*Chinese Journal of*
**POLYMER SCIENCE**

RESEARCH ARTICLE

https://doi.org/10.1007/s10118-024-3237-y
*Chinese J. Polym. Sci.* **2024**, *42*, 2059–2068

# Pre-trained Mol2Vec Embeddings as a Tool for Predicting Polymer Properties

Ivan Zlobin[a,b*], Nikita Toroptsev, Gleb Averochkin[a], and Alexander Pavlov[a]

[a] *Center NTI "Digital Materials Science: New Materials and Substances", Bauman Moscow State Technical University, 2S1 Baumanskaya St., 5/1, 105005 Moscow, Russia*
[b] *A. N. Nesmeyanov Institute of Organoelement Compounds of Russian Academy of Sciences, Vavilova St., 28, Bld. 1, 119334 Moscow, Russia*

Electronic Supplementary Information

**Abstract** Machine learning-assisted prediction of polymer properties prior to synthesis has the potential to significantly accelerate the discovery and development of new polymer materials. To date, several approaches have been implemented to represent the chemical structure in machine learning models, among which Mol2Vec embeddings have attracted considerable attention in the cheminformatics community since their introduction in 2018. However, for small datasets, the use of chemical structure representations typically increases the dimensionality of the input dataset, resulting in a decrease in model performance. Furthermore, the limited diversity of polymer chemical structures hinders the training of reliable embeddings, necessitating complex task-specific architecture implementations. To address these challenges, we examined the efficacy of Mol2Vec pre-trained embeddings in deriving vectorized representations of polymers. This study assesses the impact of incorporating Mol2Vec compound vectors into the input features on the efficacy of a model reliant on the physical properties of 214 polymers. The results will hopefully highlight the potential for improving prediction accuracy in polymer studies by incorporating pre-trained embeddings or promote their utilization when dealing with modestly sized polymer databases.

**Keywords** Properties prediction; High dimensional embeddings; Machine learning; Mol2Vec

**Citation:** Zlobin, I.; Toroptsev, N.; Averochkin, G.; Pavlov, A. Pre-trained Mol2Vec embeddings as a tool for predicting polymer properties. *Chinese J. Polym. Sci.* **2024**, *42*, 2059–2068.

## INTRODUCTION

The selection and development of new polymer materials with outstanding physical parameters remain an extremely difficult task, often associated with large amounts of synthetic and experimental work.[1,2] The required parameters are usually achieved through a large number of trial syntheses and measurements of the physical parameters of the resulting polymers, which leads to high costs and slows down the process of optimizing the synthesis conditions.[3,4] The high complexity of predicting the properties of polymer materials is due to the wide variation of the characteristics of the polymer depending on the synthetic strategy.[5,6] Thus, the chemical structure of the monomer, molecular weight of the polymer, presence and amount of cross-links, addition of copolymers and plasticizers, and synthesis conditions have a direct impact on the physical parameters of the material.[7–9] The structure of the monomer affects the conformational mobility of chains, nature and strength of secondary interactions between chains, regularity of the polymer structure and its crystallinity, and solubility of the polymer.[10] At the same time, in practice, studies of relationships between the structure and properties of polymers are carried out within a narrowly defined classes of materials, since extrapolation of the results to other systems does not provide sufficient accuracy.[11–14] Strategies such as quantum chemical calculations, molecular dynamics simulations, and computational continuum mechanics have made it possible to accurately describe the behavior of amorphous and semicrystalline polymers at different scales, from the atomic to the macroscopic scale.[15–17] At the same time, the main challenge in this area remains the integration of all these models for a comprehensive and consistent assessment of the material at all levels of the structural organization.[18,19] Despite the high accuracy and continuous development of computer methods for calculating the properties of polymers, the computational complexity of such methods remains too high to search through a sufficiently large number of structures.[20,21]

Machine learning provides the capability to leverage multiscale polymer modeling, and has become an increasingly popular strategy in materials science.[22,23] Machine learning is used in the design of new materials for membranes[24] and bioplastics,[25] optimization of 3D printing processes,[26] and much more. [27] Quantitative structure-property relationship (QSPR) models are used to predict the thermodynamic pro-

* Corresponding author, E-mail: zlobin.is@phystech.edu
Received June 27, 2024; Accepted September 11, 2024; Published online November 11, 2024

perties of materials with high accuracy and to identify the most important structural features affecting the desired property.[28,29] Using QSPR models, the glass transition temperature of polymers,[30,31] their refractive index[32] and many other indicators have been predicted. In addition to monomer structure descriptors, modern machine learning models use molecular modeling results to parameterize datasets across the atomic, micro- and macro-scales.[33–35] This strategy allows for the accurate approximation of polymer parameters and in-depth analysis of the driving forces of the material assembly. To use a chemical structure in a machine learning algorithm, the structure must be encoded numerically. Typically, chemical structures are stored in databases as strings of text (SMILES[36,37], InChI,[38] SMARTS[39]) or files containing atomic coordinates (mol, sdf, xyz, *etc.*). In this work, we used a simplified molecular-input line-entry system (SMILES), which is a text-line notation of the topology and stereochemistry of the molecule structure. Chemical structures are used to generate molecular fingerprints (such as Morgan fingerprints[40]), molecular descriptors (using RDKit[41] or computational algorithms), and graph representations combined with graph convolutional networks (GCN). All of these methods are widely used in machine learning models for polymer informatics.[42–49] However, the use of Morgan fingerprints does not provide a sufficient description of polymer structure, and GCNs rely significantly on a specific neural network architecture and typically require a high amount of input data for correct training, making them difficult to use for small datasets.

An advanced technique for encoding molecular structures is Mol2Vec[50] which is an unsupervised method. It was initially required to train it on unlabeled data from large chemical datasets (ZINC[51] and ChEMBL[52]) to obtain substructure embeddings. Subsequently, the compounds are encoded as high-dimensional vectors by summing up the pre-trained embeddings of individual substructures, giving information-rich representations of molecules. Although it is mainly designed for drug design and molecular QSPR tasks, Mol2Vec has recently been used in several machine learning models for polymer informatics.[45,53,54] However, this method is rarely used to transfer structural information into models with the physical properties of polymers as inputs. Despite the significant increase in feature dimensionality, this approach may allow better accounting of the chemical diversity of the structures to increase the quality of prediction models. In this study, we aimed to investigate the effect of adding Mol2Vec pre-trained compound vectors on the predictive ability of regression models for different calculated physical properties on a small dataset of approximately 200 samples.

