Modelling and Simulation in Materials Science and Engineering

PAPER

Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures

To cite this article: Anurag Jha *et al* 2019 *Modelling Simul. Mater. Sci. Eng.* **27** 024002

IOP Publishing Modelling and Simulation in Materials Science and Engineering
Modelling Simul. Mater. Sci. Eng. **27** (2019) 024002 (9pp) https://doi.org/10.1088/1361-651X/aaf8ca

Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures

Anurag Jha, Anand Chandrasekaran, Chiho Kim and Rampi Ramprasad

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, United States of America

E-mail: rampi.ramprasad@mse.gatech.edu

Received 20 October 2018, revised 9 December 2018
Accepted for publication 17 December 2018
Published 17 January 2019

Abstract

Over the past decade, there has been a resurgence in the importance of data-driven techniques in materials science and engineering. The utilization of state-of-the art algorithms, coupled with the increased availability of experimental and computational data, has led to the development of surrogate models offering the promise of rapid and accurate predictions of materials' properties based solely on their structure or composition. Such machine learning (ML) models are trained on available past data and are thus susceptible to the intrinsic uncertainties/errors associate with these past measurements. The glass transition temperature (T_g) of polymers, a property of paramount interest in polymer science, is one strong example of a material property that can show widespread variation in the final reported value as a result of a variety of intrinsic and extrinsic factors that occur during the experimental measurement process. In the current work, we curate a large database of T_g measurements from a variety of data sources and proceed to investigate the statistical nature of the inherent uncertainties in the database. Through the partitioning of the dataset using statistically relevant measures, we investigate the effect of variations in the dataset on the performance of the final ML model. We demonstrate that the measure of central tendency, median is a valid approximation when dealing with multiple reported values for T_g when dealing with multiple reported values of T_g for the same polymeric material. Moreover, the Bayesian model noise/uncertainty that emerges from our machine-learning pipeline is able to represent quantitatively the underlying noise/uncertainties in the experimental measurement of T_g.

Keywords: machine learning, polymers, glass transition temperature

(Some figures may appear in colour only in the online journal)

1. Introduction

Data-driven methods are seeing a revival within materials science [1–6], and are deeply influencing multiple aspects of materials research [7–13]. Materials property data, obtained either from computational or experimental efforts, are being utilized to create surrogate models using machine learning (ML) techniques [1, 14–21]. These models can be utilized to provide rapid predictions of the properties of new materials at a minuscule fraction of the cost involved in actual experimentation or computation [7, 22–26]. Moreover, a variety of techniques are being explored to invert the property prediction pipeline so as to allow for designing materials that display a desired target set of property values [4–6, 17, 27–32].

The quality of the developed surrogate model, though, depends on the quality (and quantity) of the dataset used in the model training step. Often, different experimental studies may report different values for the same property of the same material. This may be because of variations in measurement techniques, measurement conditions, and sample quality among others. How should one treat such uncertainties in data during surrogate model development? And, what is the impact of such uncertainties on the surrogate model performance?

The present contribution attempts to address the above questions for the specific case of the glass transition temperature (T_g) of polymeric materials. T_g is a complex property, the measurement of which is affected by a variety of intrinsic and extrinsic factors. The intrinisic factors include material properties such as morphology, crystallinity, tacticity, cross-linking, molecular weight and density. Extrinsic factors such as the method of measurement and the rate of cooling/heating also significantly impact the reported T_g value. Above and beyond such measurement uncertainties, the surrogate modeling pipeline (illustrated in figure 1) also entails the introduction of additional uncertainties at various stages of the process [33]. For example, the fingerprint stage necessitates the numerical representation of the polymer repeat unit in terms of a fixed dimensional vector [19] and this representation may result in a loss of chemical information. Thus, the final prediction obtained from the ML model naturally contains contributions from the errors propagated from the dataset and fingerprinting step. In the present article, we primarily address the role of the initial dataset uncertainties in the final model performance.

This paper is organized as follows. In section 2, we describe in detail the expansive dataset accumulated for T_g of polymers, highlighting specifically the statistical nature of the uncertainties in the T_g measurements. In section 3, we discuss the fingerprinting and ML techniques [19] that we employ to develop surrogate models for the prediction of T_g. In section 4, the results are summarized and the sensitivity of the model performance as a result of statistical variations in the dataset is demonstrated.

