Journal of Applied Physics ARTICLE scitation.org/journal/jap

Refractive index prediction models for polymers using machine learning

Cite as: J. Appl. Phys. **127**, 215105 (2020); doi: 10.1063/5.0008026
Submitted: 17 March 2020 · Accepted: 16 May 2020 ·
Published Online: 2 June 2020

Jordan P. Lightstone, Lihua Chen, Chiho Kim, Rohit Batra, and Rampi Ramprasad[a)]

AFFILIATIONS

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, USA

Note: This paper is part of the special collection on Machine Learning for Materials Design and Discovery
[a)]**Author to whom correspondence should be addressed:** rampi.ramprasad@mse.gatech.edu

ABSTRACT

The refractive index (RI) is an important material property and is necessary for making informed materials selection decisions when optical properties are important. Acquiring accurate empirical measurements of RI is time consuming, and while semi-empirical and computational determination of RI is generally faster than empirical determination, predictions are less accurate. In this work, we utilized experimentally measured RI data of polymers to build a machine learning model capable of making accurate near-instantaneous predictions of RI. The Gaussian process regression model is trained using data of 527 unique polymers. Feature engineering techniques were also used to optimize model performance. This new model is one of the most chemically diverse and accurate RI prediction models to date and improves upon our previous work. We also concluded that the model is capable of providing insights about structure–property relationships important for estimating the RI when designing new polymer backbones.

I. INTRODUCTION

The refractive index (RI) is a material property directly related to optical, electrical, and magnetic behavior of a material.[1] In light scattering measurements of dilute polymer solutions, the refractive index increment is an essential parameter for determining the molecular weight, size, and shape of the polymer in solution.[2] Additionally, the RI serves as an important property when designing and selecting polymeric materials used as waveguides, optical films, and optical fibers.[3] Furthermore, high refractive index polymers (HRIPs), RI $\geq$1.5, are attractive materials for substrates in advanced display devices, optical encapsulants and adhesives in organic light emitting diodes, image sensors, and anti-reflective coatings.[4–7]

Driven by pragmatic and technological needs, efforts to calculate RI have been made since the mid-19th century. Early theoretical methods proposed by Lorentz and Lorenz as well as Gladstone and Dale are accurate but limited by the lack of available molar refraction and molecular volume (V) data for new polymer materials.[8] In the 1970s, group contribution based methods emerged[2] and more recently, semi-empirical methods materialized, enabling multiple pathways for RI estimation. However, these methods are constrained by limitations of the Lorentz–Lorenz equation[9] and disregard the three-dimensional structural arrangements of polymers.[2,9] To fully incorporate physical and chemical structure effects on the RI, density functional perturbation theory (DFPT) was used to compute RI.[10] However, this method is computationally expensive and has inherent limitations arising from practical assumptions made to model polymers (e.g., highly crystalline structures).[11,12]

Regression based prediction methods appeared in the polymer science and engineering community in the 1970s, providing powerful means for rapidly predicting polymer properties.[2,10,13] Since the conception of these techniques, quantitative structural property relationship (QSPR) methods[8,14–16] and other hand-crafted feature sets were utilized to numerically represent polymer structures for infusion into machine learning (ML) workflows similar to the one seen in Fig. 1. In our previous works, we developed a set of hierarchical descriptors to numerically represent polymers.[10] Using this method, the chemical and structural features of a polymer at different length scales—atomistic, block, and morphological—are generated. Our unique fingerprinting scheme combined with available polymer property datasets, either empirical or computational, was used to build ML models that rapidly predict various polymer properties including the glass transition temperature, tensile strength, and density.[10] In this previous work, a predictive RI