## DATA AND METHODS

### Dataset Preparation

In this study, we selected a relatively small Chemical Retrieval on the Web (CROW) polymer properties database,[55] containing calculated thermophysical properties such as the molecular weight of a repeating unit ($g \cdot mol^{-1}$), molar volume ($mL \cdot mol^{-1}$), density ($g \cdot mL^{-3}$), Hildebrand solubility parameter ($MPa^{1/2}$), molar cohesive energy ($kJ \cdot mol^{-1}$), glass transition temperature (K), molar heat capacity ($J \cdot mol^{-1} \cdot K^{-1}$), refractive index, and van der Waals volume ($mL \cdot mol^{-1}$). As the original web page may no longer be available, the original data are presented in the repository linked to this work (https://github.com/iszlobin/Polymer-DataBase/). This dataset has already been used for several polymer informatics tasks, providing easily accessible and characteristic information for different polymer classes.[48,56,57] Because the properties under consideration also significantly depend on the macromolecular structural parameters of polymers (such as molecular weight and degree of polymerization), which are not represented in the database, our work uses values marked as "preferred" in the dataset.

While larger databases, such as PolyInfo[58] may record more polymer properties and have actual experimental data, they do not permit the same level of comprehensive data collection and distribution.

By selecting only data points containing SMILES representations, we obtained a dataset of 218 polymers represented by SMILES strings. The list of polymer classes in the database is presented in Table 1 (the average of eight polymer chemical structures per class). The number of samples for most classes, however, is too small to be considered an additional categorical parameter for prediction models or to provide any significant information about accuracy across classes. The characteristics of the initial dataset are listed in Table S1 (in the electronic supplementary information, ESI). The physical data present in the chosen dataset, except for the refractive

**Table 1** The occurrence of polymers classes in the CROW polymer properties database.

| Polymers class | Number of samples | Polymers class | Number of samples |
|---|---|---|---|
| Polyacrylamides | 6 | Polyketones | 2 |
| Polyacrylates | 22 | Polyethersulfones | 4 |
| Polynitriles | 2 | Polyhalodienes | 2 |
| Polyesters | 25 | Polyhalostyrenes | 6 |
| Polyolefins | 16 | Polyurethanes | 4 |
| Polyamides | 22 | Polymethacrylamides | 5 |
| Polyanhydrides | 5 | Polymethacrylates | 23 |
| Polydienes | 6 | Polyethaconates | 4 |
| Polyfumarates | 4 | Polystyrenes | 7 |
| Celluloses | 6 | Polyphenylenes | 4 |
| Polyhaloolefins | 7 | Polyphenyl ethers | 4 |
| Polycyanoacrylates | 5 | Polythioethers | 3 |
| Polyglycols | 15 | Poly(vinyl sulfides) | 5 |
| Polyetherketones | 4 | | |

index and glass transition temperature, demonstrate a distribution close to the log-normal (Fig. S1 in ESI). The log-normal distribution of the properties signifies heterogeneity of the observed data, and it could also be a sign that the dataset sufficiently generalizes the statistical population.

Because 11 samples from the dataset have missing values (with only one property missing for each sample), we utilized the HyperImpute[59] iterative imputation procedure to fill in these gaps. The samples with imputed values were used only in the training procedure.

To detect and remove outliers, the Isolation Forest algorithm was applied,[60,61] leaving 207 samples in the final dataset for further statistical analysis and prediction model construction. The characteristics of the final dataset are presented in Table S2 (in ESI).

## Representation of Polymer Structures

Pre-trained embeddings of Mol2Vec substructures were used to obtain digital representations of the chemical structures of the polymers. To avoid increasing the processing time and computational complexity, we used the DeepChem implementation[62] and weights for this algorithm, based on the training methodology described in the original study[50] on the Zinc v15[51] and ChEMBL v23[52] corpora. As a result, a featurizer was prepared to convert the chemical structure from SMILES notation to a compound vector of size 300, obtained as the sum of the Mol2Vec substructure vector embeddings. The typical polymer SMILES includes asterisks showing polymerization sites, which distinguish it from the SMILES of the molecular compounds present in the training corpus. The asterisks were ignored when preparing the polymer vector representations because the polymers in the dataset had significantly varying chemical structures, rendering additional differentiation based on the polymerization site position unnecessary.

## Workflow

A brief overview of the workflow is shown in Fig. 1. The first step represents featurizer pre-training to obtain the compound vectors discussed in the section above. Step 2 involves the training of predictive models on the data from the selected dataset of polymer properties. For each experiment, the physical properties were sequentially selected as the prediction targets. A regression model is built using different sets of input features. The features of the regression model are compound vectors repre-

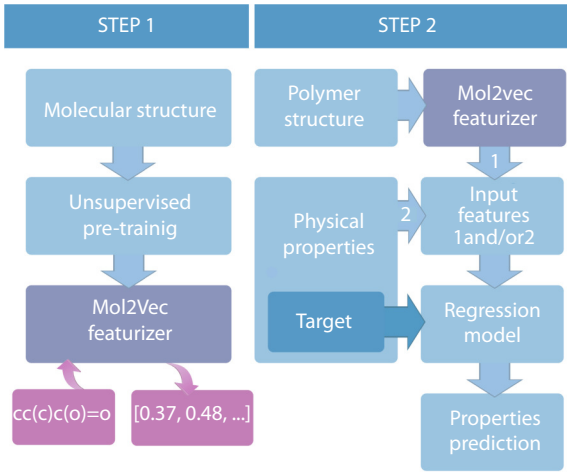


**Fig. 1** The workflow for predicting the physical properties of polymers using different input features. Step 1 represents featurizer pre-training, step 2 refers to the construction and evaluation of prediction models.

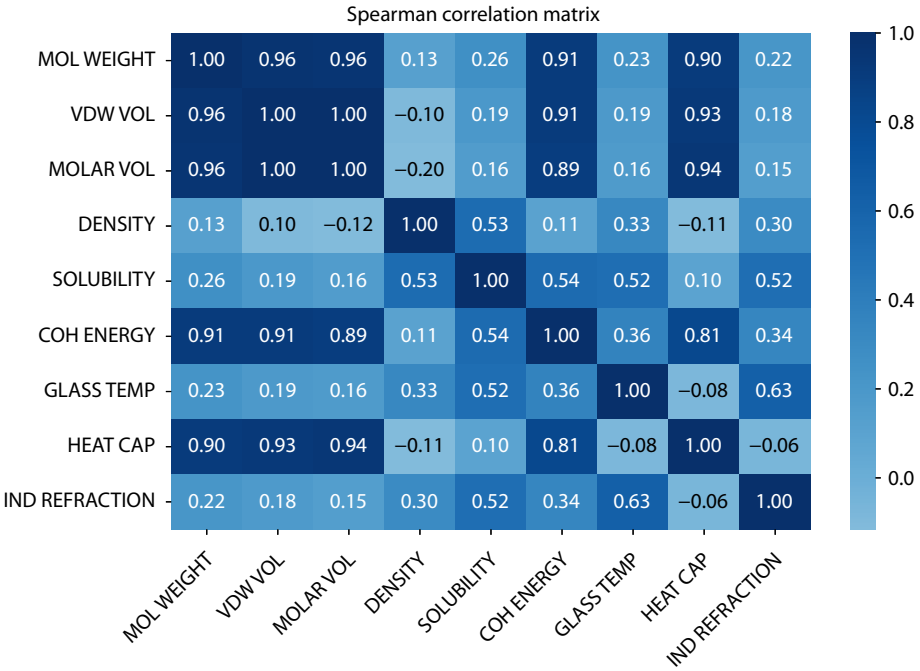


**Fig. 2** Spearman correlation coefficient matrix for physical data. For clarity, property labels have been abbreviated (MOL WEIGHT: molecular weight of a repeating unit, VDW VOL: van der Waals volume, MOLAR VOL: molar volume, SOLUBILITY: Hildebrand solubility parameter, COH ENERGY: molar cohesive energy, GLASS TEMP: glass transition temperature, HEAT CAP: molar heat capacity, IND REFRACTION: index of refraction).

