Prediction of the coefficient of linear thermal expansion for the amorphous homopolymers based on chemical structure using machine learning

Ekaterina Gracheva, Guillaume Lambard, Sadaki Samitsu, Keitaro Sodeyama and Ayako Nakata

ABSTRACT
The coefficient of thermal expansion (CTE) is an industrially crucial macroscopic property of polymers. Yet, there is no structure-based model expressing it with sufficient accuracy. In this work, we present two data-driven predictive models for the linear CTE of amorphous homopolymers in the glassy state based solely on chemical structure, showing consistent predictions. The first model is built with the SMILES-X software and is based on the simplified molecular-input line-entry system (SMILES) of polymer’s repeating unit as input. The second model is built with a random forest trained on extended-connectivity fingerprints of repeating units. Both models are trained on 106 experimental data samples taken from the PolyList database. The out-of-sample prediction shows a root-mean-square error of 2.65 ± 0.09 × 10^{-5} K^{-1} (2.58 ± 0.09 × 10^{-5} K^{-1}), a mean absolute error of 1.71 ± 0.06 × 10^{-5} K^{-1} (1.61 ± 0.06 × 10^{-5} K^{-1}) and a coefficient of determination of 0.62 ± 0.03 (0.64 ± 0.03) for SMILES-X (random forest). Additionally, the models are validated experimentally using a lab-prepared sample with good agreement (p-value >> both models). The attention mechanism, incorporated into SMILES-X, points out salient SMILES substructures, and the resulting maps suggest that the model takes decisions on a chemically interpretable basis.

Abbreviations: SMILES; CTE; CLTE; CVTE

1. Introduction
The coefficient of thermal expansion (CTE) is a property reflecting dimensional stability of a material under changing thermal conditions. Among several formal definitions of the CTE [1], one of the most popular is the linear CTE (CLTE):

\[
\text{CLTE} = \frac{1}{L(T)} \left( \frac{\partial L}{\partial T} \right)_p,
\]

where \(L(T)\) is the length of the material at temperature \(T\), \(\partial L\) is the change in length given \(\partial T\) difference in temperature at constant pressure \(p\). The higher the CTE value is, the more a given material expands with increasing temperature. CTE is an industrially crucial property since mismatches in thermal expansivities between different materials lead to internal stress, and eventually to a failure, of a manufactured part.
Nevertheless, up until now there was no general theoretical or empirical model able to accurately predict CTE for homopolymers based on their chemical composition. In the case of ceramics or metals, which have well-defined rigid atomic structure, it is possible to estimate the CTE using first principles calculations [2,3]. Amorphous polymers, on the other hand, are composed of entangled macromolecules, showing much more complex dynamics, making the investigation via first principles calculations extremely difficult [4]. Recently, it has been possible to evaluate CTE for cross-linked epoxy polymers through accelerated ReaxFF, but the process relies considerably on human experience and yields values largely underestimated CTE values [5].

The main factors influencing the CTE of a polymer are cohesive forces between molecules, topological and geometrical arrangement of atoms, chain stiffness and bond flexibility [6]. However, even though this understanding helps researchers to develop polymers with lower CTE values, such evaluation remains qualitative. Quantitatively, it was only possible, until now, to roughly relate the CTE of amorphous polymers to other measured macroscopic properties, such as the glass transition temperature (T_g) [7] or van der Waals volume (V_w) [1,8]. For some of crystalline polymers, the CTE can be evaluated through the morphology of crystals [9]. In the case of copolymers or composite materials, the CTE is computed as a combination of individual components’ CTE, and is not based on a whole chemical structure [10].

Material properties also largely depend on processing and experimental setup. In an attempt to unify multiple properties and include experimental context, a single graph neural network was presented in [11]. This network allows training on multiple datasets and multiple formats (molecular fingerprints, process parameters, textual information, audio etc.) simultaneously. However, this method fails to fit small dataset (< 100 samples), and its performance on glass expansion has a coefficient of determination of R^2 = 0.36 ± 0.43 when trained on 14 databases.

To the best of our knowledge, the only quantitative structure-based estimation of the CTE for homopolymers has been given by van Krevelen [1] (group composition method). Following his arguments, we deduce a model for the CLTE, and compare it to our results.

In this work, the CLTE is predicted by means of the machine learning tool SMILES-X [12]. Machine learning has an ever-increasing role in modern polymer science. It can be used for individual properties predictions [13–15], or for predicting molecular dynamics trajectories [16]. However, most of the methods make predictions based on fingerprints – a hand-made numerical representation of a molecule, and require big amounts of data for training. On the other hand, SMILES-X builds its predictions based directly on the chemical structure. It implements the latest advances from the field of natural language processing, treating the simplified molecular-input line-entry system (SMILES) [17] input as a character sequence, and translating it into a physicochemical property. Also, SMILES-X has a relatively compact neural architecture at its core and can successfully learn from small datasets. It employs a so-called attention mechanism [18], which is nowadays widely used in neural machine translation. This allows the model to grasp relationships between symbols composing a SMILES (e.g. atoms, bonds, etc.), regardless of their relative distance within the string. In addition, in order to prove the validity of our machine learning model, we test it experimentally with a lab-prepared sample.

The paper is structured as follows. We start with the model-building process description: Section 2.1 presents the machine learning tool SMILES-X. Section 2.2 briefly describes random forest and fingerprints, Section 2.3 summarises the data and Section 2.4 gives details on the model training process. Then, an expression for the CLTE based on the van Krevelen model is given in Section 2.5, with its derivation detailed in Appendix A. We describe the test sample preparation process in Section 2.6. The results are given and discussed in Section 3. The paper is concluded in Section 4.