2. Dataset

2.1. Sources of data

Data for this work was gathered from two books: the *Polymer Handbook* [34] and *Prediction of Polymer Properties* [35] and also from an online repository of polymer properties [36]. From these publicly available data sources, we extracted the polymer name, SMILES string [37, 38] and T_g measurements of 751 polymers. A subset of this dataset has been utilized in an earlier work [19].

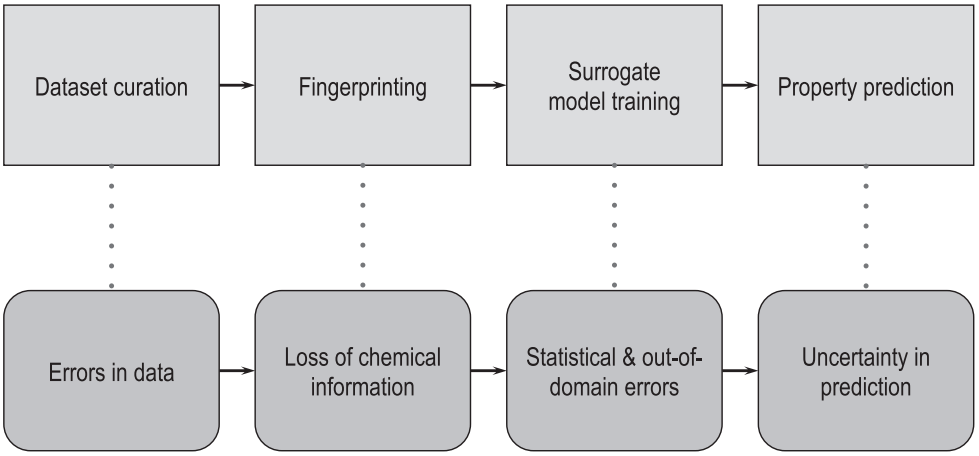

Figure 1. Workflow of data-driven surrogate modeling and property prediction with error sources associated with each step.

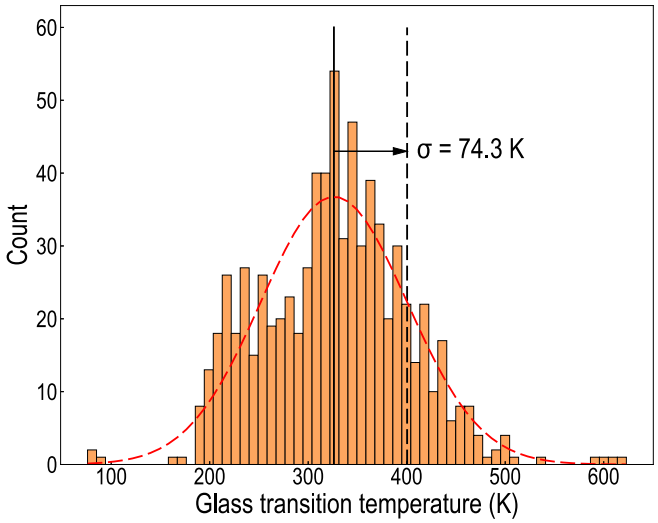

Figure 2. Distribution of the T_g values for all polymers.

The chemical space that comprises all these three datasets consists of nine elements (C, H, O, N, S, F, Cl, Br and I). We only considered T_g values reported for neat polymers (not composites) without any additives. However, we did not differentiate between tacticity, isomerism, crystallinity, different measurement techniques or different heating/cooling rates. The type of polymers that were considered as part of this dataset include main chain acyclic polymers such as poly(acrylics) and poly(methacrylics), poly(alkenes), poly(dienes), poly (styrenes), poly(vinyl alchohols) and poly(ketones), poly(vinyl esters), poly(vinyl ethers) and poly(thioethers) and some others, as well as main chain carbocylic polymers such as poly (phenylenes) and others. As can be seen from figure 2, there is a wide distribution of T_g for polymers which ranges from 76 K to 773 K. The frequency plot looks to be roughly normal with a standard deviation of 74 K with a mean of 326.1 K. The glass transition temperature is

