A Random Forest Predictor for Diblock Copolymer Phase Behavior

Akash Arora, Tzyy-Shyang Lin, Nathan J. Rebello, Sarah Av-Ron, Hidenobu Mochigase, and Bradley D. Olsen*

Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

ABSTRACT: Physics-based models are the primary approach for modeling the phase behavior of block copolymers. However, the successful use of self-consistent field theory (SCFT) for designing new materials relies on the correct chemistry- and temperature-dependent Flory-Huggins interaction parameter χ_{AB} that quantifies the incompatibility between the two blocks A and B, as well as accurate estimation of the ratio of Kuhn lengths (b_A/b_B) and block densities. This work uses machine-learning to model the phase behavior of AB diblock copolymers by using the chemical identities of blocks directly, obviating the need for measurement of χ_{AB} and b_A/b_B. The random forest approach employed predicts the phase behavior with almost 90% accuracy after training on a data set of 4,768 data points, almost twice the accuracy obtained using SCFT employing χ_{AB} from group contribution theory. The machine-learning model is notably sensitive towards the uncertainty in measuring molecular parameters; however, its accuracy still remains at least 60% even for highly uncertain experimental measurements. Accuracy is substantially reduced when extrapolating to chemistries outside the training set. This work demonstrates that a random forest phase predictor performs remarkably well in many scenarios, providing an opportunity to predict self-assembly without measurement of molecular parameters.

KEYWORDS: Block Copolymers, Machine Learning, Phase Behavior, Self-assembly, SCFT

Block copolymers self-assemble to form a variety of nanostructures useful in applications as diverse as advanced plastics,[1] drug delivery,[2] ultra-filtration membranes,[3] and photonic crystals.[4] The simplest block copolymers, AB diblocks, serve as excellent models to understand the physics of self-assembly; therefore, they have garnered exceptional interest.[5,6] The phase behavior of AB diblocks depends on five factors: the chemical identities of the A and B blocks, temperature (T), volume fraction of the A block (f_A), and the overall molar mass of the polymer (M_n), with pressure generally having a negligible effect.[7–9] Identifying the combination of block chemistries, molecular parameters (f_A and M_n), and temperature that produces a desired morphology for a required application is central to diblock copolymer materials design.

Traditionally, self-consistent field theory (SCFT) is used as the primary approach for modeling the phase behavior of block copolymers, and it has enjoyed remarkable success in discovering new structures and understanding the origins of experimentally-observed morphologies.[10–17] SCFT is a coarse-grained theory in which the phase behavior of AB diblocks is described by only three parameters: f_A, segregation strength $\chi_{AB}N$, and conformational asymmetry b_A/b_B, where b_A and b_B are the Kuhn lengths of the A and B polymer blocks, respectively. The quantity N is the total number of Kuhn segments in the block copolymer chain, and χ_{AB} is the Flory-Huggins interaction parameter. The most common functional form used in literature is $\chi_{AB} = \alpha/T + \beta$, where α and β are constants specific to the pair of polymers (A, B).[18] Since χ_{AB} is a coarse-grained parameter accounting for all types of interactions between A and B blocks, it is difficult to define a general functional form that can capture the phase behavior of all polymer pairs.[18–21] Even for the most studied diblock copolymer in literature, poly(isoprene)-*b*-poly(styrene) (IS), estimates of χ_{IS} can vary up to 200% due to experimental uncertainty or variation in methods.[22] In addition to χ_{AB}, reasonably accurate estimations of the conformational asymmetry (b_A/b_B) and block densities (ρ_A and ρ_B) that are used to calculate the volume fraction f_A are also required. Even in the best-case scenario with all parameters

known, SCFT cannot yet capture effects such as hydrogen bonding, local liquid structure,[23] and monomer shapes[24,25] that are prevalent below the size of a Kuhn segment.

As an alternative to physics-based models, data-driven techniques can model phase behavior by directly using the blocks' chemical identities, obviating the need for χ_{AB}. Historically, such approaches were used to predict a broad range of polymer properties,[26–29] as well as for modeling quantitative structure property relationships (QSPR) for small molecules.[30,31] Owing to the development of new, efficient algorithms, machine-learning methods have lately enjoyed a renaissance[32–35] and have been used to discover new materials such as advanced polymer dielectrics,[36] high thermal-conductivity polymers,[37] and exceptional gas-separation polymer membranes.[38] Recently, such methods are also used to enhance the efficiency of SCFT for constructing block copolymer phase diagrams using data from simulations.[39,40] Here, a data-driven approach is taken to model the phase behavior of neat diblock copolymers directly from an experimental dataset, benchmarking the performance of the developed machine-learning model with SCFT using χ_{AB} from solubility-parameters estimates.