2. Methods

2.1. SMILES-X

SMILES-X is a machine learning tool devoted to the prediction of physicochemical properties of molecules based solely on their SMILES representation [12]. The SMILES input is broken down into tokens, which represent the individual atoms, bonds, branches and rings (e.g. '[N]', '[Cl]', 'c', ‘ = ', 'S', ‘3’). In the case of polymers, the wildcard asterisk symbol '*'), which is used to indicate points of attachment of repeating units, is treated as an independent token. The tokens are then transformed into float vectors (this procedure is known as embedding) and encoded by a neural network. In order to access the interpretation of the trained model’s results, SMILES-X uses the attention mechanism [18]. An attention is defined as a vector of the same length as SMILES, reflecting the significance of each token within the string towards a prediction. This vector allows the model to perceive SMILES as a whole, and thus to detect distant dependencies within it (e.g. a model can focus on the backbone part of the repeating unit, even if it contains long side branches). This also allows one to visualise which part of a given SMILES has a higher impact on the prediction, and to eventually help in the
interpretation of the model’s outputs. So-called attention maps help to confirm whether the logic discovered by a model corresponds to scientific facts and intuitions. Importantly, they may also give an insight into yet unexplored dependencies. The attention layer is finally transformed into a property of interest by a linear operator (a single fully connected layer).

Since given SMILES can be written in multiple forms when starting from different atoms within a molecule, multiple SMILES strings correspond to the same molecule introducing input invariance. To address this issue, SMILES-X implements data augmentation: given a molecule consisting of \( N \) atoms, one can write \( N_{\text{augm}} \leq N \) SMILES representations, with duplicates removed. This allows the model to deepen its understanding of structure–property relationships, by becoming agnostic to the SMILES multiple arrangements. The details on the SMILES-X software can be found in the corresponding paper [12].

Since a SMILES string is created based on a molecule’s graph, it represents the relative position of atoms within a two-dimensional space. The SMILES notation distinguishes bond types, aromaticity, isotopes and permits specification of stereoisomers. Therefore, a SMILES string is similar as an input to a density-functional theory simulation, and theoretically should be able to grasp the mechanisms responsible for a given property. In contrast, SMILES-X does not use hard-encoded quantum physics rules in order to deduce a molecule’s properties. Instead, these laws emerge from the data. This means that the quality of the prediction strongly depends on the dataset: its quality, diversity and size.

### 2.2. Random forest

Another method that we used to predict the CLTE is random forest [19] with Morgan fingerprints [20]. Random forest is a widely known machine learning algorithm based on the combination of decision trees. Its versatility and robustness with respect to noise make it an excellent predictor for diverse tasks in various fields. From another hand, the nature of decision tree prohibits extrapolation, in other words random forests cannot predict the property values out of the range of the values used for training.

The fingerprints are calculated via the RDKit python library [21] based on SMILES input. Morgan fingerprints encode the information on each atom together with its immediate neighbours located within a fixed number of bonds. The number of bonds is known as radius and is set by the user. In this work, the default value of 2 is used.

### 2.3. Data

Since CLTE is easier to measure, comparing to the coefficient of volumetric thermal expansion (CVTE), it is the most often reported thermal expansion measure. Even though in the case of anisotropic polymers single dimension measurement is not representative of the overall material behaviour, the linear expansion reported in an arbitrary direction is generally considered representing the behaviour of the sample as a bulk.

The data is retrieved from the NIMS polymer database PoLynfo [22] – the largest available polymer database to date. Because of the dominance of the CLTE values in the PoLynfo database (comparing to CVTE), we choose to train a predictive model on it. To keep the data consistent, we perform meticulous data selection. Out of the initial 1580 CLTE entries available within the PoLynfo, we keep only amorphous homopolymers in glassy state. Next, we remove all the samples with fillers and ultrathin films (thickness <1 \( \mu m \)). While the data selection is mainly performed automatically, many of the entries containing no state information were manually verified. We have to confirm, for example, whether a given measurement has been performed below or above the glass transition temperature, as this information is often missing. In the case of multiple measurements reported for the same monomer, the median value is used for training. The overall data selection procedure results in a total of 106 unique repeating units associated with their median CLTE values.

### 2.4. Model training

In order to assess the generalisation ability of models, we implement a 10-fold cross-validation both for SMILES-X and random forest training: the whole dataset is split into 10 groups, and at a single training step one group is set aside for testing the model’s out-of-sample prediction ability. For SMILES-X, the remaining data is split into training and validation subsets: the training set is used for parameters update during the training process, while the validation set ensures that the model achieves the lowest out-of-sample mean squared error. The overall train: validation: test data split is 7:2:1. The augmentation step in SMILES-X is performed after the data splitting, so that there is no information leakage between splits. Since the random forest training process is not based on epochs, there is no need for the validation set. Therefore, the random forest model is trained on 90\% of the data, with 10\% hold for testing (i.e. train:test split is 9:1). In order to evaluate the error related to the model initialisation, training is performed 5 times for each cross-validation group with different random seeds.
For SMILES-X, the neural architectures and related training parameters are defined independently, via a two-step semi-automatic procedure. The architecture is established via a trainless architecture search [23]. The SMILES-X’s architecture is fully defined by three hyperparameters: the size of embedding layer, and the number of units in LSTM and dense layers. Each hyperparameter can take values from [8, 16, 32, 64, 128, 256, 512, 1024], resulting in a total of 512 potential architectures. The estimation is based on the coefficient of variation of the untrained mean squared error CVerr = σerr/μerr (also known as relative standard deviation). To assess the statistics, each architecture is initialised with multiple random seeds. In the first step of estimation, we perform 10 initialisations and select 10 best architectures with the lowest CVerr. Then, pre-selected architectures are initialised 50 times to further improve the precision, and the best architecture is determined. For the CLTE, the best architecture consists of an LSTM layer of 1024 units, a dense layer of 512 units, and has an embedding depth of 8. The geometry selection procedure takes about 40 minutes per fold.