https://doi.org/10.1007/s10118-024-3237-y

senting the chemical structures of the polymers (1), physical properties excluding the target property (2), and both 1 and 2 simultaneously. The impact of selecting different input sets was assessed using regression metrics on the test data, as described below.

## Construction of Regression Models

The dataset was divided into training and test samples with stratification by polymer classes in the ratios of 60% (124 samples) and 40% (83 samples), respectively, prior to all subsequent experiments for a correct evaluation of the performance of different models. Thus, every property prediction was conducted for the same 83 polymers present in the test set. Various machine learning algorithms, including the support vector machine[63] (SVM), random forest[64] (RF), partial least squares[65] (PLS), elastic net[66] (EN), and XGBoost[67] were tested to determine the most accurate model for further analysis. The performance of these models was tested in the subsequent prediction of all physical properties using compound vectors as input features. The hyperparameters for these models were optimized for each target property using grid search optimization with successive halving[68] to achieve the maximum coefficient of determination with 3-fold cross-validation of the training data. The test set was used for the calculation of performance metrics only.

XGBoost showed the best performance (see section **RESULTS AND DISCUSSION**); therefore, it was chosen as a baseline model for the experiments determining the impact of the chosen input features. These experiments were performed for each physical property present in the dataset selected as the prediction target. The input features consisted of compound vectors obtained after Mol2Vec featurization (1), all other physical properties (2), or both simultaneously.

The mean absolute percentage error (MAPE), root mean squared error (RMSE), and coefficient of determination ($R^2$), which are common metrics for similar problems, were used to

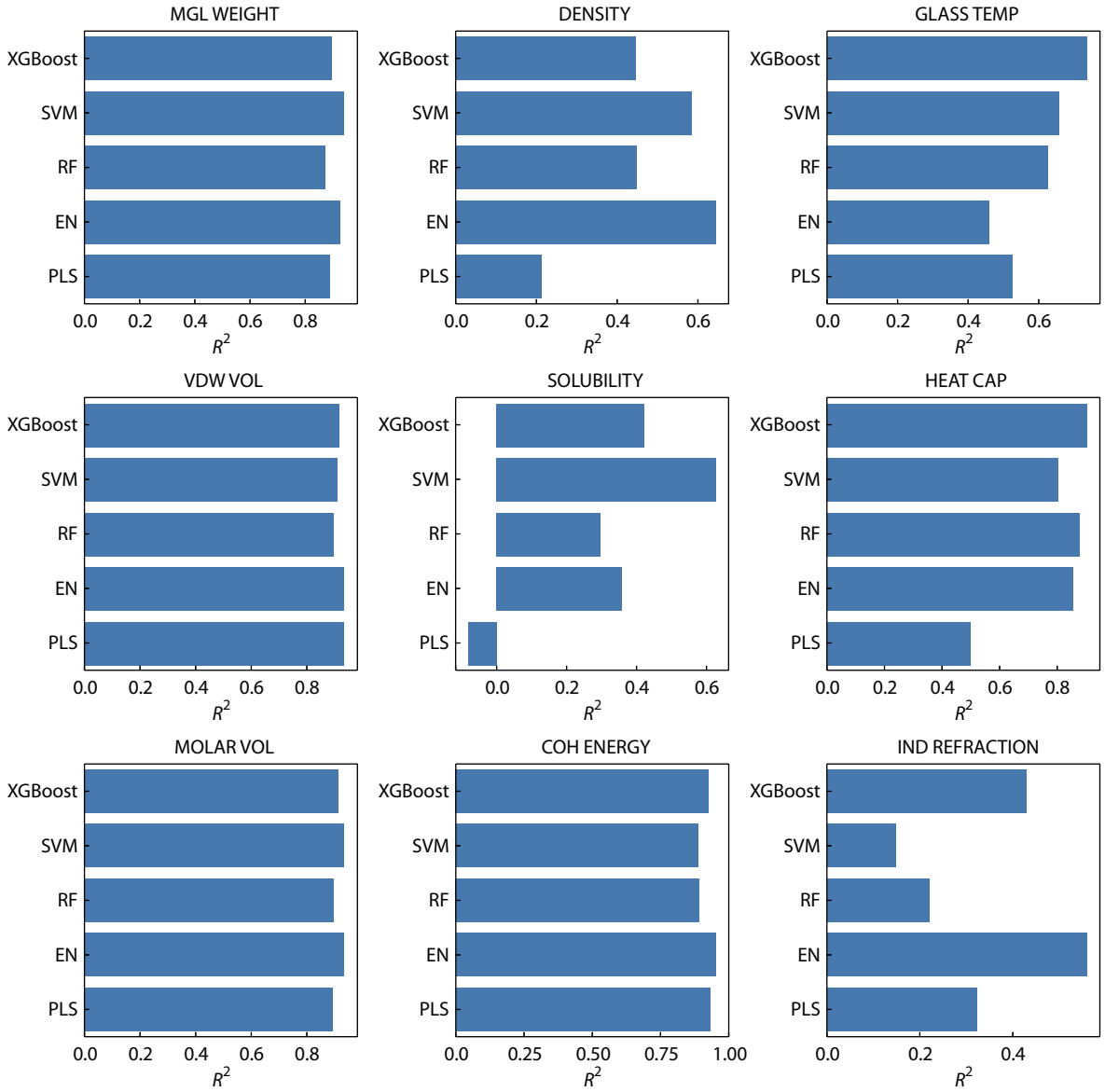


**Fig. 3** Determination coeffitients obtained for vector-based prediction of target properties using different regression algorithms.

https://doi.org/10.1007/s10118-024-3237-y

compare the quality of the models. The latter was adopted as the main model and used to visually compare the results of different models. These are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3)$$

where $y_i$ and $\hat{y}_i$ are the experimental and predicted values associated with the $i^{th}$ sample, respectively, $\bar{y}$ is the average value of all samples, and $N$ is the total number of samples.