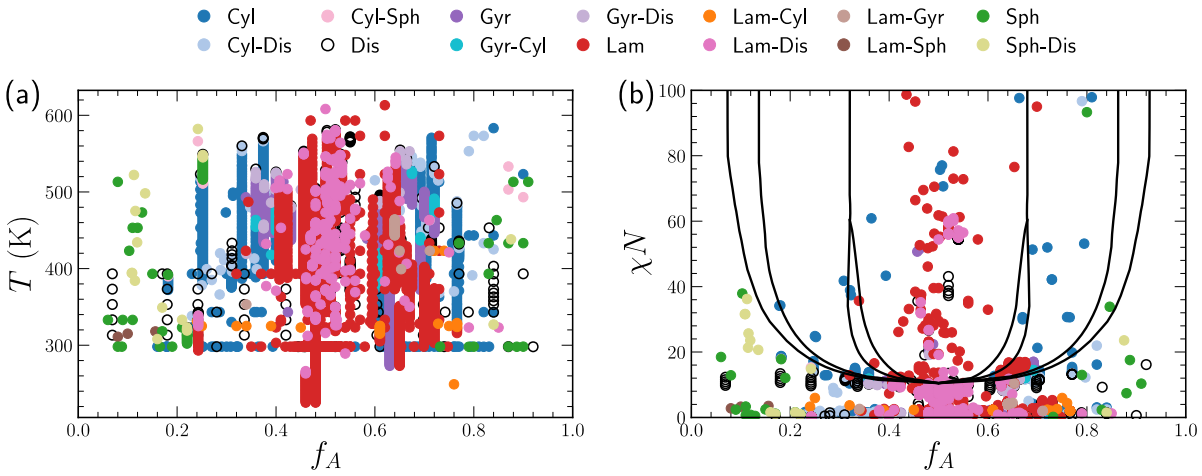

Figure 1: (a) Temperature-volume fraction ($T - f_A$) , and (b) $\chi_{AB}N - f_A$ phase diagram of diblock copolymers obtained from the data gathered by mining experimental studies in literature.[41] The solid lines in (b) denotes the SCFT -predicted phase boundaries for conformationally symmetric diblock copolymers.

A key ingredient for using a machine-learning approach is the availability of a sufficiently large and diverse set of data. An extensive set of data containing all the required information: T, f_A, M_n, blocks' chemistries, and the observed phase, has been gathered by mining the experimental literature on diblock copolymer melts.[41] Figure 1 (a) shows the T vs. f_A phase diagram of the assembled data, while Figure 1(b) shows the $\chi_{AB}N$ vs. f_A phase diagram with χ_{AB} obtained using group contribution theory (see SI for details). Overall, the dataset contains 4,768 measurements, and a detailed description of the data collection protocols, experimental methods, and uncertainty in the measured molecular parameters, is provided in Ref. [41]. Despite many newer structures,[42–44] this work focuses only on the five classical phases: lamellae (Lam), hexagonal-packed cylinders (Cyl), body-centered cubic spheres (Sph), a cubic gyroid phase (Gyr), and a disordered (Dis) phase due to the amount of available data.

To develop a machine-learning model for predicting this self-assembly behavior, chemical fingerprints are first generated for the polymers of interest. The type of fingerprinting method used is crucial for developing a machine-learning model, hence several different ways to transform polymers into fingerprint vectors are considered (see SI for details). For fingerprinting algorithms that are applicable only to finite-sized chemical compounds, ring oligomers with the first and last atoms joined together are used as representative inputs for generating fingerprints. This construct eliminates any endgroup effects and captures the high-molecular-weight nature of polymers. It is found that the number of repeat units within this ring polymer

does not have any significant effect on the quantitative fingerprint (See SI for details), so eight repeat units are used for all the results presented in the work. For each entry in the dataset, the fingerprint vectors for the individual blocks are generated separately and concatenated with each other as well as with the respective values of T, f_A, and M_n to constitute the complete feature set used to describe the block copolymer. For each entry, two data points, corresponding to the case where the two blocks are labelled AB and BA, are included in the training set to guarantee symmetry of block labeling. The conformational symmetry (b_A/b_B) depends upon the chemical groups, hence it is captured by the chemical fingerprint. The calculation of f_A involves density which depends weakly on temperature; this small variation is neglected here.