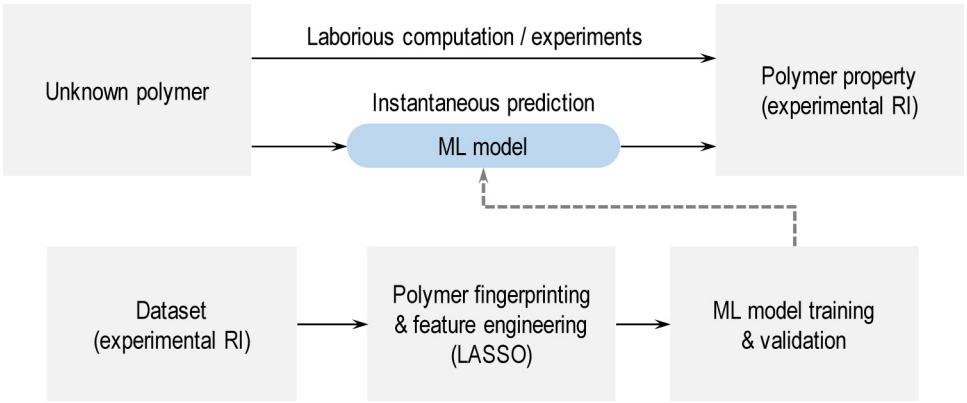


FIG. 1. Workflow for building and implementing data-driven RI prediction models for polymer design.

model was trained on DFPT computed data. There were limitations in this model due to the assumptions inherent to DFPT simulations[17] (such as the assumption of highly compact crystalline structures which led to over-estimation of the RI in the training set).

Here, we trained and optimized a ML model for predicting RI using experimentally measured RI of 500 polymers. A unique hierarchical polymer fingerprinting scheme,[10] a feature reduction technique, and the Gaussian process regression (GPR) algorithm were used to train the ML model. The performance of the developed model was bench-marked against our previous work and validated using 27 polymers entirely distinct from the training set. We believe the resulting model instantaneously and accurately predicts RI of new polymers while identifying critical features necessary for designing polymer structures to achieve specific RI values. These contributions can assist in the rational design and screening of polymer candidates for applications where optical properties, specifically RI, are crucial for design specifications.

II. DATASET AND METHODOLOGY

A. Dataset

Our dataset is comprised of experimentally measured RI of 527 polymers at room temperature. The polymers in this dataset are made of nine chemical species including H, C, N, O, S, Si, F, Cl, and Br and span multiple polymer classes, e.g., polyoxides, polyvinyls, polyolefins, polyamides, polyimides, polyureas, polyethers, etc. As illustrated in Fig. 2(a), RI values range from 1.3 to 2.0 and follow a bell-shape distribution, i.e., with the majority of data located from 1.4 to 1.8. Only a few data points are available in high (>1.8) and low (<1.4) RI ranges, respectively. Data were obtained from numerous publicly available sources, including the Polymer Handbook,[18] Handbook of Polymers,[19] and Polymer Data Handbook.[20] Data were also acquired from literature sources[8] and online repositories.[10,21] When multiple RI values were reported for the same polymer, the median RI of the set was chosen to avoid