The batch size and learning rate are estimated through Bayesian optimisation, using the GpOpt package [24]. The search region for the batch size is set to [8, 16, 32, 64, 128, 256, 512], and for the learning rate, which is defined as \(r = 10^{-y}, y \in [2, 4]\) with a step of 0.1. Bayesian optimisation evaluates performance based on validation accuracy after 150 epochs of training. First, 35 combinations are randomly sampled and evaluated for the initialization of the Bayesian optimisation process, and then another 35 combinations are selected according to the expected improvement acquisition function [25], and evaluated. Batch size and learning rate combination that have shown the best results over the folds are 16 and 10\(^{-3.8}\), respectively. We apply batch size increments starting from 1/2 of the optimal batch size, increasing twofold every 167 epochs, with a total of 500 epochs [26]. The epoch with the lowest validation error is selected for the final prediction. A single training takes about 20 minutes on a single NVIDIA TITAN V GPU.

For the random forest, hyperparameters are set to the default values.

As discussed above, the split in training and validation is performed at random, and the final model is based on the lowest validation error. In the case of large datasets, where the number of data points is significantly larger than the number of classes, training and validation sets have a high chance of containing similarly distributed data. However, in the case of the CLTE dataset, this balance cannot be guaranteed, and it often happens that poor selection of the validation set results in a trained model with poor prediction performance. To counteract this issue, for some folds training has been performed several times with different random splitting until a good learning curve was achieved (both validation and training loss functions decrease monotonically and are consistent one with another).

### 2.5. Van Krevelen model

To the best of our knowledge, the most relevant model of the thermal expansion for homopolymers has been given by van Krevelen back in 1972 [1]. Based on it, we deduce an expression for the CLTE, in Appendix A, as:

\[
\text{CLTE} \approx \frac{0.15 \times 10^{-3}}{1.3 + 0.55 \times 10^{-2} T_g + 0.45 \times 10^{-3} T_{\text{init}}},
\]

where \(T_g\) is the glass transition temperature, and \(T_{\text{init}}\) is the starting temperature of an experiment.

It is not surprising that the CLTE is related to the glass transition temperature \(T_g\), since they are both defined by the same physical mechanisms. It is important to note that the measured value of the CLTE also depends on the initial temperature \(T_{\text{init}}\) set for the experiment, which also applies to the CVTE.

In order to compare our model’s CLTE predictions against the values computed from Eq. 2, we select from PolyInfo only the entries containing the information on \(T_g\) and \(T_{\text{init}}\). The resulting comparison dataset contains 18 data points.

### 2.6. Test sample preparation

Even though SMILES-X evaluates the out-of-sample performance of the trained model, we further test it with an unseen sample prepared in our laboratory. For this, we have selected nine commercially available polymers with no CLTE entries within the PolyInfo. After processing the polymer powder into a film shape, eight samples out of nine did not meet the criteria for the measurement: five films had high air bubble content, and three of the remaining four films were too brittle to conduct the measurement. Therefore, we test the model with a single sample – poly (vinyl methyl ketone). The raw poly(vinyl methyl ketone) powder is purchased at Sigma-Aldrich (average \(M_w \approx 500.000, T_g \approx 28 ^\circ C\)).

The film is prepared using a hot press following the below procedure. About 1 g of the poly(vinyl methyl ketone) powder is disposed on a metal plate sprayed with silicon, with two metal spacers of 0.5 mm used for the thickness control. The plate is then placed onto the bottom surface of a preheated hot press (80 °C). When the powder shows signs of transition (starts melting), the second metal plate is
placed on top, and a pressure of 8 MPa is applied for 1 min. The plates are cooled down in water ice, and the sample is then extracted from between the plates. The shape of the resulting sample is nearly circular, with a diameter of about 60 mm and a thickness of 5 mm.

The CLTE measurement is performed in an extension mode using a thermomechanical analyzer Rigaku TMA8311 (Japan). A sample is applied at a constant load of 49 mN and extension was recorded while increasing sample temperature at a constant rate of 5 °C/min under nitrogen flow of 50 ml/min. Four rectangular subsamples (20 mm × 4 mm) are cut off from the original sample and measured independently by thermomechanical analysis. The detailed information on the experimental setup and temperature graphs can be found in the Supplementary Materials.

**3. Results and discussion**

**3.1. Prediction**

The out-of-sample predictions of the trained SMILES-X and random forest ensemble models on CLTE, together with their averaged predictions on the experimentally measured poly(vinyl methyl ketone) sample, are demonstrated in Figure 1. The reported prediction means and errors for SMILES-X are computed by averaging both over \( N_{augm} \) augmented SMILES and \( N_{mod} = 5 \) random initialisations as follows:

\[
\mu(\text{CLTE}_{\text{pred}}) = \frac{1}{N_{mod}} \sum_{i=1}^{N_{mod}} \sum_{j=1}^{N_{augm}} \text{CLTE}_{ij},
\]

\[
\sigma(\text{CLTE}_{\text{pred}}) = \sqrt{\frac{1}{N_{mod} N_{augm}} \sum_{i=1}^{N_{mod}} \sum_{j=1}^{N_{augm}} (\text{CLTE}_{ij} - \mu(\text{CLTE}_{\text{pred}}))^2},
\]

where \( \text{CLTE}_{ij} \) corresponds to a single model prediction over a single SMILES.