To develop a model that relates these feature vectors to the observed nanostructure formed via self-assembly, a decision-tree-based algorithm called Random Forest Classifier (RFC)[45,46] was applied. RFC captures the non-linearity of phase space while maintaining a low variance towards the uncertainty in input data, and it naturally monitors the role of different features during training. Therefore, the relative importance of different features can be extracted alongside model development. The model is developed in two stages. First an optimal set of hyperparameters is obtained using a grid-search cross-validation procedure which divides the dataset in test-train-validation sets (see SI for details). These optimal hyperparameters are then used to develop the final RFC model using a 25%-75% test-train split and a 4-fold cross-validation procedure. Because the data is highly imbalanced, cross-validation on test-train sets with stratified splits is used to evaluate model performance *in lieu* of the more commonly used out-of-bag performance. The splits are kept invariant across all the fingerprint types. All preprocessing of the data and model training is performed using the Scikit-learn library in Python[47] (see SI for model setup).

The RFC model is highly effective at predicting the self-assembled phase across the entire block copolymer database. Figure 2(a) shows the results obtained from the RFC model for various types of fingerprinting methods. The model performance is assessed by computing accuracy and $1 -$ (Hamming-loss), both evaluated on a test set that is the same for all the fingerprint types. The accuracy is simply the number of points in the test data that are predicted correctly by the trained model, while $1 -$ (Hamming-loss) takes into account the partial correctness of the prediction for the phase transition points, i.e., whether the model predicted any one of the co-existing phases correctly or not (see SI for details). Figure 2(a) demonstrates that the RFC model can predict the phase behavior with almost 92% accuracy, and if the partial correctness of phase transitions is also considered, the accuracy increases to 97%.

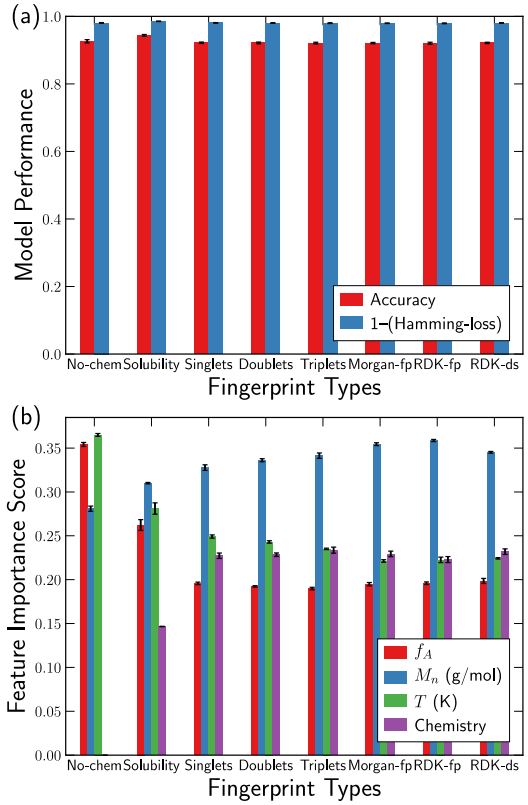

Figure 2: (a) Performance of the RFC model, characterized by accuracy and 1-(Hamming-loss), for various types of fingerprinting methods. (b) Importance score for each feature.

Owing to the binary splitting of nodes, tree-based classification methods such as RFC produce separation boundaries that are extremely sharp compared to the typical phase boundaries. Figure 3 shows the $T - f_A$ phase diagrams at constant values of M_n obtained using the RFC model trained with Morgan fingerprints. Although the phase boundaries are almost vertical, the general shape of the theoretical phase diagram is well captured for all values of M_n; the symmetric copolymers ($f_A \approx 0.5$) exhibiting the Lam phase, which transforms to curved-domain phases (Lam → Cyl → Sph) as the volume fraction deviates from 0.5. At high temperatures, there is a tendency to form a disordered phase for all values of M_n, with the lowest $M_n = 20{,}000$ g/mol phase diagram exhibiting the widest region of a pure disordered phase due to both low M_n and high T resulting in low segregation strength. In contrast, the high values of M_n do not exhibit such a wide disordered phase, instead demonstrating broad order-disorder transition regions involving a structured phase co-existing with the disordered phase. An evident feature in all the phase diagrams is the presence of empty regions which denote the parameter spaces where the RFC model predicts either no phase or more than two phases. This failure of the RFC model to yield a correct prediction is likely due to the lack of training data in those regions.

