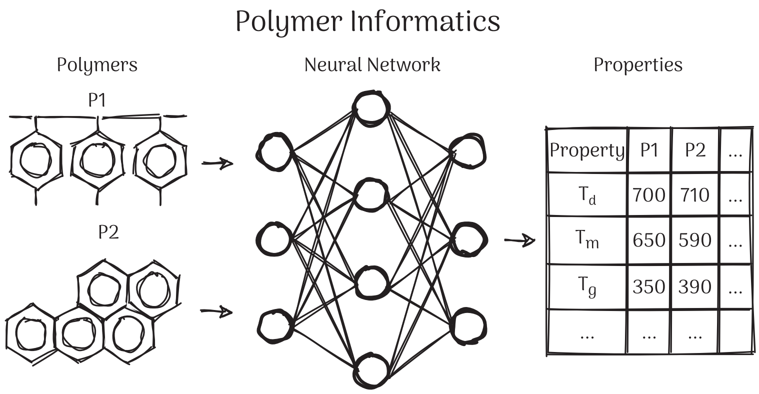
Article

Patterns

Polymer informatics with multi-task learning

Graphical abstract



Property	P1	P2	...
T_d	700	710	...
T_m	650	590	...
T_g	350	390	...
...	...	...	

Highlights

- We overcome data scarcity in polymer datasets using multi-task models
- Our approach is expected to become the preferred training method for materials data
- We derive chemical guidelines for the design of application specific polymers

Authors

Christopher Kuenneth,
Arunkumar Chitteth Rajan, Huan Tran,
Lihua Chen, Chiho Kim,
Rampi Ramprasad

Correspondence

rampi.ramprasad@mse.gatech.edu

In brief

Materials data tend to be scarce. Inherent correlation between properties in materials data sets can, however, be utilized using multi-task models. Using a combined data set of 36 polymer properties for over 13,000 polymers, we found that multi-task models not only outperform single-task models but also allows for the derivation of chemical guidelines that pave the way for the rational design of application specific polymers.

Kuenneth et al., 2021, Patterns *2*, 100238
April 9, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.patter.2021.100238

Article

Polymer informatics with multi-task learning

Christopher Kuenneth,[1] Arunkumar Chitteth Rajan,[1] Huan Tran,[1] Lihua Chen,[1] Chiho Kim,[1] and Rampi Ramprasad[1,2,*]
[1]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Lead contact
*Correspondence: rampi.ramprasad@mse.gatech.edu

https://doi.org/10.1016/j.patter.2021.100238

THE BIGGER PICTURE Polymers display extraordinary diversity in their chemistry, structure, and applications. However, finding the ideal polymer possessing the right combination of properties for a given application is non-trivial as the chemical space of polymers is practically infinite. This daunting search problem can be mitigated by surrogate models, trained using machine learning algorithms on available property data, that can make instantaneous predictions of polymer properties. In this work, we present a versatile, interpretable, and scalable scheme to build such predictive models. Our "multi-task learning" approach is used for the first time within materials informatics and efficiently, effectively, and simultaneously learns and predicts multiple polymer properties. This development is expected to have a significant impact on data-driven materials discovery.

1 2 **3** 4 5 **Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Modern data-driven tools are transforming application-specific polymer development cycles. Surrogate models that can be trained to predict properties of polymers are becoming commonplace. Nevertheless, these models do not utilize the full breadth of the knowledge available in datasets, which are oftentimes sparse; inherent correlations between different property datasets are disregarded. Here, we demonstrate the potency of multi-task learning approaches that exploit such inherent correlations effectively. Data pertaining to 36 different properties of over 13,000 polymers are supplied to deep-learning multi-task architectures. Compared to conventional single-task learning models, the multi-task approach is accurate, efficient, scalable, and amenable to transfer learning as more data on the same or different properties become available. Moreover, these models are interpretable. Chemical rules, that explain how certain features control trends in property values, emerge from the present work, paving the way for the rational design of application specific polymers meeting desired property or performance objectives.

INTRODUCTION

Polymers display extraordinary diversity in their chemistry, structure, and applications. This is reflected in the ubiquity of polymers in everyday life and technology. The vigor with which polymers are studied using both computational and experimental methods is leading to a constant flux of (mostly uncurated and heterogeneous) data. The field of polymer science and engineering is thus poised for exciting informatics-based inquiry and discovery.[1–5]

In general, materials datasets tend to be small. This presents challenges for the creation of robust and versatile machine learning (ML) models for materials property prediction. Nevertheless, the apparent data sparsity in the materials domain is somewhat compensated by the information-richness of each data point or the availability of prior physics-based knowledge of the phenomenon under inquiry. For instance, a given target property A of a material may be correlated with a different property B. If data for A is sparse but data for B is copious, effective prediction models for A may be developed by exploiting this correlation using algorithms that respect parsimony. Alternatively, imagine that property A may be measured using an accurate (but laborious or expensive) experimental procedure α and a not-so-accurate (but rapid or inexpensive) procedure β. Again, powerful models for the prediction of property A at the accuracy level of α may be developed by using sparse α-type data along with copious β-type data.

With the above in mind, let us suppose that a dataset for a particular materials sub-class involves a variety of target