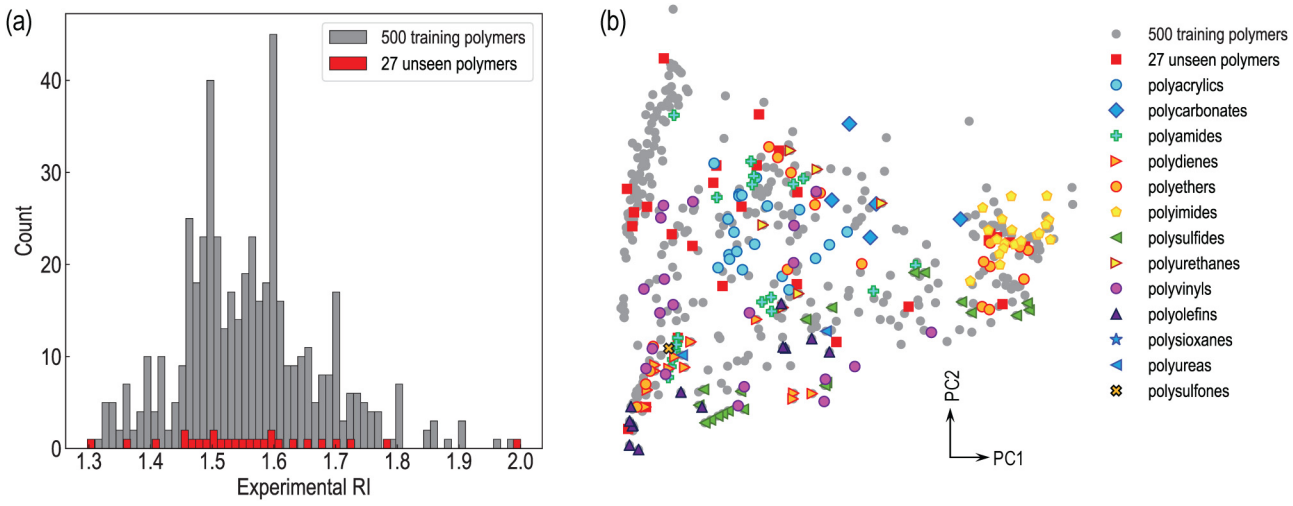


FIG. 2. (a) Refractive index (RI) dataset, including 500 training polymers and 27 unseen polymers. (b) Chemical diversity as a function of the first and second principal components (PC1 and PC2), where color symbols represent polymer class.

irregularities caused when averaging outlying values.[22] In this work, 500 of the 527 points were used to train the ML model with fivefold cross-validation (CV), while the remaining 27 polymers were withheld to validate the developed ML model.

B. Features engineering

A hierarchical fingerprinting scheme generated features that numerically represent the chemical and bonding relationships red of a polymer. It includes (1) atomic-level features that capture atomic information of "$A_iB_jC_k$" fragments (i, j, and k are the number of fold-coordinated A, B, and C atoms, respectively); (2) block-level features which describe the presence of a set of 500 predefined building blocks typically found in polymers; and (3) morphological-level features that cover information at chain-level scale, e.g., the length of the side chains and fraction of atoms that are part of rings. More detailed descriptions of our fingerprinting technique have been described previously.[10] Using this fingerprinting scheme, 388 features (denoted by X_{All}) were generated for all 527 polymers. Feature values for all 527 polymers were normalized from 0 to 1.

In addition, we performed principal component analysis (PCA) on the entire 527 polymer dataset with all 388 features to visualize the breadth of the chemical and structural diversity. In Fig. 2(b), the first (PC1) and second (PC2) principle component values for each polymer were plotted. Various polymer classes were labeled by colored symbols in Fig. 2(b), revealing the diverse chemical space in consideration.

To identify relevant features, the Least Absolute Shrinkage and Selection Operator regression (LASSO) method was used to fit the entire training set (500 polymers) and the initial 388 features with fivefold CV. By optimizing the regularization term, the model with the highest R^2 coefficient was obtained. Upon completion, 21 features (denoted by X_{LASSO}) with non-zero coefficients remained and were subsequently used to train ML models in Sec. III.

C. Gaussian process regression

Gaussian process regression (GPR) with the radial basis function (RBF) kernel was applied to train the ML models. The co-variance function between two polymers with features $\boldsymbol{x}$ and $\boldsymbol{x}'$ is expressed as

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f \exp\left(-\frac{1}{2\sigma_l^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right) + \sigma_n^2. \quad (1)$$

Here, σ_f, σ_l and σ_n denote the variance, the length-scale parameter, and the expected noise in the RI dataset, respectively. Each value was determined by maximizing the log-likelihood estimate during the model training process. In addition, fivefold CV was adopted in all ML models to avoid overfitting. The root mean squared error (RMSE) and the R^2 coefficient were the two metrics used to evaluate the performance of the GPR models.

In order to understand the effect of training set size on prediction accuracy, models were generated using increasing training set sizes. Initially, 100 polymers were randomly selected from the training set and used to train a model. The training set in subsequent models increased by 50 polymers until the entire 500 polymer dataset was used for training. 50 models were developed for each training set size and the average RMSE and standard deviation were calculated for all 50 models. The results from this process were used to build the learning curve shown in Fig. 3(a).