The accuracy of the RFC model can be put into perspective by comparing it to SCFT simulations. Here, χ_{AB} estimation using the Hilderbrand solubility parameters provides a common basis to assess SCFT accuracy for many chemistries. Details on how solubility parameters are estimated are provided in the SI. As listed in Table 1, SCFT simulations predict the phase behavior with significantly lower accuracy compared to the RFC model, even when an empirical factor of 0.34 was added to the χ_{AB} estimation. Even considering only isoprene-styrene (IS) diblock copolymers, for which accurate χ_{AB} estimations obtained from experiments were available,[48] the RFC model still outperforms SCFT. The performance of SCFT depends strongly on the estimate of χ_{AB}, so improving the accuracy of the χ_{AB} estimate could improve the accuracy

of predictions. Additionally, SCFT performance could also increase by correcting for the mean-field assumption built in the SCFT famework.[49–51] Physics-based models also yield a disordered phase when the two blocks are same; however, for a data-driven models such fundamental physics needs to be build-in by supplying additional points (see Section 6 in SI).

Table 1: Performance of the RFC model and SCFT for phase prediction

Method	**Accuracy**	**1 − (Hamming-loss)**
All diblock copolymers (4768 data points)		
SCFT with $\chi_{AB} = (\delta_A - \delta_B)^2 v_{\text{ref}}/RT$	0.188	0.688
SCFT with $\chi_{AB} = (\delta_A - \delta_B)^2 v_{\text{ref}}/RT + 0.34$	0.529	0.823
RFC with No-Chem	0.926 ± 0.005	0.980 ± 0.002
RFC with Morgan-fp	0.921 ± 0.002	0.980 ± 0.001
IS diblock copolymers (693 data points)		
SCFT with $\chi_{AB} = (\delta_A - \delta_B)^2 v_{\text{ref}}/RT$	0.118	0.667
SCFT with $\chi_{AB} = (\delta_A - \delta_B)^2 v_{\text{ref}}/RT + 0.34$	0.538	0.825
SCFT with $\chi_{AB} = 71.4/T - 0.0857$	0.538	0.825
RFC with No-Chem	0.929 ± 0.006	0.982 ± 0.002
RFC with Morgan-fp	0.915 ± 0.002	0.980 ± 0.001

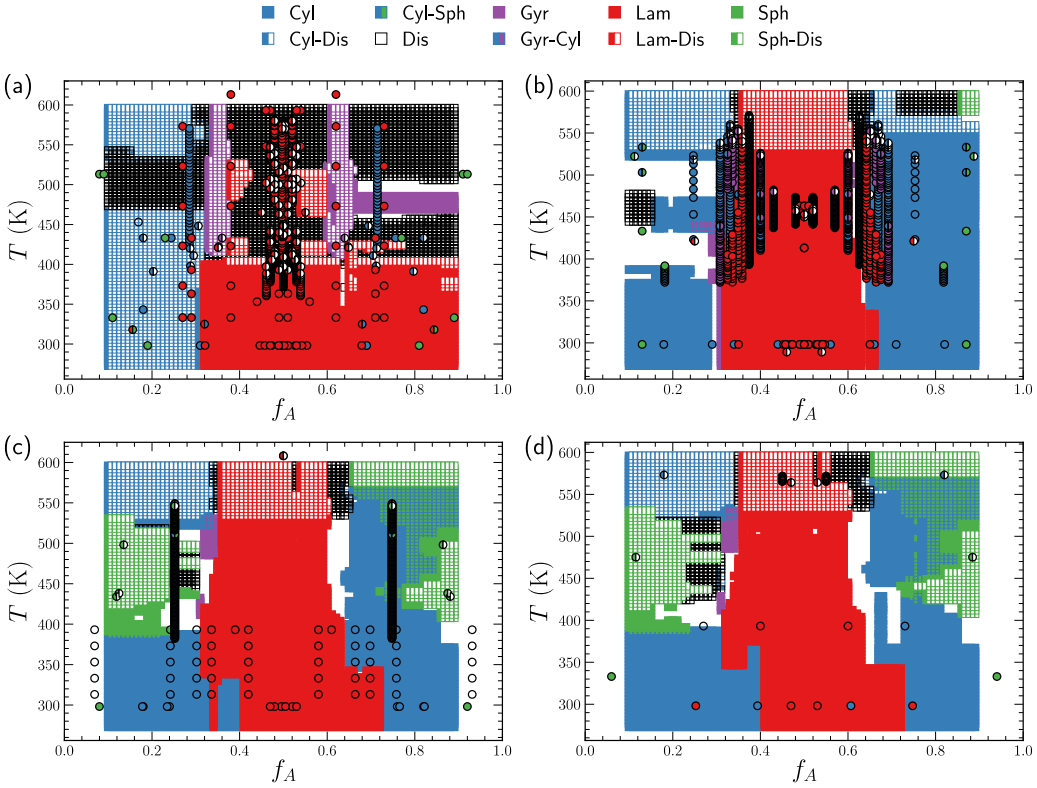

Figure 3: Phase diagrams ($T - f_A$) predicted by the RFC model, trained with all chemistries and evaluated for PI-PS, at four values of molar mass (M_n): (a) 20,000 g/mol, (b) 40,000 g/mol, (c) 60,000 g/mol, and (d) 80,000 g/mol. Points denote experimental data that have molar mass within 5000 g/mol of the specified M_n. The phase diagram is not symmetric because the training data used to build the model is asymmetric around $f_A = 0.5$, as it is selected randomly from the complete dataset.