## RESULTS AND DISCUSSION

The constructed matrix of Spearman's rank correlation coefficients between the properties (Fig. 2) revealed high correlations for the molecular weight of a repeating unit, van der Waals volume, molar volume, molar cohesive energy, and molar heat capacity. These correlations are expected because the first three parameters are directly calculated from the monomer structure, and the last two are functions of internal energy related to molar volume and thus to the other two correlated properties.[69] Oddly, the Hildebrand solubility parameter does not reveal high correlations with the molar cohesive energy and molar volume, although it is defined as the square root of the cohesive energy density.[70,71] Moreover, the quantile-quantile plot of the Hildebrand solubility parameter versus molar cohesive energy (Fig. S2 in ESI) indicated significant distribution differences between these parameters. This finding suggests the presence of substantial noise in the calculated data of this parameter within the original dataset, or a potentially significant latent correlation with other polymer parameters such as the degree of polymerization, polydispersity, and nature of end groups, as reported in comparable research investigations.[56] In addition, it is expected that the greatest difficulty in constructing a prediction model would be posed by the macroscopic target properties of poly-

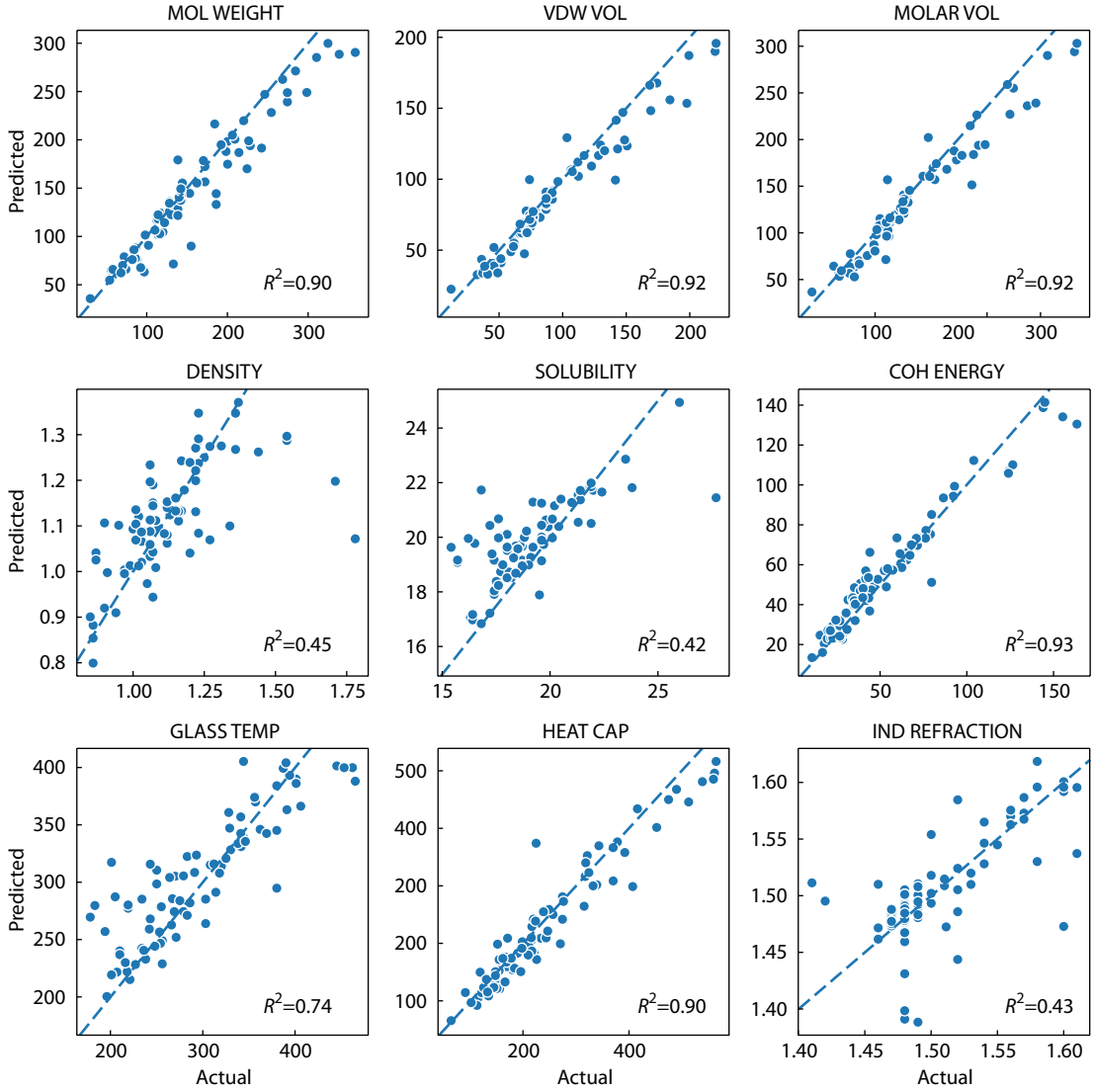


**Fig. 4** Predicted versus actual plot for XGBoost regression of different targets, based on compound vectors as input features.

mers, such as density, glass transition temperature, and refractive index.

A quick worst-case scenario study was conducted to select the baseline model for further experiments. Using several common machine learning algorithms (XGBoost, SVM, RF, EN, and PLS), the regression model was trained for each physical property as a target, using only compound vectors as input features (Fig. 3). Detailed information on the experiments is provided in Table S1 (in ESI).

EN models have shown the best performance in the case of targets with very high dependence on the polymer chemical structure, such as the molecular weight of a repeating unit, van der Waals volume, molar volume, cohesive energy, and, quite surprisingly, index of refraction. During the hyperparameter optimization, the L1 regularization ratio was set to zero, which means that all vector coordinates were present in the final statistical model. This may reflect a linear relationship between these physical parameters and the compound vector coordinates, indicating that the pre-trained featurizer performs as intended, presenting an accurate representation of the polymer structural information. Because EN is a powerful and resource-efficient regularized linear regression method, it can be used as a baseline in tasks for target properties similar to those described above. SVM demonstrated similar performance while also providing the most accurate predictions for the Hildebrand solubility parameter. These algorithms are promising for building similar regression models using vector representations of polymer structures.

However, the XGBoost models showed the most consistent and robust performance, with the highest $R^2$ for most of the targets (including those implicitly linked to the structure). Because our goal was to establish the effect of input data on model accuracy rather than to select the most accurate predictors for each case, XGBoost was chosen as the main algorithm for clarity in all subsequent comparisons. After training the regression tasks on compound vectors as input features (Fig. 4), we repeated the entire procedure using only the physical data (Fig. S3 in ESI), and combining both feature types (Fig. S4 in ESI). These three types of models will be further referred as "vectors only," "vectors added" and "physical data." By making such a comparison, we aimed to establish how the predictive power of such models is comparable and, in particular, to present a fast and computationally simple way to build regression models for calculating polymer properties (including those not directly related to the monomer structure).