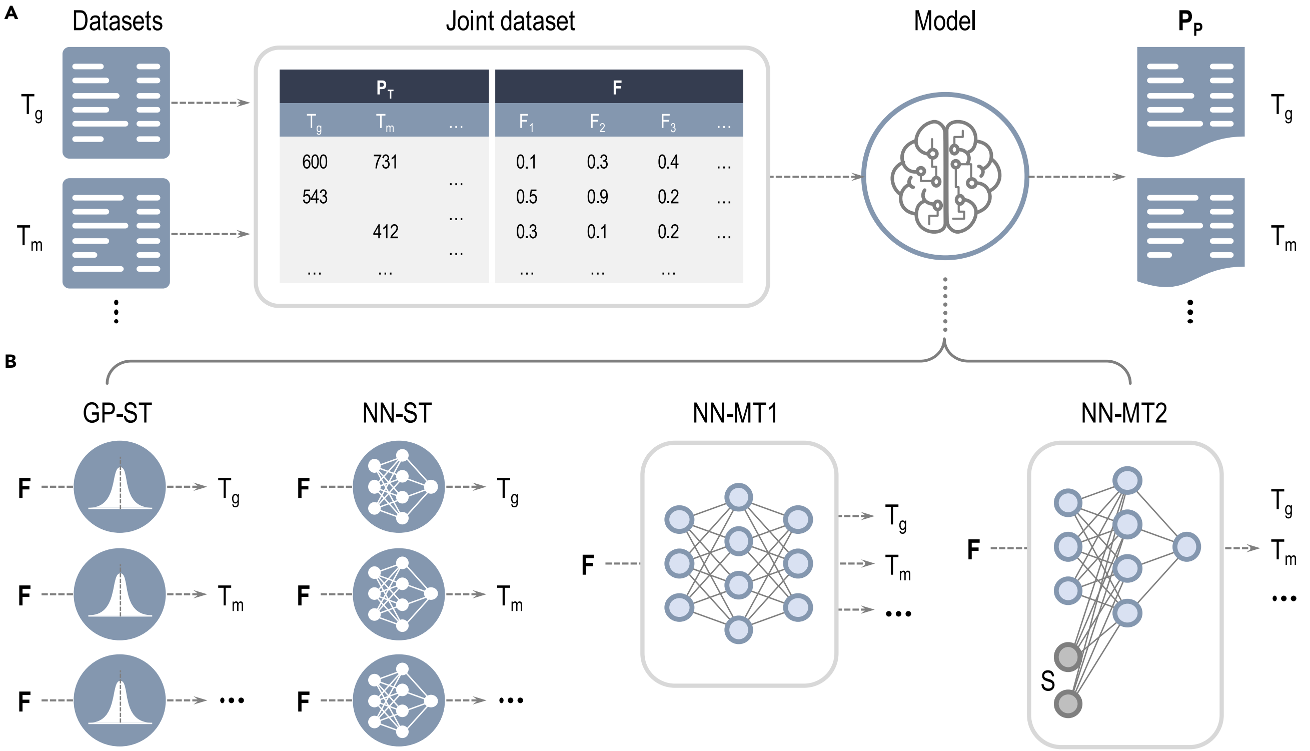


Figure 1. Data pipeline and machine learning modles.

(a) *From left to right:* Separatly collected polymer roperty datasets are merged into a joint dataset; machine learning models are trained on the joint datset with fingerprint components (**F**) as input and predicted properties (**P**$_p$) as output. **P**$_T$ are the property taret values. The loss funcrion is defined as the mean squared error of **P**$_P$ and **P**$_T$: T_g, T_m, **P** and **F** stand for galss transition temprature, melting temperature, property and fingerprint component matrix, respectively. (b) Four different machine learing models: single-task (ST). mlti-task (MT). Gaussian process (GP), neural network (NN).

properties, with each property data point potentially obtained from multiple sources or measurements. Not all property values may be available for all material cases. In other words, the dataset may contain a number of "missing values". Figure 1A shows a schematic of such a dataset for polymer properties. As mentioned above, data from subsets of property and source types may be correlated with each other. Given this scenario, our objective is to utilize a multi-task (MT) learning method that can ingest the entire dataset, recognize inherent correlations, and make predictions of all properties by effectively transferring knowledge from one property or source type to another. As a baseline to assess how MT learning performs, one may utilize learning methods that learn to predict each property individually, one at a time, implicitly disregarding correlations of the property with other properties; we call these as single-task (ST) learning methods. ST and MT learning schemes are illustrated in Figure 1B.

MT learning is an advanced data-driven learning method, which, within materials science, requires coalesced datasets of multiple properties to be effective. While materials scientists have not yet adopted MT learning, it has been effectively utilized in drug design for the classification of synthesis-related properties and has demonstrated clear advantages over other learning approaches.[6–8] Transfer learning, a related approach that has been applied in the polymer domain, likewise demonstrates advantages over traditional learning approaches.[9] Another somewhat related approach, which goes under the names of multi-fidelity learning or co-kriging, has been utilized to address some materials science problems;[10–12] nevertheless, the MT learning approach described here surpasses conventional multi-fidelity learning in terms of efficiency, scalability, dataset sizes that can be handled, and the types and number of outputs.

In the present contribution, we focus on polymers and build the first comprehensive MT model to date for the instantaneous prediction of 36 polymer properties. Data for 36 different properties of over 13,000 polymers (corresponding to over 23,000 data points) were obtained from a variety of sources.[4,12–18] Table 1 shows a synopsis of the data. All polymers are "fingerprinted", i.e., converted to a machine-readable numerical form, using methods described elsewhere[4,19,20] (and briefly in the experimental procedures section). These fingerprints (and available property values) are the inputs to our ML models. We have developed four types of learning models: two flavors of MT models and two flavors of ST models (the latter two models serve as baselines). The two MT models utilize neural network (NN) architectures and are referred to as NN-MT1 and NN-MT2 models. Once trained on the coalesced datasets corresponding to 36 polymer properties, the NN-MT1 model takes in polymer fingerprints for a new polymer and outputs all 36 properties via its last multi-head output layer. The NN-MT2 model, on the other hand, uses an architecture that receives the concatenation of the polymer fingerprint and a selector vector as input. The selector

Table 1. Synopsis of polymer properties

Property	Symbol	Unit	Source[a]	Points	Data range	Ref.
Thermal						
Melting temperature	T_m	K	Exp.	2079	[226,860]	
Glass transition temperature	T_g	K	Exp.	5072	[80,873]	15,16,4
Decomposition temperature	T_d	K	Exp.	3520	[219,11667]	
Thermal conductivity	λ	W mK^{-1}	Exp.	78	[0.1,0.49]	
Thermodynamic & physical						
Heat capacity	c_p	J gK^{-1}	Exp.	79	[0.8,2.1]	
Atomization energy	E_{at}	eV atom^{-1}	DFT	390	[–6.8,5.2]	4
Limiting oxygen index	O_i	%	Exp.	101	[13.2,70]	
Crystallization tendency (DFT)	X_c	%	DFT	432	[0.1,98.8]	
Crystallization tendency (exp.)	X_e	%	Exp.	111	[1,98.5]	
Density	ρ	g cm^{-3}	Exp.	910	[0.84,2.18]	4
Fractional free volume	V_{ff}	1	Exp.	128	[0.1,0.47]	
Electronic						
Bandgap (chain)	E_{gc}	eV	DFT	3380	[0.02,9.86]	
Bandgap (bulk)	E_{gb}	eV	DFT	561	[0.4,10.1]	12
Electron affinity	E_{ea}	eV	DFT	368	[–0.39,5.17]	
Ionization energy	E_i	eV	DFT	370	[3.56,9.84]	
Optical & dielectric						
Refractive index (DFT)	n_c	1	DFT	382	[1.48,2.95]	4
Refractive index (exp.)	n_e	1	Exp.	516	[1.29,2]	4
Dielectric constant	ε_0	1	DFT	382	[2.6,9.1]	4
Frequency dependent electric constant[b]	ε_f	1	Exp.	1187	[1.95, 10.4]	18
Mechanical						
Tensile strength	σ_{ts}	MPa	Exp.	672	[2.86,289]	
Young's modulus	Y	MPa	Exp.	629	[0.02,9.8]	
Solubility & permeability						
Hildebrand solubility parameter	δ_s	$\sqrt{\text{MPa}}$	Exp.	112	[12.3,29.2]	4,14
Gas permeability[c]	μ_g	Barrer	Exp.	2168	[0, 4.7][d]	13

The total number of single data points is 23,616, and the total number of merged data points in the joint database is 13,766.