The impact of individual features on the overall performance sheds more light on RFC-model behavior. Figure 2(a) demonstrates that the RFC model performance for predicting block copolymer phase behavior is independent of the type of fingerprinting method used. Crucially, the model performance remains almost the same even without using any descriptors for block chemistry, as depicted by the No-chem case; this effect is discussed in the SI (Section 7). Figure 2(b) depicts the rankings of different features, as indicated by respective feature importance scores. The score for chemistry is calculated as the sum of the scores for the individual entry in the fingerprint vector. For models with chemical fingerprints, molar mass is the dominant feature, followed by chemistry. For the No-chem case only, temperature is the dominant feature, further indicating the influence of high-density rheology and scattering temperature measurements. Interestingly, the score for chemistry increases as the complexity of fingerprints increases, indicating that a minimum of triplet-type fingerprints is required to capture the effect of distinct chemical groups. To further investigate the effect of chemical diversity, a smaller dataset is constructed by removing 9 chemical groups (Figure S5) totaling approximately 500 datapoints. Figure S5 shows that the importance of different features is reversed for the less chemical-rich dataset, with chemistry becoming the least important feature and temperature being the most dominant feature for the reduced dataset. This analysis clearly illustrates the need for chemical diversity in the dataset to build a reliable machine-learning model.

The RFC model described here cannot predict a new phase on which it was not trained and is limited in its ability to extrapolate to new chemistries. To examine the ability of RFC model to predict new chemistry, RFC model was trained on M-1 chemistries, and tested on the remaining one chemistry, where M is the total number of chemistries in the dataset (see section 9 in SI). Figure S6 demonstrates that there is a significant variation in the accuracy for different chemistries, indicating that the model is able to identify only a few selected chemistries while it fails to predict others. Such a behavior is not surprising considering the dataset is heavily biased towards a few extensively-studied diblock copolymers. However, this ability of RFC predictor can be increased further as the chemical diversity of the dataset grows.

Experimental data inherently contains uncertainty which may also affect the model prediction, especially if the model overfits noise in the data. The temperature measured during scattering and rheology can be reasonably accurate ($\Delta T = \pm 1$°C), while the molar mass and volume fraction measured by gel permeation chromatography can have uncertainties as high as 10%. In order to assess model sensitivity towards the uncertainties in experimental measurements, 100 new test datasets were created by adding ΔM_n, Δf_A, ΔT to M_n, f_A, and T values for each of the data points in the original dataset. These perturbations are selected randomly from Gaussian distributions having zero mean $\mu(\Delta M_n) = \mu(\Delta f_A) = \mu(\Delta T) = 0$, and specified variances $\sigma(\Delta M_n), \sigma(\Delta f_A)$ and $\sigma(\Delta T)$. Data points representing measurements from the same temperature-sweep are perturbed by the same ΔM_n and Δf_A but with different values of ΔT (see SI for details) to capture correlated error. Figure 4(a) shows the accuracy on the perturbed test sets for the No-chem model and the Morgan fingerprint model. It is evident that the RFC model is quite sensitive towards the variation in M_n, as indicated by the sharp decrease in the accuracy upon the introduction of uncertainty. Unlike accuracy, the sensitivity towards experimental uncertainty depends strongly on the fingerprinting method; Morgan fingerprint maintains 80% accuracy while for No-chem the accuracy decreases to 60% for highly varied samples. Figure 4(b) shows the accuracy as a function of uncertainty in both Δf_A and ΔM_n for models trained with Morgan fingerprints. As expected, variation in f_A affects the model performance more compared to similar variation in M_n. Compared to the RFC model, the variation in M_n and f_A has less impact on the SCFT-predictions (Figure S7). Nevertheless, the RFC model still outperforms the SCFT model even for the most varied samples.