Gain-based importances for "physical data" XGBoost models were calculated, as our goal is to interpret the relative importance of each feature. Gain implies the relative contribution of the corresponding feature to the model calculated by

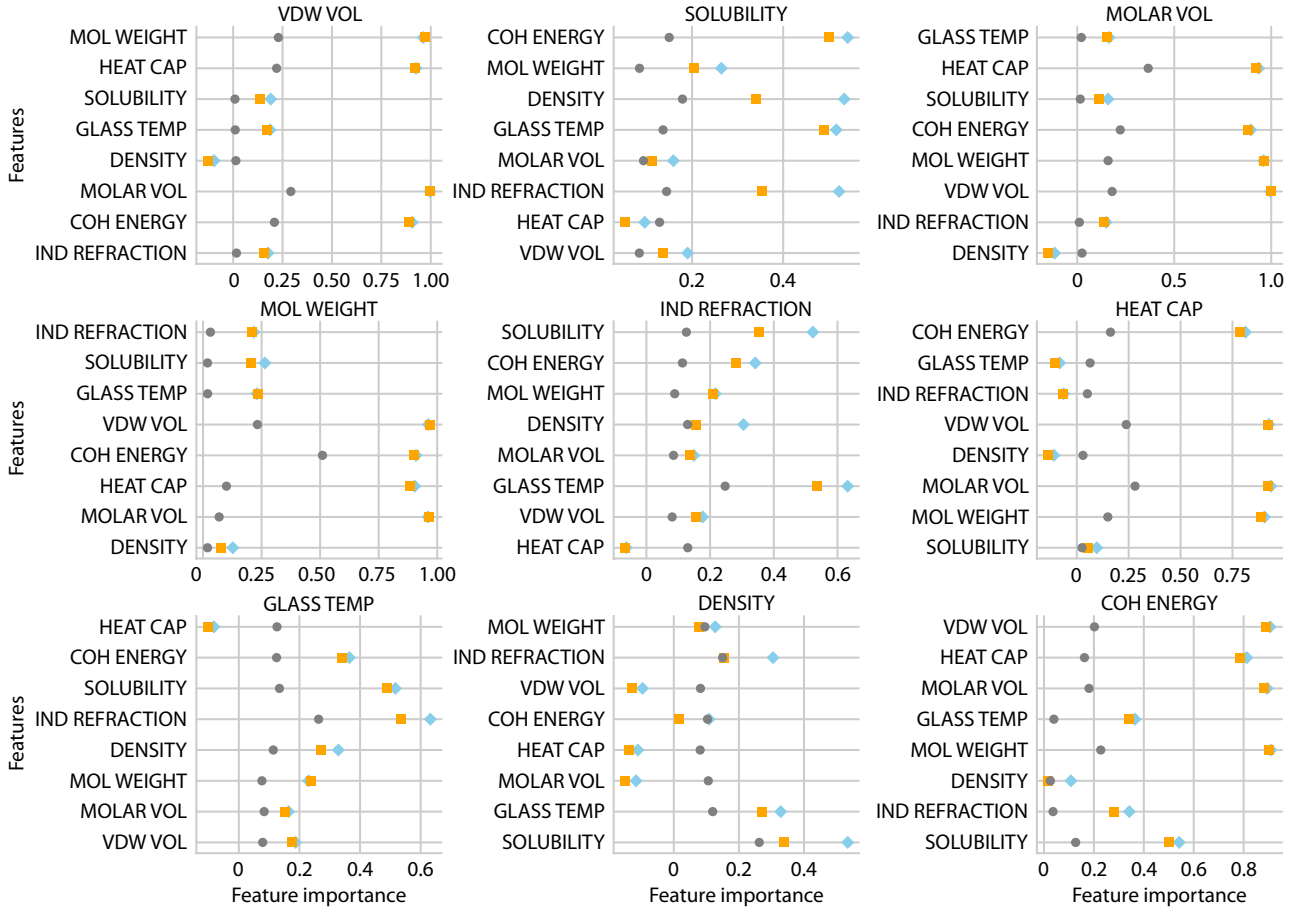


**Fig. 5** The ranking of feature importance based on the gain index (gray circle), Pearson (orange square), and Spearman (blue diamond) coefficients between features and target properties.

https://doi.org/10.1007/s10118-024-3237-y

**Table 2** Performance of prediction models depending on input data.

| Prediction target | Physical data | | | Vectors added | | | Vectors only | | |
|---|---|---|---|---|---|---|---|---|---|
| | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ |
| Index of refraction | 0.01 | 0.03 | 0.55 | 0.01 | 0.03 | 0.50 | 0.01 | 0.03 | 0.43 |
| Glass transition temperature | 0.13 | 48.24 | 0.54 | 0.09 | 33.95 | 0.77 | 0.10 | 36.58 | 0.74 |
| Density | 0.05 | 0.11 | 0.59 | 0.05 | 0.12 | 0.55 | 0.07 | 0.13 | 0.45 |
| Hildebrand solubility parameter | 0.05 | 1.34 | 0.61 | 0.05 | 1.28 | 0.65 | 0.06 | 1.64 | 0.42 |
| Molar heat capacity | 0.08 | 33.71 | 0.92 | 0.10 | 28.53 | 0.94 | 0.11 | 37.33 | 0.90 |
| Molecular weight of a repeating unit | 0.07 | 13.13 | 0.97 | 0.07 | 17.47 | 0.95 | 0.09 | 23.86 | 0.90 |
| Cohesive energy | 0.11 | 8.74 | 0.93 | 0.18 | 10.01 | 0.91 | 0.14 | 8.85 | 0.93 |
| Molar volume | 0.05 | 11.44 | 0.97 | 0.05 | 11.39 | 0.97 | 0.09 | 20.01 | 0.92 |
| Van der Waals volume | 0.06 | 8.58 | 0.96 | 0.05 | 6.57 | 0.98 | 0.09 | 13.14 | 0.92 |

taking each feature's contribution for each tree in the model; therefore, a higher value of this metric when compared to another feature implies that it is more important for generating a prediction.[72] Feature importance mainly corresponds to the parameters most correlated with the target property, with minor exceptions for parameters that do not follow the log-normal distribution (Fig. 5). Similar inconsistencies for some parameters have already been observed in other studies on predicting polymer properties, and are often found for complex models working with small and noisy datasets.[49,73] This may suggest that the present data have sufficient size and reliability for to extract strong dependences correctly and do not experience under- or overfitting in these cases.

The performance metrics obtained in all three experiments are summarized in Table 2. The mean average percentage errors for the models did not exceed 15%, which was comparable to regular experiments or calculation errors.