[a] Experiments (Exp.); density functional theory (DFT)

[b] $f \in \{1.78, 2, 3, 4, 5, 6, 7, 9, 15\}$ is the log $_{10}$ (frequency in Hz); e.g., ε_3 is the dielectric constant at a frequency of 1kHz

[c] $g \in \{\text{He}, \text{H2}, \text{CO2}, \text{O2}, \text{N2}, \text{CH4}\}$

[d] The data range is transformed by $f : \mu_g \mapsto \log_{10}(\mu_g + 1)$

vector indicates the property and instructs the NN to output just that selected property. The baseline ST models utilize either Gaussian processes (GP-ST) or a conventional NN architecture (NN-ST). The GP-ST and NN-ST models are trained independently on individual polymer datasets; there are thus 36 prediction models, one for each property, of each ST flavor. All four ML approaches developed here are shown in Figure 1B; details on the architecture of the models and training process are provided in the experimental procedures section.

RESULTS

Correlations in data

Unlike ST models, MT models learn from inherent correlations in datasets. Our polymer dataset shows such interesting (some expected, but some new) correlations between pairs of properties as illustrated in Figure 2 using Pearson correlation coefficients (PCCs). For example, the dielectric constants at different frequencies (ranging from $\varepsilon_{1.78}$ to ε_9, where the subscript indicates the frequency on log$_{10}$ scale in Hz) are highly positively correlated with each other. Understandably, the dielectric constant at optical frequency, ε_{15}, which is controlled purely by electronic polarization, is weakly correlated with the dielectric constants at low frequencies, which are related to ionic, orientational, and electronic factors. The permeabilities of gases, μ_g (where g represents one of 6 gas molecules), are highly positively correlated with each other. By contrast, gas permeabilities and dielectric constants are negatively correlated with each other, indicating that polymers with high ε_f tend to display low μ_g, and vice versa. Of note, high positive correlations can be seen between the glass transition (T_g) and melting temperatures (T_m), and large negative correlation between the electronic band gap (E_{gb}, for bulk polymers, and E_{gc}, for chains) and T_g. The important observation that should be made by the inspection of Figure 2 is that there are

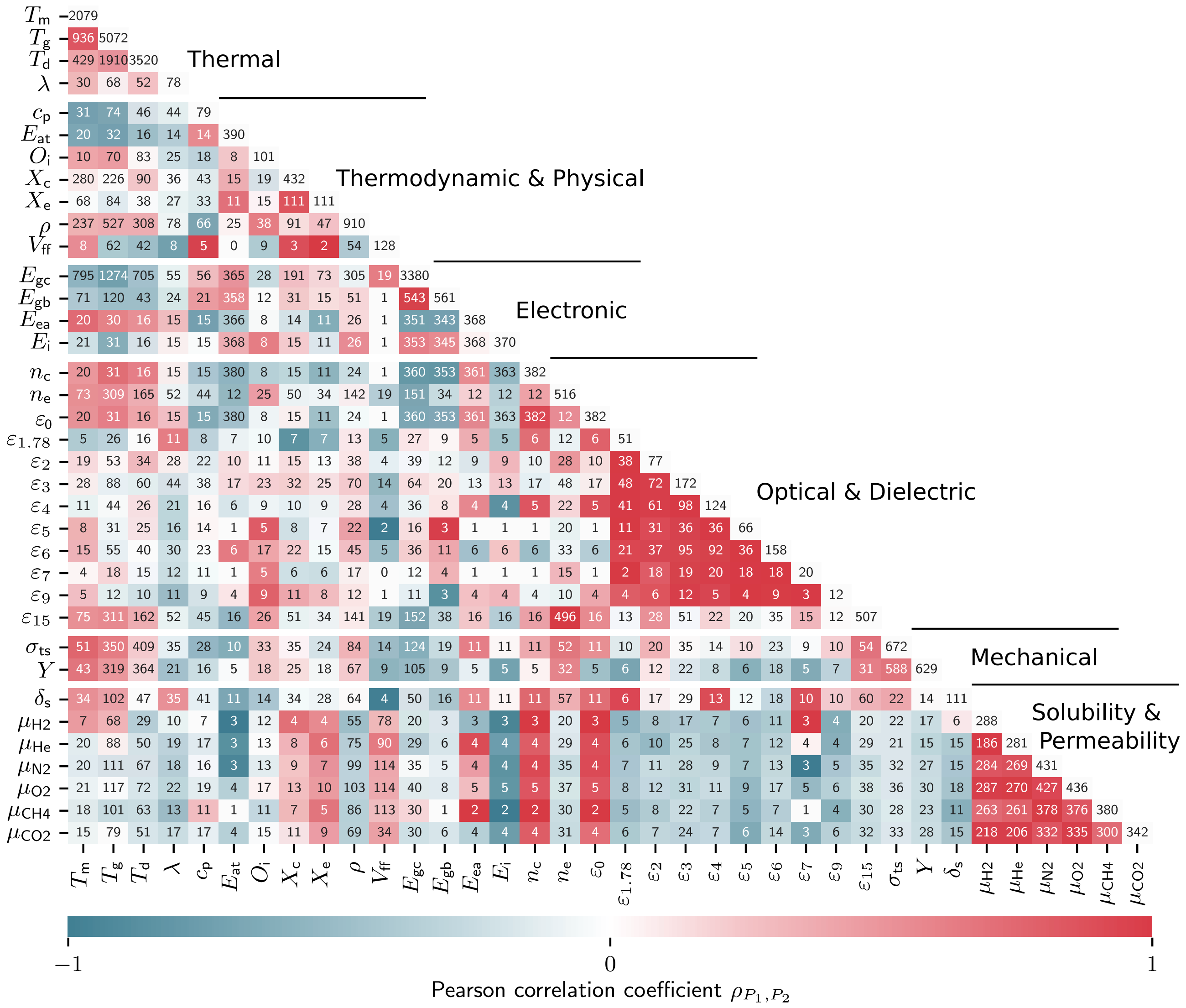


Figure 2. Polymer property heatmap of the Pearson correlation coefficients

Red patches indicate positively correlated, blue patches negatively correlated, and white patches uncorrelated Pearson correlation coefficients (PCCs). Numbers on the main diagonal indicate the total data points of a particular property in the dataset, whereas off-diagonal numbers denote the number of polymers for which both properties are available. The PCCs for less than two congruences were set to 0. Property symbols are defined in Table 1.

several examples of weak to strong positive and negative correlations between properties that can potentially be exploited in MT learning schemes.