Since random forest does not implement data augmentation, prediction means and errors are computed based on \( N_{mod} = 5 \) random initialisations only:

\[
\mu(\text{CLTE}_{\text{pred}}) = \frac{1}{N_{mod}} \sum_{i=1}^{N_{mod}} \text{CLTE}_i,
\]

\[
\sigma(\text{CLTE}_{\text{pred}}) = \sqrt{\frac{1}{N_{mod}} \sum_{i=1}^{N_{mod}} (\text{CLTE}_i - \mu(\text{CLTE}_{\text{pred}}))^2},
\]

where \( \text{CLTE}_i \) corresponds to a single model prediction.

The overall out-of-sample prediction shows a root-mean-square error of \( 2.65 \pm 0.09 \times 10^{-5} \text{ K}^{-1} \) (\( 2.58 \pm 0.9 \times 10^{-5} \text{ K}^{-1} \)), a mean absolute error of \( 1.71 \pm 0.06 \times 10^{-5} \text{ K}^{-1} \) (\( 1.61 \pm 0.06 \times 10^{-5} \text{ K}^{-1} \)) and a coefficient of determination of \( 0.62 \pm 0.03 \) (\( 0.64 \pm 0.03 \)) for SMILES-X (random forest).

Tables 1 and Tables 2 summarise overall and fold-wise performances for SMILES-X and random forest, respectively. Both models show a fair amount of the variability in the predictions depending on the fold. This can be explained by the modest size of the dataset:

![Figure 1](attachment:image.png)

*Figure 1.* The out-of-sample predictions (y-axis) for the coefficient of linear thermal expansion, CLTE \( \left( 10^{-5} \text{ K}^{-1} \right) \), by trained SMILES-X [12] (a) and random forest (b) ensemble models against the experimental data (x-axis). Grey points represent the median of the values provided within the Polylinfo [22] dataset. The black point represents the mean of the experimentally measured poly(vinyl methyl ketone) sample prepared by the authors. Error bars on the x-axis span between experimental minimum and maximum values, and error bars on the y-axis represent one standard deviation from the distribution of the predicted CLTE values. Prediction values indicate the mean over 5 trained models, as well as over the number of possible augmentations, individual for every polymer.
Table 1. $R^2$-score, root-mean-square error (RMSE) and mean absolute error (MAE) of the trained SMILES-X ensemble model on the coefficient of linear thermal expansion, CLTE ($10^{-5}$ K$^{-1}$), fold-wise on the training, validation and test sets, and overall on the test set only. The reported mean is calculated over 5 trained models, each trained with a different random number seed, as well as over the number of possible augmentations, individual for every polymer. The error corresponds to a single standard deviation.

| Fold | $R^2$-score | RMSE ($10^{-5}$ K$^{-1}$) | MAE ($10^{-5}$ K$^{-1}$) |
|------|-------------|-----------------|-----------------|
|      | train/val/test | train/val/test | train/val/test |
| 0    | 0.92 ± 0.03/0.95 ± 0.02/0.78 ± 0.07 | 1.27 ± 0.22/0.78 ± 0.15/1.55 ± 0.25 | 0.70 ± 0.07/0.53 ± 0.12/0.99 ± 0.17 |
| 1    | 0.95 ± 0.01/0.79 ± 0.02/0.63 ± 0.13 | 0.94 ± 0.10/1.89 ± 0.08/1.88 ± 0.33 | 0.54 ± 0.05/1.54 ± 0.08/3.16 ± 0.22 |
| 2    | 0.98 ± 0.01/0.69 ± 0.03/0.79 ± 0.05 | 0.69 ± 0.08/2.00 ± 0.09/1.62 ± 0.20 | 0.43 ± 0.05/1.51 ± 0.10/1.24 ± 0.17 |
| 3    | 0.98 ± 0.00/0.70 ± 0.04/0.16 ± 0.08 | 0.70 ± 0.05/1.83 ± 0.11/3.23 ± 0.15 | 0.39 ± 0.03/1.23 ± 0.08/2.16 ± 0.14 |
| 4    | 0.96 ± 0.01/0.74 ± 0.05/0.37 ± 0.14 | 0.79 ± 0.13/2.22 ± 0.20/3.59 ± 0.40 | 0.45 ± 0.05/1.45 ± 0.10/2.54 ± 0.24 |
| 5    | 0.97 ± 0.00/0.90 ± 0.01/–0.30 ± 0.15 | 0.81 ± 0.04/1.16 ± 0.07/4.49 ± 0.26 | 0.49 ± 0.03/0.83 ± 0.07/2.84 ± 0.17 |
| 6    | 0.91 ± 0.02/0.72 ± 0.05/0.81 ± 0.03 | 1.18 ± 0.10/1.99 ± 0.18/3.04 ± 0.24 | 0.73 ± 0.05/1.33 ± 0.23/1.51 ± 0.15 |
| 7    | 0.96 ± 0.01/0.74 ± 0.02/0.80 ± 0.02 | 0.87 ± 0.06/2.33 ± 0.09/1.96 ± 0.11 | 0.50 ± 0.03/1.49 ± 0.11/4.33 ± 0.11 |
| 8    | 0.95 ± 0.01/0.73 ± 0.04/0.53 ± 0.08 | 0.97 ± 0.08/1.98 ± 0.15/1.58 ± 0.14 | 0.56 ± 0.04/1.47 ± 0.10/1.21 ± 0.11 |
| 9    | 0.95 ± 0.01/0.57 ± 0.05/0.91 ± 0.04 | 1.07 ± 0.09/1.47 ± 0.08/1.26 ± 0.26 | 0.63 ± 0.05/1.10 ± 0.09/0.96 ± 0.18 |
| Overall test performance: | 0.62 ± 0.03 | 2.65 ± 0.09 | 1.71 ± 0.06 |