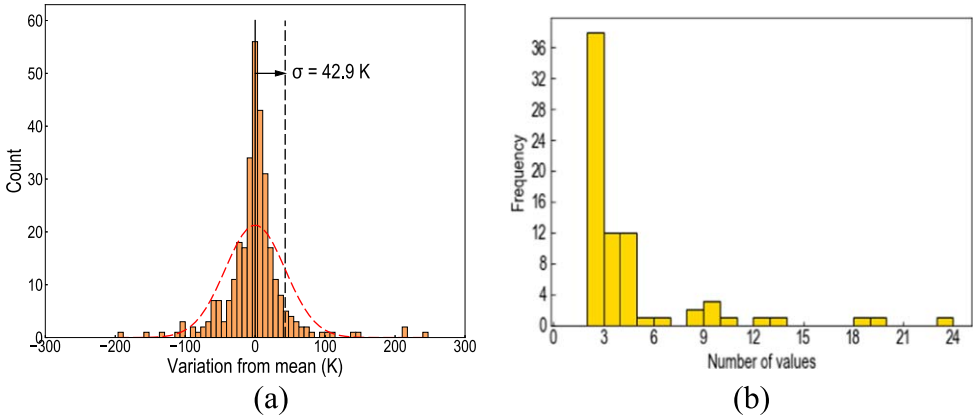

Figure 3. (a) Normal distribution fit for variation from mean for the 75 polymers with multiple reported values of T_g. (b) Frequency distribution of T_g values for the polymers with multiple reported values of T_g.

known to be higher for polymers with more polar groups and for those with high van der Waals surface area. Also, it is known to be higher for polymers that have more rigid monomers [39].

2.2. Spread in the data

T_g is a property which cannot be defined in terms of thermodynamic state variables, unlike the melting point of materials. Rather, it is considered to be a phenomenon driven by kinetics [35, 39]. Therefore, its value is highly dependent on the time scale of the measurement. Other factors that contribute to the uncertainty are the history of the sample, percent crystallinity, tacticity, isomerism, solvents and additives used, cross linking and molecular weight of the polymer. There are uncertainties also associated with the choice of the measurement method, the accuracy of the calculations due to the inherent resolution of the method and human error.

Out of the 751 polymers considered in this study, 75 polymers had multiple reported values (as a result of the aforementioned variations). For these 75 polymers, a histogram for deviations from their respective mean measurements is shown in figure 3(a). Among the 75 polymers for which multiple T_g values were reported, most of the polymers possessed two reported values as can be seen in figure 3(b). The maximum number of reported values for a particular polymer was 24 (for the case of the polymer PMMA). It appears that the measured uncertainties (or errors or noise) follow a normal distribution with a standard deviation of 40 K with peak at 0 (representing the mean value). This behavior sets this problem up as an ideal candidate for checking if a model constructed with mean measurements would yield good results. The above analysis also implies that T_g measurements have an intrinsic typical uncertainty of about 40 K.

3. Method

3.1. Fingerprinting and ML pipeline

Both the fingerprinting and ML model were described in detail in a recent work [19]. The fingerprint building process consists of three hierarchical levels of descriptors [40, 41]. The

first one is at the atomic scale wherein the number fraction of atomic triples (or a set of three contiguous atoms) was calculated. The next level deals with quantitative structure property relationship (QSPR) descriptors, such as van der Waals surface area [42], topological surface area [43, 44] and fraction of rotatable bonds [19], implemented in the RDkit cheminformatics library [45–47]. The third level and largest length scale descriptors captured morphological feautures such as the topological distance between rings, fraction of atoms that are part of side chains and length of largest side chain [19].

After the three levels of fingerprints are generated for all the polymers, we employed a recursive feature elimination technique [19] to eliminate unwanted descriptors. This feature elimination process aided in reducing over-fitting and improved the final model performance. The final model was obtained by mapping the filtered descriptors to the T_g values using Gaussian process regression (GPR) with a sum-kernel consisting of a radial basis function kernel and a white-noise kernel. The white-noise kernel provides us with the so-called noise parameter which gives us an estimate of the intrinsic noise or uncertainty in the output value (in our case, the T_g measurement).