Single- and multi-task models

To investigate whether these correlations improve the prediction performance when used in MT models, we train four different ML models. The first two models use the ST architecture, which predicts single polymer properties, and the next two models use the MT architecture, which predicts all properties (see Figure 1B). For the last architecture (NN-MT2), we present two variants: (1) trained on all properties (NN-MT2-all) and (2) trained on just the properties within a given category of Table 1 (NN-MT2-sub). There are thus a total of six NN-MT2-sub models, one for each category. Figure 3 compiles the training results of all models. The average of the five-fold cross-validation root mean squared errors (RMSEs) of the unseen validation dataset are shown, with the error bars indicating 68% confidence intervals of the RMSE averages. A more condensed overview of the training results, using the categories defined in Figure 2, is shown in Table 2. The RMSE and R^2 values of all models are documented in Table S2 of the supplemental information.

Using the training results in Figure 3 and Table 2, we first evaluate the performance of both ST models (GP-ST and NN-ST). In general, we find the NN-based ST models to perform better than their Gaussian process (GP) counterparts. This is an interesting result as both models only differ in their underlying learning algorithm but otherwise follow the same ST doctrine that learns polymer properties independently. Nevertheless, it is also known that NN models can approximate more general function classes

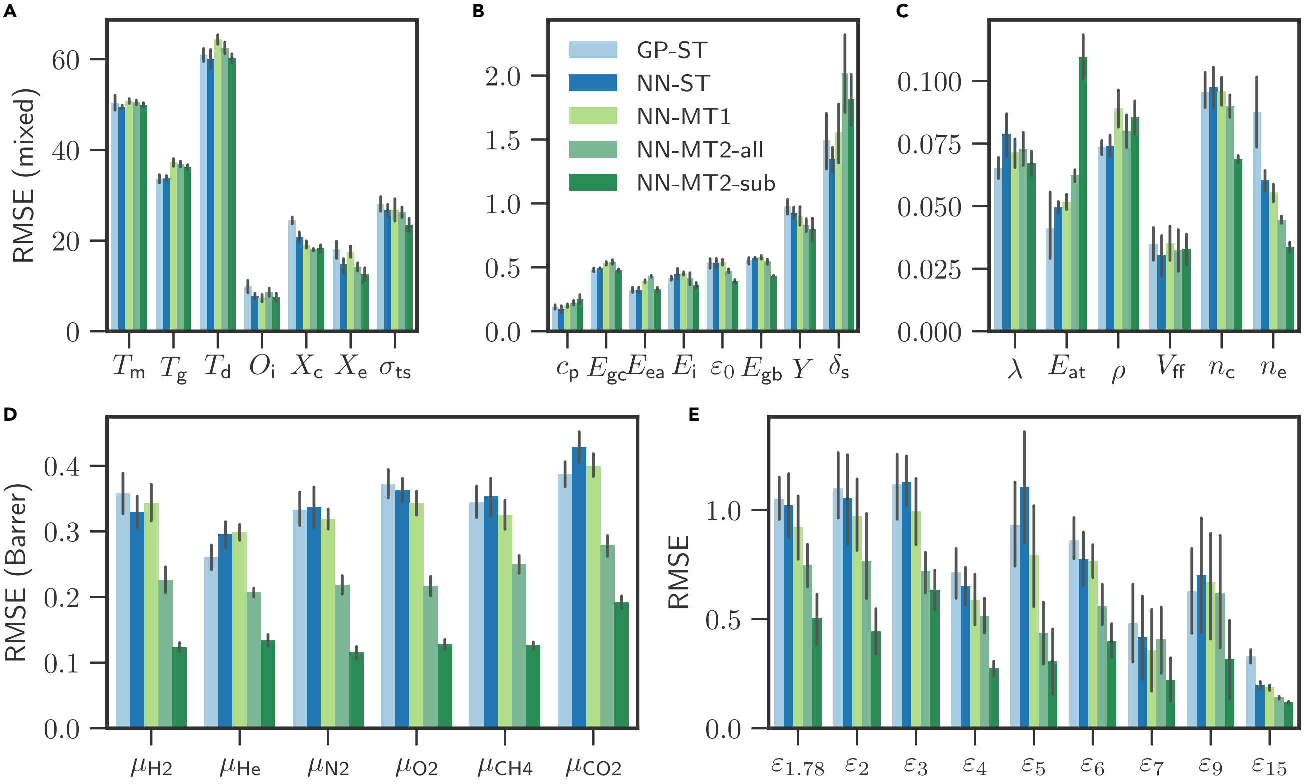


Figure 3. Five-fold cross validation root mean squared errors of four machine learning models for 36 polymer properties

The properties are arranged in sub-figures according to their magnitudes. The colored bars indicate the average of five-fold cross validation root mean squared errors (RMSEs), and the error bars are the 68% confidence intervals of the RMSE averages. Units can be found in Table 1.

better than GP models,[21] which ultimately leads to the better overall performance of the NN models.

Next, we compare the ST and MT models using the average of the normalized RMSE values in Table 2. The RMSE values were normalized so that the maximal value is 1. Overall, Table 2 shows that the MT models perform generally better than the ST models. The six NN-MT2-sub models provide the best accuracy over all the 36 properties and are followed by NN-MT2-all, which is superior to NN-MT1 and NN-ST (with GP-ST finishing up last). Comparing the first two categories (thermal and thermodynamic & physical) of Table 2, the ST models perform slightly better than the MT models. Similar observations can be made in Figures 3A–3C, which comprise the properties of these first two categories. The reason for the good performance of the ST models on these two categories is the copious amount of data that is available. This allows the optimizer to fully focus on a large single-property space during the optimization. MT models on the other hand tend to compromise their performance on these cases to also be able to provide high predictive performance for the cases where the dataset is sparse (where ST models suffer). Moreover, given that the property values are scaled to similar data ranges in the pre-processing step (see experimental procedures section), the MT models are effectively trained on data ranges with different sparsities, which present numerical challenges for the optimizer. The last four categories in Table 2 paint a different picture compared to the first two categories; the MT models clearly outperform the ST models. The last four categories comprise the highly correlated properties μ_g and ε_f (c.f., Figure 2) and also those with small datasets, which the MT models use to improve their prediction performance. We note that although Table 1 indicates that there are 1,187 points for the frequency-dependent dielectric constant, this dataset is spread across 10 frequencies.

Among the MT models, the concatenation-based conditioned NN-MT2-all and six NN-MT2-sub models display a significantly lower averaged RMSE of 0.79 and 0.65. The degraded performance of the multi-head NN-MT1 model (0.89) in comparison to NN-MT2 may be ascribed to the sparse population of our dataset (sparsity of 95%) due to missing properties for many polymers. When the optimizer computes the gradients to back-propagate over the network, it has to exclude these missing properties, effectively leaving related network parts unchanged. The architecture of the NN-MT2 eliminates this problem by using a one-hot representation of the dataset. As this representation has no missing values, the optimizer can always back-propagate over the entire network.