Table 2. $R^2$-score, root-mean-square error (RMSE) and mean absolute error (MAE) of the trained random forest ensemble model on the coefficient of linear thermal expansion, CLTE ($10^{-5}$ K$^{-1}$), fold-wise on the training and test sets, and overall on the test set only. The reported mean is calculated over five trained models, each trained with a different random number seed. The error corresponds to a single standard deviation.

| Fold | $R^2$-score | RMSE ($10^{-5}$ K$^{-1}$) | MAE ($10^{-5}$ K$^{-1}$) |
|------|-------------|-----------------|-----------------|
|      | train/test | train/test | train/test |
| 0    | 0.94 ± 0.01/0.51 ± 0.01 | 1.06 ± 0.13/0.79 ± 0.25 | 0.67 ± 0.05/1.23 ± 0.22 |
| 1    | 0.95 ± 0.01/0.52 ± 0.00 | 1.03 ± 0.11/2.13 ± 0.16 | 0.62 ± 0.05/1.43 ± 0.15 |
| 2    | 0.95 ± 0.01/0.36 ± 0.01 | 0.99 ± 0.12/2.82 ± 0.37 | 0.59 ± 0.05/1.91 ± 0.19 |
| 3    | 0.96 ± 0.01/–0.10 ± 0.01 | 0.86 ± 0.10/3.70 ± 0.21 | 0.55 ± 0.05/2.14 ± 0.15 |
| 4    | 0.95 ± 0.01/0.40 ± 0.01 | 0.95 ± 0.12/3.48 ± 0.30 | 0.60 ± 0.05/2.24 ± 0.20 |
| 5    | 0.95 ± 0.01/0.78 ± 0.01 | 1.01 ± 0.11/1.85 ± 0.41 | 0.67 ± 0.06/1.29 ± 0.26 |
| 6    | 0.94 ± 0.01/0.76 ± 0.01 | 0.97 ± 0.10/3.44 ± 0.15 | 0.59 ± 0.05/2.19 ± 0.13 |
| 7    | 0.94 ± 0.01/0.81 ± 0.00 | 1.04 ± 0.09/1.93 ± 0.21 | 0.64 ± 0.05/1.34 ± 0.18 |
| 8    | 0.96 ± 0.01/0.67 ± 0.00 | 0.93 ± 0.08/1.33 ± 0.09 | 0.60 ± 0.05/0.93 ± 0.08 |
| 9    | 0.94 ± 0.01/0.78 ± 0.01 | 1.03 ± 0.13/0.93 ± 0.63 | 0.68 ± 0.06/1.32 ± 0.32 |
| Overall test performance: | 0.64 ± 0.03 | 2.58 ± 0.09 | 1.61 ± 0.06 |

Table 3. The coefficient of linear thermal expansion, CLTE ($10^{-5}$ K$^{-1}$), experimental results for poly(vinyl methyl ketone). Overall, the mean value with one standard deviation is given.

| Sample # | CLTE ($10^{-5}$ K$^{-1}$) |
|----------|-----------------|
| 1        | 8.16            |
| 2        | 11.43           |
| 3        | 12.16           |
| 4        | 10.32           |
| Overall  | 10.52 ± 1.51    |

Table 4. Description.

|          | CLTE ($10^{-5}$ K$^{-1}$) | Experiment | Prediction | $p$-value |
|----------|-----------------|-------------|------------|----------|
| SMILES-X | 11.16 ± 1.52    | 10.52 ± 1.52 | 0.47      |
| Random forest | 10.43 ± 1.78  | 0.92        |

random splitting into training, validation and test sets does not guarantee that the three sets will have the same proportions of polymer types.

Since the CTE of a polymer varies with the molecular weight $M_W$ [27], the accuracy of our method is limited by the amount of variation of the CTE values between extreme $M_W$ values. Assuming that most of the measurements accumulated within the PolYInfo database are performed in the range of average molecular weights, it is not surprising that the models can make relatively good predictions even without the $M_W$ information.

### 3.2. Experimental validation

The resulting CLTE for each of the measurements, together with the corresponding mean and standard deviation are given in Table 3 and Figure 1 (black dot).

The CLTE prediction, CLTE$\text{pred}$, on the experimentally prepared poly(vinyl methyl ketone) sample shows a very good agreement with the measurement, CLTE$\text{exp}$, for both models (see Table 4).

### 3.3. Interpretation

The attention mechanism implemented in SMILES-X permits to have a visual comprehension of which atom or bond a trained model pays attention to when computing the CLTE (other than single bonds and hydrogen atoms, not represented by the SMILES used in this work). Molecular fingerprint-based random forest model allows a similar kind of visualisation. However, due to the nature of fingerprint computation, it is only possible to evaluate the importance of individual atoms towards the prediction, and not of branches or bonds. We present below, in Figures 2–5, attention maps for some of the studied homopolymers. These maps show out-of-sample attention, i.e. they correspond to the fold where a given polymer...
Figure 2. Attention maps built upon the SMILES-X (red) and random forest (blue) predictive ensemble models for coefficient of linear thermal expansion, CLTE ($10^{-5}$ K$^{-1}$), for homopolymers containing (a) alkyl chains (poly(decyl vinyl ether)); (b) double bonds (poly(isobutyl methacrylate)); (c) triple bonds (poly(methacrylonitrile)). For SMILES-X top and bottom rows show 1D and 2D attention maps, respectively. For random forest fingerprint similarity maps are shown. The darker the shade of the colour is, the stronger is the attention paid by the respective ensemble model. At the bottom of each figure, the experimental results are given, and the predictions are given as mean with one standard deviation.

appears in the test set and is not seen by the model during training. Similar to predictions, they are a result of averaging over $N_{mod} = 5$ random initialisations for random forest, as well as over $N_{augm}$ augmented SMILES for SMILES-X (Equations 4 and 3). Note that for SMILES-X attention maps vary between different SMILES representations for the same molecule (see Figure S1 in the Supplementary materials for an example).