III. RESULTS AND DISCUSSION

As illustrated in Fig. 3(a), the performance of developed ML models were evaluated using the learning curves, which show the average training and test RMSE as a function of training set size. The test set in this figure refers to 500 minus the training set size, all of which are distinct from the 27 polymers used for model validation. The error bars represent 1 standard deviation of the average

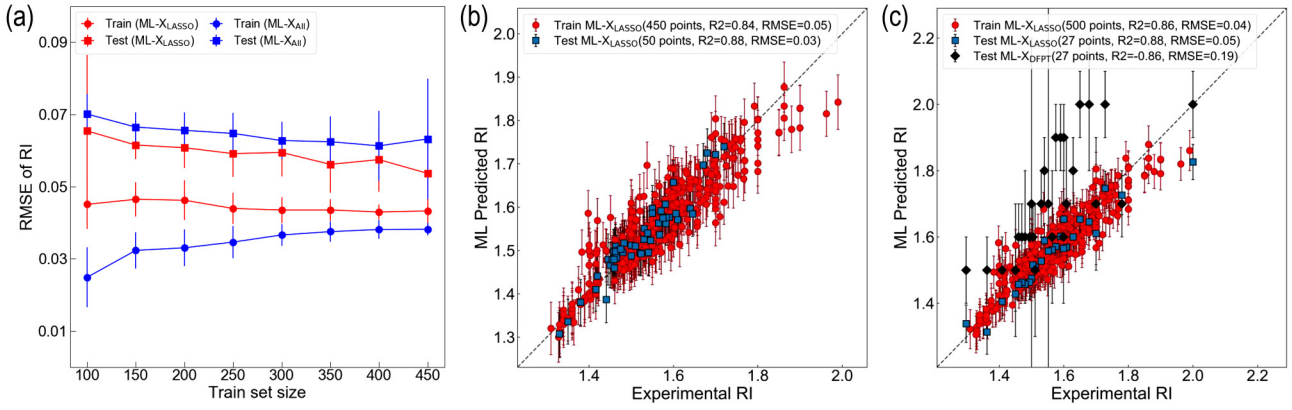


FIG. 3. (a) Prediction accuracy for ML-X_{All} and ML-X_{LASSO} models trained using different train set sizes, averaged over 50 runs. The corresponding test set sizes in (a) are equal to the difference between total training dataset (500) minus the train set size. (b) Parity plot obtained from the ML-X_{LASSO} model (21 features) with train and test set size of 450 and 50, respectively. (c) Parity plot obtained from ML-X_{LASSO} model using the entire training set, including prediction of the 27 unseen with the ML-X_{LASSO} and prediction using the ML- X_{DFPT} from our previous work.

RMSE values over 50 runs. As expected, the test RMSE of the ML model trained with all initial features (ML-X_{All}) and LASSO reduced features (ML-X_{LASSO}) decreased with increasing training set size. We also note that models trained with X_{LASSO} features, on average, led to lower test RMSE than X_{All} features, demonstrating that LASSO regression is an effective method for eliminating irrelevant features in this work. Further, ML-X_{LASSO} provides a test RMSE of 0.05 (<4% of absolute RI values), when 90% of the training set was used. A corresponding parity plot is shown in Fig. 3(b), i.e., experimental RI vs ML predicted RI using ML-X_{LASSO}. The error bars in the plot represent the GPR uncertainty.

The RIs ranging from 1.8 to 2.0 are underestimated (<10%) by the ML model. This is a result of sparse training data in this specific region [see Fig. 2(b)]. Table S1 in the supplementary material shows five HRIPs with under-predicted RI. One commonality among the repeat units is the presence of rings on the main chain. In addition, all five examples contain S and or N. Although these features correlate positively with RI (see Fig. 4), perhaps the degree to which these features contribute to high RI is not enough. It is possible other feature reduction techniques would have yielded a more accurate feature set. Features such as number or weight average molecular weight or stereochemistry not present in the current feature set could also improve prediction of HRIPs. Despite the under-prediction of HRIPs, a test RMSE of 0.05 is achieved with the ML-X_{LASSO} model.