By holistically evaluating the performance of all properties and models in Figure 3 and Table 2, it can be stated that NNs should be preferred over GPs. NNs predict not only with higher accuracy than GPs but they also scale efficiently in terms of growing dataset, training, and prediction time. MT models comprise similar accuracy as ST models and should

Table 2. Averages of the normalized root mean squared errors values

Model	All	Categories					
		Thermal	Thermod. & physical	Electronic	Optical & dielectric	Mech.	Solubility & permeability
GP-ST	0.93	0.92	0.86	0.88	0.98	1.00	0.93
NN-ST	0.89	0.95	0.76	0.91	0.91	0.95	0.94
NN-MT1	0.89	0.98	0.83	0.97	0.83	0.94	0.92
NN-MT2-all	0.79	0.97	0.82	0.96	0.69	0.89	0.70
NN-MT2-sub	0.65	0.94	0.87	0.79	0.47	0.82	0.46

The root mean squared error averages were normalized for each property so that the maximal value is 1.

particularly be utilized whenever the considered data exhibit high correlations. Moreover, the NN-MT2-sub models show that MT models trained on property categories with high expected correlations outperform a MT model trained on all (possibly uncorrelated) properties. The six NN-MT2-sub models display the overall best performance among all properties for our dataset.

Deriving chemical guidelines

While our ML models learn the mapping between fingerprints of polymers and properties, they do not provide insights on how single fingerprint components relate to properties, nor how modifications of fingerprint components affect properties. To address such problems, we calculate Shapley additive explanation (SHAP) values[22,23] that measure the impact of structural polymer features, which are indicated through our fingerprint components, to polymer properties. As these SHAP values do only quantify the magnitude of the fingerprint-structure relation and not the direction, we compute special fingerprint impact values as the product of SHAP and PCC values. Using these fingerprint impact values, we derive chemical guidelines that can be compared with well-known empirical guidelines from the literature to validate our model irrespective of the used training dataset.

The most influential fingerprint component in Figure 4 is $F_{\mathrm{e,main,chain,ring}}$, which is defined as the ratio of the number of non-hydrogen atoms in rings (cycles of atoms) to the total number of atoms, is large for polymers containing many rings. $F_{\mathrm{e,main,chain,ring}}$ has a strong positive impact on T_{m}, T_{g}, T_{d}, σ_{ts}, Y, O_{i}, n_e, n_c, ε_0, and ε_{15} and strong negative impact on E_{gc}, E_{gb}, E_{at}, E_{i}, and c_{p}. This means the presence of atomic rings increases the former-mentioned properties but decreases the latter. As such, using the fingerprint impacts, we can provide chemical guidelines helpful to design future polymers. The derived chemical guidelines as impacts of the fingerprint component $F_{e,main,chain,ring}$ may be mapped to empirical guidelines that scientists have learned over the years. For instance, it is known that the presence of atomic rings stiffens polymers, which explains the increase of the mechanical properties, σ_{ts} and Y. Moreover, the atomic rings restrict chain motion, which is the reason for increased T_{m}, T_{g}, and T_{d} values in ring-rich polymers. The conjugated double bonds in atomic rings introduce agitated π-electrons, which increase n_e, n_c, and ε_f, especially at high frequencies (ε_{15}) where electronic displacements contribute significantly to optical properties. Also, the agitated π-electrons of atomic rings can participate in electrical conduction, which is why rings increase the conductivity of polymers. In contrast, properties such as E_{gc}, E_{gb}, E_{at}, and E_{i},[24] which correlate with insulating behavior or stability, are decreased as $F_{\mathrm{e,main,chain,ring}}$ has negative impact.

The second-most impactful fingerprint component is $F_{\mathrm{e,fam,acrylate}}$, which is defined to be one if the acrylate group is present in the polymer and zero otherwise. Polyacrylates are known to have T_{g} values below room temperature. Consistent with this expectation, the presence of $F_{\mathrm{e,fam,acrylate}}$ negatively impacts T_{m}, T_{d}, and T_{g}. Another interesting finding is that $F_{\mathrm{e,fam,polyamides}}$, the fourth-most impactful fingerprint component, positively impacts δ_{s} because the amide bonds in polyamides strengthen inter-molecular forces that make polymers resist dissolution. One can likewise derive useful insights from the other features identified in Figure 4.

DISCUSSION

In this work, we demonstrate how MT learning improves the property prediction of ML models in materials sciences by using inherent property correlations of coalesced datasets. Our polymer dataset includes 36 properties from over 23,000 data points of more than 13,000 polymers. The dataset is learned using four different ML models: the first two models are based on GPs and NNs and use the ST architecture. Models three and four are solely based on NNs and use two different types of MT architectures. Our analysis shows that the fourth model (NN-MT2) outperforms the other three models overall. Upon closer inspection of performance within individual property sub-classes, it is evident that MT models outperform ST models especially when correlations between properties within the subclass are high and/or when the dataset sizes within those sub-classes are small. In closing, we conclude that MT learning successfully improved the property prediction by utilizing the inherent correlations in our coalesced polymer dataset. Furthermore, we compute fingerprint impact values, which are based on SHAP and PCC values, that allow us to derive chemical guidelines for polymer design from the trained MT model and add an additional validation (and value-added) step pertaining to knowledge extraction.

Besides better performance, our MT learning approach makes fast predictions of all properties in a short time and eliminates the laborious training of many single ML models for each property. In addition, the NNs enable scalability and fast retraining of the MT models when new properties or data become available. MT models can also be developed further to include uncertainty