The main known factors that affect the CLTE of a homopolymer are cohesive forces between chains, topological and geometrical arrangement of atoms, chain stiffness and bond flexibility [6]. For example, alkyl chains are flexible and show high values of the CLTE. As demonstrated in Figure 2a, SMILES-X model has successfully learnt this feature from the data, while random forest shows that the most important atoms are sitting on the very end of the chain.

Another example is the inclusion of double or triple bonds in the homopolymer chains. These are known to be rigid structures showing little or no rotation, thus obstructing the movement of polymer chains within the bulk material and therefore lowering the CTE. However, there are only two samples containing triple bonds within our dataset, and, unsurprisingly, the model does not pay much attention to this component. Figure 2(b,Figure 2c) demonstrate performance of both ensemble models on such polymers.

Polymides, due to their inherently low thermal expansion, are of great interest to the industry and attract a lot of attention in research and development, which makes this class to be the most represented within our CLTE dataset. Accordingly, both presented machine learning models successfully associate polymides with low CLTE values, as seen in Figure 3(a,Figure 3b). While SMILES-X rather pays attention to the double bonds and nitrogens characteristic of polyimides, the random forest is focused on other structures instead.

Note also that the SMILES-X attention is rather paid to branches and bonds than to the atoms themselves, as shown in Figure 3c. This reflects the intuition that the shape of a repeating unit is one of the most important features influencing the CLTE. Nevertheless, molecular fingerprints contained hashed information about every atom’s environment including the bond information, so while random forest model is incapable to point out bonds themselves, it may point out the atom in the vicinity of an important bond.
Figure 3. Attention maps built upon SMILES-X (red) and random forest (blue) predictive ensemble models for coefficient of linear thermal expansion, CLTE \((10^{-5} \text{ K}^{-1})\), for homopolymers containing (a) PMDA group (poly\((1,1':4,1''-\text{terphenyl-4,4''-diamine})\)-alt-\((\text{pyromellitic anhydride})\)); (b) BPDA group (poly\((4,4''-\text{oxydianiline})\)-alt-\((\text{biphenyl-3,3',4,4''-tetracarboxylic anhydride})\)); (c) complex shape (poly\((1,2-\text{bis(4-aminophenyl)benzene})\)-alt-\((\text{biphenyl-3,3',4,4''-tetracarboxylic anhydride})\)). For SMILES-X top and bottom rows show 1D and 2D attention maps, respectively. For random forest fingerprint similarity maps are shown. The darker the shade of the colour is, the stronger is the attention paid by the respective ensemble model. At the bottom of each figure, the experimental results are given as median with a range of available values, and the predictions are given as mean with one standard deviation.

Figure 4. Attention maps built upon SMILES-X (red) and random forest (blue) predictive ensemble models for the coefficient of linear thermal expansion, CLTE \((10^{-5} \text{ K}^{-1})\), for homopolymers when changing a single C atom into O and S atoms in (a), (b) and (c), respectively ((poly\((\text{diaminodiphenylmethane})\)-alt-\((\text{pyromellitic anhydride})\)), poly\((\text{diaminodiphenylether})\)-alt-\((\text{pyromellitic anhydride})\)), poly\((\text{diaminodiphenylsulfide})\)-alt-\((\text{pyromellitic anhydride})\)). For SMILES-X top and bottom rows show 1D and 2D attention maps, respectively. For random forest fingerprint similarity maps are shown. The darker the shade of the colour is, the stronger is the attention paid by the respective ensemble model. At the bottom of each figure, the experimental results are given as median with a range of available values, and the predictions are given as mean with one standard deviation.
The prevailing importance of a repeating unit shape can be further confirmed when comparing homopolymers differing by a single atom: replacing a C atom in the poly[(diaminodiphenylmethane)-alt-(pyromellitic anhydride)] (Figure 4a) by an O (Figure 4b), or an S (Figure 4c) atom, shows no significant difference between the experimental values, but also between predictions and attention maps built upon SMILES-X (red) and random forest (blue) predictive ensemble models for the coefficient of linear thermal expansion, CLTE \(10^{-5} \text{ K}^{-1}\), for homopolymers when changing the placement of a branch in (a) and (c), or the alignment of a wildcard with the main chain in (b) (poly[(4-methyl-m-phenylenediamine)-alt-(biphenyl-3,3′,4,4′-tetraarboxylic dianhydride)], poly[(2-methyl-p-phenylenediamine)-alt-(biphenyl-3,3′,4,4′-tetraarboxylic dianhydride)], poly[(2-methyl-m-phenylenediamine)-alt-(biphenyl-3,3′,4,4′-tetraarboxylic dianhydride)]). For SMILES-X top and bottom rows show 1D and 2D attention maps, respectively. For random forest fingerprint similarity maps are shown. The darker the shade of the colour is, the stronger is the attention paid by the respective ensemble model. At the bottom of each figure, the experimental results are given as median with a range of available values, and the predictions are given as mean with one standard deviation.

**Figure 6.** Comparison of the coefficient of linear thermal expansion, CLTE \(10^{-5} \text{ K}^{-1}\), values computed using the formalism of van Krevelen (grey circles) against SMILES-X (black circles) and random forest (open black circles) ensemble model predictions. Error bars on the x-axis span between experimental minimum and maximum values, and error bars on the y-axis represent one standard deviation from the distribution of the predicted CLTE values.
maps. Figure 5(a, Figure 5b, Figure 5c) demonstrate that to some extent the trained model distinguishes between the placements of a branch and the alignments of a wildcard with the main chain.