The advantage of using GPR over other kernel methods (like kernel ridge regression) is that one automatically obtains both the target value (T_g in our case) and the associated uncertainty of the prediction [48]. For example, when a prediction is made for a polymer which is chemically very different from the polymer in the training set, we automatically observe a large uncertainty associated with the prediction.

3.2. Effects of variation in the dataset

As mentioned in the earlier section, a wide variation exists in the reported values of T_g for 75 polymers and therefore it is difficult to decide what is the most appropriate T_g value to be used in the input dataset while building the model. To understand the effect of these variations, we first calculate the median, mean, minimum and maximum of the T_g measurements of the 75 polymers with multiple reported values. These values were then used to create four training datasets (Datasets 1, 2, 3 and 4 respectively), thus resulting in four different models (Models 1, 2, 3 and 4 respectively), as shown in figure 4. As described in the next section, the predictions of the resulting 4 models were used to assess which strategy is most suitable when we are confronted with dataset uncertainty.

4. Results

The performance of the above four models were then evaluated based on four metrics: 1-R^2, root mean squared error, average GPR-uncertainty and GPR noise-hyperparameter. As depicted in figure 5, the model build using the median value dataset shows the best performance (lower errors). On the other hand, the dataset containing the maximum T_g values systematically exhibited the poorest performance with respect to all four error metrics. The results of the thus obtained best and worst model (i.e. the median-model and maximum-model) results are shown in more detail through the parity-plots depicted in figure 6. The model constructed from the maximum of the measurements not only shows poorer performance (in terms of prediction accuracy) but the Bayesian model uncertainty (i.e. the noise hyperparameters) for every prediction, depicted by the gray bars in figure 6, is also significantly larger. It is also worth noting that the GPR noise hyperparameter, determined here independently during the training process, is of the same magnitude as the dataset uncertainty of about 40 K noted earlier.

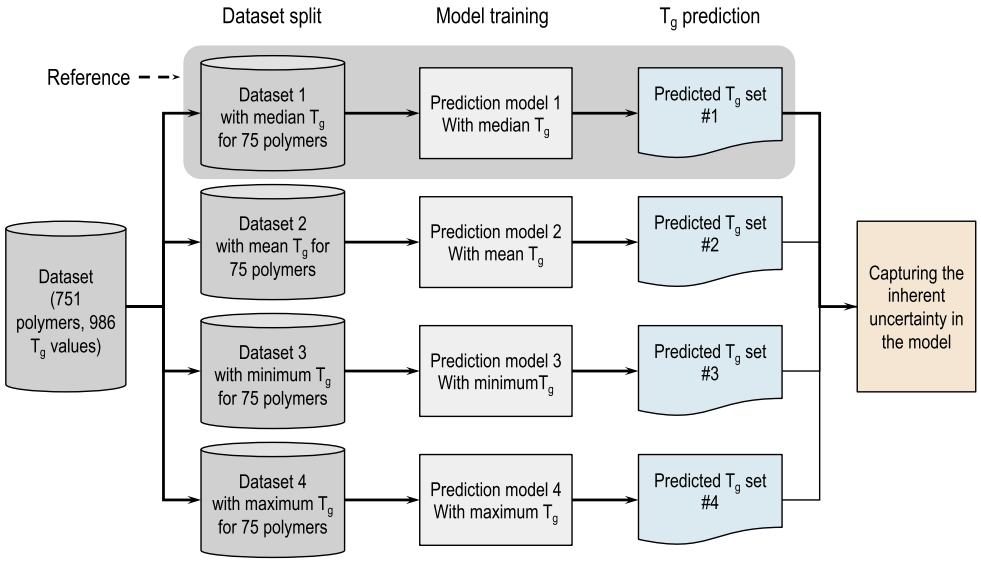

Figure 4. Workflow for building the four different datasets and models for capturing the effect of dataset variation on model performance.