A visual comparison of the acquired $R^2$ metrics for different input features is shown in Fig. 6. Because of a significant rise in input data dimensionality, the accuracy of predictions for target properties that are strongly linked to physical characteristics (such as the index of refraction, molecular weight of a repeating unit, molar cohesive energy, and density) is slightly decreased when compound vectors are included as extra features along with the physical data. This decrease in accuracy, known as "the curse of dimensionality," occurs when the input dimensionality is too high to effectively comprehend the features. However, exclusion of the physical data from input

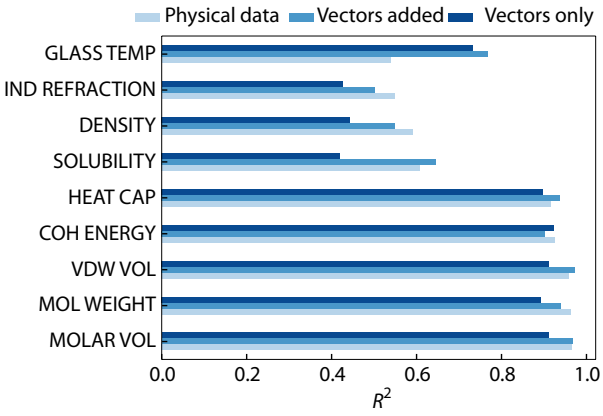


**Fig. 6** Performance of prediction models with different input data for all target physical properties.

features for these targets did not lead to a significant decrease in $R^2$, with the exception of the Hildebrand solubility parameter. As discussed above, the Hildebrand solubility parameter could contain large noise or be implicitly dependent on other structural characteristics, such as the degree of polymerization, polydispersity, and nature of the end groups, as discussed above. As "vectors added" models contained 300-dimensional compound vector and only 8 physical properties, their impact on model performance may have been insufficient to provide accuracy similar to experiments using only physical input data. The difference between the perfor-

**Table 3** Summary of predictive investigated in the literature.

| Property | Database | Number of samples | Feature | $R^2$ | Refs. |
|---|---|---|---|---|---|
| Glass transition temperature | Polymer Handbook[74] and Prediction of Polymer properties[75] | 751 | Descriptors | 0.87 | [76] |
| | PolyInfo and QM9[77] | 1321 | Multiple molecular fingerprints | 0.92 | [78] |
| | PolyInfo | 6923 | Descriptors and multiple molecular fingerprints | 0.87 | [44] |
| | PolyAskInG and QM9 | 214 real and 5000 "synthetic" samples | Molecular graph | 0.62 | [79] |
| Index of refraction | Polymer Handbook and other literature | 527 | Molecular fingerprints | 0.88 | [80] |
| | Calculated using the density functional theory | 382 | Descriptors | 0.82 | [81] |
| Density | PolyInfo | 1248 | Molecular fingerprints | 0.70 | [82] |
| Hildebrand solubility parameter | Prediction of Polymer Properties and other literature | 45 | Computational descriptors | 0.94 | [47] |
| Heat capacity | Physical properties of polymers handbook[83] | 68 | Descriptors | 0.76 | [84] |
| | PolyInfo | 124 | Descriptors | 0.95 | [85] |

https://doi.org/10.1007/s10118-024-3237-y

mances of vector-based models in these cases does not provide statistically reliable conclusions, as it may be affected by overfitting of the training data and incorrect hyperparameter optimization.

By incorporating structural information, we could significantly increase the accuracy of the glass transition temperature. It is worth noting that this parameter does not exhibit a log-normal distribution and lacks strong correlation with the physical properties of the dataset. This case indicates that the compound vector, as a polymer representation, may play a crucial role in constructing an efficient regression model.

In general, the prediction accuracy of most of the individual properties using vector representations was lower than the previously reported results for single-target predictions, with the exception of heat capacity (Table 3). However, most studies involve either significantly larger databases or extensive task-specific feature design involving manual selection of optimal descriptors calculated based on RDKit or other computational algorithms.

## CONCLUSIONS

The comparison between models constructed using physical data and compound vectors indicates that this structural representation method can enhance the predictive capability for certain properties compared to regression solely on the physical parameters of the polymer. Importantly, the ease of incorporating such featurization techniques into research workflows without significant computational resources or time-consuming setups suggests that leveraging pre-trained embeddings derived from extensive chemical datasets could benefit the development of new polymeric materials. This approach may be particularly useful in scenarios in which the transfer of polymer structure information to the model has not been previously explored. Moreover, models focusing on regression based solely on vector representations of the polymer demonstrated a prediction quality comparable to those utilizing physical parameters as features. Consequently, the integration of machine learning methods with pre-trained molecular representations, such as Mol2Vec, holds promise for advancing polymer research. In particular, this approach offers a valuable method for selecting promising monomers, estimating the parameters of existing polymers based on their structure and physical properties, and identifying potential avenues for monomer functionalization. This advancement could facilitate more efficient and informed decision-making in material design and development processes. More complex regression tasks, however, may require more accurate and information-rich representations, as well as lower vector dimensionality, to better fit small polymeric datasets. Therefore, a more accurate language model that accounts for distinctions in the chemical structure between polymers and molecular compounds, as well as the integration of polymer-specific embeddings, will be the focus of our future studies.

### Conflict of Interests

The authors declare no interest conflict.

### Electronic Supplementary Information

Electronic supplementary information (ESI) is available free of charge in the online version of this article at http://doi.org/10.1007/s10118-024-3237-y.

### Data Availability Statement

The related data (DOI/CSTR: 10.1007/s10118-024-3237-y) for this paper is available in the GitHub repository https://github.com/iszlobin/PolymerDataBase/.

### ACKNOWLEDGMENTS

This work was carried out within the framework of the program of state support for the centers of the National Technology Initiative (NTI) on the basis of educational institutions of higher education and scientific organizations (Center NTI "Digital Materials Science: New Materials and Substances" on the basis of the Bauman Moscow State Technical University). The authors are grateful to the head of the Laboratory of Molecular modeling of Materials Eugeny Alexandrov for fruitful discussions of the results.