In this way, attention maps indicate that the final trained ensemble models make predictions on an interpretable basis. The full list of attention maps for each homopolymer within our dataset can be found in Supplementary Materials, both for out-of-sample and in-sample cases.

3.4. Comparison to the group contribution method

Here we compare the prediction results of SMILES-X and random forest ensemble models against the CLTE values computed from Equation 2. Figure 6 demonstrates the prediction vs. observation performance for the three models based on 18 data points for which both $T_g$ and $T_{\text{int}}$ are known. It is obvious that the model based on van Krevelen’s assumptions does not meet the experimental CLTE values, predicted values being roughly constant. On the other hand, the ensemble machine learning models show satisfactory agreement. Thus, the machine learning approach appears to be a relevant alternative to the van Krevelen’s semi-empirical CLTE model.

4. Conclusions

In the present work, we predict the CLTE of amorphous homopolymers in glassy state from a monomer’s structure by using two different machine learning models. The first one is obtained via the machine learning tool SMILES-X and is based on the SMILES input. The second one is a random forest model based on molecular fingerprints. Both models show a very satisfactory agreement both with experimental data and each other. While the random forest shows slightly higher precision comparison to the experimental data, SMILES-X allows for more detail during the visualisation phase. The out-of-sample prediction shows a root-mean-square error of $2.65 \pm 0.09 \times 10^{-5}$ K$^{-1}$ ($2.58 \pm 0.09 \times 10^{-5}$ K$^{-1}$), a mean absolute error of $1.71 \pm 0.06 \times 10^{-5}$ K$^{-1}$ ($1.61 \pm 0.06 \times 10^{-5}$ K$^{-1}$) and a coefficient of determination of $0.62 \pm 0.03$ (0.64 ± 0.03) for SMILES-X (random forest). To exclude any possible overfitting, and qualify the potential of ensemble model at test time, we further test both models against a sample of poly(vinyl methyl ketone) prepared by the authors in laboratory, with a very good agreement: predicted SMILES-X (random forest) CLTE value of $11.16 \pm 1.52 \times 10^{-5}$ K$^{-1}$ ($10.43 \pm 1.78 \times 10^{-5}$ K$^{-1}$) has a $p$-value of 0.47 (0.92) when compared against the experimentally obtained CLTE of $10.52 \pm 1.52 \times 10^{-5}$ K$^{-1}$. Given the modest size of the dataset (106 samples), the precision of both machine learning algorithms is beyond our expectations. Moreover, the attention maps indicate that SMILES-X model develop coherent comprehension of the mechanisms responsible for the CLTE of homopolymers. We consider our data-driven model to be a relevant alternative to the existing semi-empirical CLTE models. Further improvement could be achieved with a larger and more accurate dataset.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partly supported by JSPS KAKENHI Grant Number [18H01143].

ORCID

Ekaterina Gracheva http://orcid.org/0000-0002-9704-5939
Guillaume Lambard http://orcid.org/0000-0003-0275-4079
Sadaki Samitsu http://orcid.org/0000-0002-4139-1656
Keitaro Sodeyama http://orcid.org/0000-0002-9228-0729
Ayako Nakata http://orcid.org/0000-0002-3311-6283

References

[1] Van Krevelen DW, Te Nijenhuis K. Properties of polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions. Amsterdam, Netherlands: Elsevier; 2009.
[2] Moruzzi V, Janak J, Schwarz K. Calculated thermal properties of metals. Phys Rev B. 1988;37(2):790.
[3] Ruffa A. Thermal expansion in insulating materials. J Mater Sci. 1980;15(9):2258–2267.
[4] Shenogina NB, Tsige M, Patnaik SS, et al Molecular modeling approach to prediction of thermo-mechanical behavior of thermoset polymer networks. Macromolecules. 2012;45(12):5307–5315.
[5] Vashisth A, Ashraf C, Bakis CE, et al. Effect of chemical structure on thermo-mechanical properties of epoxy polymers: comparison of accelerated reaxff simulations and experiments. Polymer. 2018 [cited 2021 Aug];158:354–363. Available from: https://www.sciencedirect.com/science/article/pii/S0032386118310279
[6] Bicerano J. Prediction of polymer properties. Boca Raton, Florida: CRC Press; 2002.
[7] Simha R, Wilson PS. Thermal expansion of amorphous polymers at atmospheric pressure. ii. theoretical considerations. Macromolecules. 1973;6(6):908–914.
[8] Askadskii AA. Computational materials science of polymers. Great Abington, Cambridge: Cambridge Int Science Publishing; 2003.
[9] Kardos J, Raison J, Piccarolo S, et al Prediction and measurement of the thermal expansion coefficient of crystalline polymers. Polym Eng Sci. 1979;19 (14):1000–1009.
[10] Schapery RA. Thermal expansion coefficients of composite materials based on energy principles. J Compos Mater. 1968;2(3):380–404.
[11] Hatakeyama-Sato K, Oyazku K. Integrating multiple materials science projects in a single neural network. Commun Mater. 2020;1(1):1–10.
[12] Lambard G, Gracheva E. SMILES-X: autonomous molecular compounds characterization for small datasets without descriptors. Mach Learn. 2020;2(2):025004.
[13] Mannodi-Kanakkithodi A, Pilania G, Huan TD, et al Machine learning strategy for accelerated design of polymer dielectrics. Sci Rep. 2016;6 (1):1–10.
[14] Kim C, Chandrasekaran A, Huan TD, et al Polymer genome: a data-powered polymer informatics platform for property predictions. J Phys Chem C. 2018;122 (31):17575–17585.
[15] Barnett JW, Bilchak CR, Wang Y, et al Designing exceptional gas-separation polymer membranes using machine learning. Sci Adv. 2020;6(20):eaaz4301.
[16] Bhattacharya D, Patra TK. dpyol: deep learning of polymer phases and phase transition. Macromolecules. 2021;54(7):3065–3074.
[17] Weininger D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28 (1):31–36.
[18] Vaswani A, Shazeer N, Parmar N, et al Attention is all you need. arXiv preprint arXiv:170603762. 2017.
[19] Breiman L. Random forests. Mach Learn. 2001 Oct;45 (1):5–32. Available from.
[20] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754.
[21] Landrum G. Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling; 2013.
[22] Otsuka S, Kuwajima I, Junko H, et al Polyinfo: polymer database for polymeric materials design. In: 2011 International Conference on Emerging Intelligent Data and Web Technologies, Tirana, Albania; 2011. p. 22–29.
[23] Gracheva E. Trainless model performance estimation for neural architecture search. arXiv preprint arXiv:210308312. 2021.
[24] The GPyOpt authors. Gpyopt: a Bayesian optimization framework in python. 2016 [cited 2021 Apr]. Available from: http://github.com/SheffieldML/GPyOpt
[25] Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. J Global Optim. 1998;13(4):455–492.
[26] Smith SL, Kindermans PJ, Ying C, et al. Don’t decay the learning rate, increase the batch size. arXiv preprint arXiv:171100489. 2017.
[27] González-Benito J, Castillo E, Cruz-Caldisio J. Determination of the linear coefficient of thermal expansion in polymer films at the nanoscale: influence of the composition of eva copolymers and the molecular weight of pmma. Phys Chem Chem Phys. 2015;17(28):18495–18500.
[28] Simha R, Boyer R. On a general relation involving the glass temperature and coefficients of expansion of polymers. J Chem Phys. 1962;37(5):1003–1007.
[29] Bondi A, Sons JW. Physical properties of molecular crystals, liquids, and glasses. Hoboken, NJ: Wiley. 1968;Wiley series on the science and technology of materials. Available from: https://books.google.co.jp/books?id=oSzCAAAAIAAJ