To validate the generality and accuracy of the developed ML models, the RI of 27 unseen polymers was predicted using the ML-X_{LASSO} model trained with the entire dataset (500 data points). These 27 unseen polymers were entirely unique structures from the training set, and their RI uniformly spanned the range (1.3–2.0) of the training set, as shown in Fig. 2. In addition, we compared the predicted RI of 27 unseen polymers using ML-X_{LASSO} and our previous ML model trained on 400 DFPT computed data (ML-X_{DFPT}).[10]

Figure 3(c) shows ML-X_{LASSO} predicted RI of 500 training polymers, ML-X_{LASSO} predicted RI of 27 unseen polymers, and ML- X_{DFPT} predicted RI of the same 27 unseen polymers. We note that the ML-X_{LASSO} model can accurately predict RI of 27 unseen polymers, with the test RMSE of 0.05 and R^2 of 0.88. Our previous ML-X_{DFPT} had a test RMSE of 0.19 and an R^2 of 0.86, which indicates that the present ML-X_{LASSO} model has better RI prediction capabilities (higher R^2 and lower RMSE), when compared with the ML-X_{DFPT} model. The experimentally measured RI (EXP) and the ML-X_{LASSO} and ML-X_{DFPT} predicted RI values of the 27 unseen polymers are summarized in Table S1 in the supplementary material. All 27 polymers predicted using the ML-X_{DFPT} were over-predicted. In fact, of the entire 527 experimental dataset, only 19 RI were not over-predicted. There are two main reasons for the improved results. First, in the ML-X_{LASSO}, experimental values used for training overcome the accuracy problem of the DFPT training data mentioned earlier. Second, a more diverse chemical space of polymers is present in the ML-X_{LASSO} training dataset. Figure 3(c) indicates that the ML-X_{LASSO} model accurately predicted the RI of new polymers and could act as a tool for predicting the RI of novel polymer structures.

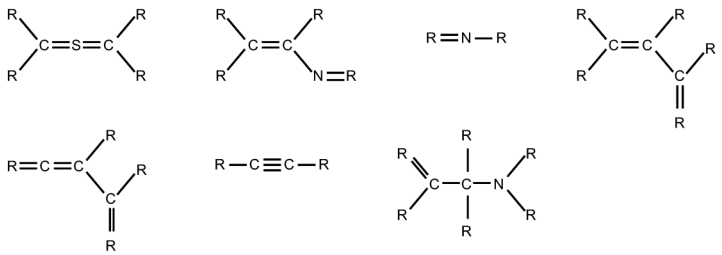
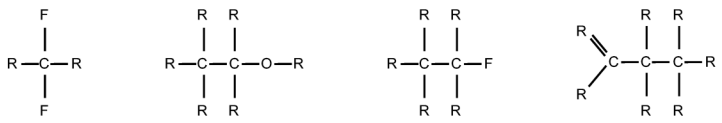
Correlation with RI		Representative features
Positive correlation	Atomic & block level	
	Chain level	Number of rings in monomer, number of ring atoms/total atoms in monomer, polyamides, length (number of atoms) of largest side chain
Negative correlation	Atomic & block level	
	Chain level	Number of 3-vertex carbon atoms, acrylate, linear chain, distance between rings, Number of 3-membered rings/total atoms in monomer

FIG. 4. LASSO selected features having strong positive or negative correlations with RI. R denotes an arbitrary chemical group of C, O, H, N elements.

It is also worth analyzing the LASSO reduced features (X_{LASSO}). Figure 4 lists the 21 features correlated with RI. The positive and negative coefficients from the LASSO method indicate positive and negative correlations with RI. RI arises from the electronic polarization of materials, i.e., the electron cloud displacement under an electric field. We note that carbon double and triple bonds (atomic and block-level features, respectively) and the number of rings (chain-level feature) have a positive correlation with RI. This is because the double and triple bonds have a high mobility of π electrons, leading to high electronic polarization and thus high RI. The introduction of C–F or C–O bonds, on the other hand, can decrease the electronic polarization by strongly binding electrons, due to the high electronegativity of F and O atoms. Moreover, there is a negative correlation between RI and the chain-level features including number of three-vertex carbon atoms and distance between rings, since these features can introduce large volumes, resulting in low polarization density.

Using correlation information, it is reasonable to create guidelines to assist the polymer design process when RI as a target property. To achieve a high RI select monomers and polymerization pathways that maintain numerous carbon double and triple bonds, rings, and S and N atoms. Alternatively, if a slightly lower RI is desired, prioritize halogen or ether groups, for example. Coupling these guidelines with instantaneous ML prediction capabilities could allow accelerated screening and synthesis of novel polymers with tuned RI.