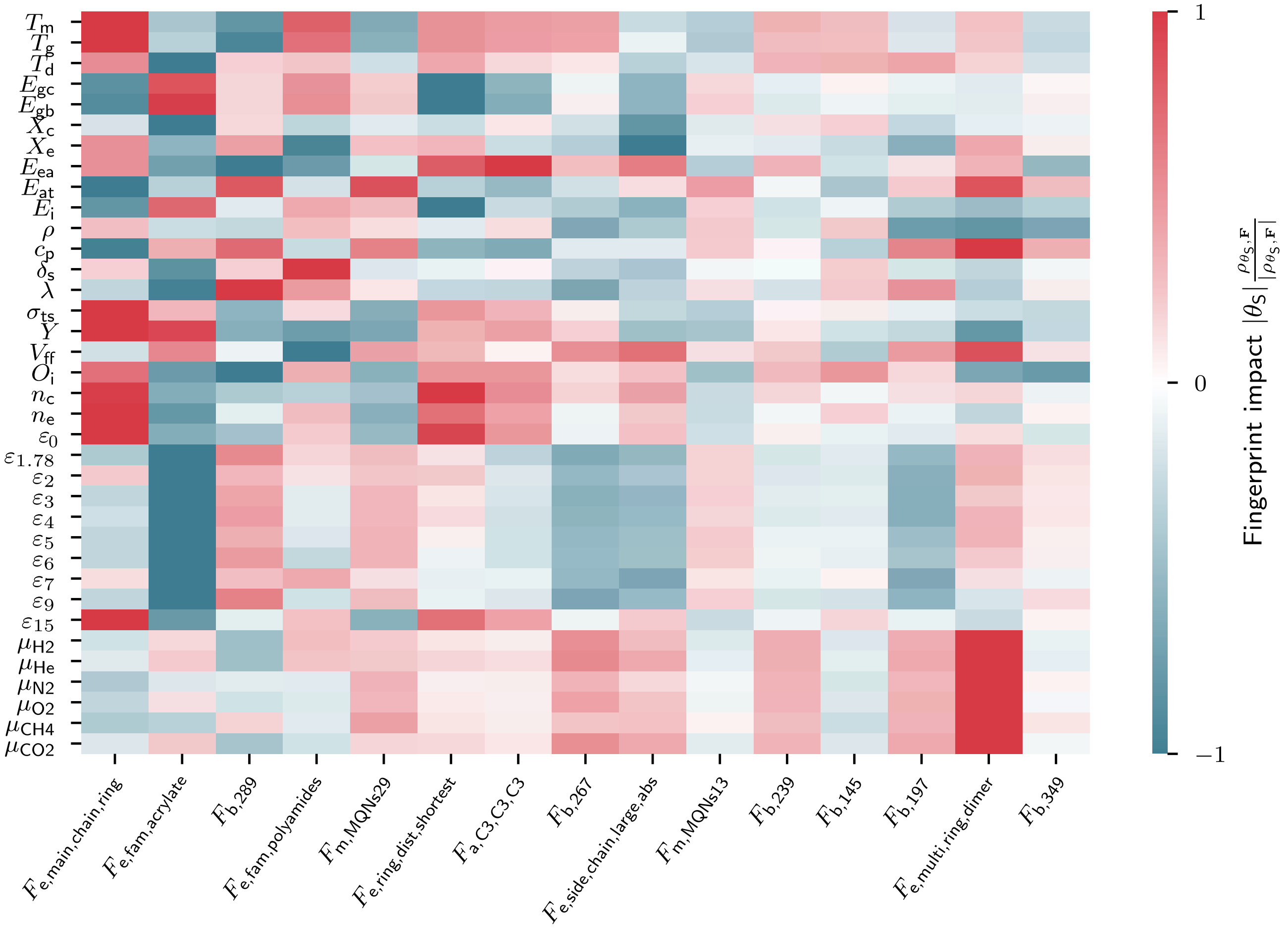


Figure 4. Fingerprint impact values of the 15 most important fingerprint components

Positive fingerprint impact values (red) indicate that a positive value of a fingerprint will potentially increase a property value, and vice versa. Small impact values, however, suggest little or no change of the property value. The fingerprint component names are defined in Table S1 of the supplemental information.

quantifications, which is often helpful for end-users. Also, it is important to note that our MT learning approach and fingerprint impact analysis are not limited to polymeric materials; in fact, they can easily be modified to handle any material. Given all these factors and the good performance, we believe that MT models should be the preferred method for property predictive ML in materials informatics. All ML models developed in this work will be made available on the Polymer Genome platform at https://www.polymergenome.org/.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Rampi Ramprasad (rampi.ramprasad@mse.gatech.edu).

Materials availability

There are no physical materials associated with this study.

Data and code availability

All data points of DFT computed properties of Table 1 that support the findings of this study are openly available at https://khazana.gatech.edu/. The code for training the ML models is available at https://github.com/Ramprasad-Group/multi-task-learning.

Data and preparation

The polymer database used in this work comprises 36 individual polymer properties, which are meticulously collected and curated from two main sources: (i) in-house high-accuracy and high-throughput density functional theory (DFT) based computations,[12,20,25,26] (ii) experimental measurements reported in the literature (as referenced in Table 1), printed handbooks[27–29] and online databases.[30,31] Both sources come with distinct uncertainties and should not be mixed together under the same property; while DFT contains systematic uncertainties introduced through the approximations of the density functional or chosen convergence parameters, experimental uncertainties arise from sample and measurement conditions. However, along the lines of multi-fidelity learning approaches, the concurrent use of properties of different sources in separate columns of the dataset may help to lower the total generalization error. An overview of the used 36 polymer properties, their symbols, units, sources, and data ranges can be found in Table 1. It should be noted that some of the individual property datasets have already been used in other publications (see references in Table 1). However, this work marks the first time that these single property datasets have been fused for the holistic training of the MT models.

MT architectures take in the fingerprint of a polymer and use the same NN to predict on all properties. The property prediction happens either simultaneously, as in our multi-head MT model (NN-MT1), or iteratively, as in our concatenation-based MT model (NN-MT2). MT models thus need a coalesced dataset that lists all properties for one polymer per row, see Figure 1A. To construct such a coalesced dataset, we merge the 36 single-property datasets using our polymer fingerprints (see next Section). Moreover, to ease the work of the optimizer and accelerate the training, we scale all 36 property values to a comparable data range using Scikit-learn's Robust Scaler.[32] Additionally, the gas permeabilities (μ_g) are logartihmically pre-processed by $f: \mu_g \mapsto \log_{10}(\mu_g + 1)$ to narrow down their large data range. Ultimately, for computing error measurements and production, the original metrics are restored by inversely transforming the predictions. Apart from scaling the polymer properties, fingerprint components are normalized to the range of $[0, 1]$.

Fingerprinting

Fingerprinting converts geometric and chemical information of polymers to machine-readable numerical representations. Polymer chemical structures are represented using SMILES[33] strings that follow the SMILES syntax but use two stars to indicate the two endpoints of the repetitive unit of the polymers.

Our polymer fingerprints capture key features of polymers at three hierarchical length scales.[19] At the atomic-scale, our fingerprints track the occurrence of a fixed set of atomic fragments (or motifs).[20,34] For example, the fragment "O1-C3-C4" is made up of three contiguous atoms, namely, a one-fold coordinated oxygen, a 3-fold coordinated carbon, and a 4-fold coordinated carbon, in this order. A vector of such triplets form the fingerprint components at the lowest hierarchy. The next level uses the quantitative structure-property relationship (QSPR) fingerprints[4,35] to capture features on larger length-scales. QSPR fingerprints are often used in chemical and biological sciences, and implemented in the cheminformatics toolkit RDKit.[36] Examples of such fingerprints are the van der Waals surface area,[37] the topological polar surface area (TPSA),[38,39] the fraction of atoms that are part of rings (i.e., the number of atoms associated with rings divided by the total number of atoms in the formula unit), and the fraction of rotatable bonds. The highest length-scale fingerprint components in our polymer fingerprints deal with "morphological descriptors." They include features such as the shortest topological distance between rings, fraction of atoms that are part of side-chains, and the length of the largest side-chain. Eventually, the used polymer fingerprint vector (**F**) of a polymer in this study has 953 components of which 371 are from the first, 522 from the second and 60 from the third level.