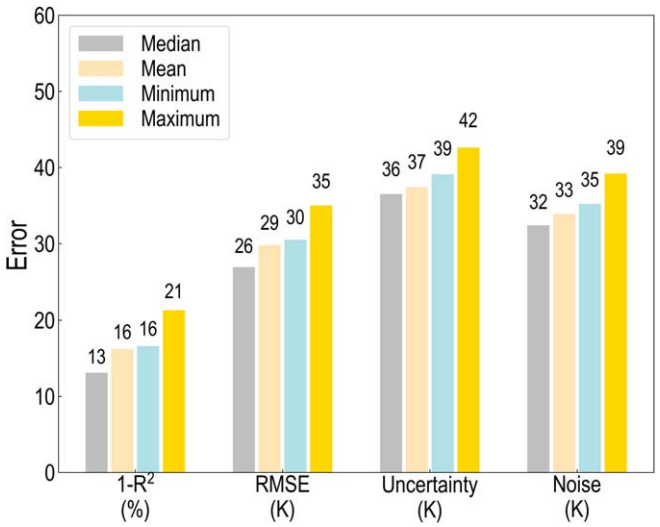

Figure 5. Comparison between the four models built from four different datasets using metrics 1-R^2, RMSE, average GPR uncertainty and GPR noise-parameters.

These results indicate that using the median of a distribution of property values would be the best approach to tackle uncertainty in data sources. The calculated mean and median is found to be very close for most polymers as can be seen from figure 3(a) which shows the symmetric nature of the deviations from mean T_g values. The median-model, however, is less sensitive to extremely low and high T_g values and this could be the likely cause of the median-model outperforming the mean-model.

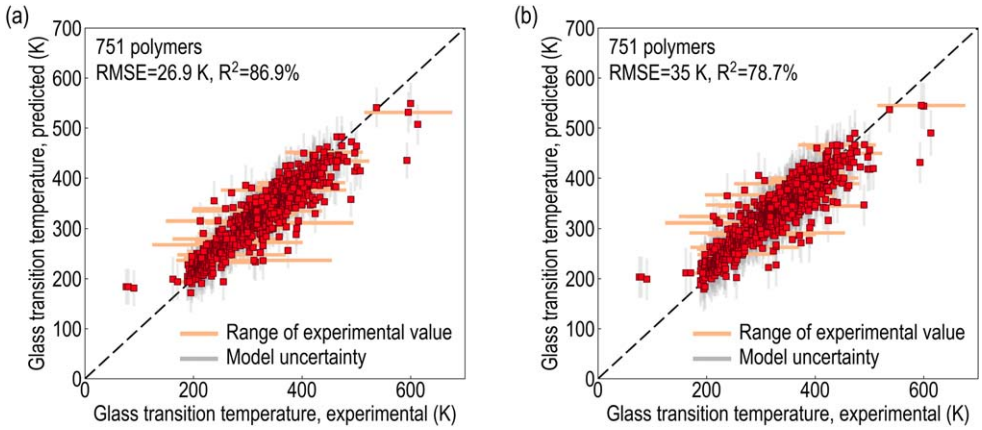

Figure 6. Parity plot for experimental T_g measurement versus GPR predicted T_g values for models constructed from (a) median (b) maximum value data sets.

Across the four models, we observe that the 1-R^2 ranges from 13% to 21% indicating that variations in small portions of the dataset can have a significant impact on the performance of the model.

5. Conclusions and future work

Measurement uncertainty in data is common. In the current work, we have investigated which value from a range of reported values would be ideal for model development for the example of the glass transition temperature (T_g) of polymers. Using different statistical measures, we have constructed four different datasets and evaluated the performance of four models trained on these corresponding datasets. The model constructed with the median of the experimental measurements was found to outperform other models. Also, there was a large change in model accuracy for changes in the data and therefore it can be said that the model is highly dependent on the input dataset.

As seen from figure 5, both the average uncertainty and the noise parameter of the models approximately range between 30 and 40 K. The standard deviation of the T_g measurements of the 75 polymers is, as mentioned earlier, about 40 K. The similarity of these two values likely indicates the ability of our surrogate modeling pipeline (hierarchical fingerprinting + GPR) to capture the intrinsic noise in T_g measurement across the dataset.