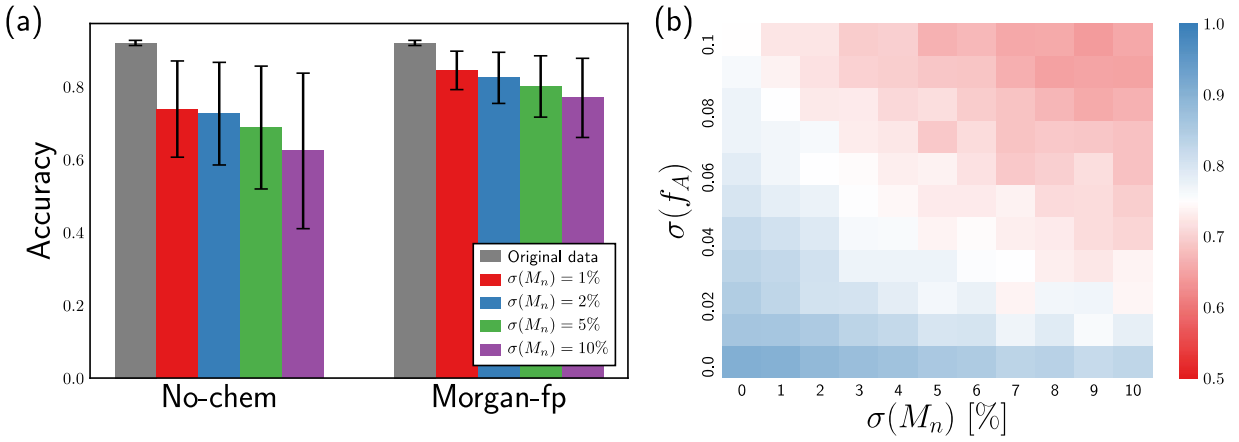

Figure 4: Sensitivity of the RFC-model prediction towards the uncertainty in experimentally measured molecular parameters. (a) Mean accuracy of 100 new samples created by adding Gaussian distributed perturbations ($\Delta T, \Delta f_A, \Delta M_n$) to each data point in the original dataset. For various values of ΔM_n, $\Delta f_A = 0.05$, and $\Delta T = 1$ K. (b) Heatmap for model performance with Morgan fingerprints for variations in both Δf_A and ΔM_n, with $\Delta T = 1$ K.

In summary, this work demonstrates that machine learning is promising for modeling the phase behavior of block copolymers while eliminating the need to estimate the Flory-Huggins parameter and other physiochemical parameters. Due to an *a priori* structure present in the data, the model performance is independent of the type of chemical fingerprinting method used with all different methods demonstrating 90% accuracy. However, the performance of model built without considering chemical descriptors is significantly more sensitive to experimental uncertainty than models containing such descriptors. More sophisticated methods such as boosting algorithms or deep neural networks may further increase the accuracy of predictions. Overall, the analyses presented in this work highlight the importance of model interpretability and domain knowledge in assessing the robustness and reliability of models built on practical chemical datasets which typically possess strong internal structure and chemical bias.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website. Code repository (link to the Github Repository). Details about group contribution theory calculations, machine-learning algorithm and hyperparameter optimization, numerical values of SCFT and RFC models accuracy, model sensitivity analysis, and data reduction protocols.

AUTHOR INFORMATION

Corresponding Author

* Bradley D. Olsen (bdolsen@mit.edu)

Funding Sources

This work is funded by the Community Resource for Innovation in Polymer Technology, a project supported by the National Science Foundation (NSF) Convergence Accelerator program (Convergence Accelerator Research-2040636).

ACKNOWLEDGMENT
This work is funded by the Community Resource for Innovation in Polymer Technology, a project supported by the National Science Foundation (NSF) Convergence Accelerator program (Convergence Accelerator Resrch-2040636). Fruitful discussions with Wengong Jin and Tommi S. Jaakkola are gratefully acknowledged.