Machine learning models

To allow for comparison our four ML models, we consistently chose the loss function being the mean squared error (MSE) of predicted and true values for five different training datasets, generated by five-fold cross-validation. The five-fold cross validation means along with the 68% confidence intervals are reported in Figure 3.

Single-task learning with Gaussian process regression (GP-ST)

Scikit-learn's[32] implementation of GP regression was used as the baseline model, denoted by GP-ST. The kernel function was chosen as the parameterized radial basis functions plus a white kernel contribution to capture noise. GP predicts probability distributions from which prediction values are derived as the means of the distributions, and confidence intervals of the distributions define the uncertainties. GP's limiting factor is the inversion of the kernel matrix, which grows squared ($\mathbf{F}^2$) with the number of used features (**F**), rendering GP unsuitable for big-data learning problems. NNs eliminate this problem.

Learning with neural networks (NN-ST, NN-MT1, NN-MT2)

All three NN models were implemented using the Python API of Tensorflow.[40] We used the Adam optimizer with a learning rate of 10^{-3} to minimize the MSE of the prediction and target polymer property. Early stopping combined with a learning rate scheduler was deployed. All hyper-parameters such as the initial learning rate, number of layers and neurons were optimized with respect to the generalization error using the Hyperband method[41] of the Python package Keras-Tuner.

The NN-ST model takes in the fingerprint vector and outputs one polymer property. Just as the GP-ST model, we train an ensemble of 36 independent NN-ST models to predict all 36 properties. The NN-MT1 model has a multi-head MT architecture that takes in the fingerprint vector (**F**) and outputs 36 properties (**P**) at the same time. On the other hand, the NN-MT2 model uses a concatenation-based MT architecture that takes in the fingerprint vector and a selector vector **S**, outputting only the selected polymer property. The selector vector has 36 components where one component is 1 and the rest 0. Each of the three NN models has two dense layers, followed by a parameterized ReLU activation function and a dropout layer with rate 0.5. The Hyperband method optimized the two dense layers to 480 and 224 neurons for the NN-ST model, 480 and 416 neurons for the NN-MT1 model, and 224 and 160 neurons for NN-MT2 model. An additional dense layer was added with 1 neuron for the NN-ST and NN-MT2 model and 36 for the NN-MT2 model to resize the output layer.

SHAP

The Shapley's cooperative game theory-based SHAP (SHapley Additive exPlanations)[22,23] analysis is a unified framework for interpreting predictions of ML models by assigning impact values to input features. To establish the interpretability of fingerprint components and polymer properties in our work, we intially compute SHAP values for the prediction on the validation dataset using the best NN-MT1 model. Since these raw SHAP values (θ_S) indicate a fingerprint's ability to amend certain polymer properties, mean sums of absolute SHAP values $|\theta_S|$ may be used to measure the total fingerprint component impact on each property. However, $|\theta_S|$ does not measure the proportionality of fingerprint and property, that is to say, the positive or negative change of the property owing to the fingerprint. This is why we compute the PCCs of SHAP and fingerprint components, $\rho_{\theta_S,\mathbf{F}}$, and multiply these PCC with the mean sum of the absolute SHAP values, finally leading to our definition of the fingerprint impact values as $|\theta_S| \cdot \frac{\rho_{\theta_S,\mathbf{F}}}{|\rho_{\theta_S,\mathbf{F}}|}$. SHAP values were computed using the GradientExplainer class of the SHAP Python package (https://github.com/slundberg/shap).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100238.

ACKNOWLEDGMENTS

C. Kuenneth thanks the Alexander von Humboldt Foundation for financial support. This work is financially supported by the Office of Naval Research through a Multi-University Research Initiative (MURI) grant (N00014-17-1-2656) and a regular grant (N00014-20-2175).

AUTHOR CONTRIBUTIONS

C. Kuenneth designed, trained, and evaluated the machine learning models and wrote this paper. L.C., H.T., A.C.R., and C. Kim collected and curated the polymer property database. The work was conceived and guided by R.R. All authors discussed results and commented on the manuscript.

DECLARATION OF INTERESTS

R.R. is a founder of Matmerize, a company that intends to provide materials informatics services. The following patent has been filed: Systems and methods for prediction of polymer properties, Rampi Ramprasad, Anand Chandrasekaran, Chiho Kim (PCT/US2020/028449).

Received: January 7, 2021
Revised: February 12, 2021
Accepted: March 16, 2021
Published: April 9, 2021

REFERENCES

1. Batra, R., Song, L., and Ramprasad, R. (2020). Emerging materials intelligence ecosystems propelled by machine learning. Nat. Rev. Mater. https://doi.org/10.1038/s41578-020-00255-y.

2. Doan Tran, H., Kim, C., Chen, L., Chandrasekaran, A., Batra, R., Venkatram, S., Kamal, D., Lightstone, J.P., Gurnani, R., Shetty, P., et al. (2020). Machine-learning predictions of polymer properties with Polymer Genome. J. Appl. Phys. *128*, 171104, https://doi.org/10.1063/5.0023759.

3. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. npj Computational Materials *3*, https://doi.org/10.1038/s41524-017-0056-5.

4. Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., and Ramprasad, R. (2018). Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. J. Phys. Chem. C *122*, 17575–17585, https://doi.org/10.1021/acs.jpcc.8b02913.

5. Pilania, G., Iverson, C.N., Lookman, T., and Marrone, B.L. (2019). Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. J. Chem. Inf. Model. *59*, 5013–5025, https://doi.org/10.1021/acs.jcim.9b00807.

6. Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively Multitask Networks for Drug Discovery (ICML).

7. Ramsundar, B., Liu, B., Wu, Z., Verras, A., Tudor, M., Sheridan, R.P., and Pande, V. (2017). Is Multitask Deep Learning Practical for Pharma? J. Chem. Inf. Model. *57*, 2068–2076, https://doi.org/10.1021/acs.jcim.7b00146.

8. Wenzel, J., Matter, H., and Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. J. Chem. Inf. Model. *59*, 1253–1268, https://doi.org/10.1021/acs.jcim.8b00785.