Going forward, we intend to continue to expand the list of polymers in our T_g dataset to not only improve the model performance but also to obtain a better understanding of the statistical nature of T_g measurements. Moreover, the development of additional descriptors (capturing, for example, tacticity and isomerism) would lead to better performing models and also simultaneously reduce uncertainty propagation in our model development pipeline.

Acknowledgments

The authors acknowledge support of this work by the Toyota Research Institute through the Accelerated Materials Design and Discovery program.

ORCID iDs

Anand Chandrasekaran https://orcid.org/0000-0002-2794-3717

References

[1] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 Machine learning in materials informatics: recent applications and prospects *npj Comput. Mater.* **3** 1–13
[2] Kim C, Huan T D, Krishnan S and Ramprasad R 2017 A hybrid organic–inorganic perovskite dataset *Sci. Data* **4** 170057
[3] Kim S *et al* 2015 Pubchem substance and compound databases *Nucleic Acids Res.* **44** D1202–13
[4] Zhu Q, Sharma V, Oganov A R and Ramprasad R 2014 Predicting polymeric crystal structures by evolutionary algorithms *J. Chem. Phys.* **141** 154102
[5] Lorenzini R G, Kline W M, Wang C C, Ramprasad R and Sotzing G A 2013 The rational design of polyurea & polyurethane dielectric materials *Polymer* **54** 3529–33
[6] Mannodi-Kanakkithodi A, Pilania G, Ramprasad R, Lookman T and Gubernatis J E 2016 Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers *Comput. Mater. Sci.* **125** 92–9
[7] Mueller T, Kusne A G and Ramprasad R 2016 Machine learning in materials science: recent progress and emerging applications *Reviews in Computational Chemistry* vol 29 (New York: Wiley) ch 4 pp 186–273
[8] Hattrick-Simpers J, Wen C and Lauterbach J 2015 The materials super highway: integrating high-throughput experimentation into mapping the catalysis materials genome *Catal. Lett.* **145** 290–8
[9] Hill J, Mannodi-Kanakkithodi A, Ramprasad R and Meredig B 2018 Materials data infrastructure and materials informatics *Computational Materials System Design* (Berlin: Springer) pp 193–225
[10] Mannodi-Kanakkithodi A, Huan T D and Ramprasad R 2017 Mining materials design rules from data: the example of polymer dielectrics *Chem. Mater.* **29** 9001–10
[11] Huan T D, Batra R, Chapman J, Krishnan S, Chen L and Ramprasad R 2017 A universal strategy for the creation of machine learning-based atomistic force fields *npj Comput. Mater.* **3** 1–8
[12] Huan T D, Batra R, Chapman J, Krishnan S, Chen L and Ramprasad R 2017 A universal strategy for the creation of machine learning-based atomistic force fields *npj Comput. Mater.* **3** 27
[13] Botu V, Batra R, Chapman J and Ramprasad R 2017 Machine learning force fields: construction, validation, and outlook *J. Phys. Chem.* C **121** 511
[14] Mannodi-Kanakkithodi A, Chandrasekaran A, Kim C, Huan T D, Pilania G, Botu V and Ramprasad R 2018 Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond *Mater. Today* **21** 785–96
[15] Mannodi-Kanakkithodi A, Pilania G and Ramprasad R 2016 Critical assessment of regression-based machine learning methods for polymer dielectrics *Comput. Mater. Sci.* **125** 123–35
[16] Hautier G, Fischer C C, Jain A, Mueller T and Ceder G 2010 Finding nature's missing ternary oxide compounds using machine learning and density functional theory *Chem. Mater.* **22** 3762–7
[17] 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
[18] Botu V, Batra R, Chapman J and Ramprasad R 2016 Machine learning force fields: construction, validation, and outlook *J. Phys. Chem.* C **121** 511–22
[19] 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** 17575–85
[20] Yu X 2010 Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers *Fibers Polym.* **11** 757–66
[21] Mattioni B E and Jurs P C 2002 Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks *J. Chem. Inf. Comput. Sci.* **42** 232–40
[22] Liu Y, Zhao T, Ju W and Shi S 2017 Materials discovery and design using machine learning *J. Materiomics* **3** 159–77
[23] Olson G B 2000 Designing a new material world *Science* **288** 993–8
[24] Materials Genome Initiative https://mgi.gov/