IV. CONCLUSION

In conclusion, we have developed a machine learning (ML) model capable of instantaneous refractive index (RI) prediction of polymers and provided a set of design criteria for creating new polymers with highly tuned RI. This model is trained using a dataset of experimental RI of 500 polymers, a hierarchy of polymer features and Gaussian process regression algorithm. The performance of the developed ML model was validated with 27 unseen polymers and proven to have greater accuracy and precision compared with our previous work. Key chemo-structural features that correlate to high and low RI values were identified and can be used as means for guiding design of new polymer structures where tailored RI is desired. If this model will be used to design polymers with large RI, the Gaussian process regression uncertainty will provide the reliability of the predicted values. Novel polymer structures and polymers with high RI (>1.8) may have high uncertainties and these uncertainties may provide useful guidance for next experiments via active learning, with newly generated data aiding in model improvement. A final model was trained using all 527 polymers and the 21 least absolute shrinkage and selection operator regression features. This model is hosted on the Polymer Genome platform (https://www.polymergenome.org) and can be utilized to rapidly predict the RI of desired polymers.

SUPPLEMENTARY MATERIAL

See the supplementary material for five HRIPs with under-predicted RI from Fig. 3(c) (Table S1) and 27 unseen polymers with corresponding experimental, ML-X_{LASSO} predicted, and ML-X_{DFT} predicted RI values (Table S2).

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research through Grant No. N0014-17-1-2656, a Multi-disciplinary University Research Initiative (MURI) grant.

The authors declare no competing financial interest.

DATA AVAILABILITY

The refractive index dataset will be made available upon reasonable request for academic use.

REFERENCES

[1]W. Knoll, Annu. Rev. Phys. **49**, 569 (1998).
[2]D. W. van Krevelen, Angew. Chem. **85**, 465 (1973).
[3]E. M. Pearce, J. Polym. Sci. Polym. Lett. Ed. **15**, 56 (1977).
[4]T. Nakamura, H. Fujii, N. Juni, and N. Tsutsumi, Opt. Rev. **13**, 104 (2006).
[5]J. Liang, L. Li, X. Niu, Z. Yu, and Q. Pei, Nat. Photonics **7**, 817 (2013).
[6]Y.-W. Wang and W.-C. Chen, Compos. Sci. Technol. **70**, 769 (2010).
[7]Q. Chen, D. Das, D. Chitnis, K. Walls, T. Drysdale, S. Collins, and D. Cumming, Plasmonics **7**, 695 (2012).
[8]A. R. Katritzky, S. Sild, and M. Karelson, J. Chem. Inf. Comput. Sci. **38**, 1171 (1998).
[9]A. Askadskii, *Computational Materials Science of Polymers* (Cambridge International Science Publishing Ltd, 2002), Vol. 7, pp. 65–66.
[10]C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, J. Phys. Chem. C **122**, 17575 (2018).
[11]M. A. F. Afzal and J. Hachmann, Phys. Chem. Chem. Phys. **21**, 4452 (2019).
[12]T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, Sci. Data **3**, 160012 (2016).
[13]R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, NPJ Comput. Mater. **3**, 54 (2017).
[14]J. Bicerano, *Prediction of Polymer Properties*, Plastics Engineering, 3rd ed. (Marcel Dekker, New York, 2002), Vol. 65.
[15]R. García-Domenech and J. V. de Julián-Ortiz, J. Phys. Chem. B **106**, 1501 (2002).
[16]P. R. Duchowicz, S. E. Fioressi, D. E. Bacelo, L. M. Saavedra, A. P. Toropova, and A. A. Toropov, Chemometr. Intell. Lab. **140**, 86 (2015).
[17]S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515–562 (2001).
[18]G. Brandrup and E. H. Immergut, *Polymer Handbook* (Wiley, New York, 1999).
[19]G. Wypych, *Handbook of Polymers* (Elsevier, Oxford, 2012), p. ii.
[20]J. Mark, *Polymer Data Handbook* (Oxford University Press, 1999).
[21]S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, 2011 International Conference on Emerging Intelligent Data and Web Technologies (2011), pp. 22–29.
[22]A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan, and R. Ramprasad, Comput. Mater. Sci. **172**, 109286 (2020).