks for Tg, Tm and ρ using the data from PoLyInfo, as well as 1000 models for Cv 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.

SCREENING

- From the perspective of further developments and industrial applications, we target liquid-crystalline polymers (LCPs). 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. The ordered stacked orientation along one direction of LCPs significantly increases their thermal conductivity in the direction of the molecular orientation

- To assist in the selection of synthetic targets, we imposed screening steps on the 1000 designed candidates

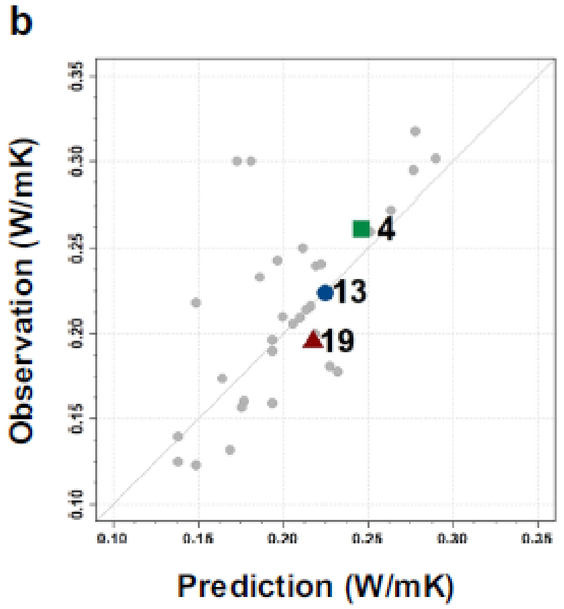
- First, to identify LCP-like structures, candidates that exhibited one or more LCP like components were moved forward.

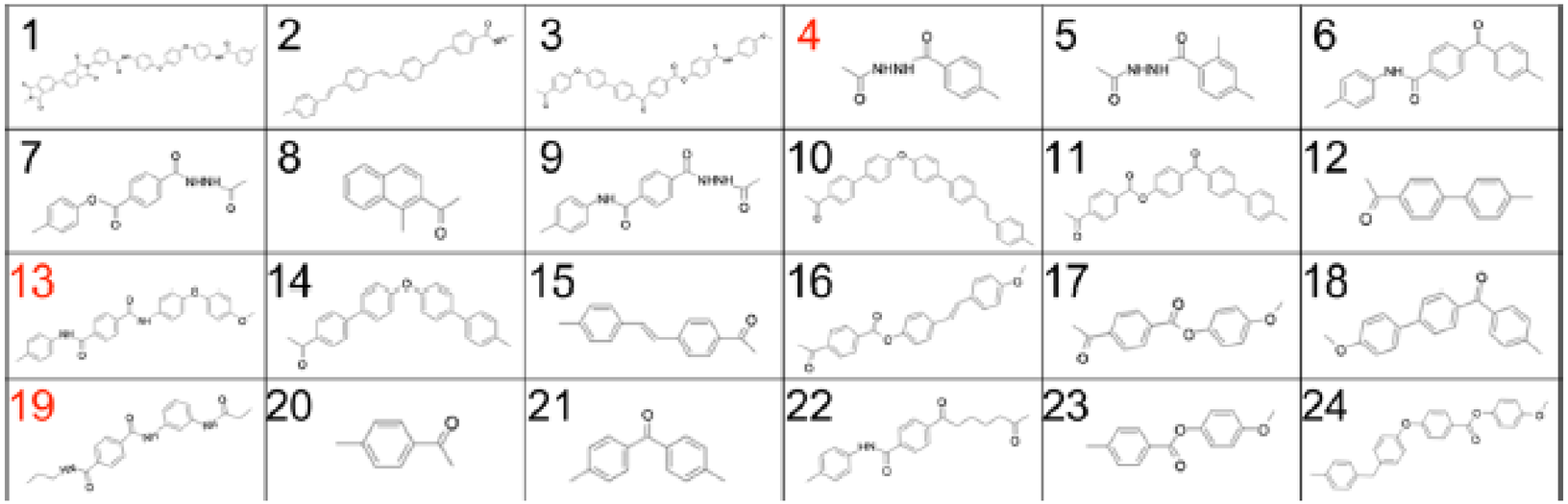
- Evaluated their synthesizability using Schuffenhauer's SA scores.

- 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

RESULTS





MAE was reduced by 40% compared with that of a random forest model trained directly using the 28 data points

24 candidate monomers (finally synthesised polymers are marked in red)

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

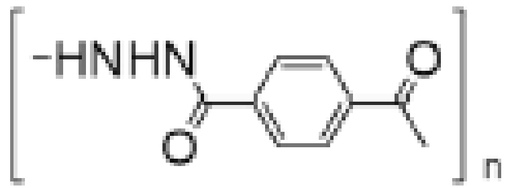
Polymer	4 (pre)	4 (obs)	4 (anneal)	13 (pre)	13 (obs)	13 (anneal)	19 (pre)	19a (obs)
T_g (°C) (DSC)	286	N/A[a]	–	228	N/A[a]	–	121	194
T_g (°C) (FSC)	286	221	–	228	226	–	121	191
T_m (°C) (FSC)	404	513	–	426	494	–	321	303
λ (W/mK)	0.246	0.261	0.408[b]	0.225	0.224	0.387	0.218	0.195
Xc	–	0.16	–	–	0.30	0.30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

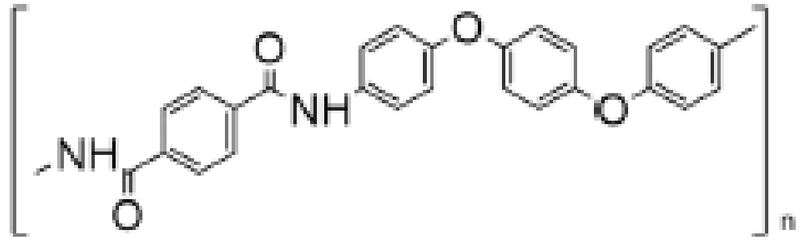
WHY IT WORKS?

- Though the connection between λ and these surrogate properties has not yet been fully understood, there is some evidence to support our strategy.
- The rigidity of polymer chains can increase the values of Tg and Tm, consequently leading to high values of λ.
- The strength of intermolecular forces affects thermal conductivity. Therefore, we expect to see some correlation, either directly or indirectly, between thermal conductivity and Tg and Tm, which are also strongly affected by the strength of intermolecular forces, as transition fundamentally involves the breaking of bonds or a cooperative mode change.
- The observed data also showed weak positive correlations between Tg, Tm and λ

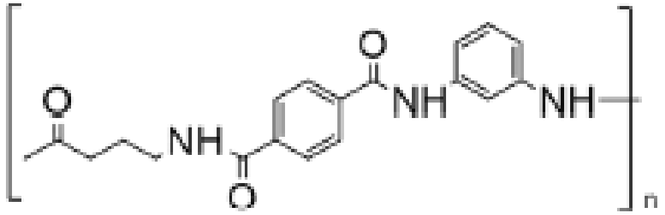
SYNTHESIS



4



13



19

- Polymers 4 and 13 were prepared by the reaction between dicarboxylic acids (dicarboxyl chloride) and diamines.
- Polymer 19 was prepared starting from a self condensation AB type monomer.
- Among the three synthesized polyamides (4, wholly aromatic polyamide; 13, aromatic polyhydrazide; 19 or 19a, aliphatic–aromatic polyamide), 19 is a completely new substance.

THANK YOU!