Appendix A. Van Krevelen model

The van Krevelen’s model is based on the molar thermal expansion E:

\[ E = \left( \frac{\partial V}{\partial T} \right)_T \]  

(A1)

where \( V \) is the molar volume.

Observing the experimental data, van Krevelen recognises a linear trend between van der Waals volumes \( V \), computed by group contribution methods, and the measured values of molar thermal expansion \( E \), with different slope angles for glassy \( (E_g) \) and rubbery (or "liquid", \( E_l \)) states:

\[ E_g \approx 0.45 \times 10^{-3} V_w \]
\[ E_l \approx 1.00 \times 10^{-3} V_w. \]  

(A2)

Since our model is based on linear thermal expansion in glassy state, in order to compare the prediction results, we need to express it in terms of \( E_g \). First, let us define the coefficient of volumetric expansion, \( \text{CVTE} \):

\[ \text{CVTE} = \frac{1}{V_{\text{init}}} \left( \frac{\partial V}{\partial T} \right)_p. \]  

(A3)

where \( V_{\text{init}} \) is the initial volume of the material, \( \partial V \) is the change in volume given \( \partial T \) difference in temperatures at constant pressure \( p \). Then, combining Equation A3 with definitions of \( \beta \) and \( E \), in Equations 1 and A1, respectively:

\[ \text{CLTE} = \frac{1}{3} \text{CVTE} = \frac{1}{3} \frac{E_g}{V_{\text{init}}}. \]  

(A4)

where the latter equation holds for stereoisomers.

The initial volume \( V_{\text{init}} \) depends on the experimental setup, specifically on the starting temperature. In the case where the volume increases linearly with temperature (which is not always a good approximation), \( V_{\text{init}} \) can be expressed as:

\[ V_{\text{init}} = V(T_{\text{init}}) = V_g(0) + E T_{\text{init}}, \]  

(A5)

where \( V_g(0) \) is the molar volume of a polymer at 0 K.

The value of \( V_g(0) \) is estimated by van Krevelen through a model based on the Simha and Boyer’s concept of thermal expansion [28]. There, two important assumptions are made. The first one is that the molar thermal expansions are approximately equal between glassy and crystalline states. In other words,

\[ V_g(T_g) - V_c(T_g) = V_g(0) - V_c(0). \]  

(A6)

The second assumption is that the hypothetical occupied volume of a polymer in crystalline state at 0 K, \( V_c(0) \), is equal to that of an undercooled liquid (which does not seem to be correct in general):

\[ V_c(T_c) - V_c(0) = V_c(0) - V_c(0). \]  

(A6)
Equations A6, A7 allow to estimate \( V_g(0) \) as:

\[
V_g(0) = V_c(0) + (E_i - E_g) T_g. \tag{A8}
\]

Remaining unknown \( V_c(0) \) is closely related to the \( V_W \), and van Krevelen gives the following approximation, referring to Bondi [29]:

\[
V_c(0) \simeq 1.3 V_W. \tag{A9}
\]

Substituting Equations A2, A5, A8 and A9 into Equation A4, the CLTE for polymers in glassy state can be approximated as:

\[
\text{CLTE} \simeq \frac{0.15 \times 10^{-3}}{1.3 + 0.55 \times 10^{-3} T_g + 0.45 \times 10^{-3} T_{\text{init}}}. \tag{A10}
\]

If linear relations with van der Waals volume \( V_W \) in Equations A2 and A9 hold, \( V_W \) in the numerator and denominator cancel out, so that ultimately CLTE does not depend on \( V_W \), and therefore is not directly related to the group contribution method.