## REFERENCES

1 Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer informatics: current status and critical next steps. *Mater. Sci. Eng.: R: Reports* **2021**, *144*, 100595.
2 Fang, W.; Mu, Z.; He, Y.; Kong, K.; Jiang, K.; Tang, R.; Liu, Z. Organic-inorganic covalent-ionic molecules for elastic ceramic plastic. *Nature* **2023**, *619*, 293–299.
3 Shi, Q.; Deng, Z.; Hou, M.; Hu, X.; Liu, S. Engineering precise sequence-defined polymers for advanced functions. *Prog. Polym. Sci.* **2023**, *141*, 101677.
4 Shao, Y.; Gao, Y.; Sun, R.; Zhang, M.; Min, J. A versatile and low-cost polymer donor based on 4-chlorothiazole for highly efficient polymer solar cells. *Adv. Mater.* **2023**, *35*, 2208750.
5 Aristova, V. A.; Bezlepkina, K. A.; Klokova, K. S.; Ardabevskaia, S. N.; Drozdov, F. V.; Cherkaev, G. V.; Milenin, S. A. Environmentally friendly synthesis and self-catalytic hydrolysis of triazole-modified organosilanes for polysiloxane production. *Chemistry Select* **2023**, *8*, e202303431.
6 Aysin, R. R.; Galkin, K. I. Impact of backbone substitution on organocatalytic activity of sterically encumbered NHC in benzoin condensation. *Molecules* **2024**, *29*, 1704.
7 Matyjaszewski, K. Architecturally complex polymers with controlled heterogeneity. *Science* **2011**, *333*, 1104–1105.
8 Vasilyeva, A. A.; Ryzhkov, A. I.; Cherkaev, G. V.; Drozdov, F. V.; Muzafarov, A. M. Silicone films with azo dyes moieties based on eugenol with response to $Cu^{2+}$ metal ions. *Mater. Chem. Phys.* **2024**, *318*, 129248.
9 Balenko, N.; Shibaev, V.; Bobrovsky, A. Polymer dispersed cholesteric liquid crystals with combined photo- and mechanochromic response. *J. Molecular Liquids* **2024**, *401*, 124637.
10 Jr, C. E. C.; Seymour, R. B. *Structure—Property Relationships in Polymers*; Springer Science & Business Media, **2012**.
11 Hart, L. F.; Hertzog, J. E.; Rauscher, P. M.; Rawe, B. W.; Tranquilli, M. M.; Rowan, S. J. Material Properties and Applications of Mechanically Interlocked Polymers. *Nat. Rev. Mater.* **2021**, *6*, 508–530.

12 Zheng, B.; Huo, L.; Li, Y. Benzodithiophenedione-based polymers: recent advances in organic photovoltaics. *NPG Asia Mater.* **2020**, *12*, 1–22.
13 Zheng, Y.; Pan, P. Crystallization of Biodegradable and biobased polyesters: polymorphism, cocrystallization, and structure-property relationship. *Prog. Polym. Sci.* **2020**, *109*, 101291.
14 Hu, R.; Qin, A.; Tang, B. Z. AIE polymers: synthesis and applications. *Prog. Polym. Sci.* **2020**, *100*, 101176.
15 Binder, K. *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*; Oxford University Press, **1995**.
16 Bergstrom, J. S. *Mechanics of Solid Polymers: Theory and Computational Modeling*; William Andrew, **2015**.
17 Ruipérez, F. Application of quantum chemical methods in polymer chemistry. *Int. Rev. Phys. Chem.* **2019**, *38*, 343–403.
18 Schmid, F. Understanding and modeling polymers: the challenge of multiple scales. *ACS Polym. Au* **2023**, *3*, 28–58.
19 Giessen, E. van der; Schultz, P. A.; Bertin, N.; Bulatov, V. V.; Cai, W.; Csányi, G.; Foiles, S. M.; Geers, M. G. D.; González, C.; Hütter, M.; Kim, W. K.; Kochmann, D. M.; LLorca, J.; Mattsson, A. E.; Rottler, J.; Shluger, A.; Sills, R. B.; Steinbach, I.; Strachan, A.; Tadmor, E. B. Roadmap on multiscale materials modeling. *Modelling Simul. Mater. Sci. Eng.* **2020**, *28*, 043001.
20 Nguyen, D.; Tao, L.; Li, Y. Integration of machine learning and coarse-grained molecular simulations for polymer materials: physical understandings and molecular design. *Front. Chem.* **2022**, *9*. DOI:10.3389/fchem.2021.820417.
21 Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a transformer-based language model for polymer property predictions. *npj Comput. Mater.* **2023**, *9*, 64.
22 Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
23 S. Clegg, P. Characterising soft matter using machine learning. *Soft Matter* **2021**, *17*, 3991–4005.
24 Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.* **2020**, *6*, eaaz4301.
25 Kuenneth, C.; Lalonde, J.; Marrone, B. L.; Iverson, C. N.; Ramprasad, R.; Pilania, G. Bioplastic design using multitask deep neural networks. *Commun. Mater.* **2022**, *3*, 1–10.
26 Bone, J. M.; Childs, C. M.; Menon, A.; Póczos, B.; Feinberg, A. W.; LeDuc, P. R.; Washburn, N. R. Hierarchical machine learning for high-fidelity 3D printed biopolymers. *ACS Biomater. Sci. Eng.* **2020**, *6*, 7021–7031.
27 Martin, T. B.; Audus, D. J. Emerging trends in machine learning: a polymer perspective. *ACS Polym. Au* **2023**, *3*, 239–258.
28 Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810.
29 Agrawal, A.; Choudhary, A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Materials* **2016**, *4*, 053208.
30 Karuth, A.; Alesadi, A.; Xia, W.; Rasulev, B. Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. *Polymer* **2021**, *218*, 123495.
31 Li, D.; Dong, C.; Chen, Z.; Dong, Y.; Liu, J. A Combinatorial machine-learning-driven approach for predicting glass transition temperature based on numerous molecular descriptors. *Molecular Simulation* **2023**, *49*, 617–627.
32 Khan, P. M.; Rasulev, B.; Roy, K. QSPR modeling of the refractive index for diverse polymers using 2D descriptors. *ACS Omega* **2018**, *3*, 13374–13386.
33 Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. *Materials Today* **2018**, *21*, 785–796.
34 Kopal, I.; Harničárová, M.; Valíček, J.; Krmela, J.; Lukáč, O. Radial basis function neural network-based modeling of the dynamic thermo-mechanical response and damping behavior of thermoplastic elastomer systems. *Polymers* **2019**, *11*, 1074.
35 Xu, P.; Lu, T.; Ju, L.; Tian, L.; Li, M.; Lu, W. Machine learning aided design of polymer with targeted band gap based on DFT computation. *J. Phys. Chem. B* **2021**, *125*, 601–611.
36 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
37 Warr, W. A. Representation of chemical structures. *WIREs Comput. Mol. Sci.* **2011**, *1*, 557–579.
38 Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J Cheminform* **2013**, *5*, 7.
39 *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2024-06-01).
40 Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
41 Landrum, G. RDKit Documentation.
42 Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U. F.; Tang, S.; Bi, J.; Chen, W.; Li, Y. Machine-learning-assisted *De Novo* design of organic molecules and polymers: opportunities and challenges. *Polymers* **2020**, *12*, 163.
43 Sun, L. W.; Li, H.; Zhang, X. Q.; Gao, H. B.; Luo, M. B. Identifying conformation states of polymer through unsupervised machine learning. *Chinese J. Polym. Sci.* **2020**, *38*, 1403–1408.
44 Tao, L.; Chen, G.; Li, Y. Machine learning discovery of high-temperature polymers. *Patterns* **2021**, *2*, 100225.
45 Tao, L.; Varshney, V.; Li, Y. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *J. Chem. Inf. Model.* **2021**, *61*, 5395–5413.
46 Aldeghi, M.; W. Coley, C. A Graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **2022**, *13*, 10486–10498.
47 Liu, T. L.; Liu, L. Y.; Ding, F.; Li, Y. Q. A machine learning study of polymer-solvent interactions. *Chinese J. Polym. Sci.* **2022**, *40*, 834–842.
48 Zhao, Y.; J. Mulder, R.; Houshyar, S.; C. Le, T. A review on the application of molecular descriptors and machine learning in polymer design. *Polym. Chem.* **2023**, *14*, 3325–3346.
49 Ding, F.; Liu, L. Y.; Liu, T. L.; Li, Y. Q.; Li, J. P.; Sun, Z. Y. Predicting the mechanical properties of polyurethane elastomers using machine learning. *Chinese J. Polym. Sci.* **2023**, *41*, 422–431.
50 Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35.
51 Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
52 Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.
53 Shibayama, S.; Funatsu, K. Industrial case study: identification of important substructures and exploration of monomers for the rapid design of novel network polymers with distributed representation. *BCSJ* **2021**, *94*, 112–121.
54 Huang, X.; Ma, S.; Zhao, C. Y.; Wang, H.; Ju, S. Exploring high thermal conductivity polymers via interpretable machine