REFERENCES

(1) Ruzette, A.-V.; Leibler, L. Block Copolymers in Tomorrow's Plastics. *Nat. Mater.* **2005**, *4* (1), 19–31.
(2) Kataoka, K.; Harada, A.; Nagasaki, Y. Block Copolymer Micelles for Drug Delivery: Design, Characterization and Biological Significance. *Adv. Drug Deliv. Rev.* **2012**, *64*, 37–48.
(3) Jackson, E. A.; Hillmyer, M. A. Nanoporous Membranes Derived from Block Copolymers: From Drug Delivery to Water Filtration. *ACS Nano* **2010**, *4* (7), 3548–3553.
(4) Urbas, A. M.; Maldovan, M.; DeRege, P.; Thomas, E. L. Bicontinuous Cubic Block Copolymer Photonic Crystals. *Adv. Mater.* **2002**, *14* (24), 1850–1853.
(5) Bates, F. S.; Fredrickson, G. H. Block Copolymer Thermodynamics: Theory and Experiment. *Annu. Rev. Phys. Chem.* **1990**, *41* (1), 525–557.
(6) Bates, C. M.; Bates, F. S. 50th Anniversary Perspective: Block Polymers–Pure Potential. *Macromolecules* **2017**, *50* (1), 3–22.
(7) Hajduk, D. A.; Gruner, S. M.; Erramilli, S.; Register, R. A.; Fetters, L. J. High-Pressure Effects on the Order- Disorder Transition in Block Copolymer Melts. *Macromolecules* **1996**, *29* (5), 1473–1481.
(8) Ruzette, A.-V.; Mayes, A.; Pollard, M.; Russell, T.; Hammouda, B. Pressure Effects on the Phase Behavior of Styrene/n-Alkyl Methacrylate Block Copolymers. *Macromolecules* **2003**, *36* (9), 3351–3356.
(9) Ryu, D. Y.; Lee, D. J.; Kim, J. K.; Lavery, K. A.; Russell, T. P.; Han, Y. S.; Seong, B. S.; Lee, C. H.; Thiyagarajan, P. Effect of Hydrostatic Pressure on Closed-Loop Phase Behavior of Block Copolymers. *Phys. Rev. Lett.* **2003**, *90* (23), 235501.
(10) Helfand, E. Block Copolymer Theory. III. Statistical Mechanics of the Microdomain Structure. *Macromolecules* **1975**, *8* (4), 552–556.
(11) Helfand, E.; Wasserman, Z. R. Block Copolymer Theory. 4. Narrow Interphase Approximation. *Macromolecules* **1976**, *9* (6), 879–888.
(12) Leibler, L. Theory of Microphase Separation in Block Copolymers. *Macromolecules* **1980**, *13* (6), 1602–1617.
(13) Semenov, A. Contribution to the Theory of Microphase Layering in Block-Copolymer Melts. *Zh Eksp Teor Fiz* **1985**, *88* (4), 1242–1256.
(14) Matsen, M. W. The Standard Gaussian Model for Block Copolymer Melts. *J. Phys. Condens. Matter* **2001**, *14* (2), R21.
(15) Matsen, M. W.; Bates, F. S. Unifying Weak-and Strong-Segregation Block Copolymer Theories. *Macromolecules* **1996**, *29* (4), 1091–1098.
(16) Drolet, F.; Fredrickson, G. H. Combinatorial Screening of Complex Block Copolymer Assembly with Self-Consistent Field Theory. *Phys. Rev. Lett.* **1999**, *83* (21), 4317.
(17) Arora, A.; Qin, J.; Morse, D. C.; Delaney, K. T.; Fredrickson, G. H.; Bates, F. S.; Dorfman, K. D. Broadly Accessible Self-Consistent Field Theory for Block Polymer Materials Discovery. *Macromolecules* **2016**, *49* (13), 4675–4690.
(18) Hiemenz, P. C.; Lodge, T. P. *Polymer Chemistry*; CRC press, 2007.