9. Ma, R., Liu, Z., Zhang, Q., Liu, Z., and Luo, T. (2019). Evaluating Polymer Representations via Quantifying Structure-Property Relationships. J. Chem. Inf. Model. *59*, 3110–3119, https://doi.org/10.1021/acs.jcim.9b00358.

10. Pilania, G., Gubernatis, J.E., and Lookman, T. (2017). Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput. Mater. Sci. *129*, 156–163, https://doi.org/10.1016/j.commatsci.2016.12.004.

11. Batra, R., Pilania, G., Uberuaga, B.P., and Ramprasad, R. (2019). Multifidelity Information Fusion with Machine Learning: A Case Study of Dopant Formation Energies in Hafnia. ACS Appl. Mater. Interfaces *11*, 24906–24918, https://doi.org/10.1021/acsami.9b02174.

12. Patra, A., Batra, R., Chandrasekaran, A., Kim, C., Huan, T.D., and Ramprasad, R. (2020). A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. Comput. Mater. Sci. *172*, 109286, https://doi.org/10.1016/j.commatsci.2019.109286.

13. Zhu, G., Kim, C., Chandrasekarn, A., Everett, J.D., Ramprasad, R., and Lively, R.P. (2020). Polymer genome-based prediction of gas permeabilities in polymers. Journal of Polymer Engineering *40*, 451–457, https://doi.org/10.1515/polyeng-2019-0329.

14. Venkatram, S., Kim, C., Chandrasekaran, A., and Ramprasad, R. (2019). Critical Assessment of the Hildebrand and Hansen Solubility Parameters for Polymers. J. Chem. Inf. Model. *59*, 4188–4194, https://doi.org/10.1021/acs.jcim.9b00656.

15. Jha, A., Chandrasekaran, A., Kim, C., and Ramprasad, R. (2019). Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures. Model. Simul. Mater. Sci. Eng. *27*, 24002, https://doi.org/10.1088/1361-651X/aaf8ca.

16. Kim, C., Chandrasekaran, A., Jha, A., and Ramprasad, R. (2019). Active-learning and materials design: The example of high glass transition temperature polymers. MRS Commun. *9*, 860–866, https://doi.org/10.1557/mrc.2019.78.

17. Chen, L., Tran, H., Batra, R., Kim, C., and Ramprasad, R. (2019). Machine learning models for the lattice thermal conductivity prediction of inorganic materials. Comput. Mater. Sci. *170*, 109155, https://doi.org/10.1016/j.commatsci.2019.109155.

18. Chen, L., Kim, C., Batra, R., Lightstone, J.P., Wu, C., Li, Z., Deshmukh, A.A., Wang, Y., Tran, H.D., Vashishta, P., Sotzing, G.A., Cao, Y., and Ramprasad, R. (2020). Frequency-dependent dielectric constant prediction of polymers using machine learning. npj Computational Materials *6*, https://doi.org/10.1038/s41524-020-0333-6.

19. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T.D., Lookman, T., and Ramprasad, R. (2016). Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. Sci. Rep. *6*, 20952, https://doi.org/10.1038/srep20952.

20. Huan, T.D., Mannodi-Kanakkithodi, A., and Ramprasad, R. (2015). Accelerated materials property predictions and design using motif-based fingerprints. Phys. Rev. B *92*, 1–10, https://doi.org/10.1103/PhysRevB.92.014106.

21. Csáji, B. (2001). Approximation with artificial neural networks. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.2647&rep=rep1&type=pdf.

22. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems https://arxiv.org/pdf/1705.07874.pdf.

23. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning, *70*, pp. 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html.

24. van Krevelen, D.W., and te Nijenhuis, K. (2009). Properties of Polymers: Their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions (Elsevier Science).

25. Huan, T.D., Mannodi-Kanakkithodi, A., Kim, C., Sharma, V., Pilania, G., and Ramprasad, R. (2016). A polymer dataset for accelerated property prediction and design. Sci. Data *3*, 160012, https://doi.org/10.1038/sdata.2016.12.

26. Sharma, V., Wang, C., Lorenzini, R.G., Ma, R., Zhu, Q., Sinkovits, D.W., Pilania, G., Oganov, A.R., Kumar, S., Sotzing, G.A., et al. (2014). Rational design of all organic polymer dielectrics. Nat. Commun. *5*, 4845, https://doi.org/10.1038/ncomms5845.

27. Wiley. (1999). Polymer Handbook, 2 Volumes Set, Fourth Edition, J. Bandrup, E.H. Immergut, and E.A. Grulke, eds. (John Wiley & Sons).

28. Barton, A.F.M. (1991). CRC handbook of solubility parameters and other cohesion parameters (CRC Press).

29. Bicerano, J. (2002). Prediction of polymer properties (CRC Press). https://www.routledge.com/Prediction-of-Polymer-Properties/Bicerano/p/book/9780824708214.

30. Crow Polymer Properties Database. http://polymerdatabase.com/.

31. PolyInfo. https://polymer.nims.go.jp/en/.

32. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., and Mueller, A. (2015). Scikit-learn: Machine Learning Without Learning the Machinery. GetMobile: Mobile Computing and Communications *19*, 29–33, https://doi.org/10.1145/2786984.2786995.

33. Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. *28*, 31–36, https://doi.org/10.1021/ci00057a005.

34. Mannodi-Kanakkithodi, A., Huan, T.D., and Ramprasad, R. (2017). Mining Materials Design Rules from Data: The Example of Polymer Dielectrics. Chem. Mater. *29*, 9001–9010, https://doi.org/10.1021/acs.chemmater.7b02027.

35. Le, T., Epa, V.C., Burden, F.R., and Winkler, D.A. (2012). Quantitative structure-property relationship modeling of diverse materials properties. Chem. Rev. *112*, 2889–2919, https://doi.org/10.1021/cr200066h.

36. Landrum, G.. RDKit. http://www.rdkit.org.

37. Iler, N., Rowitch, D.H., Echelard, Y., McMahon, A.P., and Abate-Shen, C. (1995). A single homeodomain binding site restricts spatial expression of Wnt-1 in the developing brain. Mech. Dev. *53*, 87–96, https://doi.org/10.1016/0925-4773(95)00427-0.

38. Ertl, P., Rohde, B., and Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J. Med. Chem. *43*, 3714–3717, https://doi.org/10.1021/jm000942e.

39. Prasanna, S., and Doerksen, R.J. (2009). Topological polar surface area: a useful descriptor in 2D-QSAR. Curr. Med. Chem. *16*, 21–41, https://doi.org/10.2174/092986709787002817.

40. Martin, A., Ashish, A., Paul, B., Eugene, B., Zhifeng, C., Craig, C., Greg, S.C., Andy, D., Jeffrey, D., Matthieu, D., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation https://www.tensorflow.org/.

41. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. J. Mach. Learn. Res. *18*, 1–52.