learning with physical descriptors. *npj Comput. Mater.* **2023**, *9*, 191.
55 Chemical Retrieval on the Web (CROW). Available online: http://www.polymerdatabase.com/ (accessed on 12 September 2023).
56 Chi, M.; Gargouri, R.; Schrader, T.; Damak, K.; Maâlej, R.; Sierka, M. Atomistic descriptors for machine learning models of solubility parameters for small molecules and polymers. *Polymers* **2022**, *14*, 26.
57 Hatakeyama-Sato, K.; Watanabe, S.; Yamane, N.; Igarashi, Y.; Oyaizu, K. Using GPT-4 in parameter selection of polymer informatics: improving predictive accuracy amidst data scarcity and 'ugly duckling' dilemma. *Digital Discovery* **2023**, *2*, 1548–1557.
58 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*; **2011**; pp 22–29. DOI:10.1109/EIDWT.2011.13.
59 Jarrett, D.; Cebere, B.; Liu, T.; Curth, A.; van der Schaar, M. HyperImpute: generalized iterative imputation with automatic model selection. arXiv June 15, 2022. DOI:10.48550/arXiv.2206.07769.
60 Liu, F. T.; Ting, K. M.; Zhou, Z. H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*; **2008**; pp 413–422. DOI:10.1109/ICDM.2008.17.
61 Borrohou, S.; Fissoune, R.; Badir, H. Data cleaning survey and challenges – improving outlier detection algorithm in machine learning. *J. Smart Cities and Society* **2023**, *2*, 125–140.
62 *Molecular Machine Learning with DeepChem - ProQuest*. https://www.proquest.com/openview/9c0e06a343233b48d962991d19873ed8/1?pq-origsite=gscholar&cbl=18750&diss=y (accessed 2024-08-25).
63 Cortes, C.; Vapnik, V. Support-vector networks. *Mach Learn* **1995**, *20*, 273–297.
64 Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.
65 Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **1986**, *185*, 1–17.
66 Zou, H.; Hastie, T. Regularization and variable selection *via* the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.
67 Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, **2016**; pp. 785–794.
68 Jamieson, K.; Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*; PMLR, **2016**; pp 240–248.
69 Saeki, S.; Yamaguchi, T. Semiempirical examination of the van Der Waals and Tonks models for polymers and simple liquids. *Polymer* **1987**, *28*, 484–488.
70 Barton, A. F. M. *Solubility parameters*. ACS Publications. DOI:10.1021/cr60298a003.
71 Venkatram, S.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Critical assessment of the hildebrand and hansen solubility parameters for polymers. *J. Chem. Inf. Model.* **2019**, *59*, 4188–4194.
72 Shi, X.; Wong, Y. D.; Li, M. Z. F.; Palanisamy, C.; Chai, C. A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Anal. Prevention* **2019**, *129*, 170–179.
73 Liu, T.; Liu, L.; Cui, F.; Ding, F.; Zhang, Q.; Li, Y. Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning. *J. Mater. Chem. A* **2020**, *8*, 21862–21871.
74 *Polymer Handbook*, 4. ed.; Brandrup, J., Ed.; A Wiley-Interscience publication; Wiley: New York Weinheim, **1999**.
75 Bicerano, J. *Prediction of Polymer Properties*, 3rd Ed.; CRC Press: Boca Raton, **2002**.
76 Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Modelling Simul. Mater. Sci. Eng.* **2019**, *27*, 024002.
77 Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
78 Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **2019**, *5*, 1–11.
79 Volgin, I. V.; Batyr, P. A.; Matseevich, A. V.; Dobrovskiy, A. Yu.; Andreeva, M. V.; Nazarychev, V. M.; Larin, S. V.; Goikhman, M. Ya.; Vizilter, Y. V.; Askadskii, A. A.; Lyulin, S. V. Machine learning with enormous "synthetic" data sets: predicting glass transition temperature of polyimides using graph convolutional neural networks. *ACS Omega* **2022**, *7*, 43678–43691.
80 Lightstone, J. P.; Chen, L.; Kim, C.; Batra, R.; Ramprasad, R. Refractive index prediction models for polymers using machine learning. *J. Appl. Phys.* **2020**, *127*, 215105.
81 Najeeb, J.; Shah, S. S. A.; Tahir, M. H.; I. Hanafy, A.; M. El-Bahy, S.; M. El-Bahy, Z. Machine learning assisted designing of polymers and refractive index prediction: easy and fast screening of polymers from chemical space. *Mater. Chem. Phys.* **2024**, *324*, 129685.
82 Lee, F. L.; Park, J.; Goyal, S.; Qaroush, Y.; Wang, S.; Yoon, H.; Rammohan, A.; Shim, Y. Comparison of machine learning methods towards developing interpretable polyamide property prediction. *Polymers* **2021**, *13*, 3653.
83 Mark, J. *Physical Properties of Polymers*; Cambridge University Press, **2004**.
84 Bhowmik, R.; Sihn, S.; Pachter, R.; Vernon, J. P. Prediction of the specific heat of polymers from experimental data and machine learning methods. *Polymer* **2021**, *220*, 123558.
85 Kazemi-Khasragh, E.; González, C.; Haranczyk, M. Toward diverse polymer property prediction using transfer learning. *Comput. Mater. Sci.* **2024**, *244*, 113206.