(19) Schwahn, D.; Hahn, K.; Streib, J.; Springer, T. Critical Fluctuations and Relaxation Phenomena in the Isotopic Blend Polystyrene/Deuteropolystyrene Investigated by Small Angle Neutron Scattering. *J. Chem. Phys.* **1990**, *93* (11), 8383–8391.
(20) Scheffold, F.; Eiser, E.; Budkowski, A.; Steiner, U.; Klein, J.; Fetters, L. J. Surface Phase Behavior in Binary Polymer Mixtures. I. Miscibility, Phase Coexistence, and Interactions in Polyolefin Blends. *J. Chem. Phys.* **1996**, *104* (21), 8786–8794.
(21) Maurer, W. W.; Bates, F. S.; Lodge, T. P.; Almdal, K.; Mortensen, K.; Fredrickson, G. H. Can a Single Function for χ Account for Block Copolymer and Homopolymer Blend Phase Behavior? *J. Chem. Phys.* **1998**, *108* (7), 2989–3000.
(22) Arora, A.; Pillai, N.; Bates, F. S.; Dorfman, K. D. Predicting the Phase Behavior of ABAC Tetrablock Terpolymers: Sensitivity to Flory–Huggins Interaction Parameters. *Polymer* **2018**, *154*, 305–314.
(23) Fredrickson, G. H.; Liu, A. J.; Bates, F. S. Entropic Corrections to the Flory-Huggins Theory of Polymer Blends: Architectural and Conformational Effects. *Macromolecules* **1994**, *27* (9), 2503–2511.
(24) Dudowicz, J.; Freed, K. F. Effect of Monomer Structure and Compressibility on the Properties of Multicomponent Polymer Blends and Solutions: 1. Lattice Cluster Theory of Compressible Systems. *Macromolecules* **1991**, *24* (18), 5076–5095.
(25) Honnell, K. G.; Curro, J. G.; Schweizer, K. S. Local Structure of Semiflexible Polymer Melts. *Macromolecules* **1990**, *23* (14), 3496–3505.
(26) Fedors, R. F. A Method for Estimating Both the Solubility Parameters and Molar Volumes of Liquids. *Polym. Eng. Sci.* **1974**, *14* (2), 147–154.
(27) Oishi, T.; Prausnitz, J. M. Estimation of Solvent Activities in Polymer Solutions Using a Group-Contribution Method. *Ind. Eng. Chem. Process Des. Dev.* **1978**, *17* (3), 333–339.
(28) Van Krevelen, D. W.; Te Nijenhuis, K. *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*; Elsevier, 2009.
(29) Bicerano, J. *Prediction of Polymer Properties*; cRc Press, 2002.
(30) Cheng, A.; Merz, K. M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure- Property Relationships. *J. Med. Chem.* **2003**, *46* (17), 3572–3580.
(31) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546.
(32) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6* (10), 1078–1082.
(33) Jackson, N. E.; Webb, M. A.; de Pablo, J. J. Recent Advances in Machine Learning towards Multiscale Soft Materials Design. *Curr. Opin. Chem. Eng.* **2019**, *23*, 106–114.
(34) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *Npj Comput. Mater.* **2017**, *3* (1), 1–13.
(35) Doi, H.; Takahashi, K. Z.; Tagashira, K.; Fukuda, J.; Aoyagi, T. Machine Learning-Aided Analysis for Complex Local Structure of Liquid Crystal Polymers. *Sci. Rep.* **2019**, *9* (1), 1–12.
(36) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
(37) Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; others. Machine-Learning-Assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. *Npj Comput. Mater.* **2019**, *5* (1), 1–11.
(38) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv.* **2020**, *6* (20), eaaz4301.
(39) Aoyagi, T. Deep Learning Model for Predicting Phase Diagrams of Block Copolymers. *Comput. Mater. Sci.* **2021**, *188*, 110224.

(40) Zhao, S.; Cai, T.; Zhang, L.; Li, W.; Lin, J. Autonomous Construction of Phase Diagrams of Block Copolymers by Theory-Assisted Active Machine Learning. *ACS Macro Lett.* **2021**, *10* (5), 598–602.
(41) DataPaper. Block Copolymers Data. *Prep.* **2020**.
(42) Lee, S.; Bluemle, M. J.; Bates, F. S. Discovery of a Frank-Kasper σ Phase in Sphere-Forming Block Copolymer Melts. *Science* **2010**, *330* (6002), 349–353.
(43) Li, W.; Duan, C.; Shi, A.-C. Nonclassical Spherical Packing Phases Self-Assembled from AB-Type Block Copolymers. *ACS Macro Lett.* **2017**, *6* (11), 1257–1262.
(44) Kim, K.; Schulze, M. W.; Arora, A.; Lewis, R. M.; Hillmyer, M. A.; Dorfman, K. D.; Bates, F. S. Thermal Processing of Diblock Copolymer Melts Mimics Metallurgy. *Science* **2017**, *356* (6337), 520–523.
(45) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd international conference on document analysis and recognition*; IEEE, 1995; Vol. 1, pp 278–282.
(46) Liaw, A.; Wiener, M.; others. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18–22.
(47) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(48) Khandpur, A. K.; Foerster, S.; Bates, F. S.; Hamley, I. W.; Ryan, A. J.; Bras, W.; Almdal, K.; Mortensen, K. Polyisoprene-Polystyrene Diblock Copolymer Phase Diagram near the Order-Disorder Transition. *Macromolecules* **1995**, *28* (26), 8796–8806.
(49) Vorselaars, B.; Stasiak, P.; Matsen, M. W. Field-Theoretic Simulation of Block Copolymers at Experimentally Relevant Molecular Weights. *Macromolecules* **2015**, *48* (24), 9071–9080.
(50) Fredrickson, G. H.; Helfand, E. Fluctuation Effects in the Theory of Microphase Separation in Block Copolymers. *J. Chem. Phys.* **1987**, *87* (1), 697–705.
(51) Delaney, K. T.; Fredrickson, G. H. Recent Developments in Fully Fluctuating Field-Theoretic Simulations of Polymer Melts and Solutions. *J. Phys. Chem. B* **2016**, *120* (31), 7615–7634.