[25] Kim C, Pilania G and Ramprasad R 2016 From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown *Chem. Mater.* **28** 1304–11
[26] Jain A, Shin Y and Persson K A 2016 Computational predictions of energy materials using density functional theory *Nat. Rev. Mater.* **1** 15004
[27] Patra T K, Meenakshisundaram V, Hung J-H and Simmons D S 2017 Neural-network-biased genetic algorithms for materials design: evolutionary algorithms that learn *ACS Comb. Sci.* **19** 96–107
[28] Huan T D, Mannodi-Kanakkithodi A and Ramprasad R 2015 Accelerated materials property predictions and design using motif-based fingerprints *Phys. Rev.* B **92** 014106
[29] Sanchez-Lengeling B and Aspuru-Guzik A 2018 Inverse molecular design using machine learning: generative models for matter engineering *Science* **361** 360–5
[30] Mannodi-Kanakkithodi A, Treich G M, Huan T D, Ma R, Tefferi M, Cao Y, Sotzing G A and Ramprasad R 2016 Rational co-design of polymer dielectrics for energy storage *Adv. Mater.* **28** 6277–91
[31] Treich G M, Tefferi M, Nasreen S, Mannodi-Kanakkithodi A, Li Z, Ramprasad R, Sotzing G A and Cao Y 2017 A rational co-design approach to the creation of new dielectric polymers with high energy density *IEEE Trans. Dielectr. Electr. Insul.* **24** 732–43
[32] Baldwin A F, Huan T D, Ma R, Mannodi-Kanakkithodi A, Tefferi M, Katz N, Cao Y, Ramprasad R and Sotzing G A 2015 Rational design of organotin polyesters *Macromolecules* **48** 2422–8
[33] Smith R C *Uncertainty Quantification: Theory, Implementation, and Applications* vol 12 (Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM))
[34] Brandrup J, Immergut E H, Grulke E A, Abe A and Bloch D R 1999 *Polymer Handbook* vol 2 (New York: Wiley)
[35] Bicerano J 2002 *Prediction of Polymer Properties* (Boca Raton, FL: CRC Press)
[36] Crow Polymer Database http://polymerdatabase.com/index.html (Accessed: 1 February 2018)
[37] Weininger D 1988 SMILES, a chemical language and information system: I. Introduction to methodology and encoding rules *J. Chem. Inf. Comput. Sci.* **28** 31–6
[38] Lowe D M, Corbett P T, Murray-Rust P and Glen R C 2011 Chemical name to structure: OPSIN, an open source solution *J. Chem. Inf. Modeling* **51** 739–53
[39] Painter P C and Coleman M M 2008 *Essentials of Polymer Science and Engineering* (Lancaster, PA: DEStech Publications, Inc.)
[40] Pankajakshan P, Sanyal S, de Noord O E, Bhattacharya I, Bhattacharyya A and Waghmare U 2017 Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights *Chem. Mater.* **29** 4190–201
[41] Baldwin A F, Ma R, Huan T D, Cao Y, Ramprasad R and Sotzing G A 2014 Effect of incorporating aromatic and chiral groups on the dielectric properties of poly (dimethyltin esters) *Macromol. Rapid Commun.* **35** 2082–8
[42] Labute P 2000 A widely applicable set of descriptors *J. Mol. Graph. Modelling* **18** 464–77
[43] 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** 3714–7
[44] Prasanna S and Doerksen R J 2009 Topological polar surface area: a useful descriptor in 2D-QSAR *Curr. Med. Chem.* **16** 21–41
[45] Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T and Prachayasittikul V 2009 A practical overview of quantitative structure-activity relationship *EXCLI J.* **8** 74–88
[46] Nantasenamat C, Isarankura-Na-Ayudhya C and Prachayasittikul V 2010 Advances in computational methods to predict the biological activity of compounds *Expert Opin. Drug Discovery* **5** 633–54
[47] Rdkit, open source toolkit for cheminformatics http://rdkit.org/ (Accessed: 10 February 2018)
[48] Williams C K and Rasmussen C E 2006 *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press)